MEAN-FIELD LANGEVIN DYNAMICS FOR TWO-LAYER RELU NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mean-field Langevin dynamics provide a framework for analyzing convergence of wide two-layer neural networks while capturing feature learning. But standard uniform log-Sobolev arguments do not apply to ReLU activations with weak weight decay, where the associated Gibbs potentials may be non-coercive. For square loss, under boundedness, regularity of the activation indicator, and a mild initialization condition, we prove existence of global weak solutions and establish exponential convergence of the free energy to its global minimum for every positive weight-decay and temperature parameter. The proof uses a positive coercivity threshold to separate two regimes. In the low-coercivity regime, activation regularity and two-homogeneity yield a uniform lower bound on the dissipation; in the coercive regime, radial rescaling yields a log-Sobolev inequality. Together, these estimates imply a Polyak–Łojasiewicz inequality on every free-energy sublevel. For fixed input dimension and other problem parameters, the resulting convergence rate deteriorates at most polynomially in the inverse temperature.

1 INTRODUCTION

Mean-field Langevin dynamics (MFLD) provide a framework for analyzing gradient-based training of wide two-layer neural networks while capturing the evolution of their learned features. In the mean-field limit, a probability distribution describes the neuron parameters, and the network output is the average of the corresponding neuron features. Training changes this distribution and hence the representation itself (Mei et al., 2018; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2018; Mei et al., 2019). This differs from the neural tangent kernel regime, in which training is approximated by a linearization around initialization (Jacot et al., 2018; Du et al., 2019; Allen-Zhu et al., 2019).

A central question of MFLD is whether the resulting nonlinear feature-learning dynamics admits a quantitative global convergence guarantee. For objectives that are convex in the parameter distribution, a standard route to quantitative convergence of MFLD for networks with bounded smooth activation is through log-Sobolev inequalities (LSIs) for the associated family of proximal Gibbs distributions. When these distributions are well-defined and satisfy a uniform log-Sobolev inequality, the entropy-regularized objective value converges exponentially fast to its global minimum (Nitanda et al., 2022; Chizat, 2022; Suzuki et al., 2023a). Existing analyses establish such control under suitable smoothness or confinement assumptions (Hu et al., 2021; Zhang et al., 2022; Suzuki et al., 2023b; Takakura & Suzuki, 2024).

However, this standard LSI route becomes problematic for jointly trained ReLU networks under weak quadratic weight decay (Chizat, 2022). Beyond the nonsmoothness of ReLU, its unbounded growth creates a more fundamental difficulty: the feature function is positively homogeneous of degree two in the joint parameters, so bounded- or Lipschitz-perturbation arguments for establishing an LSI no longer apply. Moreover, under weak weight decay, the potential term can dominate the quadratic confinement along some directions, causing the potential to lose coercivity and the associated proximal Gibbs distribution to become non-normalizable. Coercivity could be enforced by sufficiently strong weight decay, but such a requirement would exclude the weakly regularized regime of practical interest. Even when coercivity remains positive, the LSI estimates deteriorate as the coercivity margin approaches zero. Thus, a basic question remains open:

Can MFLD for exact ReLU networks converge globally at a quantitative rate?

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

In this paper, we provide an affirmative answer to this question for square loss and jointly trained weights and biases under any fixed positive weight decay and temperature. Under boundedness assumptions, an activation-indicator regularity condition (Wojtowytsch, 2020, P4), and mild initialization conditions, we prove the existence of global weak trajectories and show that, along every such trajectory, the objective value converges exponentially to its global minimum (Theorem 1).

The proof involves a new geometric analysis that establishes positive dissipation in the regime where the coercivity is small or negative (Proposition 3). This, along with a log-Sobolev inequality in the coercive regime, yields a Polyak–Łojasiewicz (PŁ) inequality on every free-energy sublevel and thus exponential convergence of the free energy. In particular, with the input dimension, positive weight decay, and data-dependent constants fixed, the guaranteed convergence rate deteriorates only polynomially as the Langevin temperature decreases. This contrasts with the exponential inverse-temperature deterioration of guarantees for smooth bounded features (Nitanda et al., 2022; Chizat, 2022). For a more detailed discussion of related work, see Section A.

2 Mean-field preliminaries and the ReLU obstruction

Notation. We write $\|\cdot\|$ for the Euclidean norm and $\mathrm{Lip}(f)$ for the Lipschitz constant with respect to the Euclidean metric on the stated domain. We use $\lesssim$, $\gtrsim$, and $\asymp$ for bounds up to positive multiplicative constants. For a Euclidean space V and $k \geq 1$, $\mathcal{P}_k(V)$ denotes the Borel probability distributions with finite kth moment, and W_2 denotes the 2-Wasserstein distance. A curve (μ_t) is narrowly continuous if $t \mapsto \int \varphi \, \mathrm{d}\mu_t$ is continuous for every bounded continuous function φ.

2.1 The mean-field network model

Let $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, $d \geq 1$, follow an underlying data distribution with $\mathbb{E}[\|X\|^2 + Y^2] < \infty$. Write P for the input distribution and $m(x) = \mathbb{E}[Y \mid X = x]$ for the regression function. For $x \in \mathbb{R}^d$, define the lifted vector $\bar{x} = (x, 1) \in \mathbb{R}^{d+1}$, and let $\bar{X} = (X, 1)$ denote the lifted random input. Set

$$\Theta := \mathbb{R} \times \mathbb{R}^{d+1} = \mathbb{R}^p, \qquad p := d + 2. \tag{2.1}$$

Thus p is the total number of parameters of one neuron: one output coefficient, d input weights, and one bias. All distributions, Euclidean gradients, and diffusion operators below are defined on this parameter space. With $v = (w, b)$, a trainable parameter $\theta = (a, v) \in \Theta$ specifies an output-layer coefficient a, an input weight w, and a bias b. Its ReLU feature is

$$\Phi(\theta, x) := a(v^\top \bar{x})_+ = a(w^\top x + b)_+. \tag{2.2}$$

The mean-field network averages these features over a parameter distribution $\mu \in \mathcal{P}_2(\Theta)$,

$$f_\mu(x) := \int_\Theta \Phi(\theta, x) \, \mathrm{d}\mu(\theta). \tag{2.3}$$

2.2 Mean-field Langevin dynamics for ReLU networks

Entropy-regularized objective. Let $\tau > 0$ denote the Langevin temperature, equivalently the entropy regularization coefficient, and $\lambda > 0$ the quadratic regularization strength. The latter induces the weight-decay drift $-\lambda\theta$. For $\mu \in \mathcal{P}_2(\Theta)$ absolutely continuous with respect to Lebesgue measure, define the entropy-regularized objective (free energy)[1]

$$\mathcal{F}_{\tau,\lambda}(\mu) := \frac{1}{2}\|f_\mu - m\|^2_{L^2(P)} + \frac{\lambda}{2}\int_\Theta \|\theta\|^2 \, \mathrm{d}\mu(\theta) + \tau \int_\Theta \log\left(\frac{d\mu}{d\theta}(\theta)\right) \, \mathrm{d}\mu(\theta) + C_{\tau,\lambda}, \tag{2.4}$$

where $C_{\tau,\lambda} := \frac{\tau p}{2}\log(2\pi\tau/\lambda)$. This constant normalizes the combined entropy and confinement terms to have minimum zero. We set $\mathcal{F}_{\tau,\lambda}(\mu) = +\infty$ when μ is not absolutely continuous with respect to Lebesgue measure on Θ.

First variation and homogeneity. The prediction residual for the square loss is $q_\mu := f_\mu - m$. For a fixed $q \in L^2(P)$, the associated potential is defined as

$$U_q(\theta) := \Psi_q(\theta) + \frac{\lambda}{2}\|\theta\|^2, \quad \Psi_q(\theta) = \mathbb{E}[q(X)\Phi(\theta, X)]. \tag{2.5}$$

[1] Equivalently, $\mathcal{F}_{\tau,\lambda}(\mu) = \frac{1}{2}\|f_\mu - m\|^2_{L^2(P)} + \tau \, \mathrm{KL}(\mu\|\gamma_{\tau,\lambda})$ with $\gamma_{\tau,\lambda} := \mathcal{N}(0, (\tau/\lambda)I_p)$.

Since $\mu \mapsto f_\mu$ is linear, the interaction potential Ψ_{q_μ} is a first variation of the square-loss term at μ. Adding the contribution from quadratic confinement gives U_{q_μ} as a first variation of the sum of the loss and confinement terms, excluding entropy. As both the ReLU feature and quadratic confinement are positively two-homogeneous in the full parameter θ, we have

$$\Phi(s\theta, x) = s^2\Phi(\theta, x), \qquad U_q(s\theta) = s^2 U_q(\theta), \qquad \forall s \geq 0. \tag{2.6}$$

This homogeneity means that U_q is determined by its restriction to the unit sphere $\mathbb{S}^{p-1} \subset \Theta$ (Chizat & Bach, 2020; Akiyama & Suzuki, 2021). Defining the angular coefficient $c_q(u) := U_q(u)$ and writing $\theta = su$, we obtain the polar representation

$$U_q(su) = s^2 c_q(u), \qquad \forall s \geq 0, u \in \mathbb{S}^{p-1}. \tag{2.7}$$

The minimum of c_q over the sphere is the *coercivity margin*.

Dynamics and dissipation. The mean-field Langevin dynamics are formally described by the nonlinear McKean–Vlasov SDE

$$d\Theta_t = -\nabla U_{q_{\mu_t}}(\Theta_t)\, dt + \sqrt{2\tau}\, dW_t, \qquad \mu_t = \mathrm{Law}(\Theta_t), \tag{2.8}$$

where W_t is a standard Brownian motion in $\mathbb{R}^p$. The law $\mu_t = \mathrm{Law}(\Theta_t)$ of (2.8) formally evolves according to the nonlinear Fokker–Planck equation, which is also the formal Wasserstein gradient flow of $\mathcal{F}_{\tau,\lambda}$:

$$\partial_t \mu_t = \nabla \cdot \left(\mu_t \nabla U_{q_{\mu_t}}\right) + \tau \Delta \mu_t. \tag{2.9}$$

The nonlinearity comes from the dependence of q_{μ_t} on the evolving distribution μ_t. Parameter gradients are weak gradients; Appendix C specifies a Borel representative for the exact ReLU interaction. To formulate the evolution and energy dissipation for weak trajectories, we introduce the following probability current and dissipation functional.

Definition 1 (Probability current and dissipation)**.** Let $\mu \in \mathcal{P}_2(\Theta)$ have finite free energy. Define the probability current J_μ, in the sense of distributions, and the associated dissipation $\mathcal{D}(\mu)$ by

$$J_\mu := \tau \nabla \left(\frac{d\mu}{d\theta}\right) + \frac{d\mu}{d\theta} \nabla U_{q_\mu}, \tag{2.10}$$

$$\mathcal{D}(\mu) := \begin{cases} \displaystyle\int_\Theta \|\xi_\mu\|^2 \, d\mu, & \text{if } J_\mu = \xi_\mu \mu \text{ for some } \xi_\mu \in L^2(\mu; \mathbb{R}^p), \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.11}$$

With this sign convention, $\partial_t \mu_t = \nabla \cdot J_{\mu_t}$; the transport velocity in the continuity equation is $-\xi_{\mu_t}$.

2.3 Standard LSI arguments to convergence

A powerful tool for establishing exponential convergence of MFLD is to prove a PŁ inequality that controls the free-energy gap by the dissipation.

Definition 2 (Polyak–Łojasiewicz inequality)**.** Let $\mathcal{S} \subseteq \mathcal{P}_2(\Theta)$ consist of finite-energy measures. The free energy $\mathcal{F}_{\tau,\lambda}$ satisfies a *Polyak–Łojasiewicz inequality* on $\mathcal{S}$ with constant $\kappa > 0$ if

$$\mathcal{D}(\mu) \geq \kappa \left(\mathcal{F}_{\tau,\lambda}(\mu) - \inf_{\pi \in \mathcal{P}_2(\Theta)} \mathcal{F}_{\tau,\lambda}(\pi)\right), \qquad \mu \in \mathcal{S}. \tag{2.12}$$

A standard route to establishing this PŁ inequality is to use uniform LSIs for the associated proximal Gibbs distributions (Nitanda et al., 2022; Chizat, 2022). For $\mu \in \mathcal{P}_2(\Theta)$, whenever $Z_{q_\mu} < \infty$, we define the *proximal Gibbs distribution* associated with μ as

$$\nu_{q_\mu}(\mathrm{d}\theta) = \frac{1}{Z_{q_\mu}} \exp\left(-\tfrac{\Psi_{q_\mu}(\theta)}{\tau}\right) \gamma_{\tau,\lambda}(\mathrm{d}\theta), \quad \text{where } Z_{q_\mu} := \int_\Theta \exp\left(-\tfrac{\Psi_{q_\mu}(\theta)}{\tau}\right) \mathrm{d}\gamma_{\tau,\lambda}(\theta). \tag{2.13}$$

Its density with respect to Lebesgue measure is proportional to $\exp(-U_{q_\mu}/\tau)$. For probability measures $\mu \ll \nu$ on Θ, with ν having a positive density, the relative entropy and relative Fisher information are

$$\mathrm{KL}(\mu\|\nu) := \int_\Theta \log\left(\tfrac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu, \qquad I(\mu\|\nu) := 4 \int_\Theta \left\|\nabla \sqrt{\tfrac{\mathrm{d}\mu}{\mathrm{d}\nu}}\right\|^2 \mathrm{d}\nu. \tag{2.14}$$

The gradient is understood weakly, and $I(\mu\|\nu) = +\infty$ when the required weak regularity fails. If $\mu \not\ll \nu$, both quantities are $+\infty$.

Definition 3 (Log-Sobolev inequality)**.** A probability measure ν on Θ with a positive density satisfies a log-Sobolev inequality with constant $\alpha > 0$, denoted $\mathrm{LSI}(\alpha)$, if $\mathrm{KL}(\mu\|\nu) \le \frac{1}{2\alpha} I(\mu\|\nu)$ for every probability measure μ on Θ.

For a finite-energy measure μ with $\mathcal{D}(\mu) < \infty$, the probability-current formulation in Definition 1 gives $\mathcal{D}(\mu) = \tau^2 I(\mu\|\nu_{q_\mu})$ whenever ν_{q_μ} exists. If this Gibbs distribution satisfies $\mathrm{LSI}(\alpha)$, convexity of the square loss and the Gibbs variational formula yield the desired PŁ inequality

$$\mathcal{F}_{\tau,\lambda}(\mu) - \inf_{\pi \in \mathcal{P}_2(\Theta)} \mathcal{F}_{\tau,\lambda}(\pi) \le \tau\, \mathrm{KL}(\mu\|\nu_{q_\mu}) \le \frac{\mathcal{D}(\mu)}{2\tau\alpha}. \tag{2.15}$$

2.4 Why the standard route does not apply to ReLU networks

The standard route in Section 2.3 requires the proximal Gibbs distributions to be normalizable and to satisfy a log-Sobolev inequality with a positive constant. For jointly trained ReLU networks, quadratic regularization with $\lambda > 0$ does not by itself guarantee these conditions. Indeed, the perturbation relative to the Gaussian reference measure is Ψ_q/τ. Unless $\Psi_q \equiv 0$, two-homogeneity implies that this perturbation is neither bounded nor globally Lipschitz. The bounded perturbation argument of Chizat (2022) and the Lipschitz perturbation criterion of Suzuki et al. (2023a, Lemma 6) do not apply directly. As discussed in Chizat (2022, Section 5.1), standard ReLU features also violate the smoothness assumptions of that framework. A structural obstruction to the standard LSI argument is that the required proximal Gibbs distribution may fail to exist. Its existence depends on the competition between quadratic confinement and the first variation of the square-loss term, as captured by $c_q(u) = \lambda/2 + \Psi_q(u)$. Under weak weight decay, $\Psi_q(u)$ can be smaller than $-\lambda/2$ in some direction. Continuity then gives an open cone on which U_q tends to $-\infty$ quadratically, implying $Z_q = \infty$. Even when the proximal Gibbs distribution is well defined, a further challenge is to derive a quantitative LSI for it.

3 Standing assumptions and main result

3.1 Data assumptions

Assumption 1 (Bounded inputs and regression function)**.** *There exist finite constants $R, B_m > 0$ such that the following hold almost surely:*

$$\|X\| \le R, \qquad |m(X)| \le B_m. \tag{3.1}$$

Assumption 2 (Lipschitz continuity of activation indicators)**.** *There exists $L_{\mathrm{act}} < \infty$ such that*

$$\mathbb{E}\left[\|\bar{X}\| \left|\mathbf{1}_{\{v^\top \bar{X} > 0\}} - \mathbf{1}_{\{v'^\top \bar{X} > 0\}}\right|\right] \le L_{\mathrm{act}} \|v - v'\|, \qquad v, v' \in \mathbb{S}^d. \tag{3.2}$$

This is condition (P4) of Wojtowytsch (2020). It requires the activation indicator to vary Lipschitz-continuously with the normalized affine parameter in $L^1(\|\bar{x}\|\,\mathrm{d}P(x))$. In particular, it rules out mass on activation hyperplanes,

$$P(v^\top \bar{X} = 0) = 0 \qquad \text{for every } v \neq 0,$$

and ensures differentiability of the averaged ReLU feature at every nonzero affine parameter. Bounded, compactly supported input densities satisfy this condition and more examples are given in Wojtowytsch (2020, Theorem 2.1).

3.2 Initialization and weak trajectories

Assumption 3 (Finite-energy initialization with a fourth moment)**.** *The initial law $\mu_0 \in \mathcal{P}_4(\Theta)$ satisfies $\mathcal{F}_{\tau,\lambda}(\mu_0) \le F_0$ for some $F_0 \in (0, \infty)$. Thus*

$$M_4(\mu_0) := \int_\Theta \|\theta\|^4 \,\mathrm{d}\mu_0(\theta) < \infty. \tag{3.3}$$

Every nondegenerate Gaussian initialization satisfies the moment and finite-energy conditions. The fourth moment is used in the construction of the nonlinear evolution; the static convergence estimates below require only the second-moment bound on a free-energy sublevel.

For a fixed energy level $F_0 > 0$, write

$$\mathcal{S}_{F_0} := \{\mu \in \mathcal{P}_2(\Theta) : \mathcal{F}_{\tau,\lambda}(\mu) \le F_0\}. \tag{3.4}$$

We now define weak trajectories through the weak formulation of the Fokker-Planck equation and the corresponding energy-dissipation inequality.

Definition 4 (Global energy-dissipating weak trajectory)**.** A narrowly continuous curve $(\mu_t)_{t\ge 0} \subset \mathcal{P}_2(\Theta)$ of finite-energy measures is a global energy-dissipating weak trajectory if $t \mapsto \mathcal{D}(\mu_t)$ is measurable and there is a measurable vector field $\xi_t(\theta)$ such that $J_{\mu_t} = \xi_t \mu_t$ with $\xi_t \in L^2(\mu_t; \Theta)$ for almost every t, and, for every $0 \le s \le t < \infty$,

$$\int_\Theta \phi \,\mathrm{d}\mu_t - \int_\Theta \phi \,\mathrm{d}\mu_s = -\int_s^t \int_\Theta \langle \nabla\phi, \xi_r\rangle \,\mathrm{d}\mu_r \,\mathrm{d}r, \qquad \phi \in C_c^\infty(\Theta), \tag{3.5}$$

and

$$\mathcal{F}_{\tau,\lambda}(\mu_t) + \int_s^t \mathcal{D}(\mu_r)\,\mathrm{d}r \le \mathcal{F}_{\tau,\lambda}(\mu_s). \tag{3.6}$$

The energy-dissipation inequality keeps every such trajectory with initial energy at most F_0 in $\mathcal{S}_{F_0}$ for all $t \ge 0$.

Proposition 1 (Global existence and energy dissipation)**.** *Suppose Assumptions 1 and 3 hold and $\tau, \lambda > 0$. There exists a global energy-dissipating weak trajectory starting from μ_0. It is continuous in W_2, has a locally bounded fourth moment, and satisfies the energy-dissipation inequality* (3.6) *with equality. Moreover, with $M_2(\mu) := \int_\Theta \|\theta\|^2 \,\mathrm{d}\mu$,*

$$M_2(\mu_t) \le e^{-2\lambda t} M_2(\mu_0) + \frac{2\tau p + \|m\|^2_{L^2(P)}}{2\lambda}(1 - e^{-2\lambda t}). \tag{3.7}$$

Appendix C constructs the trajectory by a contraction in $C([0,T]; L^2(P))$ on short time intervals and continues it using moment bounds. The entropy estimate of Lacker (2023, Lemma 4.4 and Remark 4.5) and the time-reversal identities of Cattiaux et al. (2023, Propositions 4.6 and 4.8, Theorem 4.9) supply the probability current and the entropy chain rule.

3.3 Main result

We first establish the existence and uniqueness of the free-energy minimizer in Proposition 2. Theorem 1 then establishes exponential convergence of the free energy to the corresponding minimum. The proofs are given in Appendix F.

Proposition 2 (Existence and uniqueness of the minimizer)**.** *Suppose Assumptions 1 and 2 hold, and fix $\tau, \lambda > 0$. The functional $F_{\tau,\lambda}$ admits a unique minimizer $\mu_\star \in \mathcal{P}_2(\Theta)$, and its optimal value satisfies*

$$F_{\tau,\lambda}(\mu_\star) = F_{\inf} := \inf_{\pi \in \mathcal{P}_2(\Theta)} F_{\tau,\lambda}(\pi) \in [0, \infty). \tag{3.8}$$

Theorem 1 (Global exponential convergence in free energy)**.** *Suppose Assumptions 1, 2, and 3 hold and fix $\tau, \lambda, F_0 > 0$. There is a constant $\kappa_{F_0} > 0$, explicitly given in* (F.4)*, such that every $\mu \in \mathcal{S}_{F_0}$ satisfies*

$$\mathcal{D}(\mu) \ge \kappa_{F_0}\big(\mathcal{F}_{\tau,\lambda}(\mu) - \mathcal{F}_{\tau,\lambda}(\mu_\star)\big). \tag{3.9}$$

A global energy-dissipating weak trajectory starting from μ_0 exists. Every trajectory in Definition 4 with this initial law satisfies

$$\mathcal{F}_{\tau,\lambda}(\mu_t) - \mathcal{F}_{\tau,\lambda}(\mu_\star) \le e^{-\kappa_{F_0}(t-s)}\big(\mathcal{F}_{\tau,\lambda}(\mu_s) - \mathcal{F}_{\tau,\lambda}(\mu_\star)\big), \qquad 0 \le s \le t. \tag{3.10}$$

Moreover, for fixed $d, F_0, \lambda, R, B_m, L_{\mathrm{act}}$, there are constants $c, \tau_0 > 0$ such that

$$\kappa_{F_0}(\tau) \ge c\,\tau^{(d+4)(d+8)/6}, \qquad 0 < \tau \le \tau_0. \tag{3.11}$$

Theorem 1 establishes exponential convergence for jointly trained ReLU networks with standard quadratic weight decay, addressing in this setting open problems raised by Zhang et al. (2022); Chizat (2022). The guaranteed convergence rate has polynomial temperature dependence with exponent $(d+4)(d+8)/6$ for each fixed input dimension d, while general analyses of mean-field

Langevin dynamics for convex functionals typically yield convergence-time bounds with exponential dependence on inverse temperature (Chizat, 2022; Nitanda et al., 2022). Polynomial temperature dependence has been obtained under near-optimal initialization (Wang et al., 2024), under a positive lower bound on temperature at fixed dimension (Mousavi-Hosseini et al., 2025), or through an annealing scheme tailored to classification losses (Suzuki et al., 2023b). Our result broadens the scope of such guarantees in the present ReLU setting, allowing arbitrarily small fixed positive temperatures with a mild initialization condition.

4 Proof of the main result

Proposition 1 provides global weak trajectories and their energy identity. We now sketch the proof of exponential convergence of the free-energy gap along these trajectories. Let $\underline{c}_\mu := \min_{u\in\mathbb{S}^{p-1}} c_{q_\mu}(u)$ denote the coercivity margin. In Sections 4.1 and 4.2, we establish PŁ inequalities in the low-coercivity regime $\underline{c}_\mu \le c_*$ and the high-coercivity regime $\underline{c}_\mu > c_*$, respectively, where $c_* > 0$ is a threshold to be chosen. Section 4.3 then combines the energy-dissipation inequality (3.6) and the PŁ inequalities to establish global convergence.

Throughout this section, Assumptions 1 and 2 hold, and $\tau, \lambda, F_0 > 0$ are fixed. Appendix B provides the following bounds, uniformly over $\mu \in \mathcal{S}_{F_0}$:

$$M_2(\mu) \le M, \qquad \|q_\mu\|_\infty \le Q_0, \qquad \|c_{q_\mu}\|_\infty \le C_0, \qquad \mathrm{Lip}(c_{q_\mu}) \le L_0. \tag{4.1}$$

In particular, the free-energy sublevel bound controls the second moment even when U_{q_μ} has low coercivity.

4.1 Low-coercivity regime: Positive dissipation

Since the free-energy gap is bounded on $\mathcal{S}_{F_0}$, a uniform positive lower bound on the dissipation suffices to establish the PŁ inequality in the low-coercivity regime. In this subsection, we derive the following result.

Proposition 3 (Positive dissipation at low coercivity)**.** *There exist constants $c_* \in (0, \lambda/4]$ and $d_* > 0$, uniform over $\mathcal{S}_{F_0}$, such that*

$$\underline{c}_\mu \le c_* \quad \Longrightarrow \quad \mathcal{D}(\mu) \ge d_*. \tag{4.2}$$

Moreover, as $\tau \downarrow 0$ with $p, F_0, \lambda, R, B_m, L_{\mathrm{act}}$ fixed, the choices below satisfy

$$c_* \asymp \tau^{(p+2)/3}, \qquad d_* \gtrsim \tau^{2+\max\left\{\frac{p}{2}, \frac{2(p-1)}{3}\right\}}, \tag{4.3}$$

with implicit constants independent of τ.

We next outline the proof of Proposition 3 in three steps. In Step 1, we express dissipation in terms of the spatial variation of a transformation of the probability density. On regions where the potential is low, this representation allows comparisons of probability mass to yield a lower bound on dissipation. Step 2 uses low coercivity to construct such a region. Step 3 combines the resulting comparison with a mass estimate near the origin and the second-moment bound to obtain the claimed uniform positive lower bound on dissipation.

Step 1. A gradient representation of dissipation. We first relate the dissipation to the spatial distribution of probability. To do so, we introduce a Gibbs-relative square-root density h_μ, by which probabilities are then represented as weighted integrals of h_μ^2, and dissipation as a weighted integral of $\|\nabla h_\mu\|^2$. This allows us to compare probabilities in different regions in terms of dissipation.

Definition 5 (Gibbs-relative square-root density)**.** For $\mu \in \mathcal{S}_{F_0}$, its *Gibbs-relative square-root density* is defined as

$$h_\mu := \left(\frac{\mathrm{d}\mu}{\mathrm{d}\theta}\right)^{1/2} e^{U_{q_\mu}/(2\tau)}.$$

This definition uses the unnormalized Gibbs weight and remains valid when the proximal Gibbs distribution is not normalizable.

Lemma 1 (Probability and dissipation representations)**.** *For every $\mu \in \mathcal{S}_{F_0}$ with $\mathcal{D}(\mu) < \infty$, the Gibbs-relative square-root density h_μ belongs to $H^1_{\mathrm{loc}}(\Theta)$ and has weak gradient*

$$\nabla h_\mu = \frac{h_\mu}{2\tau}\xi_\mu.$$

The probability of any measurable set $B \subset \Theta$ and the dissipation can be represented by h_μ:

$$\mu(B) = \int_B h_\mu^2 e^{-U_{q_\mu}/\tau}\,\mathrm{d}\theta, \qquad \mathcal{D}(\mu) = 4\tau^2 \int_\Theta \|\nabla h_\mu\|^2 e^{-U_{q_\mu}/\tau}\,\mathrm{d}\theta. \tag{4.4}$$

If a region $\Omega \subseteq \Theta$ satisfies $U_{q_\mu} \le C\tau$ for a fixed constant $C > 0$, the dissipation identity in Lemma 1 gives

$$\int_\Omega \|\nabla h_\mu\|^2\,\mathrm{d}\theta \le e^C \int_\Omega \|\nabla h_\mu\|^2 e^{-U_{q_\mu}/\tau}\,\mathrm{d}\theta \le \frac{e^C}{4\tau^2}\mathcal{D}(\mu).$$

The left-hand side is the *local Dirichlet energy* of h_μ on Ω. Hence, a lower bound on the local Dirichlet energy of h_μ over a low-potential region yields a lower bound on the dissipation. We next construct such a region.

Step 2. Constructing a low-potential tube. Along the ray of a minimum coercivity direction $u_0 = (\alpha_0, v_0) \in \arg\min_{u\in\mathbb{S}^{p-1}} c_{q_\mu}(u)$, we construct the tube

$$\mathcal{T}_L = \{su_0 + z : b_0/2 < s < L,\ z \perp u_0,\ \|z\| < \delta\}. \tag{4.5}$$

Here δ is the cross-sectional radius, and b_0 determines the entrance location at $s = b_0/2$. The tube extends along u_0 to $s = L$, with total length $L - b_0/2$. The potential remains uniformly small in a sufficiently narrow neighborhood of the minimizing ray.

Lemma 2 (Potential bound near a minimizing ray)**.** *Let $\mu \in \mathcal{S}_{F_0}$ and $u_0 \in \arg\min_{u\in\mathbb{S}^{p-1}} c_{q_\mu}(u)$. Suppose that $c_{q_\mu}(u_0) \le c_*$, where $0 < c_* \le \lambda/4$. For $r_0 = \frac{1}{8}, H = 4Q_0\big(\sqrt{R^2+1} + L_{\mathrm{act}}\big)$, the potential along the tube satisfies*

$$U_{q_\mu}(su_0 + z) \le 2c_* s^2 + H\|z\|^2, \forall s > 0, \quad z \perp u_0, \|z\| \le r_0 s. \tag{4.6}$$

The constants r_0 and H are uniform over $\mathcal{S}_{F_0}$.

See Appendix D.1 for the proof. We can choose the following width, entrance, and length of the tube to make it a low-potential region with $U_{q_\mu} \lesssim \tau$.

Lemma 3 (Choice of the tube parameters)**.** *Let r_0, H be as in Lemma 2. Choose the radius, length, entrance, and coercivity threshold as follows:*

$$\begin{aligned} b_0 &= \frac{1}{64}\min\left\{2\sqrt{M},\ \sqrt{\frac{\tau}{C_0}}\right\},\ \ \delta = \frac{1}{128}\min\left\{b_0,\ b_0 r_0,\ \sqrt{\frac{\tau}{H + C_0}}\right\}, \\ L &= K_p\sqrt{M}\left(\frac{\sqrt{M}}{\delta}\right)^{(p-1)/3},\ \ c_* = \min\left\{\frac{\lambda}{4},\ \frac{\tau}{128L^2}\right\}, \end{aligned} \tag{4.7}$$

where $K_p > 0$ is a sufficiently large constant depending only on p. For every $\mu \in \mathcal{S}_{F_0}$ with $\underline{c}_\mu \le c_*$, the tube $\mathcal{T}_L$ satisfies

$$U_{q_\mu}(\theta) \le 2c_* L^2 + H\delta^2 \le \frac{\tau}{16}, \qquad \theta \in \mathcal{T}_L. \tag{4.8}$$

Inside the low-potential tube, we therefore have the desired Dirichlet-energy-dissipation inequality,

$$\int_{\mathcal{T}_L} \|\nabla h_\mu\|^2\,\mathrm{d}\theta \le \frac{e^{1/16}}{4\tau^2}\mathcal{D}(\mu). \tag{4.9}$$

Step 3. A lower bound on the dissipation. We next derive a lower bound on the local Dirichlet energy within the tube $\mathcal{T}_L$. For $b_0/2 \le a < b \le L$, we define the tube segment as

$$E(a,b) = \{su_0 + z : a < s < b,\ z \perp u_0,\ \|z\| < \delta\}.$$

Since the Dirichlet energy controls the variation of h_μ along the tube, a discrepancy between the probability mass near the entrance and that in the terminal segment yields a lower bound on the Dirichlet energy. We first relate these two masses to the local Dirichlet energy, and then bound the entrance mass from below and the terminal mass from above.

Lemma 4 (Local Dirichlet energy bound from mass comparison)**.** *Let b_0, δ, L, c_* be chosen as in Lemma 3. Suppose that $\mu \in \mathcal{S}_{F_0}$ satisfies $\underline{c}_\mu \le c_*$ and $\mathcal{D}(\mu) < \infty$. There exist constants $b_p, B_p > 0$, depending only on p, such that*

$$B_p L^2 \int_{\mathcal{T}_L} \|\nabla h_\mu\|^2 \,\mathrm{d}\theta \ge b_p \frac{L}{\delta} \mu\left(E(\tfrac{b_0}{2}, \tfrac{b_0}{2} + \delta)\right) - \mu\left(E(\tfrac{L}{2}, L)\right). \tag{4.10}$$

Appendix D.7 establishes the following entrance segment probability bound.

Lemma 5 (Entrance probability bound)**.** *Under the assumptions of Lemma 4, there exist constants $c_p, C_p > 0$, depending only on p, such that*

$$\mu\left(E(\tfrac{b_0}{2}, \tfrac{b_0}{2} + \delta)\right) \ge c_p \left(\frac{\delta}{\sqrt{M}}\right)^p - C_p \frac{M}{\tau^2} \mathcal{D}(\mu). \tag{4.11}$$

Moreover, as the second moment is uniformly bounded, the terminal segment probability is upper bounded.

$$\frac{L^2}{4} \mu\left(E(\tfrac{L}{2}, L)\right) \le \int_{E(L/2,L)} \|\theta\|^2 \,\mathrm{d}\mu \le M_2(\mu) \le M. \tag{4.12}$$

Rearranging (4.11), (4.10), and (4.12), for some constant $\kappa_p > 0$ depending only on p, we arrive at the dissipation lower bound by the Dirichlet-energy-dissipation inequality (4.9).

$$\mathcal{D}(\mu) \ge \kappa_p \frac{\tau^2}{M + \delta L} \left(\frac{\delta}{\sqrt{M}}\right)^p =: d_* > 0. \tag{4.13}$$

Finally, with $p, F_0, \lambda, R, B_m, L_{\mathrm{act}}$ fixed, $M, C_0, H \asymp 1$ as $\tau \downarrow 0$. The scale choices give

$$b_0 \asymp \delta \asymp \tau^{1/2}, \qquad L \asymp \tau^{-(p-1)/6}, \qquad c_* \asymp \tau^{(p+2)/3}. \tag{4.14}$$

Substituting into (4.13) yields

$$d_* \gtrsim \frac{\tau^2}{M} \min\left\{ \left(\frac{\delta}{\sqrt{M}}\right)^p, \left(\frac{\delta}{\sqrt{M}}\right)^{4(p-1)/3} \right\} \gtrsim \tau^{2+\max\left\{\frac{p}{2}, \frac{2(p-1)}{3}\right\}}, \tag{4.15}$$

proving (4.3).

4.2 Coercive regime: radial rescaling LSI

We now turn to the complementary regime $\min_{u \in \mathbb{S}^{p-1}} c_{q_\mu}(u) > c_*$. In this regime, two-homogeneity gives $U_{q_\mu}(ru) = r^2 c_{q_\mu}(u)$, so that each direction of the Gibbs weight $e^{-U_{q_\mu}/\tau}$ has a Gaussian radial tail with a direction-dependent quadratic coefficient. We normalize these coefficients by the radial map

$$T_{q_\mu}(ru) := r\sqrt{2c_{q_\mu}(u)}\, u, \qquad U_{q_\mu}(\theta) = \frac{1}{2}\|T_{q_\mu}(\theta)\|^2.$$

Under this transformation, the Gibbs weight becomes isotropic Gaussian, while the Jacobian contributes the angular factor $c_{q_\mu}(u)^{-p/2}$. Since $c_* \le c_{q_\mu} \le C_0$, the angular factor is uniformly bounded above and below. The Holley–Stroock criterion therefore gives a log-Sobolev inequality for the transformed law (Holley & Stroock, 1987). To recover an inequality for the proximal Gibbs distribution ν_{q_μ}, we push this estimate forward through $T_{q_\mu}^{-1}$. The lower and Lipschitz bounds on c_{q_μ} give $\mathrm{Lip}(T_{q_\mu}^{-1}) \lesssim c_*^{-1/2} + L_0 c_*^{-3/2}$, and Lipschitz stability of log-Sobolev inequalities then gives $\mathrm{LSI}(\alpha_{\mathrm{hc}})$ for ν_{q_μ}, with

$$\alpha_{\mathrm{hc}} \ge \frac{c_p}{\tau} \left(\frac{c_*}{C_0}\right)^{p/2} \left(c_*^{-1/2} + L_0 c_*^{-3/2}\right)^{-2}. \tag{4.16}$$

This log-Sobolev inequality provides the desired control of the free-energy gap. The comparison with the proximal Gibbs distribution, followed by the log-Sobolev inequality and the identity $\mathcal{D}(\mu) = \tau^2 I(\mu\|\nu_{q_\mu})$, gives

$$\mathcal{F}_{\tau,\lambda}(\mu) - \inf_{\pi \in \mathcal{P}_2(\Theta)} \mathcal{F}_{\tau,\lambda}(\pi) \le \tau\, \mathrm{KL}(\mu\|\nu_{q_\mu}) \le \frac{\tau}{2\alpha_{\mathrm{hc}}} I(\mu\|\nu_{q_\mu}) = \frac{\mathcal{D}(\mu)}{2\tau\alpha_{\mathrm{hc}}}. \tag{4.17}$$

Hence, throughout the high-coercivity regime, we have

$$\mathcal{D}(\mu) \geq 2\tau\alpha_{\mathrm{hc}} \left(\mathcal{F}_{\tau,\lambda}(\mu) - \inf_{\pi \in \mathcal{P}_2(\Theta)} \mathcal{F}_{\tau,\lambda}(\pi) \right). \tag{4.18}$$

Using the temperature scaling of c_* from Section 4.1 and the uniform bounds on C_0 and L_0, we have

$$2\tau\alpha_{\mathrm{hc}} \gtrsim \frac{c_*^{p/2+3}}{(c_* + L_0)^2} \gtrsim \tau^{\frac{(p+2)(p+6)}{6}}, \tag{4.19}$$

for sufficiently small $\tau > 0$, with the remaining parameters fixed.

4.3 PŁ INEQUALITY AND EXPONENTIAL CONVERGENCE

We now prove Theorem 1 by combining the two estimates. Since $F_{\inf} \geq 0$ and $F_{\inf} \leq \mathcal{F}_{\tau,\lambda}(\mu)$, the gap on $\mathcal{S}_{F_0}$ lies in $[0, F_0]$. Proposition 3 therefore gives the PŁ coefficient d_*/F_0 in the low-coercivity regime, while Section 4.2 gives the coefficient $2\tau\alpha_{\mathrm{hc}}$ in the coercive regime. Taking their minimum yields

$$\kappa_{F_0} = \min\left\{ \frac{d_*}{F_0}, 2\tau\alpha_{\mathrm{hc}} \right\} > 0, \tag{4.20}$$

and proves (3.9) on the whole sublevel. This argument compares objective values directly with $F_{\inf}$.

Let $(\mu_t)_{t\geq 0}$ be a global energy-dissipating weak trajectory with initial energy at most F_0. The energy-dissipation inequality keeps the trajectory in $\mathcal{S}_{F_0}$. Combining it with the PŁ inequality and using $\mathcal{F}_{\tau,\lambda}(\mu_\star) = F_{\inf}$ from Proposition 2, we obtain

$$\mathcal{F}_{\tau,\lambda}(\mu_t) - \mathcal{F}_{\tau,\lambda}(\mu_\star) + \kappa_{F_0} \int_s^t (\mathcal{F}_{\tau,\lambda}(\mu_r) - \mathcal{F}_{\tau,\lambda}(\mu_\star))\,\mathrm{d}r \leq \mathcal{F}_{\tau,\lambda}(\mu_s) - \mathcal{F}_{\tau,\lambda}(\mu_\star), \quad 0 \leq s < t. \tag{4.21}$$

The integral Gronwall inequality then yields $\mathcal{F}_{\tau,\lambda}(\mu_t) - \mathcal{F}_{\tau,\lambda}(\mu_\star) \leq e^{-\kappa_{F_0}(t-s)}(\mathcal{F}_{\tau,\lambda}(\mu_s) - \mathcal{F}_{\tau,\lambda}(\mu_\star))$, establishing (3.10) for the stated weak trajectories. Appendix F gives the complete argument.

The positive dissipation bound also quantifies the time spent at low coercivity. Integrating it and using the energy-dissipation inequality gives

$$\int_0^\infty \mathbf{1}_{\{\underline{c}_{\mu_t} \leq c_*\}}\,\mathrm{d}t \leq \frac{\mathcal{F}_{\tau,\lambda}(\mu_0) - \mathcal{F}_{\tau,\lambda}(\mu_\star)}{d_*} \leq \frac{F_0}{d_*}. \tag{4.22}$$

Thus, the total time spent in the low-coercivity regime is bounded by F_0/d_*, counting all visits to this regime. Finally, (4.15) and (4.19) give the temperature dependence. The smaller of the two stated rate lower bounds satisfies

$$\kappa_{F_0}(\tau) \geq c\tau^{(p+2)(p+6)/6}, \qquad 0 < \tau \leq \tau_0, \tag{4.23}$$

for constants $c, \tau_0 > 0$ independent of τ when d and the remaining problem parameters are fixed.

5 CONCLUDING REMARKS

We establish global exponential free-energy convergence for mean-field Langevin dynamics of jointly trained, biased two-layer ReLU networks. Assuming bounded inputs, a bounded regression function, and the activation-indicator Lipschitz condition, a direct dissipation estimate at low coercivity and a log-Sobolev inequality in the coercive regime give a PŁ inequality relative to the free-energy infimum. Finite-energy initial laws with a finite fourth moment admit global weak trajectories with exact energy dissipation. The PŁ estimate then controls their free-energy gaps, for every positive weight-decay coefficient. Exact homogeneity gives a rate with polynomial temperature dependence at each fixed input dimension.

AI USE DISCLOSURE

We used generative AI tools to assist in editing and polishing the writing, improving presentation, discovering relevant literature, exploring research ideation, and drafting standard intermediate arguments. All AI-generated outputs were thoroughly reviewed and verified by the authors, who take full responsibility for the final content of the manuscript.

REFERENCES

Shunta Akiyama and Taiji Suzuki. On learnability via gradient method for two-layer relu neural networks in teacher-student setting. In *International Conference on Machine Learning*, 2021.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, 2019.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*. Springer, 2014.

Herm Jan Brascamp and Elliott H Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366–389, 1976.

Patrick Cattiaux, Giovanni Conforti, Ivan Gentil, and Christian Léonard. Time reversal of diffusion processes under a finite entropy condition. In *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, volume 59, pp. 1844–1881. Institut Henri Poincaré, 2023.

Zhengdao Chen, Eric Vanden-Eijnden, and Joan Bruna. On feature learning in neural networks with global convergence guarantees. In *International Conference on Learning Representations*, 2022.

Lénaïc Chizat. Mean-field Langevin dynamics: Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, 2018.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, 2020.

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

Simon S Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

Richard Holley and Daniel W Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46:1159–1194, 1987.

Kaitong Hu, Zhenjie Ren, David Šiška, and Łukasz Szpruch. Mean-field Langevin dynamics and energy landscape of neural networks. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 57(4):2043–2065, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Daniel Lacker. Hierarchies, entropy, and quantitative propagation of chaos for mean field diffusions. *Probability and mathematical physics*, 4(2):377–432, 2023.

Arvind Mahankali, Haochen Zhang, Kefan Dong, Margalit Glasgow, and Tengyu Ma. Beyond NTK with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time. In *Advances in Neural Information Processing Systems*, 2023.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, 2019.

Alireza Mousavi-Hosseini, Denny Wu, and Murat A Erdogdu. Learning multi-index models with neural networks via mean-field langevin dynamics. In *International Conference on Learning Representations*, 2025.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field Langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Atsushi Nitanda, Anzelle Lee, Damian Tan Xing Kai, Mizuki Sakaguchi, and Taiji Suzuki. Propagation of chaos for mean-field Langevin dynamics and its application to model ensemble. In *International Conference on Machine Learning*, 2025.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Lawrence E Payne and Hans F Weinberger. An optimal Poincaré inequality for convex domains. *Archive for Rational Mechanics and Analysis*, 5(1):286–292, 1960.

Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*. Springer Science & Business Media, 2013.

Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Alexander Shevchenko, Vyacheslav Kungurtsev, and Marco Mondelli. Mean-field analysis of piecewise linear solutions for wide ReLU networks. *Journal of Machine Learning Research*, 23 (130):1–55, 2022.

Taiji Suzuki, Denny Wu, and Atsushi Nitanda. Convergence of mean-field Langevin dynamics: time-space discretization, stochastic gradient, and variance reduction. In *Advances in Neural Information Processing Systems*, 2023a.

Taiji Suzuki, Denny Wu, Kazusato Oko, and Atsushi Nitanda. Feature learning via mean-field langevin dynamics: classifying sparse parities and beyond. In *Advances in Neural Information Processing Systems*, 2023b.

Shokichi Takakura and Taiji Suzuki. Mean-field analysis on two-layer neural networks from a kernel perspective. In *International Conference on Machine Learning*, 2024.

Berk Tinaz, Changzhi Xie, and Mahdi Soltanolkotabi. When both layers learn: Training dynamics of representing linear models via ReLU networks. In *Conference on Learning Theory*, 2026.

Guillaume Wang, Alireza Mousavi-Hosseini, and Lénaïc Chizat. Mean-field langevin dynamics for signed measures via a bilevel approach. In *Advances in Neural Information Processing Systems*, 2024.

Stephan Wojtowytsch. On the convergence of gradient descent training for two-layer ReLU-networks in the mean field regime. *arXiv preprint arXiv:2005.13530*, 2020.

Jingwei Zhang, Xunpeng Huang, and Jincheng Yu. Mean-field analysis of two-layer neural networks: Global optimality with linear convergence rates. *arXiv preprint arXiv:2205.09860*, 2022.

A Related work

Deterministic mean-field training. Foundational analyses derive the distributional evolution and establish asymptotic convergence or optimality under structural and initialization conditions (Mei et al., 2018; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2018). Quantitative results include linear decay of empirical loss under mean-field scaling and linearly independent inputs (Chen et al., 2022), and polynomial-time progress for a structured single-index problem (Mahankali et al., 2023). These results provide complementary guarantees for feature-learning dynamics. We study the entropy-regularized dynamics of jointly trained biased ReLU neurons under square loss.

Mean-field Langevin dynamics. Quantitative convergence of MFLD has been studied under conditions on regularization and functional inequalities for associated Gibbs distributions. Hu et al. (2021) establish convergence to the minimizer of an entropy-regularized objective and obtain exponential convergence under conditions satisfied by sufficiently strongly regularized problems. Nitanda et al. (2022) use a proximal Gibbs distribution and uniform log-Sobolev control to derive convergence rates in continuous and discrete time. Chizat (2022) proves exponential convergence under a uniform log-Sobolev inequality for an associated Gibbs family. Zhang et al. (2022) extend quantitative analyses to unbounded two-layer features by estimating log-Sobolev constants along the dynamics. Their assumptions require twice-differentiable activations, including smoothed ReLU, and a confining regularizer tailored to the growth of the features. Further work quantifies finite-particle approximation, time-discretization, and stochastic-gradient errors in general mean-field Langevin frameworks (Suzuki et al., 2023a; Nitanda et al., 2025). Takakura & Suzuki (2024) study feature learning through a two-timescale kernel formulation. A central distinction in our analysis is the control of states for which the Gibbs weight is non-normalizable under standard quadratic confinement.

Convergence for exact ReLU networks. The convergence theory of exact ReLU networks includes qualitative mean-field results, equilibrium characterizations, and quantitative guarantees for specific dynamics and data models. Wojtowytsch (2020) studies qualitative convergence to minimum Bayes risk for deterministic ReLU mean-field training under omni-directional initialization. The analysis relates risk convergence to limits of the first-variation potential and uses the activation-indicator condition (P4), which we also adopt. We study the entropy-regularized dynamics with quadratic weight decay and derive a global quantitative rate. Shevchenko et al. (2022) analyze noisy SGD for univariate regularized ReLU regression. Their analysis characterizes the predictor associated with the Gibbs-form free-energy minimizer and establishes its piecewise linear structure in a vanishing-noise limit. More recently, Chizat et al. (2026) derive an explicit polynomial local convergence rate for ReLU networks under a Wasserstein–Fisher–Rao flow on nonnegative measures over a sphere. Tinaz et al. (2026) prove linear convergence of gradient descent from suitably small random initialization when both layers of a finite-width ReLU network are trained on Gaussian inputs with a planted linear target. Our result concerns global quantitative convergence for entropy-regularized square loss with the exact biased feature $a(w^\top x + b)_+$. All parameters (a, w, b) are trained jointly with standard quadratic confinement, under assumptions stated directly on the data distribution.

B Uniform bounds on free-energy sublevels

We collect estimates on $\mathcal{S}_{F_0}$ that will be used in both the low-coercivity and log-Sobolev arguments. Throughout this appendix, Assumption 1 holds, and $\bar{R} := \sqrt{R^2 + 1}$.

Second-moment bound. For $\mu \in \mathcal{S}_{F_0}$, the entropy variational inequality for the Gaussian reference law gives

$$\frac{\lambda}{4\tau} \int_\Theta \|\theta\|^2 \, \mathrm{d}\mu \le \mathrm{KL}(\mu \| \gamma_{\tau,\lambda}) + \frac{p}{2} \log 2 \le \frac{F_0}{\tau} + \frac{p}{2} \log 2,$$

where the last inequality uses $\tau \, \mathrm{KL}(\mu \| \gamma_{\tau,\lambda}) \le \mathcal{F}_{\tau,\lambda}(\mu)$. This yields the uniform second-moment bound

$$M_2(\mu) := \int_\Theta \|\theta\|^2 \, \mathrm{d}\mu \le M := \frac{4F_0}{\lambda} + \frac{2p\tau \log 2}{\lambda}. \tag{B.1}$$

Prediction and residual bounds. For $x, x' \in \mathbb{R}^d$ with $\|x\|, \|x'\| \le R$, the ReLU feature satisfies

$$|\Phi(\theta, x)| \le \frac{\bar{R}}{2}\|\theta\|^2, \qquad |\Phi(\theta, x) - \Phi(\theta, x')| \le \frac{1}{2}\|\theta\|^2\|x - x'\|. \tag{B.2}$$

These inequalities ensure that f_μ is well defined and Lipschitz on $\{x \in \mathbb{R}^d : \|x\| \le R\}$ for every $\mu \in \mathcal{P}_2(\Theta)$. For $\mu \in \mathcal{S}_{F_0}$, the moment bound and $q_\mu = f_\mu - m$ give

$$\|f_\mu\|_\infty \le B_f, \qquad \|q_\mu\|_\infty \le Q_0, \tag{B.3}$$

where

$$B_f := \frac{\bar{R}M}{2}, \qquad Q_0 := B_f + B_m.$$

Angular coefficient bounds. For $u, u' \in \mathbb{S}^{p-1}$ and $\|x\| \le R$, we also have $|\Phi(u, x) - \Phi(u', x)| \le 2\bar{R}\|u - u'\|$. Since $c_{q_\mu}(u) = \lambda/2 + \mathbb{E}[q_\mu(X)\Phi(u, X)]$, the residual bound gives

$$|c_{q_\mu}(u)| \le C_0 := \frac{\lambda}{2} + \frac{\bar{R}Q_0}{2}, \qquad \mathrm{Lip}(c_{q_\mu}) \le L_0 := 2\bar{R}Q_0. \tag{B.4}$$

The Lipschitz constant is taken on $\mathbb{S}^{p-1}$ with the Euclidean distance.

Proposition 4 (Uniform bounds along the flow)**.** *Under Assumption 1, every global energy-dissipating weak trajectory in Definition 4 satisfying Assumption 3 obeys*

$$\mu_t \in \mathcal{S}_{F_0} \qquad \textit{for every } t \ge 0. \tag{B.5}$$

All these bounds are uniform in time. Their constants depend only on F_0, τ, λ, p, $\bar{R}$, and B_m.

Proof. The initial energy bound and the energy-dissipation inequality give

$$\mathcal{F}_{\tau,\lambda}(\mu_t) + \int_0^t \mathcal{D}(\mu_r)\,\mathrm{d}r \le \mathcal{F}_{\tau,\lambda}(\mu_0) \le F_0.$$

Since $\mathcal{D} \ge 0$, we have $\mu_t \in \mathcal{S}_{F_0}$ for every $t \ge 0$. Applying the preceding estimates to μ_t gives the uniform bounds. Their parameter dependence follows from the explicit formulas for the constants. □

C Proof of Proposition 1

We prove Proposition 1 under Assumptions 1 and 3. For notational convenience, we use the notation $\mathcal{H} = L^2(P)$ and $\gamma = \gamma_{\tau,\lambda}$. A short-time contraction constructs the self-consistent diffusion. Two-homogeneity gives global moment bounds, and finite path-space entropy supplies the current and entropy chain rule.

C.1 The diffusion with prescribed output

For $q \in \mathcal{H}$, define the Borel weak-gradient representative

$$G_q(a, v) := \begin{pmatrix} \mathbb{E}[q(X)(v^\top \bar{X})_+] \\ a\mathbb{E}[q(X)\bar{X}\mathbf{1}_{\{v^\top \bar{X} > 0\}}] \end{pmatrix}. \tag{C.1}$$

The scalar ReLU weak derivative, Fubini's theorem, and Cauchy–Schwarz give $G_q = \nabla\Psi_q$ Lebesgue-almost everywhere and

$$\begin{aligned} \|G_q(\theta)\| &\le \bar{R}\|q\|_{\mathcal{H}}\|\theta\|, \\ \|G_q(\theta) - G_{\widetilde{q}}(\theta)\| &\le \bar{R}\|q - \widetilde{q}\|_{\mathcal{H}}\|\theta\|, \\ \langle \theta, G_q(\theta)\rangle &= 2\Psi_q(\theta). \end{aligned} \tag{C.2}$$

Let $\Phi_{\mathcal{H}}(\theta) = \Phi(\theta, \cdot) \in \mathcal{H}$. The feature bounds imply

$$\begin{aligned} \|\Phi_{\mathcal{H}}(\theta)\|_{\mathcal{H}} &\le \frac{\bar{R}}{2}\|\theta\|^2, \\ \|f_\mu - f_\nu\|_{\mathcal{H}} &\le \bar{R}\sqrt{2(M_2(\mu) + M_2(\nu))}\, W_2(\mu, \nu). \end{aligned} \tag{C.3}$$

Fix $I = [s, s+h]$, $\nu \in \mathcal{P}_4(\Theta)$, and $g \in C(I; \mathcal{H})$ with $\sup_{t\in I} \|g_t\|_{\mathcal{H}} \le B$.

Consider

$$\mathrm{d}\Theta_t = -\big(\lambda\Theta_t + G_{g_t - m}(\Theta_t)\big)\,\mathrm{d}t + \sqrt{2\tau}\,\mathrm{d}W_t, \qquad \mathrm{Law}(\Theta_s) = \nu. \tag{C.4}$$

Here W is standard Brownian motion in $\mathbb{R}^p$. Continuity in q in (C.2) and Borel measurability in θ make the drift jointly Borel. It has linear growth with coefficient $K_B := \lambda + \bar{R}(B + \|m\|_{\mathcal{H}})$. A parameter cutoff gives a bounded perturbation of the Ornstein–Uhlenbeck drift. Girsanov's theorem in both directions gives weak existence and uniqueness for the cutoff equation (Revuz & Yor, 2013, Chapter VIII, §1). The pathwise bound

$$\sup_{s\le t\le s+h} \|\Theta_t\| \le e^{K_B h}\left(\|\Theta_s\| + \sqrt{2\tau}\sup_{s\le t\le s+h}\|W_t - W_s\|\right) \tag{C.5}$$

follows from Gronwall's inequality and is uniform in the cutoff. It gives nonexplosion and a finite fourth moment of the path supremum. Uniqueness before exit from each ball identifies the stopped laws and removes the cutoff. The same construction works from every starting time and point. Conditioning and weak uniqueness show that the resulting law is time-inhomogeneous Markov and that the prescribed-output equation is well posed in the sense of Lacker (2023, Definition 2.1).

C.2 A CONTRACTION IN NETWORK-OUTPUT SPACE

For $j = 2, 4$, put $M_{j,0} = M_j(\nu) = \int_\Theta \|\theta\|^j \,\mathrm{d}\nu(\theta)$ and $M_{j,c} = M_{j,0} + 1$, and choose $B = \bar{R}M_{2,c}/2$. On the set of admissible output paths

$$\mathcal{K} = \{g \in C(I; \mathcal{H}) : g_s = f_\nu,\ \sup_{t\in I}\|g_t\|_{\mathcal{H}} \le B\},$$

define $(\mathcal{T}g)_t = f_{\mu_t^g}$, where μ_t^g is the marginal law of (C.4). Itô's formula and (C.2) give

$$\frac{\mathrm{d}}{\mathrm{d}t}M_2(\mu_t^g) \le 2K_B M_2(\mu_t^g) + 2\tau p, \qquad \frac{\mathrm{d}}{\mathrm{d}t}M_4(\mu_t^g) \le 4K_B M_4(\mu_t^g) + 4\tau(p+2)M_2(\mu_t^g). \tag{C.6}$$

Stopping and (C.5) justify these estimates. Choose $0 < h \le 1$ such that

$$\begin{aligned} e^{2K_B h}(M_{2,0} + 2\tau p h) &\le M_{2,c}, \\ e^{4K_B h}(M_{4,0} + 4\tau(p+2)M_{2,c}h) &\le M_{4,c}. \end{aligned} \tag{C.7}$$

Then $M_j(\mu_t^g) \le M_{j,c}$ for $j = 2, 4$ and $t \in I$, and $\mathcal{T}(\mathcal{K}) \subset \mathcal{K}$. Continuity in time follows from the SDE increment estimate and (C.3).

For any two laws $\eta, \widetilde{\eta}$ with fourth moments at most $M_{4,c}$, a weighted Hellinger estimate yields

$$\|f_\eta - f_{\widetilde{\eta}}\|_{\mathcal{H}}^2 \le \bar{R}^2 M_{4,c}\,\mathrm{KL}(\eta\|\widetilde{\eta}). \tag{C.8}$$

Indeed, relative to a common dominating measure, factor the density difference as $(\sqrt{r} - \sqrt{\widetilde{r}})(\sqrt{r} + \sqrt{\widetilde{r}})$ and use Cauchy–Schwarz. The squared weighted factor is bounded by $2\int \|\Phi_{\mathcal{H}}\|_{\mathcal{H}}^2\,\mathrm{d}(\eta + \widetilde{\eta}) \le \bar{R}^2 M_{4,c}$; the squared Hellinger distance is at most $\mathrm{KL}(\eta\|\widetilde{\eta})$.

Let $\mathbb{P}^g_{[s,t]}$ be the prescribed-output path law on $[s, t]$. We use the notation KL for relative entropy on path space. The diffusion entropy estimate of Lacker (2023, Lemma 4.4(i) and Remark 4.5) gives

$$\mathrm{KL}\Big(\mathbb{P}^g_{[s,t]}\,\Big\|\,\mathbb{P}^{\widetilde{g}}_{[s,t]}\Big) \le \frac{\bar{R}^2 M_{2,c}}{4\tau}\int_s^t \|g_u - \widetilde{g}_u\|_{\mathcal{H}}^2\,\mathrm{d}u. \tag{C.9}$$

Both equations are weakly well posed and their drift difference has linear growth, so its squared time integral is finite on every continuous path, as required by that estimate. Data processing and (C.8) give

$$\|(\mathcal{T}g)_t - (\mathcal{T}\widetilde{g})_t\|_{\mathcal{H}}^2 \le \frac{\bar{R}^4 M_{2,c}M_{4,c}}{4\tau}\int_s^t \|g_u - \widetilde{g}_u\|_{\mathcal{H}}^2\,\mathrm{d}u. \tag{C.10}$$

Reducing h so that $h\bar{R}^4 M_{2,c}M_{4,c}/(4\tau) \le 1/2$ makes $\mathcal{T}$ a contraction on the complete set $\mathcal{K}$. Banach's fixed-point theorem yields a local self-consistent diffusion.

C.3 GLOBAL CONTINUATION

At a fixed point, two-homogeneity gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}M_2(\mu_t) &= 2\tau p - 2\lambda M_2(\mu_t) - 4\langle f_{\mu_t} - m, f_{\mu_t}\rangle_{\mathcal{H}} \\ &\le 2\tau p + \|m\|_{\mathcal{H}}^2 - 2\lambda M_2(\mu_t).\end{aligned} \tag{C.11}$$

The inequality uses $-4\langle f - m, f\rangle = -4\|f - m/2\|_{\mathcal{H}}^2 + \|m\|_{\mathcal{H}}^2$. It proves (3.7) before any energy identity is invoked. Set

$$M_* := \max\left\{M_2(\mu_0), \frac{2\tau p + \|m\|_{\mathcal{H}}^2}{2\lambda}\right\}, \qquad Q_* := \frac{\bar{R}M_*}{2} + B_m, \qquad K_* := \lambda + \bar{R}Q_*.$$

Writing $q_t = f_{\mu_t} - m$, we have $M_2(\mu_t) \le M_*$ and $\|q_t\|_\infty \le Q_*$ on the constructed intervals. The fourth-moment estimate gives

$$M_4(\mu_t) \le e^{4K_* t}\big(M_4(\mu_0) + 4\tau(p+2)M_* t\big). \tag{C.12}$$

On every fixed finite horizon these bounds give a common positive restart time in (C.7) and (C.10). Successive local solutions therefore produce a global W_2-continuous curve with locally bounded fourth moment. Its output $g_t = f_{\mu_t}$ is deterministic and continuous. Solving (C.4) once more with this output on $[0, T]$ and using weak uniqueness interval by interval identifies the constructed marginals with a single Markov path law $\mathbb{P}_T$.

C.4 RELATIVE ENTROPY AND THE PROBABILITY CURRENT

Let $\mathbb{R}_T$ be stationary Ornstein–Uhlenbeck on $[0, T]$, with invariant law γ and diffusion matrix $2\tau I_p$. The preceding estimates imply

$$\int_0^T \int_\Theta \|G_{q_t}\|^2 \,\mathrm{d}\mu_t \,\mathrm{d}t \le \bar{R}^2 Q_*^2 M_* T. \tag{C.13}$$

Applying Lacker (2023, Lemma 4.4(i) and Remark 4.5) with this reference gives

$$\mathrm{KL}(\mathbb{P}_T \,\|\, \mathbb{R}_T) \le \mathrm{KL}(\mu_0\|\gamma) + \frac{1}{4\tau}\int_0^T \int_\Theta \|G_{q_t}\|^2 \,\mathrm{d}\mu_t \,\mathrm{d}t < \infty. \tag{C.14}$$

The drift difference has linear growth under both path laws. Data processing gives $\mathrm{KL}(\mu_t\|\gamma) < \infty$ and hence a density for each $t \in [0, T]$.

The smooth quadratic reference potential and the constant positive diffusion matrix satisfy Hypotheses 1.10 of Cattiaux et al. (2023). Their Propositions 4.6 and 4.8 and Theorem 4.9 apply to the Markov law $\mathbb{P}_T$ with finite entropy. For $r_t = \mathrm{d}\mu_t/\mathrm{d}\gamma$, they give

$$\int_0^T \int_\Theta \|\nabla \log r_t\|^2 \,\mathrm{d}\mu_t \,\mathrm{d}t < \infty, \qquad \xi_t = G_{q_t} + \tau \nabla \log r_t. \tag{C.15}$$

Here the relative score is measurable and satisfies $\nabla r_t = r_t \nabla \log r_t$ distributionally for almost every t. Since $\tau\nabla\gamma = -\lambda\theta\gamma$, these identities give $J_{\mu_t} = \xi_t \mu_t$, $\xi \in L^2(\mathrm{d}t\,\mathrm{d}\mu_t)$, and (3.5). The forward and backward relative drifts are

$$\beta_t^+ = -\frac{G_{q_t}}{2\tau}, \qquad \beta_t^- = \nabla \log r_t + \frac{G_{q_t}}{2\tau}.$$

On each interval $[s, t]$, Proposition 4.6 of the cited work equates their entropy decompositions. Subtracting the corresponding squared drifts, with diffusion matrix $2\tau I_p$, gives

$$\tau\big[\mathrm{KL}(\mu_t\|\gamma) - \mathrm{KL}(\mu_s\|\gamma)\big] = -\int_s^t \int_\Theta \tau \nabla \log r_u \cdot \xi_u \,\mathrm{d}\mu_u \,\mathrm{d}u. \tag{C.16}$$

C.5 THE SQUARE-LOSS CHAIN RULE AND COMPLETION

For a random input $X \sim P$, the weak parameter derivative satisfies the following almost surely:

$$\nabla_\theta \Phi((a,v),x) = \begin{pmatrix} (v^\top \bar{x})_+ \\ a\bar{x}\mathbf{1}_{\{v^\top \bar{x}>0\}} \end{pmatrix}, \qquad \|\nabla_\theta \Phi(\theta,x)\| \le \bar{R}\|\theta\| \quad \text{for almost every } \theta. \tag{C.17}$$

Test the continuity equation with a mollification of $\chi_L(\theta)\Phi(\theta,x)$, where χ_L is a smooth cutoff equal to one on B_L and zero outside B_{2L}, with $\|\nabla \chi_L\|_\infty \le C/L$. Absolute continuity of μ_u and (C.15) permit removal of the mollification. The gradient error from subsequently removing the cutoff is at most $C\bar{R}\|\theta\|\mathbf{1}_{\{\|\theta\|>L\}}$. Its pairing with the current tends to zero by Cauchy–Schwarz:

$$\int_s^t \int_{\{\|\theta\|>L\}} \|\theta\|\|\xi_u\| \,\mathrm{d}\mu_u \,\mathrm{d}u \le \left(\int_s^t \int_{\{\|\theta\|>L\}} \|\theta\|^2 \,\mathrm{d}\mu_u \,\mathrm{d}u\right)^{1/2} \left(\int_s^t \int \|\xi_u\|^2 \,\mathrm{d}\mu_u \,\mathrm{d}u\right)^{1/2} \longrightarrow 0.$$

The endpoint errors vanish by finite second moments. Define

$$k_u(x) = -\int_\Theta \nabla_\theta \Phi(\theta,x)\cdot \xi_u(\theta)\,\mathrm{d}\mu_u(\theta).$$

The integrand is jointly measurable, and $\|k_u\|_{\mathcal{H}} \le \bar{R}\sqrt{M_*}\|\xi_u\|_{L^2(\mu_u)}$. Thus $k \in L^1([s,t];\mathcal{H})$. Pairing the preceding tests with a countable dense subset of the separable space $\mathcal{H}$, using Fubini, and passing through the same bounds yields

$$f_{\mu_t} - f_{\mu_s} = \int_s^t k_u \,\mathrm{d}u \quad \text{in } \mathcal{H}. \tag{C.18}$$

The Hilbert-space chain rule therefore gives

$$\frac{1}{2}\|f_{\mu_t} - m\|_{\mathcal{H}}^2 - \frac{1}{2}\|f_{\mu_s} - m\|_{\mathcal{H}}^2 = -\int_s^t \int_\Theta G_{q_u}\cdot \xi_u \,\mathrm{d}\mu_u \,\mathrm{d}u. \tag{C.19}$$

Adding (C.16) proves the energy identity, since $G_{q_u} + \tau \nabla \log r_u = \xi_u$.

Finally, for any finite-energy law, duality gives

$$\mathcal{D}(\mu) = \sup_{\zeta \in C_c^\infty(\Theta;\Theta)} \left\{ -2\tau \int \nabla\cdot\zeta \,\mathrm{d}\mu + 2\int (\lambda\theta + G_{q_\mu})\cdot\zeta \,\mathrm{d}\mu - \int \|\zeta\|^2 \,\mathrm{d}\mu \right\}. \tag{C.20}$$

Finiteness of this supremum makes the current a bounded linear functional on $L^2(\mu)$ by scaling the test field; Riesz representation identifies its density. Conversely, completing the square gives the formula. Take the supremum over a countable family dense in C^1 on each compact support. Each expression is Borel along the trajectory, because q_t is continuous in $\mathcal{H}$ and G_{q_t} is jointly Borel and locally bounded. Hence $t \mapsto \mathcal{D}(\mu_t)$ is measurable and equals $\int \|\xi_t\|^2 \,\mathrm{d}\mu_t$ almost everywhere. This completes Proposition 1.

D PROOF OF THE LOW-COERCIVITY ESTIMATE

We prove a positive lower bound on dissipation when the angular coercivity margin is small. The proof uses the probability current and a second-moment bound, and does not require a normalized associated Gibbs distribution. Throughout this appendix, Assumptions 1 and 2 hold. We use the constants M, B_f, Q_0, C_0, L_0 from Appendix B.

We use the same tube parameters b_0, δ, L, and c_* as in Section 4.1. Section D.1 proves Lemma 2, and Section D.3 proves Lemma 1. The entrance estimate proved in Appendix D.7 proves Lemma 5. Lemma 9 proves the main-text local Dirichlet-energy bound from mass comparison. We combine these two mass estimates in Section D.6 to obtain the dissipation lower bound.

D.1 BALANCE AND THE LOCAL QUADRATIC UPPER BOUND

Set $\eta_0 = \lambda/4$ and

$$\chi_u(z) = \frac{u+z}{\sqrt{1+\|z\|^2}}, \qquad u \in \mathbb{S}^{p-1}, \quad z \in u^\perp. \tag{D.1}$$

We first establish the following proposition, from which Lemma 2 follows.

Proposition 5 (Local quadratic upper bound at a balanced minimizer). *For every $\mu \in \mathcal{S}_{F_0}$ and every minimizer $u_0 = (\alpha_0, v_0) \in \arg\min_{u \in \mathbb{S}^{p-1}} c_{q_\mu}(u)$ with $c_{q_\mu}(u_0) \le \eta_0$, we have $|\alpha_0| = \|v_0\| = 1/\sqrt{2}$ and*

$$c_{q_\mu}(\chi_{u_0}(z)) \le c_{q_\mu}(u_0) + \frac{H}{2}\|z\|^2, \qquad z \perp u_0, \quad \|z\| \le r_0, \tag{D.2}$$

where

$$r_0 := \frac{1}{8}, \qquad H := 4Q_0(\bar{R} + L_{\mathrm{act}}). \tag{D.3}$$

Proof. Write $q = q_\mu$, $H_q(v) = \mathbb{E}[q(X)(v^\top \bar{X})_+]$, and $A_q = \max_{\omega \in \mathbb{S}^d} |H_q(\omega)|$. For $\alpha^2 + \|v\|^2 = 1$, one-homogeneity gives $\alpha H_q(v) \ge -|\alpha|\|v\|A_q \ge -A_q/2$. Both bounds are attained by a direction maximizing $|H_q|$ and a choice of output sign. Therefore we have $\min_{\mathbb{S}^{p-1}} c_q = \frac{\lambda}{2} - \frac{A_q}{2}$. The low-coercivity condition implies $A_q > 0$, so equality forces $|\alpha_0| = \|v_0\| = 1/\sqrt{2}$.

Assumption 2 implies that H_q is continuously differentiable away from zero, with

$$\nabla H_q(v) = \mathbb{E}[q(X)\bar{X}\mathbf{1}_{\{v^\top \bar{X} > 0\}}].$$

For $\|v\|, \|v'\| \ge 1/2$, its degree-zero gradient satisfies

$$\|\nabla H_q(v) - \nabla H_q(v')\| \le Q_0 L_{\mathrm{act}} \left\| \frac{v}{\|v\|} - \frac{v'}{\|v'\|} \right\| \le 4Q_0 L_{\mathrm{act}} \|v - v'\|. \tag{D.4}$$

If $z = (h, k)$ and $\|z\| \le 1/8$, every point on the segment from v_0 to $v_0 + k$ has norm at least $1/2$. Integrating (D.4) along this segment and expanding $\Psi_q(\alpha, v) = \alpha H_q(v)$ gives

$$|\Psi_q(u_0 + z) - \Psi_q(u_0) - \langle \nabla \Psi_q(u_0), z \rangle| \le Q_0 \left(2L_{\mathrm{act}} + \frac{\bar{R}}{2} \right) \|z\|^2. \tag{D.5}$$

Here $|H_q(v_0 + k) - H_q(v_0)| \le \bar{R} Q_0 \|k\|$ bounds the cross term $h[H_q(v_0 + k) - H_q(v_0)]$. For $z \perp u_0$ the tangent derivative vanishes, and two-homogeneity implies

$$c_q(\chi_{u_0}(z)) - c_q(u_0) = \frac{\Psi_q(u_0 + z) - \Psi_q(u_0) - \Psi_q(u_0)\|z\|^2}{1 + \|z\|^2}.$$

Since $|\Psi_q(u_0)| \le \bar{R} Q_0/2$, the last two displays prove (D.2) with (D.3). □

Now, we prove Lemma 2.

Proof of Lemma 2. Let $s > 0$ and $z \perp u_0$ satisfy $\|z\| \le r_0 s$. Since $c_{q_\mu}(u_0) \le c_* \le \eta_0$, Proposition 5 applies. By two-homogeneity and (D.2),

$$\begin{aligned} U_{q_\mu}(s u_0 + z) &= (s^2 + \|z\|^2)\, c_{q_\mu}(\chi_{u_0}(z/s)) \\ &\le (s^2 + \|z\|^2) \left(c_* + \frac{H\|z\|^2}{2s^2} \right) \\ &\le 2c_* s^2 + H\|z\|^2. \end{aligned}$$

The last inequality uses $c_* > 0$ and $\|z\| \le r_0 s \le s$. □

D.2 Tube scales and the dissipation estimate

For the tube construction, we use the radius $S := 2\sqrt{M}$. Set

$$\widehat{H} := H + C_0, b_0 := \frac{1}{64} \min\left\{ S, \sqrt{\frac{\tau}{C_0}} \right\}, \delta := \frac{1}{128} \min\left\{ b_0, b_0 r_0, \sqrt{\frac{\tau}{\widehat{H}}} \right\}, \beta := \left(\frac{\delta}{S} \right)^p. \tag{D.6}$$

The radius δ controls the width of the tube, while β is the volume ratio used in the entrance estimate. For a sufficiently large constant $\widehat{K}_p = 2^{-p/3} K_p$ depending only on p, let

$$L := \widehat{K}_p \left(\frac{M\delta}{\beta} \right)^{1/3}, \qquad c_* := \min\left\{ \eta_0, \frac{\tau}{128 L^2} \right\}. \tag{D.7}$$

Since $S = 2\sqrt{M}$ and $\beta = (\delta/S)^p$, this choice gives $L = K_p\sqrt{M}\left(\frac{\sqrt{M}}{\delta}\right)^{(p-1)/3}$, which agrees with the main-text tube length. We choose $\widehat{K}_p$ so that $L \geq 4S$ and the absorption step in the final proof holds. For a sufficiently small $c_p > 0$ depending only on p, set

$$e_{\mathrm{lc}} := c_p \frac{\beta}{S^2 + 4\delta L}, \qquad d_* := 4\tau^2 e_{\mathrm{lc}}. \tag{D.8}$$

With $S = 2\sqrt{M}$ and $\beta = (\delta/S)^p$, this agrees with the main-text definition of d_* upon choosing $\kappa_p = 2^{-p}c_p$, where c_p is the coefficient fixed in the preceding display. The constants c_p, C_p in the estimates below depend only on p and may change between estimates.

Theorem 2 (Positive dissipation at low coercivity)**.** *Under Assumptions 1 and 2, every $\mu \in \mathcal{S}_{F_0}$ satisfies*

$$\min_{u\in\mathbb{S}^{p-1}} c_{q_\mu}(u) \leq c_* \quad \Longrightarrow \quad \mathcal{D}(\mu) \geq d_*. \tag{D.9}$$

Moreover,

$$d_* \geq \frac{c_p\tau^2}{M} \min\left\{ \left(\frac{\delta}{S}\right)^p, \left(\frac{\delta}{S}\right)^{4(p-1)/3} \right\}. \tag{D.10}$$

All scales above use the common constants from Appendix B and the explicit local constants (D.3). They therefore depend only on $F_0, \tau, \lambda, p, \bar{R}, B_m, L_{\mathrm{act}}$. For fixed values of these parameters other than τ, the reciprocal of the lower bound in (D.10) grows at most polynomially in τ^{-1} as $\tau \downarrow 0$.

D.3 Probability current and weighted Dirichlet energy

The conclusion of Theorem 2 is immediate when $\mathcal{D}(\mu) = +\infty$. For the remainder of the proof, fix $\mu = \rho\,\mathrm{d}\theta \in \mathcal{S}_{F_0}$ with $\mathcal{D}(\mu) < \infty$ and $\min c_{q_\mu} \leq c_*$. We abbreviate $q = q_\mu, U = U_q, c = c_q$. The transformation $w = e^{-U/\tau}, h = \sqrt{\rho e^{U/\tau}}$ gives $\rho = h^2 w$. Here $h = h_\mu$, with h_μ as defined in Definition 5. The weight w is used without normalization.

Lemma 6 (Dirichlet energy representation of dissipation)**.** *For any finite-free-energy μ with $\mathcal{D}(\mu) < \infty$, the Gibbs-relative square-root density $h = \sqrt{\rho e^{U_{q_\mu}/\tau}}$ satisfies $h \in H^1_{\mathrm{loc}}(\Theta)$ and*

$$\nabla h = \frac{h}{2\tau}\xi_\mu \qquad \textit{almost everywhere}. \tag{D.11}$$

Consequently,

$$\mathcal{E}(h) := \int_\Theta \|\nabla h\|^2 w\,\mathrm{d}\theta = \frac{\mathcal{D}(\mu)}{4\tau^2}. \tag{D.12}$$

Proof. Let $g = \rho e^{U/\tau}$. Since $U \in W^{1,\infty}_{\mathrm{loc}}(\Theta)$, the probability-current identity gives

$$\nabla\rho = \frac{\rho}{\tau}(\xi_\mu - \nabla U) \in L^1_{\mathrm{loc}}(\Theta).$$

The weak product rule then yields

$$\nabla g = \frac{1}{\tau} g\xi_\mu$$

in distributions. Both g and $g\|\xi_\mu\|$ are locally integrable, because $e^{U/\tau}$ is locally bounded and $\xi_\mu \in L^2(\mu)$. Thus $g \in W^{1,1}_{\mathrm{loc}}(\Theta)$. For $\varepsilon > 0$, the Sobolev chain rule gives

$$\nabla\sqrt{g+\varepsilon} = \frac{g}{2\tau\sqrt{g+\varepsilon}}\xi_\mu.$$

On each compact set $K \subset \Theta$,

$$\int_K \left\|\nabla\sqrt{g+\varepsilon}\right\|^2 w\,\mathrm{d}\theta \leq \frac{1}{4\tau^2}\int_K \|\xi_\mu\|^2\,\mathrm{d}\mu. \tag{D.13}$$

The weight is bounded above and below on compact sets. Weak compactness in H^1, together with $\sqrt{g+\varepsilon} \to \sqrt{g}$ in L^2_{loc}, shows that $h = \sqrt{g} \in H^1_{\mathrm{loc}}(\Theta)$. On $\{g > 0\}$, the chain rule identifies the gradient as $h\xi_\mu/(2\tau)$. On $\{g = 0\}$, the gradient of the nonnegative Sobolev function h vanishes almost everywhere. This proves (D.11). Squaring the identity, multiplying by w, and integrating gives (D.12). □

This proof uses finite dissipation and local Lipschitz regularity of U_{q_μ}, and applies at every coercivity level. It proves Lemma 1 as well.

D.4 Mass at the tube entrance

Choose a minimizer u_0 as in Proposition 5. Recall that the entrance segment is $E_0 := E\left(\frac{b_0}{2}, \frac{b_0}{2} + \delta\right)$.

Lemma 7 (Entrance mass)**.** *The entrance segment satisfies*

$$\mu(E_0) \geq c_p\beta - C_p S^2 \mathcal{E}(h). \tag{D.14}$$

The proof is given in Appendix D.7. It uses the moment bound to place at least $3/4$ of the mass in $\{\theta : \|\theta\| \leq S\}$, then combines the weighted radial Hardy-type inequality in Lemma 10 with one-dimensional and local Poincaré inequalities to transfer mass into E_0. The resulting contribution is the volume ratio $\beta = (\delta/S)^p$. The local comparisons near the entrance take place where $|U|$ is bounded by a numerical multiple of τ. With $S = 2\sqrt{M}$ and (D.12), this is the entrance probability bound of Lemma 5, after adjusting constants depending only on p.

D.5 Estimates along the tube

Recall the tube $\mathcal{T}_L$ defined in Section 4.1:

$$\mathcal{T}_L := \left\{ tu_0 + z : \frac{b_0}{2} < t < L, \quad z \in u_0^\perp, \quad \|z\| < \delta \right\}. \tag{D.15}$$

The terminal segment is $E_1 := E\left(\frac{L}{2}, L\right)$.

Lemma 8 (Potential along the tube)**.** *For every $\theta \in \mathcal{T}_L$, we have $U(\theta) \leq \frac{\tau}{16}$. In particular, $w \geq e^{-1/16}$ on $\mathcal{T}_L$.*

Proof. Write $\theta = tu_0 + z$. Since $t > b_0/2$ and $\delta \leq b_0 r_0/128$, we have

$$\frac{\|z\|}{t} \leq \frac{r_0}{64}, \qquad \frac{\theta}{\|\theta\|} = \chi_{u_0}\left(\frac{z}{t}\right).$$

Proposition 5 gives

$$c\left(\frac{\theta}{\|\theta\|}\right) \leq c(u_0) + \frac{H\|z\|^2}{2t^2}.$$

Using two-homogeneity, $c(u_0) \leq c_*$, and $\delta \leq t < L$, we obtain

$$\begin{aligned} U(\theta) &\leq c_*(t^2 + \|z\|^2) + \frac{H}{2}(t^2 + \|z\|^2)\frac{\|z\|^2}{t^2} \\ &\leq 2c_* L^2 + H\delta^2 \leq \frac{\tau}{64} + \frac{\tau}{16384} < \frac{\tau}{16}. \end{aligned} \tag{D.16}$$

The last inequality follows from the definitions of c_* and δ. □

Lemma 9 (Local Dirichlet energy bound from mass comparison)**.** *If $L \geq 4S$, then we have*

$$B_p L^2 \int_{\mathcal{T}_L} \|\nabla h\|^2 \, \mathrm{d}\theta \geq b_p \frac{L}{\delta}\mu(E_0) - \mu(E_1), \tag{D.17}$$

where $b_p = e^{-1/8}/4$ and $B_p = e^{-1/16}/2$.

Proof. The bounds $b_0 \leq S/64$, $\delta \leq b_0/128$, and $L \geq 4S$ imply $b_0/2 + \delta < L/2$. Set

$$Z_\delta := \{z \in u_0^\perp : \|z\| < \delta\}, \qquad I_0 := (b_0/2, b_0/2 + \delta), \qquad I_1 := (L/2, L).$$

By Lemma 6, $h \in H^1_{\mathrm{loc}}(\Theta)$, and hence $h \in H^1(\mathcal{T}_L)$. Sobolev slicing gives an absolutely continuous representative of $r \mapsto h(ru_0 + z)$ for almost every $z \in Z_\delta$. For such z, $s \in I_0$, and $t \in I_1$, the fundamental theorem of calculus and Cauchy–Schwarz inequality give

$$\begin{aligned} |h(su_0 + z)|^2 &\leq 2|h(tu_0 + z)|^2 + 2\left|\int_s^t \langle \nabla h(ru_0 + z), u_0\rangle \, \mathrm{d}r\right|^2 \\ &\leq 2|h(tu_0 + z)|^2 + 2L\int_{b_0/2}^L \|\nabla h(ru_0 + z)\|^2 \, \mathrm{d}r. \end{aligned}$$

We integrate the preceding inequality over $(s,t,z) \in I_0 \times I_1 \times Z_\delta$ and divide by $|I_1| = L/2$. Since $|I_0| = \delta$ and the coordinate map $(r,z) \mapsto ru_0 + z$ has unit Jacobian, we obtain

$$\int_{E_0} h^2 \,\mathrm{d}\theta \le \frac{4\delta}{L} \int_{E_1} h^2 \,\mathrm{d}\theta + 2\delta L \int_{\mathcal{T}_L} \|\nabla h\|^2 \,\mathrm{d}\theta.$$

Since $\delta \le b_0/128$, we have $E_0 \subset B_{b_0}$. Two-homogeneity, $\|c\|_\infty \le C_0$, and the choice of b_0 give the following

$$|U(\theta)| \le C_0 b_0^2 \le \frac{\tau}{4096} \le \frac{\tau}{16}, \qquad \theta \in E_0.$$

Lemma 8 also gives $U \le \tau/16$ on E_1. Using $h^2 \,\mathrm{d}\theta = e^{U/\tau} \,\mathrm{d}\mu$, we have

$$\int_{E_0} h^2 \,\mathrm{d}\theta \ge e^{-1/16} \mu(E_0), \qquad \int_{E_1} h^2 \,\mathrm{d}\theta \le e^{1/16} \mu(E_1).$$

Consequently, we have

$$e^{-1/16} \mu(E_0) \le 4e^{1/16} \frac{\delta}{L} \mu(E_1) + 2\delta L \int_{\mathcal{T}_L} \|\nabla h\|^2 \,\mathrm{d}\theta.$$

Multiplying by $L/(4\delta e^{1/16})$ and rearranging proves (D.17) with the stated constants. □

D.6 Completion of the proof

Proof of Theorem 2. First, a choice of $\widehat{K}_p$ depending only on p ensures $L \ge 4S$. Indeed, $M = S^2/4$ and $\beta = (\delta/S)^p$ give

$$\frac{L}{S} = \widehat{K}_p 4^{-1/3} \left(\frac{\delta}{S}\right)^{(1-p)/3}. \tag{D.18}$$

Since $0 < \delta/S \le 1$ and $p \ge 2$, it is enough to take $\widehat{K}_p \ge 4^{4/3}$.

On E_1, we have $\|\theta\| \ge L/2$, so the second-moment bound gives

$$\frac{L^2}{4} \mu(E_1) \le \int_{E_1} \|\theta\|^2 \,\mathrm{d}\mu \le M. \tag{D.19}$$

By Lemma 8, we have

$$\int_{\mathcal{T}_L} \|\nabla h\|^2 \,\mathrm{d}\theta \le e^{1/16} \mathcal{E}(h).$$

Multiplying (D.17) by δ/L gives

$$b_p \mu(E_0) \le \frac{\delta}{L} \mu(E_1) + B_p e^{1/16} \delta L \mathcal{E}(h).$$

Combining this estimate with (D.14) and (D.19), and adjusting constants depending only on p, yields

$$c_p \beta \le C_p \frac{M\delta}{L^3} + C_p (S^2 + \delta L) \mathcal{E}(h). \tag{D.20}$$

By the definition of L, $M\delta/L^3 = \beta/\widehat{K}_p^3$. Increase $\widehat{K}_p$, still depending only on p, so that the first term on the right-hand side of (D.20) is at most $c_p\beta/2$. It follows that

$$\mathcal{E}(h) \ge c_p \frac{\beta}{S^2 + \delta L}.$$

Choose the coefficient in (D.8) no larger than the one obtained here. Then $\mathcal{E}(h) \ge e_{\mathrm{lc}}$, and (D.12) yields

$$\mathcal{D}(\mu) = 4\tau^2 \mathcal{E}(h) \ge d_*.$$

For the explicit bound, (D.18) gives

$$\frac{\delta L}{S^2} = \widehat{K}_p 4^{-1/3} \left(\frac{\delta}{S}\right)^{(4-p)/3}.$$

Using $(1+x)^{-1} \geq \frac{1}{2}\min\{1, x^{-1}\}$ for $x > 0$, we obtain

$$\frac{\beta}{S^2+\delta L} \geq \frac{c_p}{S^2}\min\left\{\left(\frac{\delta}{S}\right)^p, \left(\frac{\delta}{S}\right)^{4(p-1)/3}\right\}. \tag{D.21}$$

The definition of d_* and the inequality $S^2 + 4\delta L \leq 4(S^2+\delta L)$ give

$$d_* = 4c_p\tau^2\frac{\beta}{S^2+4\delta L} \geq c_p\tau^2\frac{\beta}{S^2+\delta L}.$$

Combining this with the preceding bound on $\beta/(S^2+\delta L)$ and using $S^2 = 4M$ proves the claimed explicit lower bound, after adjusting the constant depending only on p. □

D.7 Proof of the entrance localization estimate

We prove Lemma 7. The moment bound places most of the mass in B_S. The weighted radial Hardy-type inequality in Lemma 10 controls the mass in directions where c is strongly negative. On the remaining directions, a radial Poincaré inequality gives a lower bound on the mass in B_{b_0}. A local Poincaré inequality then transfers this bound to the entrance segment E_0.

We use the state and tube scales from Appendix D, with

$$\rho = h^2 w, \qquad w(ru) = e^{-r^2 c(u)/\tau}, \qquad \mathcal{E}(h) = \int_\Theta \|\nabla h\|^2 w\,\mathrm{d}\theta < \infty, \qquad M_2(\mu) \leq M. \tag{D.22}$$

Here $\mu = \rho\,\mathrm{d}\theta$, $h \geq 0$, and $S = 2\sqrt{M}$. The weight w is used without normalization. For $r > 0$, B_r denotes the ball of radius r centered at the origin in Θ. As in Appendix D, $p \geq 2$, $\|c\|_\infty \leq C_0$, $b_0 \leq S/64$, $C_0 b_0^2 \leq \tau/4096$, $\delta \leq b_0/128$, and $\beta = (\delta/S)^p$. The constants $c_p, C_p > 0$ depend only on p and may change from one estimate to the next.

D.7.1 Mass in strongly negative directions

Lemma 10 (Weighted radial Hardy-type inequality)**.** *Let $a_0 > 0$ and*

$$\mathrm{d}\nu_{a_0}(r) = r^{p-1}e^{a_0 r^2}\,\mathrm{d}r, \qquad r > 0. \tag{D.23}$$

Every $f \in H^1(\nu_{a_0})$ satisfies

$$\int_0^\infty |f'|^2\,\mathrm{d}\nu_{a_0} \geq pa_0\int_0^\infty f^2\,\mathrm{d}\nu_{a_0} + a_0^2\int_0^\infty r^2 f^2\,\mathrm{d}\nu_{a_0}. \tag{D.24}$$

Here $H^1(\nu_{a_0})$ consists of locally Sobolev functions with $\int_0^\infty (f^2+|f'|^2)\,\mathrm{d}\nu_{a_0} < \infty$.

Proof. First suppose that f has compact support at infinity. For $\varepsilon > 0$, expand

$$\int_\varepsilon^\infty |f'(r) + a_0 r f(r)|^2 r^{p-1}e^{a_0 r^2}\,\mathrm{d}r. \tag{D.25}$$

Integration by parts in the cross term produces the nonpositive boundary term $-a_0\varepsilon^p e^{a_0\varepsilon^2} f(\varepsilon)^2$. Discarding this term and letting $\varepsilon \downarrow 0$ gives

$$0 \leq \int_0^\infty |f'|^2\,\mathrm{d}\nu_{a_0} - pa_0\int_0^\infty f^2\,\mathrm{d}\nu_{a_0} - a_0^2\int_0^\infty r^2 f^2\,\mathrm{d}\nu_{a_0}. \tag{D.26}$$

Local Sobolev regularity justifies the integration by parts on every interval bounded away from zero.

For general $f \in H^1(\nu_{a_0})$, apply this estimate to $\chi_R f$, where $\chi_R = 1$ on $[0,R]$, $\chi_R = 0$ on $[2R,\infty)$, and $|\chi_R'| \leq C/R$. Then $\chi_R f \to f$ in $H^1(\nu_{a_0})$. Fatou's lemma handles the term containing $r^2 f^2$ and proves (D.24). □

Separate the angular directions at the threshold

$$\delta_- := \frac{\tau}{128S^2}, \qquad N := \{u \in \mathbb{S}^{p-1} : c(u) < -\delta_-\}, \qquad G := \mathbb{S}^{p-1}\setminus N. \tag{D.27}$$

The cones generated by N and G are denoted by C_N and C_G.

Lemma 11 (Mass in the negative cone)**.** *There is a constant $C_p > 0$, depending only on p, such that*

$$\mu(C_N) \le C_p S^2 \mathcal{E}(h). \tag{D.28}$$

Proof. For almost every $u \in N$, let $a_u = -c(u)/\tau > \delta_-/\tau$. Fubini's theorem gives $r \mapsto h(ru) \in H^1(\nu_{a_u})$. Applying Lemma 10 along these rays yields

$$\mathcal{E}(h) \ge \int_N \int_0^\infty |\partial_r h(ru)|^2 r^{p-1} e^{a_u r^2} \,\mathrm{d}r \,\mathrm{d}\omega(u) \tag{D.29}$$

$$\ge \frac{p\delta_-}{\tau} \int_{C_N} h^2 w \,\mathrm{d}\theta = \frac{p\delta_-}{\tau} \mu(C_N). \tag{D.30}$$

The choice $\delta_- = \tau/(128S^2)$ proves the claim. □

D.7.2 Mass in the inner ball

Two elementary estimates will convert the mass in $B_S \cap C_G$ into a lower bound on $\mu(B_{b_0})$. The first controls the mass of a subset through the variance. The second provides the required radial variance bound.

Lemma 12 (Lower bound using variance)**.** *Let ν be a probability measure and A a measurable set with $\nu(A) = \alpha > 0$. For every $f \in L^2(\nu)$, we have*

$$\int_A f^2 \,\mathrm{d}\nu \ge \frac{\alpha}{2} \int f^2 \,\mathrm{d}\nu - \frac{3}{2} \operatorname{Var}_\nu(f). \tag{D.31}$$

Proof. For $\bar{f} = \int f \,\mathrm{d}\nu$, the triangle inequality in $L^2(A, \nu)$ gives

$$\left(\int_A f^2 \,\mathrm{d}\nu \right)^{1/2} \ge \sqrt{\alpha}\, |\bar{f}| - \operatorname{Var}_\nu(f)^{1/2}. \tag{D.32}$$

Apply $(x-y)_+^2 \ge x^2/2 - y^2$ and use $\bar{f}^2 = \int f^2 \,\mathrm{d}\nu - \operatorname{Var}_\nu(f)$ and $\alpha \le 1$. □

Lemma 13 (Brascamp–Lieb inequality on an interval)**.** *Let ν be a probability measure on an interval with density proportional to e^{-V}. If $V'' \ge \kappa > 0$ in the interior, then*

$$\operatorname{Var}_\nu(f) \le \frac{1}{\kappa} \int |f'|^2 \,\mathrm{d}\nu \tag{D.33}$$

for every $f \in H^1(\nu)$. The estimate also holds at singular endpoints where the density is integrable.

Proof. On a compact subinterval with smooth positive density, the Brascamp–Lieb inequality (Brascamp & Lieb, 1976) gives $\operatorname{Var}_\nu(f) \le \int |f'|^2/V'' \,\mathrm{d}\nu \le \kappa^{-1} \int |f'|^2 \,\mathrm{d}\nu$. Apply this bound to normalized restrictions on increasing compact subintervals. The constant is independent of the restriction, and convergence of the integrals proves the estimate at integrable singular endpoints. Approximation gives the result for $f \in H^1(\nu)$. □

Lemma 14 (Mass in the inner ball)**.** *For the radius b_0 from the tube construction,*

$$\mu(B_{b_0}) \ge c_p \left(\frac{b_0}{S} \right)^p - C_p S^2 \mathcal{E}(h). \tag{D.34}$$

Proof. The second-moment bound and $S = 2\sqrt{M}$ give

$$\mu(B_S) \ge 1 - \frac{M}{S^2} = \frac{3}{4}. \tag{D.35}$$

For $u \in G$, consider the radial probability measure on $[0, S]$

$$\mathrm{d}\lambda_u(r) = r^{p-1} e^{-c(u) r^2/\tau} \,\mathrm{d}r, \qquad \nu_u = \frac{\lambda_u}{\lambda_u([0,S])}. \tag{D.36}$$

Its potential is

$$V_u(r) = \frac{c(u)}{\tau} r^2 - (p-1)\log r. \tag{D.37}$$

Since $c(u) \geq -\delta_-$ and $p \geq 2$, we have

$$V_u''(r) = \frac{2c(u)}{\tau} + \frac{p-1}{r^2} \geq \frac{1}{2S^2}, \qquad 0 < r \leq S. \tag{D.38}$$

Lemma 13 then gives

$$\mathrm{Var}_{\nu_u}(f) \leq 2S^2 \int_0^S |f'|^2 \,\mathrm{d}\nu_u. \tag{D.39}$$

We also need a lower bound on $\nu_u([0, b_0])$. If $c(u) \geq 0$, the numerator is at least $e^{-C_0 b_0^2/\tau} b_0^p/p$, and the denominator is at most S^p/p. If $-\delta_- \leq c(u) < 0$, these bounds become b_0^p/p and $e^{\delta_- S^2/\tau} S^p/p$. Using $C_0 b_0^2/\tau \leq 1/4096$ and $\delta_- S^2/\tau = 1/128$, both cases give

$$\nu_u([0, b_0]) \geq c_p \left(\frac{b_0}{S}\right)^p. \tag{D.40}$$

Apply Lemma 12 to $f(r) = h(ru)$ and $A = [0, b_0]$ under ν_u. Combine (D.39) and (D.40), multiply by $\lambda_u([0, S])$, and integrate over $u \in G$. This gives

$$\mu(B_{b_0}) \geq c_p \left(\frac{b_0}{S}\right)^p \mu(B_S \cap C_G) - C_p S^2 \mathcal{E}(h). \tag{D.41}$$

The bounds on $\mu(B_S)$ and $\mu(C_N)$ imply

$$\mu(B_S \cap C_G) \geq \frac{3}{4} - C_p S^2 \mathcal{E}(h). \tag{D.42}$$

Substitution into (D.41), together with $(b_0/S)^p \leq 1$, proves the claim. □

D.7.3 Mass at the tube entrance

Proof of Lemma 7. We transfer the mass bound from B_{b_0} to E_0. On B_{b_0}, two-homogeneity and the choice of b_0 give

$$|U(\theta)| \leq C_0 b_0^2 \leq \frac{\tau}{4096}. \tag{D.43}$$

In particular, $e^{-1/4096} \leq w \leq e^{1/4096}$ on B_{b_0}. Let ν_{b_0} be the probability measure proportional to $\mathbf{1}_{B_{b_0}} w \,\mathrm{d}\theta$. The Neumann Poincaré inequality on a ball (Payne & Weinberger, 1960), together with these bounds on w, yields

$$\mathrm{Var}_{\nu_{b_0}}(f) \leq C_p b_0^2 \int_{B_{b_0}} \|\nabla f\|^2 \,\mathrm{d}\nu_{b_0}. \tag{D.44}$$

Since $\delta \leq b_0/128$, the entrance segment $E_0 = \{tu_0 + z : b_0/2 < t < b_0/2 + \delta,\ z \perp u_0,\ \|z\| < \delta\}$ lies in B_{b_0}. The same bounds on w give

$$\nu_{b_0}(E_0) \geq c_p \left(\frac{\delta}{b_0}\right)^p. \tag{D.45}$$

Apply Lemma 12 to h under ν_{b_0}, use (D.44), and multiply by $\int_{B_{b_0}} w \,\mathrm{d}\theta$. Lemma 14 then gives

$$\mu(E_0) \geq c_p \left(\frac{\delta}{b_0}\right)^p \mu(B_{b_0}) - C_p b_0^2 \mathcal{E}(h) \tag{D.46}$$

$$\geq c_p \left(\frac{\delta}{S}\right)^p - C_p S^2 \mathcal{E}(h) \tag{D.47}$$

$$= c_p \beta - C_p S^2 \mathcal{E}(h). \tag{D.48}$$

This proves (D.14). □

E A PŁ INEQUALITY IN THE COERCIVE REGIME

When $c_{q_\mu} \geq \eta > 0$, the associated Gibbs distribution exists. We obtain a log-Sobolev bound by rescaling the radius in each direction, then combine it with the probability-current identity and the convexity of the square loss to prove a PŁ inequality. Throughout this section, $\mathcal{F}$ denotes $\mathcal{F}_{\tau,\lambda}$.

E.1 A LOG-SOBOLEV BOUND BY RADIAL RESCALING

Let $c : \mathbb{S}^{p-1} \to (0, \infty)$ satisfy

$$0 < \eta \leq c(u) \leq C_0, \qquad \mathrm{Lip}(c) \leq L_0, \tag{E.1}$$

where the Lipschitz constant uses the Euclidean distance on the sphere. The associated Gibbs law is

$$\nu_c(\mathrm{d}\theta) = \frac{1}{Z_c} \exp\left(-\frac{\|\theta\|^2 c(\theta/\|\theta\|)}{\tau}\right) \mathrm{d}\theta. \tag{E.2}$$

Lemma 15 (Log-Sobolev inequality by radial rescaling)**.** *Under* (E.1)*, the law ν_c satisfies* $\mathrm{LSI}(\alpha)$ *with*

$$\alpha = \alpha(\eta, C_0, L_0) := \frac{c_p}{\tau C_0^{p/2}} \frac{\eta^{p/2+3}}{(\eta + L_0)^2}, \tag{E.3}$$

where $c_p > 0$ is sufficiently small and depends only on p. It also satisfies

$$W_2^2(\mu, \nu_c) \leq \frac{2}{\alpha} \mathrm{KL}(\mu \| \nu_c) \qquad \textit{for every } \mu \in \mathcal{P}_2(\Theta). \tag{E.4}$$

Proof. The radial map

$$T_c(ru) := r\sqrt{2c(u)}\, u, \qquad r \geq 0, \quad u \in \mathbb{S}^{p-1}, \tag{E.5}$$

turns the exponent in (E.2) into $-\|T_c(ru)\|^2/(2\tau)$. Integration in polar coordinates gives the density of $\widetilde{\nu}_c := (T_c)_\# \nu_c$,

$$\widetilde{\nu}_c(\mathrm{d}z) = \frac{1}{\widetilde{Z}_c} c(z/\|z\|)^{-p/2} e^{-\|z\|^2/(2\tau)}\, \mathrm{d}z. \tag{E.6}$$

The angular weight lies between $C_0^{-p/2}$ and $\eta^{-p/2}$. The Gaussian log-Sobolev inequality and the Holley–Stroock criterion (Holley & Stroock, 1987; Bakry et al., 2014) give an LSI constant for $\widetilde{\nu}_c$ of at least

$$\widetilde{\alpha} := \frac{1}{\tau} \left(\frac{\eta}{C_0}\right)^{p/2}. \tag{E.7}$$

To transfer this bound back to ν_c, we estimate the Lipschitz constant of the inverse map,

$$T_c^{-1}(ru) = r\, b(u) u, \qquad b(u) := (2c(u))^{-1/2}. \tag{E.8}$$

The bounds on c imply

$$\|b\|_\infty \leq (2\eta)^{-1/2}, \qquad \mathrm{Lip}(b) \leq C L_0 \eta^{-3/2}.$$

For $x = ru$ and $y = sv$ with $r \leq s$, we have $s - r \leq \|x - y\|$ and $r\|u - v\| \leq \|x - y\|$. It follows that

$$\begin{aligned} \|T_c^{-1}(x) - T_c^{-1}(y)\| &\leq r\, \|b(u)u - b(v)v\| + (s - r)|b(v)| \\ &\leq \big(2\|b\|_\infty + \mathrm{Lip}(b)\big)\|x - y\| \\ &\leq C\big(\eta^{-1/2} + L_0 \eta^{-3/2}\big)\|x - y\|. \end{aligned} \tag{E.9}$$

Since $\nu_c = (T_c^{-1})_\# \widetilde{\nu}_c$, it inherits a log-Sobolev constant of at least $\widetilde{\alpha}/\,\mathrm{Lip}(T_c^{-1})^2$. This proves (E.3) after choosing c_p sufficiently small. The log-Sobolev inequality implies the quadratic transport bound (E.4) (Otto & Villani, 2000). □

E.2 Dissipation and relative Fisher information

The next identity allows us to apply the log-Sobolev bound to the dissipation $\mathcal{D}(\mu)$.

Lemma 16 (Dissipation and relative Fisher information)**.** *Let $\mu \in \mathcal{P}_2(\Theta)$ have finite free energy. If $\mathcal{D}(\mu) < \infty$ and ν_{q_μ} exists, then*

$$\mathcal{D}(\mu) = \tau^2 I(\mu \| \nu_{q_\mu}). \tag{E.10}$$

Proof. Let $q = q_\mu$, $w = e^{-U_q/\tau}$, and $\widehat{Z}_q = \int_\Theta w \, \mathrm{d}\theta < \infty$. Applying the identity in Lemma 6 to $h = \sqrt{\rho/w}$ gives

$$\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu_q}} = \sqrt{\widehat{Z}_q}\, h, \qquad I(\mu\|\nu_q) = 4\int_\Theta \|\nabla h\|^2 w \, \mathrm{d}\theta = \frac{\mathcal{D}(\mu)}{\tau^2}.$$

□

E.3 From relative entropy to the free-energy gap

The linear prediction map and the square loss give a comparison with an arbitrary finite-energy competitor.

Lemma 17 (Free-energy gap bound)**.** *Let $\mu \in \mathcal{P}_2(\Theta)$ have finite free energy. If ν_{q_μ} exists, then*

$$\mathcal{F}(\mu) - F_{\inf} \le \tau \, \mathrm{KL}(\mu \| \nu_{q_\mu}). \tag{E.11}$$

Proof. Fix a finite-energy $\pi \in \mathcal{P}_2(\Theta)$ and write $q = q_\mu$. Expanding the square loss gives

$$\frac{1}{2}\|f_\pi - m\|^2_{L^2(P)} = \frac{1}{2}\|f_\mu - m\|^2_{L^2(P)} + \int_\Theta \Psi_q \, \mathrm{d}(\pi - \mu) + \frac{1}{2}\|f_\pi - f_\mu\|^2_{L^2(P)}.$$

The Gibbs density satisfies $\mathrm{d}\nu_q/\mathrm{d}\gamma_{\tau,\lambda} = Z_q^{-1} e^{-\Psi_q/\tau}$. Since Ψ_q has at most quadratic growth, it is integrable under both μ and π, and their relative entropies with respect to ν_q are finite. Substituting the density gives

$$\tau\big[\mathrm{KL}(\mu\|\gamma_{\tau,\lambda}) - \mathrm{KL}(\pi\|\gamma_{\tau,\lambda})\big] = \tau\big[\mathrm{KL}(\mu\|\nu_q) - \mathrm{KL}(\pi\|\nu_q)\big] - \int_\Theta \Psi_q \, \mathrm{d}(\mu - \pi).$$

Combining the two identities yields

$$\begin{aligned} \mathcal{F}(\mu) - \mathcal{F}(\pi) &= \tau\big[\mathrm{KL}(\mu\|\nu_q) - \mathrm{KL}(\pi\|\nu_q)\big] - \frac{1}{2}\|f_\mu - f_\pi\|^2_{L^2(P)} \\ &\le \tau \, \mathrm{KL}(\mu\|\nu_q). \end{aligned} \tag{E.12}$$

Taking the supremum of the left-hand side over finite-energy π proves (E.11). □

For $\mu \in \mathcal{S}_{F_0}$ with $c_{q_\mu} \ge \eta > 0$, Appendix B supplies the constants C_0, L_0 required in Lemma 15. Combining the comparison with that lemma and the identity (E.10) gives, whenever $\mathcal{D}(\mu) < \infty$,

$$\mathcal{F}(\mu) - F_{\inf} \le \tau \, \mathrm{KL}(\mu\|\nu_{q_\mu}) \le \frac{\mathcal{D}(\mu)}{2\tau\alpha(\eta, C_0, L_0)}. \tag{E.13}$$

Thus we have

$$\mathcal{D}(\mu) \ge 2\tau\alpha(\eta, C_0, L_0)\big(\mathcal{F}(\mu) - F_{\inf}\big), \tag{E.14}$$

which also holds when $\mathcal{D}(\mu) = +\infty$. Taking $\eta = c_*$ proves (4.18).

F Proof of global exponential convergence

Throughout this appendix, Assumptions 1 and 2 hold, $\mathcal{F}$ denotes $\mathcal{F}_{\tau,\lambda}$, and $F_{\inf}$ is defined in (3.8). Fix $\tau, \lambda, F_0 > 0$. We first establish existence and uniqueness of the minimizer, then prove the global PŁ inequality and exponential decay.

F.1 Existence and uniqueness of the minimizer

Proof of Proposition 2. We first note that $\mathcal{F}$ is lower semicontinuous in W_2. The Lipschitz property of ReLU, followed by Cauchy–Schwarz under an optimal coupling, gives

$$\|f_\mu - f_\nu\|_{L^2(P)} \le \bar{R}\big(\sqrt{M_2(\mu)} + \sqrt{M_2(\nu)}\big) W_2(\mu,\nu).$$

Thus the square-loss term is continuous in W_2. Since relative entropy is narrowly lower semicontinuous, $\mathcal{F}$ is lower semicontinuous in W_2.

Let $\gamma = \gamma_{\tau,\lambda}$. Then $\gamma \in \mathcal{P}_4(\Theta)$, $f_\gamma = 0$, and $\mathcal{F}(\gamma) = \frac{1}{2}\|m\|^2_{L^2(P)} < \infty$. Choose $\widehat{F}_0 > \mathcal{F}(\gamma)$ and let $(\mu_t)_{t\ge 0}$ be the trajectory starting from γ provided by Proposition 1. For the rest of this proof, the constants c_*, d_*, C_0, L_0 are those associated with $\mathcal{S}_{\widehat{F}_0}$. By (3.6), $\mu_t \in \mathcal{S}_{\widehat{F}_0}$ for all $t \ge 0$ and

$$\int_0^\infty \mathcal{D}(\mu_t)\,\mathrm{d}t \le \mathcal{F}(\gamma) - F_{\inf} < \infty.$$

We may therefore choose $t_n \to \infty$ such that $\delta_n := \mathcal{D}(\mu_{t_n}) \to 0$. Set $\mu_n = \mu_{t_n}$. Theorem 2 implies that, for all sufficiently large n,

$$\min c_{q_{\mu_n}} > c_*,$$

since otherwise $\delta_n \ge d_* > 0$.

Discarding finitely many terms, let $\nu_n = \nu_{q_{\mu_n}}$ and $\alpha_0 = \alpha(c_*, C_0, L_0) > 0$. The log-Sobolev and transport inequalities in Lemma 15, the Fisher-information identity in Lemma 16, and the free-energy gap bound in Lemma 17 give the following:

$$W_2^2(\mu_n,\nu_n) \le \frac{2}{\alpha_0}\,\mathrm{KL}(\mu_n\|\nu_n) \le \frac{\delta_n}{\tau^2\alpha_0^2},$$
$$0 \le \mathcal{F}(\mu_n) - F_{\inf} \le \tau\,\mathrm{KL}(\mu_n\|\nu_n) \le \frac{\delta_n}{2\tau\alpha_0}.$$

Moreover, $c_* \le c_{q_{\mu_n}} \le C_0$ implies the uniform Gaussian bound

$$\frac{d\nu_n}{d\theta}(\theta) \le \left(\frac{C_0}{\pi\tau}\right)^{p/2} \exp\left(-\frac{c_*}{\tau}\|\theta\|^2\right).$$

Hence (ν_n) has uniformly bounded fourth moments and is relatively compact in W_2. Along a subsequence, $\nu_n \to \mu_\star$ in W_2 for some $\mu_\star \in \mathcal{P}_2(\Theta)$. The first estimate above gives $\mu_n \to \mu_\star$ in W_2, and the second estimate and lower semicontinuity yield

$$\mathcal{F}(\mu_\star) \le \liminf_{n\to\infty} \mathcal{F}(\mu_n) = F_{\inf}.$$

Thus $\mathcal{F}(\mu_\star) = F_{\inf}$ and $\mu_\star$ is a minimizer.

Finally, because $\mu \mapsto f_\mu$ is linear, the square loss is convex under linear mixtures of probability measures. Relative entropy is strictly convex on its effective domain. Since $\tau > 0$, $\mathcal{F}$ is strictly convex on its effective domain, and the minimizer is unique. □

F.2 A PŁ inequality on the free-energy sublevel

In the following, all sublevel-dependent constants are chosen for $\mathcal{S}_{F_0}$. Let $\alpha_{\mathrm{hc}} := \alpha(c_*, C_0, L_0) > 0$, where α is defined in (E.3) and c_* in (D.7).

Proof of the PŁ inequality in Theorem 1. Fix $\mu \in \mathcal{S}_{F_0}$. Since $\mathcal{F} \ge 0$ and $F_{\inf} \le \mathcal{F}(\mu)$,

$$0 \le \mathcal{F}(\mu) - F_{\inf} \le F_0.$$

The conclusion is immediate if $\mathcal{D}(\mu) = +\infty$, so suppose $\mathcal{D}(\mu) < \infty$.

If $\min c_{q_\mu} \le c_*$, Theorem 2 gives $\mathcal{D}(\mu) \ge d_*$, and therefore

$$\mathcal{D}(\mu) \ge \frac{d_*}{F_0}\big(\mathcal{F}(\mu) - F_{\inf}\big). \tag{F.1}$$

If $\min c_{q_\mu} > c_*$, (E.14) applies with $\eta = c_*$ and gives

$$\mathcal{D}(\mu) \geq 2\tau\alpha_{\mathrm{hc}}\big(\mathcal{F}(\mu) - F_{\mathrm{inf}}\big). \tag{F.2}$$

Combining the two cases proves

$$\mathcal{D}(\mu) \geq \kappa_{F_0}\big(\mathcal{F}(\mu) - F_{\mathrm{inf}}\big), \qquad \forall \mu \in \mathcal{S}_{F_0}, \tag{F.3}$$

where

$$\kappa_{F_0} := \min\left\{\frac{d_*}{F_0}, 2\tau\alpha_{\mathrm{hc}}\right\} > 0. \tag{F.4}$$

Since $F_{\mathrm{inf}} = \mathcal{F}(\mu_\star)$, this is (3.9). □

F.3 Exponential decay along weak trajectories

Lemma 18 (Exponential decay from an integral inequality)**.** *Let $E : [0,\infty) \to [0,\infty)$ be measurable and nonincreasing. Suppose that, for some $\kappa > 0$, the following holds*

$$E(t) + \kappa \int_s^t E(r)\,\mathrm{d}r \leq E(s), \qquad 0 \leq s < t. \tag{F.5}$$

Then

$$E(t) \leq e^{-\kappa(t-s)}E(s), \qquad 0 \leq s \leq t. \tag{F.6}$$

Proof. For $h > 0$, monotonicity gives $\int_t^{t+h} E(r)\,\mathrm{d}r \geq hE(t+h)$. Applying (F.5) on $[t, t+h]$ yields

$$(1 + \kappa h)E(t+h) \leq E(t). \tag{F.7}$$

For $s < t$, dividing $[s,t]$ into n equal intervals and iteratively using (F.7) yield

$$E(t) \leq \left(1 + \frac{\kappa(t-s)}{n}\right)^{-n} E(s). \tag{F.8}$$

Letting $n \to \infty$ proves (F.6) for $s < t$. The case $s = t$ is immediate. □

Proof of trajectory existence and (3.10). Under Assumption 3, Proposition 1 supplies a global energy-dissipating weak trajectory starting from μ_0. Let $(\mu_t)_{t\geq 0}$ be any trajectory in Definition 4 with this initial law. Define $E(t) := \mathcal{F}(\mu_t) - F_{\mathrm{inf}}$. By (3.6), E is nonnegative and nonincreasing, and $\mu_t \in \mathcal{S}_{F_0}$ for every $t \geq 0$. Combining that inequality with (F.3) gives

$$E(t) + \kappa_{F_0} \int_s^t E(r)\,\mathrm{d}r \leq E(s), \qquad 0 \leq s < t. \tag{F.9}$$

Lemma 18 now yields

$$E(t) \leq e^{-\kappa_{F_0}(t-s)}E(s), \qquad 0 \leq s \leq t. \tag{F.10}$$

Using $F_{\mathrm{inf}} = \mathcal{F}(\mu_\star)$ proves (3.10). □

F.4 Time spent below the coercivity threshold

Let $(\mu_t)_{t\geq 0}$ be a global energy-dissipating weak trajectory with initial free energy at most F_0. The time set in (4.22) is measurable. Indeed, for each fixed u, $t \mapsto c_{q_{\mu_t}}(u)$ is measurable by truncation of the continuous quadratic-growth feature integrals, and continuity in u makes the sphere minimum an infimum over a countable dense subset. The low-coercivity estimate gives, for every $T > 0$,

$$\begin{aligned} d_* \int_0^T \mathbf{1}_{\{\underline{c}_{\mu_t} \leq c_*\}}\,\mathrm{d}t &\leq \int_0^T \mathcal{D}(\mu_t)\,\mathrm{d}t \\ &\leq \mathcal{F}(\mu_0) - \mathcal{F}(\mu_T) \\ &\leq \mathcal{F}(\mu_0) - F_{\mathrm{inf}}. \end{aligned} \tag{F.11}$$

Letting $T \to \infty$ proves (4.22). The bound controls cumulative time, including repeated visits to the low-coercivity regime. The exponential decay argument uses the pointwise PŁ inequality in both regimes.

The temperature estimate (3.11) follows by tracking the constants in (F.4). The calculation is given in Appendix G.2.

G Dependence on dimension and temperature

We track how the lower bound on κ_{F_0} depends on the input dimension, temperature, and confinement strength. The fixed-confinement limit proves the low-temperature bound in Theorem 1.

G.1 Dependence on dimension

The estimates use the full parameter dimension $p = d + 2$ from (2.1). The second-moment bound $M = (4F_0 + 2p\tau \log 2)/\lambda$ and the geometry of the parameter sphere both depend on this dimension. The local radius $r_0 = 1/8$ is independent of dimension. The constant $H = 4Q_0(\bar{R} + L_{\mathrm{act}})$ depends on dimension only through the stated sublevel and data bounds.

Two geometric factors make the dimensional deterioration explicit. The entrance mass estimate contains the volume ratio $\beta = \left(\frac{\delta}{S}\right)^p$. The radial rescaling argument in Lemma 15 contains the bounded-perturbation factor $\left(\frac{\eta}{C_0}\right)^{p/2}$. For fixed ratios below one, both factors decay exponentially in p. The ball-volume ratio used in the entrance estimate satisfies

$$\frac{|B_1^{p-1}|}{|B_1^p|} = \frac{\Gamma(p/2+1)}{\sqrt{\pi}\,\Gamma((p+1)/2)} \asymp \sqrt{p}. \tag{G.1}$$

The diameter constant in the Neumann Poincaré inequality and the Gaussian log-Sobolev constant are dimension-independent.

G.2 Vanishing temperature at fixed confinement

Keep $p, F_0, \lambda, \bar{R}, B_m, L_{\mathrm{act}}$ fixed and let $\tau \downarrow 0$. The local radius is $r_0 = 1/8$, and $H = 4Q_0(\bar{R}+L_{\mathrm{act}})$ is bounded independently of τ. The constants in Appendices B and D satisfy

$$M \asymp 1, \qquad S \asymp 1, \qquad B_f, Q_0, C_0, L_0 \asymp 1, \qquad r_0, H \asymp 1. \tag{G.2}$$

For sufficiently small τ, the scales in (D.6) give

$$b_0 \asymp \tau^{1/2}, \qquad \delta \asymp \tau^{1/2}, \qquad \beta \asymp \tau^{p/2}. \tag{G.3}$$

The tube length is

$$L = \widehat{K}_p \left(\frac{M\delta}{\beta}\right)^{1/3} \asymp \tau^{-(p-1)/6}. \tag{G.4}$$

Using the threshold in (D.7),

$$c_* = \min\left\{\eta_0, \frac{\tau}{128L^2}\right\} \gtrsim \tau^{(p+2)/3}. \tag{G.5}$$

The low-coercivity estimate (D.10) gives

$$d_* \gtrsim \tau^2 \min\left\{\tau^{p/2}, \tau^{2(p-1)/3}\right\} = \tau^{2+r_p}, \tag{G.6}$$

where

$$r_p := \max\left\{\frac{p}{2}, \frac{2(p-1)}{3}\right\}. \tag{G.7}$$

For the high-coercivity regime, apply (E.3) with $\eta = c_*$. The identity $(\eta^{-1/2} + L_0\eta^{-3/2})^{-2} = \eta^3/(\eta + L_0)^2$, together with (G.2) and $c_* \le 1$ for small τ, yields

$$\alpha_{\mathrm{hc}} \gtrsim \frac{1}{\tau} c_*^{p/2+3}. \tag{G.8}$$

Substituting (G.5) gives

$$2\tau\alpha_{\mathrm{hc}} \gtrsim \tau^{A_p}, \qquad A_p := \frac{(p+2)(p+6)}{6}. \tag{G.9}$$

This exponent exceeds both low-coercivity exponents, since

$$A_p - \left(2 + \frac{p}{2}\right) = \frac{p(p+5)}{6} > 0, \tag{G.10}$$

$$A_p - \left(2 + \frac{2(p-1)}{3}\right) = \frac{(p+2)^2}{6} > 0. \tag{G.11}$$

Taking the minimum of the two rate bounds proves

$$\kappa_{F_0} = \min\left\{\frac{d_*}{F_0}, 2\tau\alpha_{\mathrm{hc}}\right\} \gtrsim \tau^{A_p}, \tag{G.12}$$

as claimed in Theorem 1. Since $p = d + 2$, the temperature exponent is $A_p = (d+4)(d+8)/6$.

G.3 Joint scaling of temperature and confinement

When $\tau = \lambda = \varepsilon$, the free energy is

$$\mathcal{F}_{\varepsilon,\varepsilon}(\mu) = \frac{1}{2}\|f_\mu - m\|^2_{L^2(P)} + \varepsilon\,\mathrm{KL}(\mu\|\mathcal{N}(0, I_p)). \tag{G.13}$$

Keep $p, F_0, \bar{R}, B_m, L_{\mathrm{act}}$ fixed and let $\varepsilon \downarrow 0$. The explicit constants satisfy

$$M \asymp \varepsilon^{-1}, \qquad S \asymp \varepsilon^{-1/2}, \qquad B_f, Q_0, C_0, L_0, H \asymp \varepsilon^{-1}, \qquad r_0 = \frac{1}{8}. \tag{G.14}$$

Both $\sqrt{\tau/C_0}$ and $\sqrt{\tau/(H + C_0)}$ are of order ε. The fixed local radius therefore gives

$$b_0 \asymp \delta \asymp \varepsilon, \qquad \frac{\delta}{S} \asymp \varepsilon^{3/2}, \qquad \beta \asymp \varepsilon^{3p/2}, \qquad L \asymp \varepsilon^{-p/2}, \qquad c_* \asymp \varepsilon^{p+1}. \tag{G.15}$$

The last relation follows from $c_* = \min\{\varepsilon/4, \varepsilon/(128L^2)\}$. The two estimates give

$$\frac{d_*}{F_0} \gtrsim \varepsilon^{3+\max\{3p/2,\, 2(p-1)\}}, \tag{G.16}$$

$$2\tau\alpha_{\mathrm{hc}} \gtrsim \varepsilon^{p/2+2} c_*^{p/2+3} \asymp \varepsilon^{(p^2+8p+10)/2}. \tag{G.17}$$

For $p \geq 2$, the exponent in (G.17) is larger than both $3 + 3p/2$ and $3 + 2(p-1)$. Taking the minimum of the two rate bounds yields

$$\kappa_{F_0}(\varepsilon) \gtrsim \varepsilon^{(p^2+8p+10)/2} = \varepsilon^{(d^2+12d+30)/2}. \tag{G.18}$$

Together with the fixed-confinement estimate, this gives polynomial upper bounds on the reciprocal guaranteed rate in both regimes.