BEYOND LINEAR AND OVERCOMPLETE REGIMES: A MEAN-FIELD ANALYSIS OF BOTTLENECK AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoencoders (AEs) learn low-dimensional representations by mapping data into a latent space while minimizing reconstruction error. Despite their empirical success, theoretical understanding remains limited and largely restricted to one of three setting: linear models, shallow non-linear models or settings without a bottleneck. In this work, we go beyond those settings and study nonlinear AEs with a fixed finite-dimensional bottleneck in the mean-field (MF) regime. We derive explicit MF learning dynamics for both encoder and decoder, providing a tractable characterization of training in the nonlinear setting. We show that, over finite time horizons, the empirical risk of finite-width networks trained with stochastic gradient descent closely tracks the MF risk trajectory with high probability. At optimality, we further establish that the finite-width risk converges to the MF optimum, demonstrating that finite networks are sufficiently expressive to approximate the infinite-width solution. As a step towards characterizing learned representations, our MF analysis shows connections between AEs, early learning of approximately linear representations, and principal curves.

1 INTRODUCTION

Representation learning is grounded in the central paradigm that the essential structure of high-dimensional data can be captured in a latent, lower-dimensional space. Classical approaches realize this idea through linear transformations, including principal component analysis (PCA), independent component analysis, factor analysis, projection pursuit, and non-negative matrix factorization [1; 2; 3]. While these methods remain widely used in practice, a common limitation is that they are inherently restricted to learning linear representations of unlabeled data. To overcome this limitation, a broad class of nonlinear methods has been developed. These include multidimensional scaling and its variants such as t-SNE [2; 4], kernel-based methods [5], and tensor factorization techniques [6]. More recently, neural network–based approaches have emerged as a dominant paradigm for representation learning, including self-organizing maps, restricted Boltzmann machines, and Autoencoders (AEs) [7; 1]. These models enable the learning of highly expressive nonlinear representations and have been successfully applied across a wide range of domains [8; 9]. And while several properties of such models have been studied empirically [10; 11; 12; 13], the increased expressivity arising from complex networks and non-nonlinearities, comes at the cost of reduced theoretical understanding [14]. This gap is especially noteworthy given the increasing deployment in system-critical applications such as healthcare [15; 16; 17].

Theoretical analysis of AEs. In this work, we take a step toward closing this gap by providing a theoretical analysis of the learning dynamics and learned representation of *nonlinear bottleneck autoencoders* (BAEs), where the latent dimension (b) is fixed and is much smaller then the data dimension (d) and both encoder and decoder are neural networks. This setting captures the essence of representation learning: compression through a nonlinear mapping. However this setting remains poorly understood due to its analytical complexity. To make progress, most prior theoretical analyses typically rely on one of the following simplifying assumptions that either removes a key aspect of representation learning or consider less expressive architectures.

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.

i) Linear autoencoders. A common simplification is to assume linear encoder and decoder mappings. In this setting, the learned representation can be directly linked to PCA [18; 19; 20; 21; 22], and a detailed theoretical understanding has been developed, including results on learning dynamics [23], regularization induced by denoising and dropout [24], and generalization [25]. However, the linearity assumption severely restricts the class of representations that can be learned, limiting these results to linear latent structures.

ii) Autoencoders without a bottleneck. Another line of work studies nonlinear autoencoders without a bottleneck constraint, i.e., in the regime $b \gg d$. While this preserves the nonlinear nature of the model, it removes the compression that is central to representation learning. In this setting, the latent representation can degenerate to a trivial embedding of the input into a higher-dimensional space. Consequently, theoretical analyses in this regime focus primarily on properties of the overall input-output mapping, including learning dynamics in the neural tangent kernel [26] and mean-field (MF) limit [27], where [27] characterizes the learned representation in absence of a fixed finite bottleneck.

iii) Shallow networks with a bottleneck and nonlinearity. This setting is characterized by a single linear encoder and decoder (which are often weight-tied) with a nonlinearity at the bottleneck[1]. For shallow networks, under additional assumptions on the data and in the high-dimensional limit, [28] derives the dynamics of representation learning and [29] describes feature learning and generalization. Under additional weight-tying assumptions, [30] provides a characterization of the risk, [31; 32] analyzes feature learning under additional data assumptions. In the regime where the input dimension scales linearly with the representation size, [33] characterizes the minimizer of the population risk. In the denoising setting, [34] derives an asymptotic characterization of low-dimensional distribution generated by the learned model. [35] proves a simplicity bias and establishes nearly sharp thresholds for the number of samples required by a simple denoiser. [36] analysis the compression of sparse data in the shallow setting and shows that increasing the the decoder depth decrease the reconstruction error.

While these results represent important steps toward a theoretical understanding of nonlinear autoencoders, the considered architecture differs significantly from those used in practice, which typically employ more expressive encoder and decoder functions. Complementary work surveys autoencoders from a manifold learning perspective [37] and establishes generalization bounds without characterizing optimization dynamics [38]. Understanding the training dynamics and limiting behavior of nonlinear bottleneck autoencoders with sufficiently expressive encoders and decoders therefore remains a central open problem.

2 Main Results and Contributions

We provide a theoretical framework for analyzing nonlinear bottleneck autoencoders using mean-field (MF) methods. Mean-field analysis has emerged as a powerful tool for studying neural networks [39; 40; 41; 42; 43] enabling the characterization of nonlinear models trained with gradient-based methods. Closest to our work is [27], which derives MF dynamics for shallow autoencoders; however, does not retain a fixed finite bottleneck between expressive encoder and decoder networks.

In this paper we introduce a MF formulation for BAE in section 3. In particular starting from the finite-width risk $\mathcal{R}_N(U, W)$ for the BAE as defined above the MF-risk is obtained by taking the limit $N \to \infty$, in which the empirical distributions of the parameters converge to probability measures ρ_U and ρ_W. In this regime, the random finite network output converges to a deterministic limit that depends only on these measures, and the risk functional correspondingly converges to $\mathcal{R}(\rho_U, \rho_W)$, which describes the reconstruction error of the infinite-width autoencoder in terms of the parameter distributions rather than individual neurons. Under mild regularity assumptions on the activation functions, learning rate, and boundedness of the data distribution we establish a rigorous link between finite-width BAE and their MF limit through the following.

Result 1a: BAE learning dynamics. Going beyond the commonly studied overcomplete (without bottleneck) setting, we derive the learning dynamics in the MF limit explicitly as given by (6).

[1]Formally, this nonlinear shallow AE is usually written as $f(x) = U^\top \sigma(Wx)$, where $W, U \in \mathbb{R}^{b \times d}$ are the linear *encoder* and *decoder*, and $\sigma(\cdot)$ is the point-wise nonlinearity. Some of the presented results also extend to the over-complete setting or include skip connections.

This expression builds both the foundation for our risk analysis as well as for the numerical analysis of the obtained latent representations.

Result 1b: Finite-Risk tracks MF-Risk during training. Let (U^t, W^t) denote the parameters of a finite-width autoencoder trained by SGD at step t with N hidden units, and let $(\rho_{U,t}, \rho_{W,t})$ denote the solution to the coupled mean-field PDEs for the encoder and decoder. For any finite time horizon $T > 0$, the empirical risk of the finite network tracks the MF risk up to an approximation error ε as $O\big(e^{CT^2}/\sqrt{N}\big)$ with high probability. This result is precisely stated in Theorem 2 and establishes that, during training, a finite-width BAE evolves approximately according to the MF gradient flow, and the finite-network risk closely follows the MF risk trajectory.

Result 2: Finite-Risk vs. MF-Risk at optimum. Extending the previous result to analyze the behavior at the optimum, we observe $|\inf_{U,W} \mathcal{R}_N(U, W) - \inf_{\rho_U, \rho_W} \mathcal{R}(\rho_U, \rho_W)| = O\left(1/N\right)$, which formalizes that finite networks are sufficiently expressive to approximate the MF optimum, with an error that vanishes as $N \to \infty$. The exact expression is stated in Theorem 3.

Result 3a: Non-linear AEs learn linear reconstructions first. Learning large directions first is known for linear models, where deep linear networks and linear autoencoders acquire singular modes in order of singular value [44; 21; 45] and have a general simplicity bias [46]. [47] shows that under certain model assumptions early-time dynamics of two-layer fully-connected non-linear networks are those of a linear model. Sequential learning occurs in shallow AEs studied with tied and untied weights [28], and in the weight-tied mean-field setting [27]. First for the linear autoencoder we re-derive similar observations using our framework in Theorem 9 (Appendix M.1). We then continue on to show, in Theorem 10 (Appendix M.2), that in general under non-linear activation, the observation that principal directions are learned first also holds. For, non-linear activations, the direction recovered is not exactly top eigenvector of the data covariance matrix but rather that of a kernel weighted covariance matrix. However, it is possible to show (c.f. Appendix M.2.3) that under mild conditions, this learned direction is indeed very close to the top eigenvector of data covariance matrix.

Result 3b: Global minimizers are self-consistent. At any global minimizer over sufficiently expressive classes, the decoder is the conditional expectation of the data given the encoding (a related notion was introduced in [48]), and the encoder is the projection index onto the decoder curve. Combining the two, for $b = 1$ the optimal curve is exactly a *principal curve* in the sense of Hastie and Stuetzle [49], and the population self-consistency error vanishes at global optima. Extending the notion to $b > 1$ links the representation more broadly to *principal surfaces* [49]. An overall notion of the connection between AEs and principal curves was discussed previously in [50] and under additional orthogonality constraints [51] relates the AE representation to weak principal manifolds. For completeness we also give a similar characterization of the blockwise-optimal encoder and decoder maps in Theorem 5.

Our theoretical analysis of the MF limit is complemented by numerical illustrations in Appendix K. The experiments show that the finite-width risk tracks the MF risk with an error scaling roughly as $O(N^{-1/2})$. They also illustrate the early emergence of linear representations and a decrease in self-consistency error during training.

Technical novelty. Our work builds on the mean-field analysis for neural-network optimization [39; 40; 42] which predominantly studies two-layer neural networks. Existing multilayer mean-field limits take every hidden layer wide, so that no finite-width layer separates the populations [52; 53; 54]. The case of infinitely wide networks with a finite bottleneck has been studied in the regime where the limit is a deep Gaussian process [55; 56]. In this setting [56] additionally derives learning dynamics and studies feature learning. In contrast we consider the particle mean-field scaling, coupled dynamics, and finite-width tracking guarantees as well as explicit characterizations fo the representation. In Mei et al. [42] the predictor is an average over one neuron population, so the risk is a quadratic functional of a single measure. Here an encoder measure ρ_W first forms the latent variable $\alpha(x; \rho_W) = \int \varphi(x; w)\, d\rho_W(w)$, and a decoder measure ρ_U then reconstructs through $\hat{x}(x; \rho_U, \rho_W) = \int \psi(\alpha(x; \rho_W); u)\, d\rho_U(u)$. For fixed ρ_W the risk is still quadratic in ρ_U, but ρ_W enters before the decoder nonlinearity, so $\mathcal{R}(\rho_U, \rho_W)$ is a genuinely coupled functional in which one population changes the input seen by the other. While coupled two-measure dynamics have been analysed for example for min–max games, [57; 58], and for multi-species McKean–Vlasov systems with pairwise interactions [59] to our knowledge, no prior work establishes a mean-field limit and

propagation of chaos for a compositionally coupled two-population system in which the output of one population is the finite-dimensional input of the other. Every result in the paper has to deal with this composition at some point, and we highlight where.

The first consequence is in the dynamics. The first variation in ρ_W must differentiate through the bottleneck and contains the averaged decoder Jacobian $\int \partial_\alpha \psi(\alpha(x;\rho_W);u)\, d\rho_U(u)$, the mean-field analogue of backpropagation, and at finite width a single encoder particle moves the bottleneck seen by every decoder particle. The propagation-of-chaos argument therefore cannot be run separately for the two layers; it requires joint Lipschitz estimates for the coupled fields (Lemma 13) and a coupled Gronwall argument (Lemma 14). To avoid the dimension-dependent rates that global weak metrics would impose, the measure-concentration step is carried out in the parametric distance $d_{\mathcal{F}}(\mu,\nu) = \sup_{f\in\mathcal{F}} \| \int f\, d\mu - \int f\, d\nu\|_2$, where $\mathcal{F}$ contains only the functions the vector fields actually evaluate (Lemma 5).

The second consequence is that we no longer assume a compact parameter domain. The sixth-power regularizer is chosen so that its inward drift dominates the outward drift of the unregularized fields, which is of order R^5 on a ball of radius R, and this yields a single deterministic ball that is invariant both for the SGD iterates (pathwise, for every width and every data sequence) and for the mean-field characteristics (Theorem 1). The same ball is what makes the coupled fixed-point argument for existence and uniqueness of the Vlasov–McKean system go through, and it makes the regularized risk coercive, so the mean-field minimum is attained (Theorem 3). In both Theorem 3 and Theorem 4 the encoder side is handled by a second-order expansion in the bottleneck variable: encoder fluctuations or perturbations are propagated through the decoder nonlinearity, while the decoder side remains additive once the bottleneck is fixed.

3 Bottleneck Autoencoder in the Mean Field Limit

Let us start by formalizing the basic idea of AEs, where one aims to map a data point $x \in \mathbb{R}^d$ into a lower dimensional latent space $\alpha \in \mathbb{R}^b$ which is called the bottleneck ($b \ll d$) and reconstruct it,

$$x \in \mathbb{R}^d \quad \overset{\phi(x;w)}{\longrightarrow} \quad \alpha \in \mathbb{R}^b \quad \overset{\psi(\alpha;u)}{\longrightarrow} \quad \hat{x} \in \mathbb{R}^d, \qquad \text{s.t. } b \ll d.$$

Here $\phi(x;w)$ and $\psi(\alpha;u)$ are the *encoder* and *decoder* functions parametrized by w and u respectively (see Appendix B for details). To pass from a finite N-sized AE to the mean-field regime, we treat the layer weights as particles where *encoder particle* and *decoder* particles are $w = (w_1, w_2)$ and $u = (u_1, u_2)$ resp.; both live in $\mathbb{R}^D$ with $D = d + b$. Their neuron maps are $\phi(x;w) = w_2\, \sigma(\langle w_1, x\rangle) \in \mathbb{R}^b$ and $\psi(\alpha;u) = u_1\, \sigma(\langle u_2, \alpha\rangle) \in \mathbb{R}^d$ resp. where $\sigma \in C_b^2(\mathbb{R})$ is the activation. For a network of width N with parameters $W = \{w_i\}_{i=1}^N$ and $U = \{u_j\}_{j=1}^N$, the layers are encoded by the empirical measures: $\hat{\rho}_W^N = \frac{1}{N}\sum_{i=1}^N \delta_{w_i}, \hat{\rho}_U^N = \frac{1}{N}\sum_{j=1}^N \delta_{u_j}$. For general probability measures $\rho_U, \rho_W \in \mathcal{P}(\mathbb{R}^D)$, the bottleneck, the reconstruction and the residual are

$$\alpha(x;\rho_W) := \int \phi(x;w)\, \rho_W(dw),\ \hat{x}(x;\rho_U,\rho_W) := \int \psi\big(\alpha(x;\rho_W);u\big)\, \rho_U(du),\ r := \hat{x} - x,$$

and the finite-width network is recovered exactly by plugging in the empirical measures: $\alpha_N(x) = \alpha(x;\hat{\rho}_W^N) = \frac{1}{N}\sum_i \phi(x;w_i)$ and $\hat{x}_N(x) = \hat{x}(x;\hat{\rho}_U^N,\hat{\rho}_W^N) = \frac{1}{N}\sum_j \psi(\alpha_N(x);u_j)$. We also record the averaged decoder Jacobian $\overline{J}(x;\rho_U,\rho_W) := \int J_\alpha \psi(\alpha(x;\rho_W);u)\, \rho_U(du) \in \mathbb{R}^{d\times b}$, which plays the role of backpropagation through the bottleneck (see Appendix B for details). The unregularized mean-field risk is $\mathcal{R}(\rho_U,\rho_W) := \frac{1}{2}\mathbb{E}_{x\sim P}\, \|r(x;\rho_U,\rho_W)\|^2$, and its sixth-power regularized version is

$$\mathcal{R}_\lambda(\rho_U,\rho_W) := \mathcal{R}(\rho_U,\rho_W) + \frac{\lambda}{6}\int \|u\|^6 \rho_U(du) + \frac{\lambda}{6}\int \|w\|^6 \rho_W(dw), \qquad \lambda > 0, \tag{1}$$

defined on $\mathcal{P}_6(\mathbb{R}^D)\times\mathcal{P}_6(\mathbb{R}^D)$, the measures with finite sixth moment. We use a sixth-order regularizer to control the dynamics of SDG and make sure that the updates lie within a fixed ball $B_D(0,R_*)$ for all particles (see Appendix F). The finite-width objective is precisely the mean-field functional evaluated at the empirical measures,

$$\mathcal{R}_\lambda^N(U,W) := \frac{1}{2}\mathbb{E}_{x\sim P}\|\hat{x}_N(x) - x\|^2 + \frac{\lambda}{6N}\sum_{j=1}^N \|u_j\|^6 + \frac{\lambda}{6N}\sum_{i=1}^N \|w_i\|^6 = \mathcal{R}_\lambda(\hat{\rho}_U^N,\hat{\rho}_W^N), \tag{2}$$

and $\mathcal{R}^N(U,W) = \mathcal{R}(\hat{\rho}_U^N, \hat{\rho}_W^N)$ for $\lambda = 0$. Expanding the square in (1) we get the external potential $V(u;\alpha) := \frac{1}{2}\mathbb{E}_x\langle x, \psi(\alpha(x);u)\rangle$ and the pairwise kernel $U(u,u';\alpha) := \frac{1}{2}\mathbb{E}_x\langle \psi(\alpha(x);u), \psi(\alpha(x);u')\rangle$, hence $\mathcal{R}_\lambda(\rho_U,\rho_W)$ becomes

$$\frac{1}{2}\mathbb{E}_x\|x\|^2 - 2\int V(u;\alpha)\,\rho_U(du) + \iint U(u,u';\alpha)\,\rho_U(du)\rho_U(du') + \frac{\lambda}{6}\int \|u\|^6\rho_U(du) + \frac{\lambda}{6}\int \|w\|^6\rho_W(dw), \tag{3}$$

with $\alpha = \alpha(\cdot;\rho_W)$. Unlike the standard single-layer mean-field setting, the encoder enters only through α inside V and U, so (3) is a genuinely coupled, non-quadratic functional of the pair (ρ_U,ρ_W) (see Appendix C for a heuristic derivation of the risk functional). The first variations of $\mathcal{R}_\lambda$ with respect to the decoder and encoder measures are (Appendix C.1)

$$\Phi_U(u;\rho_U,\rho_W) = \mathbb{E}_x\big[\langle \psi(\alpha(x);u),\, r(x;\rho_U,\rho_W)\rangle\big] + \frac{\lambda}{6}\|u\|^6, \tag{4}$$

$$\Phi_W(w;\rho_U,\rho_W) = \mathbb{E}_x\big[\langle \overline{J}(x;\rho_U,\rho_W)^\top r(x;\rho_U,\rho_W),\, \phi(x;w)\rangle\big] + \frac{\lambda}{6}\|w\|^6. \tag{5}$$

Let $V_U := -\nabla_u\Phi_U$ and $V_W := -\nabla_w\Phi_W$ denote the regularized descent fields, then the mean-field training dynamics with learning-rate schedule $\xi(t)$ are the coupled Vlasov–McKean equations

$$\partial_t\rho_{U,t} + \xi(t)\,\nabla_u\big(\rho_{U,t}\,V_U(u;\rho_{U,t},\rho_{W,t})\big) = 0, \partial_t\rho_{W,t} + \xi(t)\,\nabla_w\big(\rho_{W,t}\,V_W(w;\rho_{U,t},\rho_{W,t})\big) = 0, \tag{6}$$

i.e. a Wasserstein gradient flow of $\mathcal{R}_\lambda$ in which each population is transported by a field that depends on the other. The microscopic counterpart of (6) is one-sample SGD on (2). At iteration k, draw $x_{k+1} \sim P$, form $\alpha_k = \alpha_N(x_{k+1})$, $\hat{x}_k = \hat{x}_N(x_{k+1})$, $r_k = \hat{x}_k - x_{k+1}$ and the backpropagated bottleneck error $g_k = \overline{J}_k^\top r_k$, and update with step $s_k = \epsilon\,\xi(k\epsilon)$,

$$\begin{aligned} u_j^{k+1} &= u_j^k + s_k\Big[F_U(u_j^k;\hat{\rho}_{U,k}^N,\hat{\rho}_{W,k}^N,x_{k+1}) - \lambda\|u_j^k\|^4u_j^k\Big],\\ w_i^{k+1} &= w_i^k + s_k\Big[F_W(w_i^k;\hat{\rho}_{U,k}^N,\hat{\rho}_{W,k}^N,x_{k+1}) - \lambda\|w_i^k\|^4w_i^k\Big], \end{aligned} \tag{7}$$

where $F_U(u;\cdot) := -J_u\psi(\alpha_k;u)^\top r_k$ and $F_W(w;\cdot) := -J_w\phi(x_{k+1};w)^\top g_k$ are the single-sample unregularized descent fields (Appendix D). As both layers are normalized by $1/N$, so to make things compatible with the limit (7) uses the N-rescaled gradient $\theta^{k+1} = \theta^k - Ns_k\nabla_\theta\ell_\lambda^N$. Averaging over x_{k+1} gives $\mathbb{E}[F_U] = G_U = V_U + \lambda\|u\|^4u$ and likewise for W, so (7) is a discretization of (6).

4 Mean-Field Dynamics and Convergence

With the mean-field risk in place, we now make precise the relation between the finite-width network and its mean-field limit during training and at the optimum. Throughout we work under the following standing assumption about data, activation, learning rate, and initialization.

Assumption 1. *There exist finite constants M_x, M_ξ, L_ξ, R_0 such that:*

- *(i) the data distribution has bounded support,* $\mathrm{supp}(P) \subseteq B_d(0,M_x)$*;*
- *(ii) the activation satisfies $\sigma \in C_b^2(\mathbb{R})$, with $S_0 := \|\sigma\|_\infty$, $S_1 := \|\sigma'\|_\infty$, $S_2 := \|\sigma''\|_\infty$;*
- *(iii) the learning-rate schedule $\xi : [0,\infty) \to [0,\infty)$ is L_ξ-Lipschitz and $\|\xi\|_\infty \le M_\xi$;*
- *(iv) the initial population measures have compact support,* $\mathrm{supp}(\rho_{U,0}) \cup \mathrm{supp}(\rho_{W,0}) \subseteq B_D(0,R_0)$*, and the finite particles are initialized i.i.d. from them, so that* $\max_j \|u_j^0\|_2 \le R_0$ *and* $\max_i \|w_i^0\|_2 \le R_0$ *for every realization.*

Under Assumption 1 we define the *confinement radius* and the *admissible step size*

$$R_* := \max\Big\{R_0,\, 1,\, \tfrac{128\,C_{\mathrm{grad}}}{\lambda}\Big\}, \qquad s_* := \min\Big\{\tfrac{R_*}{2M_*},\, \tfrac{\lambda R_*^6}{64M_*^2}\Big\}, \tag{8}$$

where $C_{\mathrm{grad}} := O(M_x,S_0,S_1)$ bounds the unregularized fields by $C_{\mathrm{grad}}R^4$ on a radius-R ball, and $M_* := C_{\mathrm{grad}}R_*^4 + \lambda R_*^5$ bounds the regularized fields on $B_D(0,R_*)$ (see Appendix F.3 for details).

The first result shows that the coupled Vlasov–McKean PDEs (6) are globally well-posed and both the finite-width SGD iterates and the mean-field characteristics never leave a fixed deterministic ball.

Theorem 1 (Confinement and well-posedness of PDEs)**.** *Let Assumption 1 hold and let $\lambda > 0$.*

(a) (SGD confinement.) *If $\epsilon M_\xi \le s_*$, then for every width $N \ge 1$, every realization of the initialization and every data sequence $x_1, x_2, \cdots \in \mathrm{supp}(P)$, the SGD iterates* (7) *satisfy $\max_{1\le j\le N} \|u_j^k\|_2 \le R_*$, and $\max_{1\le i\le N} \|w_i^k\|_2 \le R_*$ for every $k \ge 0$.*

(b) (Existence, uniqueness, PDE confinement.) *There exists a unique pair $(\rho_{U,t}, \rho_{W,t})_{t\ge 0}$ of continuous distributional solutions of* (6) *with initial data $(\rho_{U,0}, \rho_{W,0})$. It is given by the push-forward of the initial measures along the characteristics $\dot u_t = \xi(t) V_U(u_t; \rho_{U,t}, \rho_{W,t})$, $\dot w_t = \xi(t) V_W(w_t; \rho_{U,t}, \rho_{W,t})$, and satisfies $\mathrm{supp}(\rho_{U,t}) \cup \mathrm{supp}(\rho_{W,t}) \subseteq B_D(0, R_*)$.*

Proof sketch. The mechanism behind both parts is a radial estimate. On a ball of radius $R \ge 1$, the bounded activation and bounded data give $\|G_U\|_2 \vee \|G_W\|_2 \le C_{\mathrm{grad}} R^4$ for the unregularized fields, so the outward radial drift $\langle u, G_U(u;\cdot)\rangle$ is at most $C_{\mathrm{grad}} R^5$, while the regularizer contributes $-\lambda\|u\|_2^6$. The choice of R_* in (8) makes the inward term dominate on the sphere $\|u\|_2 = R_*$ by a fixed margin; sixth is the smallest power that dominates an R^5 drift for *every* $\lambda > 0$, which is why we regularize with $\|\cdot\|^6$. We refer to proof of theorem 6 and theorem 7 for more details.

For uniqueness and existence (see proposition 1 and proposition 2 for details), note that on the invariant ball the fields are bounded and Lipschitz in the particle and in the pair of measures, so a Picard-type map on $C([0,T_0]; \mathcal{P}(B_D(0,R_*))^2)$ is a contraction for small T_0; since the constants do not change inside the ball, the fixed point is extended to all $t \ge 0$ by iteration. The rest of the proof is standard. Full details are in Appendices F.3 and G.3. □

The second result is a propagation-of-chaos statement: over any finite horizon, the empirical measures of the SGD particles and the finite-width regularized risk track the mean-field solution of Theorem 1(b), with an error that is polynomial in D and only logarithmic in N and ϵ.

Theorem 2 (Mean-field dynamics and propagation of chaos)**.** *Let Assumption 1 hold, let $\lambda > 0$, and let $(\rho_{U,t}, \rho_{W,t})_{t\ge 0}$ be the solution of* (6) *given by Theorem 1. Consider the SGD iterates* (7) *initialized i.i.d. from $(\rho_{U,0}, \rho_{W,0})$ with step size $s_k = \epsilon\,\xi(k\epsilon)$, $\epsilon M_\xi \le s_*$. Fix $T > 0$ and define $\mathrm{err}_{N,\epsilon}(\delta) := \sqrt{N^{-1} \vee \epsilon}\,\left(\sqrt{D + \log\left(N\left(\frac{T}{\epsilon} \vee 1\right)\right) + \delta}\right)$. There exist constants $C_1, C_2 > 0$, depending only on $b, d, M_x, S_0, S_1, S_2, \lambda, R_*, L_\xi, M_\xi$, such that for every $\delta > 0$ and every fixed test function f with $\|f\|_\infty, \|f\|_{\mathrm{Lip}} \le 1$, with probability at least $1 - 7e^{-\delta^2}$,*

$$\max_{k\in[0,T/\epsilon]\cap\mathbb{N}} \left| \frac{1}{N}\sum_{j=1}^N f(u_j^k) - \int f\, d\rho_{U,k\epsilon} \right| \le C_1(\sqrt{T} \vee T \vee 1)\, e^{C_2 T^2}\, \mathrm{err}_{N,\epsilon}(\delta), \tag{9}$$

$$\max_{k\in[0,T/\epsilon]\cap\mathbb{N}} \left| \frac{1}{N}\sum_{i=1}^N f(w_i^k) - \int f\, d\rho_{W,k\epsilon} \right| \le C_1(\sqrt{T} \vee T \vee 1)\, e^{C_2 T^2}\, \mathrm{err}_{N,\epsilon}(\delta),$$

$$\max_{k\in[0,T/\epsilon]\cap\mathbb{N}} \left| \mathcal{R}_\lambda^N(U^k, W^k) - \mathcal{R}_\lambda(\rho_{U,k\epsilon}, \rho_{W,k\epsilon}) \right| \le C_1(\sqrt{T} \vee T \vee 1)\, e^{C_2 T^2}\, \mathrm{err}_{N,\epsilon}(\delta). \tag{10}$$

Consequently, for every fixed $t \ge 0$, along any sequence $(N, \epsilon = \epsilon_N)$ with $N \to \infty$, $\epsilon_N \to 0$, $N/\log(N/\epsilon_N) \to \infty$ and $\epsilon_N \log(N/\epsilon_N) \to 0$, the empirical measures converge weakly, almost surely,

$$\hat\rho^N_{U,\lfloor t/\epsilon_N\rfloor} \Rightarrow \rho_{U,t}, \qquad \hat\rho^N_{W,\lfloor t/\epsilon_N\rfloor} \Rightarrow \rho_{W,t}.$$

Proof sketch. We couple each SGD particle to an ideal particle driven by the mean-field fields from the same initial condition: $\tilde u_j^t = u_j^0 + \int_0^t \xi(s) V_U(\tilde u_j^s; \rho_{U,s}, \rho_{W,s})\, ds$, and likewise for $\tilde w_i^t$, so that $\{\tilde u_j^t\}_j$ and $\{\tilde w_i^t\}_i$ are i.i.d. samples from $\rho_{U,t}$ and $\rho_{W,t}$. Writing the SGD recursion as an integral with the delayed index $[s] := \epsilon\lfloor s/\epsilon\rfloor$ and inserting the intermediate fields $V_U(\tilde u_j^{[s]}; \rho_{[s]})$, $V_U(u_j^{[s]}; \rho_{[s]})$, $V_U(u_j^{[s]}; \hat\rho^N_{[s]})$, the joint maximal deviation $\Delta(t) := \max_j \sup_{s\le t} \|\tilde u_j^s - u_j^{[s]}\|_2^2 + \max_i \sup_{s\le t} \|\tilde w_i^s - w_i^{[s]}\|_2^2$ is controlled by four terms: time discretization (E_{time}), spatial tracking (E_{track}), measure gap (E_{meas}) and martingale noise (E_{mart}). By Theorem 1 all particles, ideal and discrete, stay in $B_D(0,R_*)$, so every constant below is uniform.

E_{time} is $O(\epsilon^2)$ by the Lipschitz regularity of ξ and of $t \mapsto \tilde{u}_j^t$. E_{track} is bounded by $\int_0^t \Delta(s)\, ds$ using the spatial Lipschitz constant of the regularized fields on the ball. E_{mart} is a sum of bounded martingale differences (the single-sample fields F_U, F_W minus their conditional means G_U, G_W evaluated at the empirical measures); a vector Azuma–Hoeffding inequality with a union bound over particles and time steps gives an $O(T\epsilon\,[D + \log(N(T/\epsilon \vee 1)) + \delta^2])$ contribution.

The term that distinguishes the two-population bottleneck architecture from the additive two-layer case is E_{meas}. We split it through the "bridge" measures $\bar{\rho}_{U,s}^N := \frac{1}{N}\sum_j \delta_{\tilde{u}_j^s}$, $\bar{\rho}_{W,s}^N := \frac{1}{N}\sum_i \delta_{\tilde{w}_i^s}$, so that $\hat{\rho}^N \to \bar{\rho}^N$ is again tracking and $\bar{\rho}^N \to \rho$ is purely statistical. A global weak metric would bound the latter only at rate $N^{-1/D}$. Instead we show that V_U, V_W are Lipschitz in the measures with respect to parametric integral probability metrics whose test classes, $\{u' \mapsto \psi(\alpha; u'),\ J_\alpha \psi(\alpha; u')\}_\alpha$ and $\{w' \mapsto \phi(x; w')\}_x$, are exactly the functions through which each field sees the other population (α, $\hat{x}$, $\overline{J}$). These classes are finite-dimensionally parametrized, so an ε-net and Hoeffding give concentration at rate $\sqrt{(D + \log(N(T/\epsilon \vee 1)) + \delta^2)/N}$ uniformly over the time grid.

Summing the encoder and decoder estimates absorbs the cross terms into Δ, and Gronwall's lemma yields $\Delta(T) \le C_1(T \vee T^2)(\epsilon + N^{-1})[D + \log(N(T/\epsilon \vee 1)) + \delta^2]e^{C_2 T^2}$ with high probability. The bounds (9)–(10) follow by the triangle inequality through the bridge measure: $\sqrt{\Delta(T)}$ for the tracking part (the risk being Lipschitz in the particles on the ball) and Hoeffding or McDiarmid concentration of the bridge quantities for the sampling part. See Appendix J for full details. □

Now we compare the finite and the mean field risks. Because of the regularize, the risk attains its infrimum and the finite width optimum is within within $O(1/N)$ of the mean-field optimum.

Theorem 3 (Closeness between the finite-width and mean-field optimal risks)**.** *Let Assumption 1 hold and let $\lambda > 0$. Then $\mathcal{R}_\lambda$ attains its minimum on $\mathcal{P}_6(\mathbb{R}^D) \times \mathcal{P}_6(\mathbb{R}^D)$, and every minimizer (ρ_U^*, ρ_W^*) satisfies $\int \|u\|_2^6 \rho_U^*(du) + \int \|w\|_2^6 \rho_W^*(dw) \le 3M_x^2/\lambda$. There are constants $A_\lambda, B_\lambda > 0$, depending on M_x, S_0, S_1, S_2 and λ, such that for every $N \ge 1$,*

$$0 \;\le\; \inf_{U,W} \mathcal{R}_\lambda^N(U,W) \;-\; \min_{\rho_U, \rho_W \in \mathcal{P}_6(\mathbb{R}^D)} \mathcal{R}_\lambda(\rho_U, \rho_W) \;\le\; \frac{A_\lambda}{N} + \frac{B_\lambda}{N^2}. \tag{11}$$

Proof sketch. *Attainment.* The trivial pair (δ_0, δ_0) reconstructs every input as zero, so $\mathcal{R}_\lambda(\delta_0, \delta_0) = \frac{1}{2}\mathbb{E}\|x\|^2 \le \frac{1}{2}M_x^2$. Any minimizing sequence therefore has regularized risk at most $\frac{1}{2}M_x^2$, and since the reconstruction part is nonnegative, its sixth moments are bounded by $3M_x^2/\lambda$. Bounded moments give tightness, so the sequence has a weak limit. Along the sequence the regularizer can only decrease in the limit (lower semicontinuity), while the reconstruction risk passes to the limit because ϕ and ψ are bounded and Lipschitz in the particle. The limit is therefore a minimizer, and it inherits the moment bound.

Lower bound. Every finite configuration induces empirical measures with finite sixth moments and $\mathcal{R}_\lambda^N(U,W) = \mathcal{R}_\lambda(\hat{\rho}_U^N, \hat{\rho}_W^N)$ by (3), so the infimum over configurations is at least the minimum over measures.

Upper bound. Let (ρ_U^*, ρ_W^*) be a minimizer and build a random width-N network by sampling $u_j \sim \rho_U^*$ and $w_i \sim \rho_W^*$ independently. It suffices to bound the expected finite-width risk, since some realization does at least as well as the average. The regularizers have the same expectation as their population counterparts, so only the reconstruction error needs attention.

The obstacle is that the network is a Monte Carlo approximation of a composition. Averaging over the decoder particles alone reproduces the population decoder plus a variance term of order $1/N$. The encoder average, however, produces a random bottleneck $\alpha_N(x)$ that is then pushed through the nonlinear decoder, which creates a bias term. We control it by Taylor expanding the reconstruction error, as a function of the bottleneck value $\beta \in \mathbb{R}^b$, around the population bottleneck $\alpha(x)$. The first-order term vanishes because $\alpha_N(x)$ is an unbiased estimate of $\alpha(x)$, and the second-order term is at most a uniform Hessian bound times $\mathbb{E}\|\alpha_N(x) - \alpha(x)\|^2 = O(1/N)$; the Hessian bound uses only S_0, S_1, S_2 and the sixth-moment bound on ρ_U^*. Collecting the bias and the decoder variance gives $\mathbb{E}[\mathcal{R}_\lambda^N(U,W)] \le \mathcal{R}_\lambda(\rho_U^*, \rho_W^*) + A_\lambda/N + B_\lambda/N^2$, which is (11) (see Appendix H.2). □

5 Towards Characterizing the Learned Representation

Theorems 2 and 3 justify studying the mean-field problem in place of the finite network. We now ask what the mean-field optimum and early stage training looks like. First we give a necessary variational characterization, the second we characterize the encoder and decoder maps at the global minima as projection maps and conditional expectation, and at last we attempt to understand the early stage dynamics at small initializations.

The first statement is a first-order optimality condition in the spirit of Proposition 1 of [43]. Because the bottleneck couples the two measures through a nonlinear composition we obtain a one-sided result: every global minimizer is supported on the minimizers of its own first-variation potential.

Theorem 4 (Variational characterization of global minima)**.** *Let Assumption 1 hold, let $\lambda > 0$, and let (ρ_U^*, ρ_W^*) be a global minimizer of $\mathcal{R}_\lambda$ on $\mathcal{P}_6(\mathbb{R}^D)^2$. With Φ_U, Φ_W the first variations* (4)–(5)*,*

$$\phi_U^* := \inf_u \Phi_U(u; \rho_U^*, \rho_W^*) > -\infty \text{ and } \operatorname{supp}(\rho_U^*) \subseteq \arg\min_u \Phi_U(u; \rho_U^*, \rho_W^*);$$

$$\phi_W^* := \inf_w \Phi_W(w; \rho_U^*, \rho_W^*) > -\infty \text{ and } \operatorname{supp}(\rho_W^*) \subseteq \arg\min_w \Phi_W(w; \rho_U^*, \rho_W^*).$$

Proof sketch. The potentials are continuous and, because of the regularization the infimum are finite and the argmin sets are nonempty closed. We argue by contradiction with a local mass transport. Suppose ρ_U^* charges a ball $B(u_0, r_0)$ on which $\Phi_U \geq \phi_U^* + \Delta$ for some $\Delta > 0$, and let u_1 be a minimizer of Φ_U. Removing mass t from $\nu := \rho_U^*|_{B(u_0,r_0)}/\rho_U^*(B(u_0, r_0))$ and placing it at u_1 gives the probability measure $\rho_U^t = \rho_U^* + t(\delta_{u_1} - \nu)$. Since $\mathcal{R}_\lambda$ is exactly quadratic in ρ_U for fixed ρ_W (see (3)), we get $\mathcal{R}_\lambda(\rho_U^t, \rho_W^*) - \mathcal{R}_\lambda(\rho_U^*, \rho_W^*) \leq -t\Delta + Ct^2$ which is negative for small t, contradicting optimality. For the encoder the risk is *not* quadratic in ρ_W: a perturbation $\rho_W^t = \rho_W^* + t(\delta_{w_1} - \nu)$ shifts the bottleneck by $t\,\delta\alpha(x)$ with $\delta\alpha(x) = \phi(x; w_1) - \int \phi(x; w)\nu(dw)$, and this shift is pushed through the decoder. A second-order Taylor expansion of $\beta \mapsto \frac{1}{2}\| \int \psi(\beta; u)\rho_U^*(du) - x\|^2$ in the bottleneck variable, with the uniform Hessian bounds used in Theorem 3, shows that the first-order term is $t \int \Phi_W\, d(\delta_{w_1} - \nu) \leq -t\Delta$, with Φ_W carrying the backpropagated residual $\overline{J}^\top r$, and that the remainder is $O(t^2)$and we get a similar contradiction (see Appendix I). □

Classically, linear bottleneck AEs are tied to PCA, their nonlinear counterparts are tied to *principal manifolds* which are low-dimensional surfaces that are self-consistent i.e. every point is the conditional mean of the data projecting onto it [49]. We show that this structure is forced by optimality of the induced maps, for the unregularized reconstruction risk. For a mean-field pair write $f_{\rho_W}(x) := \alpha(x; \rho_W)$ for the encoder map and $g_{\rho_U}(z) := \int \psi(z; u)\rho_U(du)$ for the decoder map, so that $\mathcal{R}(\rho_U, \rho_W) = \mathcal{L}(f_{\rho_W}, g_{\rho_U})$ with $\mathcal{L}(f, g) := \frac{1}{2}\mathbb{E}\|X - g(f(X))\|_2^2$. Fix a compact latent set $\mathcal{K} \subset \mathbb{R}^b$ (by Theorem 1 the bottleneck takes values in a fixed compact set) and classes $\mathfrak{F}_W(\mathcal{K})$ of Borel encoders $f : \mathbb{R}^d \to \mathcal{K}$ and $\mathfrak{G}_U(\mathcal{K}) \subseteq C(\mathcal{K}; \mathbb{R}^d)$ of decoders. A pair $(f^\star, g^\star)$ is *blockwise optimal* if $\mathcal{L}(f^\star, g^\star) \leq \mathcal{L}(f, g^\star)$ for all $f \in \mathfrak{F}_W(\mathcal{K})$ and $\mathcal{L}(f^\star, g^\star) \leq \mathcal{L}(f^\star, g)$ for all $g \in \mathfrak{G}_U(\mathcal{K})$; every joint global minimizer is blockwise optimal. For a Borel encoder f let $Q_f := f_\# P$ and let $m_f : \mathcal{K} \to \mathbb{R}^d$ be a Borel version of the conditional mean, $m_f(f(X)) = \mathbb{E}[X \mid f(X)]$ a.s.; for a continuous decoder g let $A_g(x) := \arg\min_{z \in \mathcal{K}} \|x - g(z)\|_2^2$ be the nearest-code set. The two geometric residuals

$$\mathsf{SC}(f, g) := \int_{\mathcal{K}} \|g(z) - m_f(z)\|_2^2\, Q_f(dz), \quad \mathsf{Proj}(f, g) := \mathbb{E}\Big[\|X - g(f(X))\|_2^2 - \min_{z \in \mathcal{K}} \|X - g(z)\|_2^2\Big]$$

measure the failure of the conditional-mean and nearest-point identities. The following theorem is a generalization of self-consistency condition of Hastie and Stuetzle [49].

Theorem 5. *Let $(f^\star, g^\star) \in \mathfrak{F}_W(\mathcal{K}) \times \mathfrak{G}_U(\mathcal{K})$ be blockwise optimal and assume (i) some Borel encoder $\pi_{g^\star} \in \mathfrak{F}_W(\mathcal{K})$ satisfies $\pi_{g^\star}(x) \in A_{g^\star}(x)$ for P-a.e. x, and (ii) $m_{f^\star}$ has a continuous representative in $\mathfrak{G}_U(\mathcal{K})$. Then*

$$f^\star(x) \in A_{g^\star}(x) \text{ for } P\text{-a.e. } x, \qquad g^\star(z) = m_{f^\star}(z) \text{ for } Q_{f^\star}\text{-a.e. } z, \tag{12}$$

equivalently $\mathsf{Proj}(f^\star, g^\star) = \mathsf{SC}(f^\star, g^\star) = 0$. If, in addition, $g^\star$ is injective and the nearest point in $g^\star(\mathcal{K})$ is unique P-almost surely, then $f^\star = \pi_{g^\star}$ P-almost surely and

$$g^\star(z) = m_{\pi_{g^\star}}(z) \qquad Q_{f^\star}\text{-almost everywhere}. \tag{13}$$

If, furthermore, $\mathcal{K}$ is a compact b-dimensional C^1 manifold with boundary and $g^\star$ is a C^1 embedding, then $g^\star(\mathcal{K})$ is a self-consistent principal b-manifold in the sense of Definition 3. For $b = 1$, this is the classical principal-curve self-consistency condition of Hastie and Stuetzle [49].

Proof sketch. The argument rests on two exact decompositions of the risk, one for each block. For the decoder block, the conditional mean m_f is the L^2-projection of X onto functions of $f(X)$, so for any decoder g, $\mathcal{L}(f, g) = \mathcal{L}(f, m_f) + \frac{1}{2}\mathsf{SC}(f, g)$. For the encoder block, compactness of $\mathcal{K}$ and continuity of g make the nearest-code set $A_g(x)$ nonempty and compact, and a measurable-selection theorem provides a Borel nearest-point encoder π_g^0; comparing any encoder f with it gives $\mathcal{L}(f, g) = \mathcal{L}(\pi_g^0, g) + \frac{1}{2}\mathsf{Proj}(f, g)$, with $\mathsf{Proj}(f, g) = 0$ exactly when $f(x) \in A_g(x)$ for P-a.e. x. Now let $(f^\star, g^\star)$ be blockwise optimal. Assumption (i) makes the nearest-point encoder an admissible competitor, so optimality in the encoder block forces $\mathsf{Proj}(f^\star, g^\star) = 0$; assumption (ii) makes the conditional mean an admissible competitor, so optimality in the decoder block forces $\mathsf{SC}(f^\star, g^\star) = 0$. The principal manifold result follows trivially from the fact that $g^\star$ is an embedding and the nearest point identification is unique. For details, see Appendix L. □

Finally we consider a theorem that argues that under reasonable conditions of the initialization, the non-linear mean field AE initially recovers PCA. In the interest of space we only consider an informal discussion here, with the full theorem reserved for Appendix M. We have also performed additional experiments to verify this claim as detailed in Appendix L. Informally stated, we argue in Theorem 10 that the encoder–data correlations and the decoder's Jacobian become aligned with the same dominant data direction at a suitable early training time; moreover, the decoder's image lies approximately along that direction. The selected direction is the leading eigenvector of a kernel-weighted covariance determined by the initial encoder features. Under some additional assumptions on the features detailed in M.2.3, this direction approaches the top eigenvector of the data covariance matrix, thus giving a precise sense in which the autoencoder learns a PCA-like representation early in training.

6 Discussion and Future Work

In this work, we developed a mean-field framework for nonlinear autoencoders with a fixed finite-dimensional bottleneck, enabling analysis of both the risk and the learned representation. We derived coupled encoder and decoder dynamics, established finite-time tracking guarantees for finite-width SGD, and bounded the gap between finite-width and mean-field optimal risks. Beyond these guarantees, we characterized early PCA-like alignment under suitable initialization and connected blockwise optimality for the unregularized reconstruction risk to self-consistency and principal curves or manifolds under appropriate assumptions.

Beyond architectural generalizations, a central open question is whether, and at what rate, the learning dynamics approach the characterized optima, both in risk and in the induced encoder and decoder maps. Further directions include understanding the transition from early PCA-like behavior to nonlinear representations and the implicit bias through which initialization and training select among possible solutions.

AI USE STATEMENT

In this work, we used generative AI tools for rewrite mathematical claims, assist in the writing of proofs and implementations for the numerical simulations and generate synthetic data sets. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results, propose or refine hypotheses, design or provide feedback on research methodology or experiments and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for brainstorming and as a bouncing board for ideas. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We do not have any ethical conflicts.

REPRODUCIBILITY STATEMENT

All the results in the paper are fully reproducible using the proofs in the provided appendix and the code in the supplementary material.

REFERENCES

[1] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[2] Alan Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Modern Multivariate Statistical Techniques, 2008.

[3] Jerome H. Friedman and John W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 1974.

[4] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[5] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 1998.

[6] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 2009.

[7] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 1991.

[8] Shuangshuang Chen and Wei Guo. Auto-encoders in deep learning—a review with new perspectives. *Mathematics*, 2023.

[9] Kamal Berahmand, Fatemeh Daneshfar, Elaheh Sadat Salehi, Yuefeng Li, and Yue Xu. Autoencoders and their applications in machine learning: a survey. *Artificial Intelligence Review*, 2024.

[10] Marc-André Carbonneau, Julian Zaïdi, Jonathan Boilard, and Ghyslain Gagnon. Measuring disentanglement: A review of metrics. *IEEE Transactions on Neural Networks and Learning Systems*, 35:8747–8761, 2020. URL https://api.semanticscholar.org/CorpusID:229297600.

[11] Ikram Eddahmani, Chi-Hieu Pham, Thibault Napoléon, Isabelle Badoc, Jean-Rassaire Fouefack, and Marwa El-Bouz. Unsupervised learning of disentangled representation via autoencoding: A survey. *Sensors*, 2023.

[12] Felix Leeb, Stefan Bauer, Michel Besserve, and Bernhard Schölkopf. Exploring the latent space of autoencoders with interventional assays, 2023.

[13] Quentin Fournier and Daniel Aloise. Empirical comparison between autoencoders and traditional dimensionality reduction methods. In *International Conference on Artificial Intelligence and Knowledge Engineering*, 2019.

[14] Pascal Esser, Maximilian Fleissner, and Debarghya Ghoshdastidar. Theoretical foundations of representation learning using unlabeled data: Statistics and optimization, 2025.

[15] Jiawei Mao, Shujian Guo, Xuesong Yin, Yuanqi Chang, Binling Nie, and Yigang Wang. Medical supervised masked autoencoder: Crafting a better masking strategy and efficient fine-tuning schedule for medical image classification. *Applied Soft Computing*, 2025.

[16] Taylor W. Killian, Haoran Zhang, Jayakumar Subramanian, Mehdi Fatemi, and Marzyeh Ghassemi. An empirical study of representation learning for reinforcement learning in healthcare, 2020.

[17] Wen Xu, Jing He, and Yanfeng Shu. Deephealth: Deep representation learning with autoencoders for healthcare prediction. In *IEEE Symposium Series on Computational Intelligence*, 2020.

[18] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 1989.

[19] Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, and Cotton Seed. Loss landscapes of regularized linear autoencoders. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2019.

[20] Xuchan Bao, James Lucas, Sushant Sachdeva, and Roger B Grosse. Regularized linear autoencoders recover the principal components, eventually. In *Advances in Neural Information Processing Systems*, 2020.

[21] Arnu Pretorius, Steve Kroon, and Herman Kamper. Learning dynamics of linear denoising autoencoders. In *International Conference on Machine Learning*, 2018.

[22] Pascal Esser, Satyaki Mukherjee, Mahalakshmi Sabanayagam, and Debarghya Ghoshdastidar. Improved representation learning through tensorized autoencoders. In *International Conference on Artificial Intelligence and Statistics*, 2023.

[23] Eugene Golikov, Yaroslav Gusev, and Dmitry Yarotsky. Exact learning dynamics of a linear autoencoder through diagram expansions. In *Conference of Mathematics of AI*, 2026.

[24] Harald Steck. Autoencoders that don't overfit towards the identity. In *International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020.

[25] Jonghyun Ham, Maximilian Fleissner, and Debarghya Ghoshdastidar. Impact of bottleneck layers and skip connections on the generalization of linear denoising autoencoders, 2025.

[26] Thanh V. Nguyen, Raymond K. W. Wong, and Chinmay Hegde. Benefits of jointly training autoencoders: An improved neural tangent kernel analysis. *IEEE Transactions on Information Theory*, 2021.

[27] Phan-Minh Nguyen. Analysis of feature learning in weight-tied autoencoders via the mean field lens. *arXiv preprint arXiv:2102.08373*, 2021.

[28] Maria Refinetti and Sebastian Goldt. The dynamics of representation learning in shallow, non-linear autoencoders, 2022.

[29] Vicente Conde Mendes, Lorenzo Bardone, Cédric Koller, Jorge Medina Moreira, Vittorio Erba, Emanuele Troiani, and Lenka Zdeborová. A solvable high-dimensional model where nonlinear autoencoders learn structure invisible to pca while test loss misaligns with generalization, 2026.

[30] Hugo Cui and Lenka Zdeborova. High-dimensional asymptotics of denoising autoencoders. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[31] Andi Han, Wei Huang, Yuan Cao, and Difan Zou. On the feature learning in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[32] Thanh V Nguyen, Raymond KW Wong, and Chinmay Hegde. Autoencoders learn generative linear models. *arXiv preprint arXiv:1806.00572*, 2018.

[33] Aleksandr Shevchenko, Kevin Kögler, Hamed Hassani, and Marco Mondelli. Fundamental limits of two-layer autoencoders, and achieving them with gradient methods. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[34] Hugo Cui, Cengiz Pehlevan, and Yue M. Lu. A solvable model of learning generative diffusion: theory and insights. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=5b5wZg6Zeo.

[35] Lorenzo Bardone, Claudia Merger, and Sebastian Goldt. A theory of learning data statistics in diffusion models, from easy to hard. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=xPdpfcJ65T.

[36] Kevin Kögler, Alexander Shevchenko, Hamed Hassani, and Marco Mondelli. Compression of structured data with autoencoders: Provable benefit of nonlinearities and depth. *ArXiv*, abs/2402.05013, 2024. URL https://api.semanticscholar.org/CorpusID:267522958.

[37] Yonghyeon Lee. A geometric perspective on autoencoders, 2023.

[38] Baruch Epstein and Ron Meir. Generalization bounds for unsupervised and semi-supervised learning with autoencoders, 2019.

[39] Lénaïc Chizat and Francis R. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *ArXiv*, abs/1805.09545, 2018.

[40] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

[41] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

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

[43] 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 of the United States of America*, 2018.

[44] Gauthier Gidel, Francis R. Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:140225885.

[45] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *CoRR*, abs/1312.6120, 2013. URL https://api.semanticscholar.org/CorpusID:17272965.

[46] Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity, 2022. URL https://arxiv.org/abs/2106.15933.

[47] Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. The surprising simplicity of the early-time learning dynamics of neural networks. *Advances in Neural Information Processing Systems*, 33:17116–17128, 2020.

[48] Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *Journal of Machine Learning Research*, 15(110):3743–3773, 2014. URL http://jmlr.org/papers/v15/alain14a.html.

[49] Trevor Hastie and Werner Stuetzle. Principal curves. *Journal of the American Statistical Association*, 1989.

[50] E.C. Malthouse. Limitations of nonlinear pca as performed with generic neural networks. *IEEE Transactions on Neural Networks*, 9(1):165–173, 1998. doi: 10.1109/72.655038.

[51] Samuel Gerber. Saddlepoints in unsupervised least squares, 2021. URL https://arxiv.org/abs/2104.05000.

[52] Phan-Minh Nguyen. Mean field limit of the learning dynamics of multilayer neural networks, 2019. URL https://arxiv.org/abs/1902.02880.

[53] Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks, 2023. URL https://arxiv.org/abs/2001.11443.

[54] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 47(1):120–152, 2022. doi: 10.1287/moor.2020.1118. URL https://doi.org/10.1287/moor.2020.1118.

[55] Devanshu Agrawal, Theodore Papamarkou, and Jacob Hinkle. Wide neural networks with bottlenecks are deep gaussian processes. *Journal of Machine Learning Research*, 21(175): 1–66, 2020. URL http://jmlr.org/papers/v21/20-017.html.

[56] Etai Littwin, Omid Saremi, Shuangfei Zhai, Vimal Thilak, Hanlin Goh, Joshua M. Susskind, and Greg Yang. Implicit acceleration and feature learning in infinitely wide neural networks with bottlenecks, 2021. URL https://arxiv.org/abs/2107.00364.

[57] Carles Domingo-Enrich, Samy Jelassi, Arthur Mensch, Grant Rotskoff, and Joan Bruna. A mean-field analysis of two-player zero-sum games. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[58] Yulong Lu. Two-scale gradient descent ascent dynamics finds mixed nash equilibria of continuous games: A mean-field perspective. In *International Conference on Machine Learning*, pages 22790–22811. PMLR, 2023.

[59] Manh Hong Duong, Grigorios A. Pavliotis, and Julian Tugaut. Multi-species mckean-vlasov dynamics in non-convex landscapes, 2026. URL https://arxiv.org/abs/2507.07617.

APPENDIX

CONTENTS

A Notation **14**

B Discussion on the AE Architecture **15**

C The Mean-Field Risk Functional **18**

D SGD Updates **22**

E Preliminaries **24**

F Boundedness of the Coupled PDE and SGD **29**

G Existence and Uniqueness of the Coupled PDE **36**

H Closeness between Finite and Mean-Field Risks **41**

I Variational Characterization of Local Minima **46**

J Main Result: Mean-Field Dynamics and Propagation of Chaos **48**

K Numerical Validation and Experimental Details **62**

L Self-Consistency and Principal Manifolds **65**

M Early-Time PCA-First Dynamics for an Arbitrary Bottleneck Dimension **71**

A NOTATION

For $n \in \mathbb{N}$, we write $[n] := \{1, \dots, n\}$. Vectors are denoted by lowercase letters, such as x, u, w, and probability measures by symbols such as ρ_U, ρ_W, μ, ν. The Euclidean norm is denoted by $\|\cdot\|_2$, and $\langle \cdot, \cdot \rangle$ denotes the standard inner product. For matrices, $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. In the cases where there's no ambiguity, we use $\|\cdot\|$ for the operator norm. We use J for Jacobian matrices. In particular, $J_\alpha \psi$ denotes the Jacobian of ψ with respect to the bottleneck variable α, and $\nabla^2_\alpha \psi(\cdot)[v, v]$ denotes the corresponding second directional derivative in direction v. We use standard asymptotic notation $O(\cdot)$.

Dimensions and data. The data dimension is d, the bottleneck dimension is b, and we assume $d > b \geq 1$. The dimension of a single encoder or decoder particle is $D := d + b$. The data distribution is denoted by P, and $X \sim P$ or $x \sim P$ denotes a data point. We assume throughout that the data distribution has bounded support, $\mathrm{supp}(P) \subseteq B_d(0, M_x)$. Expectation with respect to P is denoted by $\mathbb{E}_{x \sim P}[\cdot]$, or simply by $\mathbb{E}_x[\cdot]$.

Particles and neuron maps. Encoder particles and decoder particles are denoted by $w = (w_1, w_2) \in \mathbb{R}^d \times \mathbb{R}^b$ and $u = (u_1, u_2) \in \mathbb{R}^d \times \mathbb{R}^b$. For a finite-width network of width N, we write $W = \{w_i\}_{i=1}^N$, and $U = \{u_j\}_{j=1}^N$. The activation function is $\sigma \in C_b^2(\mathbb{R})$, and we denote

$$S_0 := \|\sigma\|_\infty, \qquad S_1 := \|\sigma'\|_\infty, \qquad S_2 := \|\sigma''\|_\infty.$$

The encoder and decoder neuron maps are

$$\phi(x; w) = w_2\, \sigma(\langle w_1, x\rangle) \in \mathbb{R}^b, \qquad \psi(\alpha; u) = u_1\, \sigma(\langle u_2, \alpha\rangle) \in \mathbb{R}^d,$$

where $\alpha \in \mathbb{R}^b$ is the bottleneck variable.

Measures, bottleneck, reconstruction, and residual. The empirical measures associated with finite-width particles are denotes with a hat $\hat{\rho}_U^N$ and $\hat{\rho}_W^N$. For general probability measures $\rho_U, \rho_W \in \mathcal{P}(\mathbb{R}^D)$, the population bottleneck, reconstruction, and residual are $\alpha(x; \rho_W)$, $\hat{x}(x; \rho_U, \rho_W)$ and $r(x; \rho_U, \rho_W)$ respectively. The averaged decoder Jacobian is denotes by $\overline{J}(x; \rho_U, \rho_W)$.

Measure classes and distances. For $R < \infty$, we write $\mathcal{P}_R$ for the space of probability measure with support in $B_D(0, R)$. $\mathcal{P}_6(\mathbb{R}^D)$ denotes the space of probability measures with finite sixth moment. For two probability measures μ, ν, $W_1(\mu, \nu)$ denotes the 1-Wasserstein distance. For pairs of measures we use the product distance

$$\mathcal{W}_1\big((\mu, \nu), (\mu', \nu')\big) := W_1(\mu, \mu') + W_1(\nu, \nu').$$

Risks, first variations, and vector fields. The unregularized mean-field reconstruction risk is denoted by $\mathcal{R}(\rho_U, \rho_W)$ while the sixth order regularized risk is denotes by $\mathcal{R}_\lambda(\rho_U, \rho_W)$ where λ si the strength of the regularizer. The corresponding finite-width objective is denoted by $\mathcal{R}_\lambda^N(U, W)$. The first variations of $\mathcal{R}_\lambda$ with respect to the decoder and encoder measures are denoted by $\Phi_U(u; \rho_U, \rho_W)$, and $\Phi_W(w; \rho_U, \rho_W)$ respectively. The unregularized population descent fields are denoted by G_U and G_W, while the regularized mean-field vector fields are $V_U(u; \rho_U, \rho_W)$ and $V_W(w; \rho_U, \rho_W)$. The coupled mean-field dynamics are written in terms of the learning-rate schedule $\xi(t)$, where $\|\xi\|_\infty \le M_\xi$ and ξ is L_ξ-Lipschitz.

SGD. Discrete time is indexed by $k \ge 0$ and the step size is denoted by $s_k = \epsilon\, \xi(k\epsilon)$. At iteration k, the empirical measures are $\hat{\rho}_{U,k}^N$ and $\hat{\rho}_{W,k}^N$. The one-sample regularized loss is denoted by $\ell_{\lambda,k+1}^N$, the iteration-specific residual by r_k, and the backpropagated bottleneck error by g_k. The single-sample unregularized SGD fields are denoted by F_U and F_W.

Confinement and boundedness constants. The constant R_* denotes the deterministic confinement radius for both the finite-particle SGD iterates and the mean-field characteristics. The quantity s_* denotes the corresponding admissible step-size threshold.

Map-level representation notation. For map-level statements, $\mathcal{K} \subset \mathbb{R}^b$ denotes a compact latent set. The encoder and decoder map classes are denoted by $\mathfrak{F}_W(\mathcal{K})$, and $\mathfrak{G}_U(\mathcal{K})$. We sometimes use $\mathcal{L}(f, g) := \frac{1}{2}\mathbb{E}\|X - g(f(X))\|_2^2$ to denote the reconstruction risk whenever we want to stress the dependence on the encoder and decoder function f and g. For a Borel encoder f, we use $Q_f := f_\# P$ to denote the pushforward measure on the bottleneck. We denote Borel version of the conditional mean by m_f. For a continuous decoder g, the nearest-code correspondence is denoted by A_g. When needed, π_g^0 denotes a Borel nearest-point selector. The self-consistency and projection residuals are denoted by $\mathsf{SC}(f, g)$, and $\mathsf{Proj}(f, g)$. If M is a decoder image or a candidate principal manifold, Π_M denotes nearest-point projection onto M, whenever this projection is well-defined.

B Discussion on the AE Architecture

Let us recall the general idea of the autoencoder (AE) mapping. One aims to map a datapoint $x \in \mathbb{R}^d$ into a latent space $\alpha \in \mathbb{R}^b$ and then reconstruct it:

$$x \in \mathbb{R}^d \quad \overset{\phi(x;w)}{\longrightarrow} \quad \alpha \in \mathbb{R}^b \quad \overset{\psi(\alpha;u)}{\longrightarrow} \quad \hat{x} \in \mathbb{R}^d$$

where $\phi(x; w)$ is the *encoder* function parameterized by an encoder particle w, and $\psi(\alpha; u)$ is the *decoder* function parameterized by a decoder particle u. The encoder and decoder particles lie in $\mathbb{R}^D$ where $D = d + b$, and our bottleneck α lies in $\mathbb{R}^b$. This structure is illustrated in the following diagram:

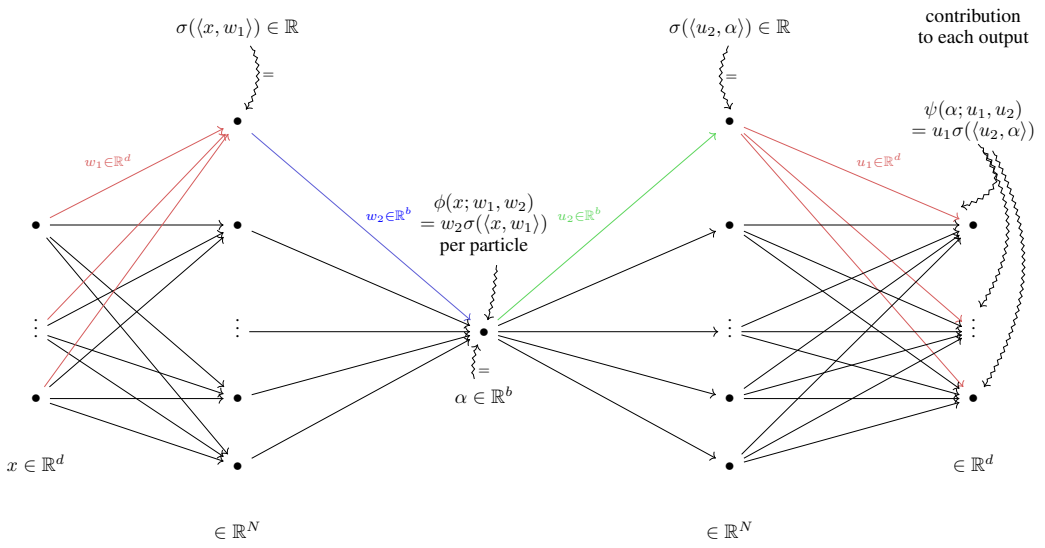


Figure 1: Architecture of the Mean-Field Bottleneck Autoencoder.

B.1 Finite and Limiting Expressions for the Autoencoder

We fix the data dimension d and bottleneck dimension b throughout this analysis. If we write the encoder and decoder particle respectively as

$$w = (w_1, w_2) \in \mathbb{R}^d \times \mathbb{R}^b, \qquad u = (u_1, u_2) \in \mathbb{R}^d \times \mathbb{R}^b.$$

then the respective neuron maps evaluated on a data point x and a bottleneck representation α are given by,

$$\phi(x; w) = w_2\, \sigma(\langle w_1, x\rangle) \in \mathbb{R}^b, \qquad \psi(\alpha; u) = u_1\, \sigma(\langle u_2, \alpha\rangle) \in \mathbb{R}^d$$

which corresponds to colored arrows in 1. We refer to Section C for heuristics and details of the quantities that follow. Let us first consider the Mean Field Case of the Autoencoder. For probability measures $\rho_U, \rho_W \in \mathcal{P}(\mathbb{R}^D)$, we define the population-level bottleneck, reconstruction, residual, and averaged Jacobian as,

$$\begin{aligned}
\alpha(x; \rho_W) &:= \int \phi(x; w)\rho_W(dw) \in \mathbb{R}^b, \\
\hat{x}(x; \rho_U, \rho_W) &:= \int \psi(\alpha(x; \rho_W); u)\rho_U(du) \in \mathbb{R}^d, \\
r(x; \rho_U, \rho_W) &:= \hat{x}(x; \rho_U, \rho_W) - x \in \mathbb{R}^d, \\
\overline{J}(x; \rho_U, \rho_W) &:= \int J_\alpha \psi(\alpha(x; \rho_W); u)\rho_U(du) \in \mathbb{R}^{d\times b}.
\end{aligned}$$

The unregularized reconstruction risk and the sixth-power regularized mean-field risk are defined as:

$$\begin{aligned}
\mathcal{R}(\rho_U, \rho_W) &:= \frac{1}{2}\mathbb{E}_{x\sim P}\|r(x; \rho_U, \rho_W)\|^2, \\
\mathcal{R}_\lambda(\rho_U, \rho_W) &:= \mathcal{R}(\rho_U, \rho_W) + \frac{\lambda}{6}\int \|u\|^6 \rho_U(du) + \frac{\lambda}{6}\int \|w\|^6 \rho_W(dw), \qquad \lambda > 0. \quad (14)
\end{aligned}$$

The natural domain of (14) is $\mathcal{P}_6(\mathbb{R}^D) \times \mathcal{P}_6(\mathbb{R}^D)$, where

$$\mathcal{P}_6(\mathbb{R}^D) := \left\{\rho \in \mathcal{P}(\mathbb{R}^D) : \int \|v\|^6 \rho(dv) < \infty\right\}.$$

Now coming to the case of finite width-N Autoencoder as illustrated in , the empirical distributions of the encoder and decoder particles are,

$$\hat{\rho}_W^N := \frac{1}{N}\sum_{i=1}^N \delta_{w_i}, \qquad \hat{\rho}_U^N := \frac{1}{N}\sum_{j=1}^N \delta_{u_j}.$$

Evaluating the population functions on these empirical measures yields the finite-width representations:

$$\alpha_N(x) := \int \phi(x;w)\hat{\rho}_W^N(dw) = \frac{1}{N}\sum_{i=1}^N \phi(x;w_i),$$

$$\hat{x}_N(x) := \int \psi(\alpha_N(x);u)\hat{\rho}_U^N(du) = \frac{1}{N}\sum_{j=1}^N \psi(\alpha_N(x);u_j).$$

Consequently, the regularized population objective is exactly the mean-field functional evaluated at the empirical measures:

$$\mathcal{R}_\lambda^N(U,W) := \frac{1}{2}\mathbb{E}_x\|\hat{x}_N(x)-x\|^2 + \frac{\lambda}{6N}\sum_{j=1}^N \|u_j\|^6 + \frac{\lambda}{6N}\sum_{i=1}^N \|w_i\|^6 = \mathcal{R}_\lambda(\hat{\rho}_U^N, \hat{\rho}_W^N).$$

B.2 Expressions for Jacobian

Before deriving the mean-field equations, we record the Jacobian expressions that will be used throughout. Since the encoder map is given by

$$\phi(x;w) = w_2\,\sigma(\langle w_1, x\rangle),$$

the partial Jacobians with respect to its particle components are

$$J_{w_1}\phi(x;w) = w_2\,\sigma'(\langle w_1,x\rangle)x^\top \in \mathbb{R}^{b\times d},$$

and

$$J_{w_2}\phi(x;w) = \sigma(\langle w_1,x\rangle)I_b \in \mathbb{R}^{b\times b}.$$

Concatenating these blocks yields the full Jacobian with respect to the encoder particle:

$$J_w\phi(x;w) = \begin{bmatrix} w_2\,\sigma'(\langle w_1,x\rangle)x^\top & \sigma(\langle w_1,x\rangle)I_b\end{bmatrix} \in \mathbb{R}^{b\times D}.$$

Similarly, since the decoder map is given by

$$\psi(\alpha;u) = u_1\,\sigma(\langle u_2,\alpha\rangle),$$

its Jacobian with respect to the bottleneck variable α is

$$J_\alpha\psi(\alpha;u) = u_1\,\sigma'(\langle u_2,\alpha\rangle)u_2^\top \in \mathbb{R}^{d\times b}.$$

The partial Jacobians with respect to the two components of the decoder particle are

$$J_{u_1}\psi(\alpha;u) = \sigma(\langle u_2,\alpha\rangle)I_d \in \mathbb{R}^{d\times d},$$

and

$$J_{u_2}\psi(\alpha;u) = u_1\,\sigma'(\langle u_2,\alpha\rangle)\alpha^\top \in \mathbb{R}^{d\times b}.$$

Therefore, the full Jacobian with respect to the decoder particle is,

$$J_u\psi(\alpha;u) = \begin{bmatrix}\sigma(\langle u_2,\alpha\rangle)I_d & u_1\,\sigma'(\langle u_2,\alpha\rangle)\alpha^\top\end{bmatrix} \in \mathbb{R}^{d\times D}.$$

Thus the averaged decoder Jacobian evaluated at the bottleneck is given by,

$$\overline{J}(x;\rho_U,\rho_W) = \int u_1\,\sigma'\left(\langle u_2,\alpha(x;\rho_W)\rangle\right)u_2^\top\,\rho_U(du) \in \mathbb{R}^{d\times b}.$$

We will need this expression during the encoder backpropagation.

C THE MEAN-FIELD RISK FUNCTIONAL

In this section, we shall start with a finite risk expression and try to motivate/obtain a mean-field risk functional. This section is mostly based on heuristics on why these should be the correct expression. The exact proof of the relations between these finite risks and mean-field risks are content of Theorem 2 and Theorem 3. We then record the derivations of the mean-field risk functionals with respect to encoder and the decoder measures, and the coupled Vlasov-McKean equation that cover the evolution of the enocder and decoder measures.

Let the data distribution be P. The population risk for the finite-width autoencoder can be written as

$$\mathcal{R}^N(U,W) = \frac{1}{2}\mathbb{E}_{x\sim P}\left[\|x - \hat{x}_N(x)\|^2\right],$$

and the regularized population risk can be written as

$$\mathcal{R}^N_\lambda(U,W) = \frac{1}{2}\mathbb{E}_{x\sim P}\left[\|x - \hat{x}_N(x)\|^2\right] + \frac{\lambda}{6N}\sum_{j=1}^N \|u_j\|^6 + \frac{\lambda}{6N}\sum_{i=1}^N \|w_i\|^6.$$

Remark 1. *Here we have taken a 6th-order regularizer. This is to ensure that the SDG iterates and the continuous mean-field dynamics remain strictly within a deterministic ball $B_D(0,R_*)$ (see Section F for description of the fixed radius R_*). As we shall see in the analysis of the bounds (see proof of Theorem 1, the unregularized outward drift is of order R^5 on product of radius R balls. Hence, sixth is the minimum power that strictly dominates the drift for every fixed $\lambda > 0$. One could use a lower order regularization but it would come at the expense of complicated proofs, unnatural assumptions on the support of the data distribution P, and various other bookkeeping.*

We now focus on the regularized risk; the formulation for the unregularized risk is perfectly analogous and can be obtained by setting $\lambda = 0$. If we change our point of view from the parameter sets (U,W) to the discrete empirical probability measures $\hat{\rho}^N_U = \frac{1}{N}\sum_{j=1}^N \delta_{u_j}$ and $\hat{\rho}^N_W = \frac{1}{N}\sum_{i=1}^N \delta_{w_i}$, we can rewrite the regularized risk as a functional of the empirical measures:

$$\mathcal{R}^N_\lambda(\hat{\rho}^N_U, \hat{\rho}^N_W) = \frac{1}{2}\mathbb{E}_{x\sim P}\left[\|x - \hat{x}_N(x)\|^2\right] + \frac{\lambda}{6}\int \|u\|^6 \hat{\rho}^N_U(du) + \frac{\lambda}{6}\int \|w\|^6 \hat{\rho}^N_W(dw).$$

We can rewrite the reconstruction part of the risk by expanding the quadratic expression within the expectation as

$$\begin{aligned}\mathcal{R}^N_\lambda(\hat{\rho}^N_U, \hat{\rho}^N_W) = \underbrace{\frac{1}{2}\mathbb{E}_x[\|x\|^2]}_{\text{Term 1}} - \underbrace{\mathbb{E}_x[\langle x, \hat{x}_N(x)\rangle]}_{\text{Term 2}} \\ + \underbrace{\frac{1}{2}\mathbb{E}_x[\|\hat{x}_N(x)\|^2]}_{\text{Term 3}} + \frac{\lambda}{6}\int \|u\|^6 \hat{\rho}^N_U(du) + \frac{\lambda}{6}\int \|w\|^6 \hat{\rho}^N_W(dw).\end{aligned}$$

Following standard mean-field limits, we would like to get hold of a risk functional $\mathcal{R}_\lambda(\rho_U, \rho_W)$ for any probability measure. We thus analyze heuristically what the function could be in the limit as one lets $N \to \infty$.

Term 1 is the baseline risk and is independent of N and the weights. We let

$$\text{Term 1} = T_1^{(N)} = \mathcal{R}_\# := \frac{1}{2}\mathbb{E}_x[\|x\|^2].$$

If we expand **Term 2** by substituting the integral form of $\hat{x}_N$ we get,

$$T_2^{(N)} = \mathbb{E}_x\left[\left\langle x, \int \psi(\alpha_N(x);u)\hat{\rho}^N_U(du)\right\rangle\right].$$

If Fubini's theorem were applicable, one would hope to swap the expectation and integral and under the limit, if the empirical measure $\hat{\rho}^N_U \to \rho_U$ and the argument $\alpha_N(x) \to \alpha(x;\rho_W)$, we can hope that

$$\lim_{N\to\infty} T_2^{(N)} \approx \int \underbrace{\mathbb{E}_x\left[\langle x, \psi(\alpha(x;\rho_W);u)\rangle\right]}_{\equiv 2V(u;\alpha)} \rho_U(du).$$

We formally define the external potential $V(u;\alpha)$ and the bottleneck potential $\alpha(x)$ (we shall suppress ρ_W dependence where clear) as

$$V(u;\alpha) := \frac{1}{2}\mathbb{E}_x\left[\langle x, \psi(\alpha(x);u)\rangle\right] \in \mathbb{R},$$
$$\alpha(x) = \int \phi(x;w)\rho_W(dw) \in \mathbb{R}^b,$$

which makes the second term

$$\text{Term } 2 = 2\int V(u;\alpha)\rho_U(du).$$

If we substitute the integral form of $\hat{x}_N$ in **Term 3**, the squared norm of an integral can be rewritten as a double integral over a product measure as follows (assuming the integrals and expectation can be switched),

$$T_3^{(N)} = \frac{1}{2}\mathbb{E}_x\left[\left\langle \int \psi(\alpha_N(x);u)\hat{\rho}_U^N(du), \int \psi(\alpha_N(x);u')\hat{\rho}_U^N(du')\right\rangle\right],$$

and hence

$$T_3^{(N)} = \frac{1}{2}\iint \mathbb{E}_x\left[\langle \psi(\alpha_N(x);u), \psi(\alpha_N(x);u')\rangle\right]\hat{\rho}_U^N(du)\hat{\rho}_U^N(du').$$

If we let $\rho_U \otimes \rho_U$ denote the limiting product measure of $\hat{\rho}_U^N \otimes \hat{\rho}_U^N$, and follow a similar heuristic as in the analysis of the second term, our prospective third term should look like

$$\begin{aligned}\text{Term } 3 \approx \lim_{N\to\infty} T_3^{(N)} &\approx \frac{1}{2}\iint \mathbb{E}_x\left[\langle \psi(\alpha(x);u), \psi(\alpha(x);u')\rangle\right]\rho_U(du)\rho_U(du') \\ &= \iint \underbrace{\frac{1}{2}\mathbb{E}_x\left[\langle \psi(\alpha(x);u), \psi(\alpha(x);u')\rangle\right]}_{\equiv U(u,u';\alpha)}\rho_U(du)\rho_U(du').\end{aligned}$$

This lets us formally define the pairwise interaction kernel $U(u,u';\alpha)$,

$$U(u,u';\alpha) := \frac{1}{2}\mathbb{E}_x\left[\langle \psi(\alpha(x);u), \psi(\alpha(x);u')\rangle\right].$$

The regularization terms are already linear functionals of the corresponding measures. Hence, under convergence of the empirical measures, they formally converge as

$$\frac{\lambda}{6}\int \|u\|^6\hat{\rho}_U^N(du) \longrightarrow \frac{\lambda}{6}\int \|u\|^6\rho_U(du),$$

and

$$\frac{\lambda}{6}\int \|w\|^6\hat{\rho}_W^N(dw) \longrightarrow \frac{\lambda}{6}\int \|w\|^6\rho_W(dw).$$

Combining the three reconstruction terms and the two regularization terms, we arrive at the continuous regularized risk functional on the space of probability measures:

$$\begin{aligned}\mathcal{R}_\lambda(\rho_U,\rho_W) = \mathcal{R}_\# &- 2\int V(u;\alpha)\rho_U(du) + \iint U(u,u';\alpha)\rho_U(du)\rho_U(du') \\ &+ \frac{\lambda}{6}\int \|u\|^6\rho_U(du) + \frac{\lambda}{6}\int \|w\|^6\rho_W(dw),\end{aligned}$$

where

$$\alpha(x) = \int \phi(x;w)\rho_W(dw).$$

We also note that the limiting mean-field output $\hat{x}$ is given by

$$\hat{x}(x) = \int \psi(\alpha(x);u)\rho_U(du),$$

and hence the residual is

$$r(x;\rho_U,\rho_W) = \hat{x}(x) - x.$$

C.1 Derivation of the Regularized Mean-Field Functional

We give a brief description of the first variations of $\mathcal{R}_\lambda$ with respect to the encoder and decoder measure. This will be used to obtain the Wasserstein gradient flow formulation.

Derivative with respect to the Decoder Measure (ρ_U)

The bottleneck potential $\alpha(x)$ holds all the ρ_W dependence and is independent of the decoder measure ρ_U; hence we can treat it as a constant for this variation. The functional derivative $\Phi_U(u;\rho_U,\rho_W)$ of the regularized risk $\mathcal{R}_\lambda$ with respect to the decoder measure ρ_U yields

$$\Phi_U(u;\rho_U,\rho_W) := \frac{\delta \mathcal{R}_\lambda}{\delta \rho_U}(u) = -2V(u;\alpha) + 2\int U(u,u';\alpha)\rho_U(du') + \frac{\lambda}{6}\|u\|^6.$$

If we substitute the expressions for V and U in the expression for Φ_U, we get

$$\begin{aligned}
\Phi_U(u;\rho_U,\rho_W) = & -\mathbb{E}_x\left[\langle x, \psi(\alpha(x);u)\rangle\right] \\
& + \int \mathbb{E}_x\left[\langle \psi(\alpha(x);u), \psi(\alpha(x);u')\rangle\right]\rho_U(du') + \frac{\lambda}{6}\|u\|^6 \\
= & \ \mathbb{E}_x\left[-\langle \psi(\alpha(x);u), x\rangle + \left\langle \psi(\alpha(x);u), \int \psi(\alpha(x);u')\rho_U(du')\right\rangle\right] + \frac{\lambda}{6}\|u\|^6 \\
= & \ \mathbb{E}_x\left[\left\langle \psi(\alpha(x);u), \int \psi(\alpha(x);u')\rho_U(du') - x\right\rangle\right] + \frac{\lambda}{6}\|u\|^6 \\
= & \ \mathbb{E}_x\left[\langle \psi(\alpha(x);u), \hat{x}(x) - x\rangle\right] + \frac{\lambda}{6}\|u\|^6.
\end{aligned}$$

Therefore,

$$\boxed{\Phi_U(u;\rho_U,\rho_W) = \mathbb{E}_x\left[\langle \psi(\alpha(x);u), r(x;\rho_U,\rho_W)\rangle\right] + \frac{\lambda}{6}\|u\|^6.}$$

One way to interpret this is to say that the first part of the gradient of the risk with respect to a decoder particle u is the expected inner product of the difference of the output and the input of the AE ("the residual") with the activation of that specific particle ψ_u, whereas the second part comes from the sixth-power confinement regularizer.

Now we differentiate Φ_U with respect to the decoder particle u. Using

$$J_u\psi(\alpha;u) = \begin{bmatrix}\sigma(\langle u_2,\alpha\rangle)I_d & u_1\,\sigma'(\langle u_2,\alpha\rangle)\alpha^\top\end{bmatrix},$$

and

$$\nabla_u\left(\frac{\lambda}{6}\|u\|^6\right) = \lambda\|u\|^4 u,$$

we obtain

$$\nabla_u\Phi_U(u;\rho_U,\rho_W) = \mathbb{E}_x\left[J_u\psi(\alpha(x);u)^\top r(x;\rho_U,\rho_W)\right] + \lambda\|u\|^4 u.$$

Thus, if G_U denotes the unregularized descent vector field and V_U the regularized descent vector field, then

$$\begin{aligned}
G_U(u;\rho_U,\rho_W) &:= -\mathbb{E}_x\left[J_u\psi(\alpha(x);u)^\top r(x;\rho_U,\rho_W)\right], \\
V_U(u;\rho_U,\rho_W) &:= -\nabla_u\Phi_U(u;\rho_U,\rho_W) \\
&= G_U(u;\rho_U,\rho_W) - \lambda\|u\|^4 u.
\end{aligned}$$

Derivative with respect to the Encoder Measure (ρ_W)

The encoder affects the reconstruction part of the risk only through α which is a function of input x and the encoder measure ρ_W. The variation of α with respect to the encoder measure is

$$\frac{\delta\alpha(x;\rho_W)}{\delta\rho_W}(w) = \phi(x;w) \in \mathbb{R}^b.$$

Since $b \geq 1$, α and its variation $\delta\alpha$ are vectors in $\mathbb{R}^b$. Therefore the Jacobian of the decoder neuron with respect to α is

$$J_\alpha \psi(\alpha; u) = u_1 \, \sigma'(\langle u_2, \alpha \rangle) u_2^\top \in \mathbb{R}^{d \times b}.$$

We apply the functional chain rule through the intermediate activation α to calculate the functional derivative $\Phi_W(w; \rho_U, \rho_W)$ of the regularized risk $\mathcal{R}_\lambda$ wrt to the encoder measure ρ_W as shown below.

We first compute the variation of the reconstruction risk with respect to α,

$$\begin{aligned}
\delta\mathcal{R}(\alpha) = & -2 \int \delta V(u; \alpha) \rho_U(du) + \iint \delta U(u, u'; \alpha) \rho_U(du) \rho_U(du') \\
= & -\int \mathbb{E}_x \left[\langle x, J_\alpha \psi(\alpha; u) \delta\alpha \rangle \right] \rho_U(du) \\
& + \frac{1}{2} \iint \mathbb{E}_x \left[\langle J_\alpha \psi(\alpha; u) \delta\alpha, \psi(\alpha; u') \rangle + \langle \psi(\alpha; u), J_\alpha \psi(\alpha; u') \delta\alpha \rangle \right] \rho_U(du) \rho_U(du') \\
= & -\int \mathbb{E}_x \left[\langle x, J_\alpha \psi(\alpha; u) \delta\alpha \rangle \right] \rho_U(du) \\
& + \iint \mathbb{E}_x \left[\langle \psi(\alpha; u), J_\alpha \psi(\alpha; u') \delta\alpha \rangle \right] \rho_U(du) \rho_U(du') \\
= & \, \mathbb{E}_x \left[\left\langle -x, \left(\int J_\alpha \psi(\alpha; u) \rho_U(du) \right) \delta\alpha \right\rangle \right] \\
& + \mathbb{E}_x \left[\left\langle \int \psi(\alpha; u) \rho_U(du), \left(\int J_\alpha \psi(\alpha; u) \rho_U(du) \right) \delta\alpha \right\rangle \right] \\
= & \, \mathbb{E}_x \left[\left\langle -x + \int \psi(\alpha; u) \rho_U(du), \left(\int J_\alpha \psi(\alpha; u) \rho_U(du) \right) \delta\alpha \right\rangle \right] \\
= & \, \mathbb{E}_x \left[\left\langle \hat{x}(x) - x, \overline{J}(x; \rho_U, \rho_W) \delta\alpha \right\rangle \right].
\end{aligned}$$

Using the identity

$$\langle r, \overline{J} \delta\alpha \rangle_{\mathbb{R}^d} = \langle \overline{J}^\top r, \delta\alpha \rangle_{\mathbb{R}^b},$$

we can write

$$\delta\mathcal{R}(\alpha) = \mathbb{E}_x \left[\left\langle \overline{J}(x; \rho_U, \rho_W)^\top r(x; \rho_U, \rho_W), \delta\alpha \right\rangle \right].$$

Putting both parts of the chain rule together gives us the first variation of the unregularized reconstruction part with respect to ρ_W:

$$\frac{\delta\mathcal{R}}{\delta\rho_W}(w) = \mathbb{E}_x \left[\left\langle \overline{J}(x; \rho_U, \rho_W)^\top r(x; \rho_U, \rho_W), \phi(x; w) \right\rangle \right].$$

The encoder regularization term is linear in ρ_W, and its functional derivative is simply

$$\frac{\delta \left[\frac{\lambda}{6} \int \|w'\|^6 \rho_W(dw') \right]}{\delta\rho_W}(w) = \frac{\lambda}{6} \|w\|^6.$$

Therefore, the first variation of the full regularized risk with respect to the encoder measure is

$$\boxed{\Phi_W(w; \rho_U, \rho_W) = \mathbb{E}_x \left[\left\langle \overline{J}(x; \rho_U, \rho_W)^\top r(x; \rho_U, \rho_W), \phi(x; w) \right\rangle \right] + \frac{\lambda}{6} \|w\|^6.}$$

This can again be written as,

$$\Phi_W(w; \rho_U, \rho_W) = \mathbb{E}_x \left[\left\langle r(x; \rho_U, \rho_W), \overline{J}(x; \rho_U, \rho_W) \phi(x; w) \right\rangle \right] + \frac{\lambda}{6} \|w\|^6.$$

This is the continuous mean-field analogue of backpropagation, combined with the contribution of the sixth-power regularizer.

To obtain the corresponding descent vector field, we differentiate with respect to w. Recall that

$$J_w\phi(x;w) = \begin{bmatrix} w_2\,\sigma'(\langle w_1, x\rangle)x^\top & \sigma(\langle w_1, x\rangle)I_b \end{bmatrix} \in \mathbb{R}^{b\times D}.$$

Therefore,

$$\nabla_w \Phi_W(w;\rho_U,\rho_W) = \mathbb{E}_x\left[J_w\phi(x;w)^\top \overline{J}(x;\rho_U,\rho_W)^\top r(x;\rho_U,\rho_W)\right] + \lambda\|w\|^4 w.$$

Thus, if G_W denotes the unregularized descent vector field and V_W the regularized descent vector field, then

$$\begin{aligned} G_W(w;\rho_U,\rho_W) &:= -\mathbb{E}_x\left[J_w\phi(x;w)^\top \overline{J}(x;\rho_U,\rho_W)^\top r(x;\rho_U,\rho_W)\right], \\ V_W(w;\rho_U,\rho_W) &:= -\nabla_w \Phi_W(w;\rho_U,\rho_W) \\ &= G_W(w;\rho_U,\rho_W) - \lambda\|w\|^4 w. \end{aligned}$$

C.2 Coupled Vlasov-McKean Equations

The limiting dynamics of $\rho_{U,t}$ and $\rho_{W,t}$ are governed by the coupled continuity equations as given below,

$$\partial_t \rho_{U,t}(u) = \xi(t)\nabla_u \cdot \left(\rho_{U,t}(u)\nabla_u \Phi_U(u;\rho_{U,t},\rho_{W,t})\right),$$

and

$$\partial_t \rho_{W,t}(w) = \xi(t)\nabla_w \cdot \left(\rho_{W,t}(w)\nabla_w \Phi_W(w;\rho_{U,t},\rho_{W,t})\right).$$

Equivalently, in terms of the regularized descent vector fields V_U and V_W, these equations can be written as:

$$\partial_t \rho_{U,t} + \xi(t)\nabla_u \cdot \left(\rho_{U,t} V_U(u;\rho_{U,t},\rho_{W,t})\right) = 0,$$

and

$$\partial_t \rho_{W,t} + \xi(t)\nabla_w \cdot \left(\rho_{W,t} V_W(w;\rho_{U,t},\rho_{W,t})\right) = 0.$$

The first variations of the regularized risk are

$$\Phi_U(u;\rho_U,\rho_W) = \mathbb{E}_x\left[\langle r(x;\rho_U,\rho_W), \psi(\alpha(x;\rho_W);u)\rangle\right] + \frac{\lambda}{6}\|u\|^6,$$

and

$$\Phi_W(w;\rho_U,\rho_W) = \mathbb{E}_x\left[\left\langle \overline{J}(x;\rho_U,\rho_W)^\top r(x;\rho_U,\rho_W), \phi(x;w)\right\rangle\right] + \frac{\lambda}{6}\|w\|^6.$$

Their corresponding gradients are

$$\nabla_u \Phi_U(u;\rho_U,\rho_W) = \mathbb{E}_x\left[J_u\psi(\alpha(x;\rho_W);u)^\top r(x;\rho_U,\rho_W)\right] + \lambda\|u\|^4 u,$$

and

$$\nabla_w \Phi_W(w;\rho_U,\rho_W) = \mathbb{E}_x\left[J_w\phi(x;w)^\top \overline{J}(x;\rho_U,\rho_W)^\top r(x;\rho_U,\rho_W)\right] + \lambda\|w\|^4 w.$$

D SGD Updates

We now describe the finite-particle forward pass and the corresponding stochastic gradient descent dynamics. Throughout, we adopt the following iteration convention. At iteration $k \geq 0$, the current particle configurations are

$$U^k = (u_1^k, \ldots, u_N^k), \qquad W^k = (w_1^k, \ldots, w_N^k),$$

with associated empirical measures

$$\hat{\rho}_{U,k}^N = \frac{1}{N}\sum_{j=1}^N \delta_{u_j^k}, \qquad \hat{\rho}_{W,k}^N = \frac{1}{N}\sum_{i=1}^N \delta_{w_i^k}.$$

We now draw a sample $x_{k+1} \sim P$ and use it to update the particles from iteration k to iteration $k+1$. Thus, every quantity/term that appears on the right-hand side of the update are for current particles U^k, W^k and at the newly sampled data x_{k+1}.

The empirical bottleneck representation is

$$\alpha_k := \alpha(x_{k+1}; \hat{\rho}^N_{W,k}) = \frac{1}{N}\sum_{i=1}^N \phi(x_{k+1}; w_i^k) \in \mathbb{R}^b. \tag{15}$$

Using this bottleneck feature, the network output is

$$\hat{x}_k := \hat{x}_N(x_{k+1}; U^k, W^k) = \frac{1}{N}\sum_{j=1}^N \psi(\alpha_k; u_j^k) \in \mathbb{R}^d.$$

We denote the corresponding reconstruction residual by

$$r_k := \hat{x}_k - x_{k+1} \in \mathbb{R}^d. \tag{16}$$

The one-sample regularized loss evaluated at iteration k is therefore

$$\ell^N_{\lambda,k+1} := \ell^N_\lambda(x_{k+1}; U^k, W^k) = \frac{1}{2}\|r_k\|^2 + \frac{\lambda}{6N}\sum_{j=1}^N \|u_j^k\|^6 + \frac{\lambda}{6N}\sum_{i=1}^N \|w_i^k\|^6. \tag{17}$$

The step size is taken to be

$$s_k = \epsilon\, \xi(k\epsilon), \qquad k \geq 0.$$

The Jacobians needed below are

$$\begin{aligned}
J_\alpha \psi(\alpha; u) &= u_1 u_2^\top \sigma'(\langle u_2, \alpha\rangle) \in \mathbb{R}^{d\times b},\\
J_u \psi(\alpha; u)^\top z &= \begin{pmatrix} \sigma(\langle u_2, \alpha\rangle) z \\ \alpha\, \sigma'(\langle u_2, \alpha\rangle)\langle u_1, z\rangle \end{pmatrix}, \qquad z \in \mathbb{R}^d,\\
J_w \phi(x; w)^\top h &= \begin{pmatrix} x\, \sigma'(\langle w_1, x\rangle)\langle w_2, h\rangle \\ \sigma(\langle w_1, x\rangle) h \end{pmatrix}, \qquad h \in \mathbb{R}^b.
\end{aligned}$$

Here $u = (u_1, u_2)$ and $w = (w_1, w_2)$ are the encoder and decoder particle.

The derivative of the empirical network output with respect to the bottleneck variable is

$$\overline{J}_k := \frac{1}{N}\sum_{j=1}^N J_\alpha \psi(\alpha_k; u_j^k) \in \mathbb{R}^{d\times b}.$$

Equivalently, using the component notation for the j-th particle $u_j^k = (u_{j,1}^k, u_{j,2}^k)$,

$$\overline{J}_k = \frac{1}{N}\sum_{j=1}^N u_{j,1}^k (u_{j,2}^k)^\top \sigma'\big(\langle u_{j,2}^k, \alpha_k\rangle\big).$$

We further define the backpropagated bottleneck error

$$g_k := \overline{J}_k^\top r_k \in \mathbb{R}^b. \tag{18}$$

For a decoder particle $u = (u_1, u_2)$, define the single-sample unregularized descent field

$$\begin{aligned}
F_U\big(u; \hat{\rho}^N_{U,k}, \hat{\rho}^N_{W,k}, x_{k+1}\big) &:= -J_u \psi(\alpha_k; u)^\top r_k\\
&= -\begin{pmatrix} \sigma(\langle u_2, \alpha_k\rangle) r_k \\ \alpha_k\, \sigma'(\langle u_2, \alpha_k\rangle)\langle u_1, r_k\rangle \end{pmatrix}.
\end{aligned} \tag{19}$$

Similarly, for an encoder particle $w = (w_1, w_2)$, define

$$\begin{aligned}
F_W\big(w; \hat{\rho}^N_{U,k}, \hat{\rho}^N_{W,k}, x_{k+1}\big) &:= -J_w \phi(x_{k+1}; w)^\top g_k\\
&= -\begin{pmatrix} x_{k+1}\, \sigma'(\langle w_1, x_{k+1}\rangle)\langle w_2, g_k\rangle \\ \sigma(\langle w_1, x_{k+1}\rangle) g_k \end{pmatrix}.
\end{aligned} \tag{20}$$

We emphasize the normalization used in the particle dynamics. Since both empirical layers are normalized by $1/N$, differentiation of (17) with respect to a single particle produces an overall factor $1/N$. In particular,

$$\nabla_{u_j}\ell_\lambda^N(x_{k+1};U^k,W^k)=\frac{1}{N}\left[J_u\psi(\alpha_k;u_j^k)^\top r_k+\lambda\|u_j^k\|^4u_j^k\right],$$
$$\nabla_{w_i}\ell_\lambda^N(x_{k+1};U^k,W^k)=\frac{1}{N}\left[J_w\phi(x_{k+1};w_i^k)^\top g_k+\lambda\|w_i^k\|^4w_i^k\right].$$

We therefore use the N-rescaled particle-gradients

$$\theta^{k+1}=\theta^k-Ns_k\nabla_\theta\ell_\lambda^N.$$

Consequently, the decoder iteration is

$$\begin{aligned}u_j^{k+1}&=u_j^k-Ns_k\nabla_{u_j}\ell_\lambda^N(x_{k+1};U^k,W^k)\\&=u_j^k+s_k\left[F_U\big(u_j^k;\hat\rho_{U,k}^N,\hat\rho_{W,k}^N,x_{k+1}\big)-\lambda\|u_j^k\|^4u_j^k\right],\qquad j=1,\dots,N.\end{aligned}$$

Using the explicit form of F_U, this becomes

$$u_j^{k+1}=u_j^k-s_k\left[\begin{pmatrix}\sigma(\langle u_{j,2}^k,\alpha_k\rangle)r_k\\ \alpha_k\sigma'(\langle u_{j,2}^k,\alpha_k\rangle)\langle u_{j,1}^k,r_k\rangle\end{pmatrix}+\lambda\|u_j^k\|^4u_j^k\right].$$

Likewise, the encoder iteration is

$$\begin{aligned}w_i^{k+1}&=w_i^k-Ns_k\nabla_{w_i}\ell_\lambda^N(x_{k+1};U^k,W^k)\\&=w_i^k+s_k\left[F_W\big(w_i^k;\hat\rho_{U,k}^N,\hat\rho_{W,k}^N,x_{k+1}\big)-\lambda\|w_i^k\|^4w_i^k\right],\qquad i=1,\dots,N.\end{aligned}$$

Equivalently, using the component notation $w_i^k=(w_{i,1}^k,w_{i,2}^k)$, this is

$$w_i^{k+1}=w_i^k-s_k\left[\begin{pmatrix}x_{k+1}\sigma'(\langle w_{i,1}^k,x_{k+1}\rangle)\langle w_{i,2}^k,g_k\rangle\\ \sigma(\langle w_{i,1}^k,x_{k+1}\rangle)g_k\end{pmatrix}+\lambda\|w_i^k\|^4w_i^k\right].$$

E PRELIMINARIES

In this section, we recall and prove some of the results that we shall use in proving the main theorems. We first talk about the bounded Lipschitz metric between two probability measures. The **bounded Lipschitz metric** between two probability measures μ and ν is defined as

$$d_{\mathrm{BL}}(\mu,\nu):=\sup_{\substack{\|h\|_\infty\le1\\\|h\|_{\mathrm{Lip}}\le1}}\left|\int h(z)\mu(dz)-\int h(z)\nu(dz)\right|.$$

Lemma 1 (Folklore)**.** *If a specific test function g is bounded by a constant A (so $\|g\|_\infty\le A$) and is L-Lipschitz (so $\|g\|_{\mathrm{Lip}}\le L$), then we can say that*

$$\left|\int g(z)d(\mu-\nu)(z)\right|\le\max(A,L)\cdot d_{\mathrm{BL}}(\mu,\nu).$$

Proof. Let $K=\max(A,L)$. We define a new, rescaled function $\tilde g(x):=\frac{g(x)}{K}$. Now, the maximum value of $\tilde g$ is $\frac{A}{\max(A,L)}\le1$, so $\|\tilde g\|_\infty\le1$. Similarly, the Lipschitz constant of $\tilde g$ is $\frac{L}{\max(A,L)}\le1$, so $\|\tilde g\|_{\mathrm{Lip}}\le1$. Because $\tilde g$ satisfies both conditions, it is a valid candidate for the supremum in the definition of d_{BL}. Therefore, we can conclude that

$$\left|\int\tilde g(z)d(\mu-\nu)(z)\right|\le d_{\mathrm{BL}}(\mu,\nu).$$

Substituting $\tilde g(x)=g(x)/K$ back in and multiplying both sides by K yields the final result. □

We next define the **Wasserstein metric** between two probability measures as

$$W_p(\mu,\nu) := \inf_{\gamma\in\mathcal{C}(\mu,\nu)} \left(\int \|x-y\|_2^p \,\gamma(dx,dy)\right)^{1/p},$$

where $\mathcal{C}(\mu,\nu)$ is the set of all couplings of μ and ν. By Kantorovich-Rubinstein, one has the following for the case $p=1$:

$$W_1(\mu,\nu) = \sup_{\|h\|_{\mathrm{Lip}}\le 1} \left|\int h(z)\mu(dz) - \int h(z)\nu(dz)\right|.$$

As the set of all functions considered in the expression for d_{BL} is a subset of that of W_1, we conclude that

$$d_{\mathrm{BL}}(\mu,\nu) \le W_1(\mu,\nu).$$

Now, we state a result about the W_1 distance of two measures when one measure is the push-forward of the other measure. For a probability measure μ on $\mathbb{R}^d$ and a map measurable map $T:\mathbb{R}^d\to\mathbb{R}^d$, let us denote the pushforward measure $T_{\#}\mu$ by ν.

Lemma 2 (Wasserstein Bound of a Push-forward Measure)**.** *The 1-Wasserstein distance between μ and the pushforward measure ν is bounded by*

$$d_{\mathrm{BL}}(\mu,\nu) \le W_1(\mu,\nu) \le \sup_{x\in\mathrm{supp}(\mu)} \|x-T(x)\|_2.$$

Proof. By definition of Wasserstein distance, we have

$$W_1(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|_2 \,d\pi(x,y),$$

where $\Pi(\mu,\nu)$ is the set of all joint probability measures $\pi(x,y)$ with marginals distributions μ and ν. Let $\hat{\pi} := (\mathrm{id},T)_{\#}\mu$. For any measurable function $f(x,y)$, integrating with respect to $\hat{\pi}$ is equivalent to evaluating $f(x,T(x))$ over the base measure μ. So, we can say that

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} f(x,y)\,d\hat{\pi}(x,y) = \int_{\mathbb{R}^d} f(x,T(x))\,d\mu(x).$$

By setting $f(x,y)=f(x)$, it is clear the first marginal is μ. By setting $f(x,y)=g(y)$, the second marginal is $\int g(T(x))d\mu(x) = \int g(y)d\nu(y) = \nu$. Thus, $\hat{\pi}\in\Pi(\mu,\nu)$ is a valid coupling.

Now, since W_1 is defined as an infimum, it must be less than or equal to the cost of our specific coupling $\hat{\pi}$. So, we can conclude that

$$\begin{aligned} W_1(\mu,\nu) &\le \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|_2\,d\hat{\pi}(x,y) \\ &= \int_{\mathbb{R}^d} \|x-T(x)\|_2\,d\mu(x). \end{aligned}$$

We know that a function is bounded by its supremum, so pulling the supremum outside the integral gives us

$$\begin{aligned} \int_{\mathbb{R}^d} \|x-T(x)\|_2\,d\mu(x) &\le \int_{\mathbb{R}^d} \left(\sup_{z\in\mathrm{supp}(\mu)} \|z-T(z)\|_2\right) d\mu(x) \\ &= \sup_{z\in\mathrm{supp}(\mu)} \|z-T(z)\|_2 \int_{\mathbb{R}^d} 1\,d\mu(x) \\ &= \sup_{z\in\mathrm{supp}(\mu)} \|z-T(z)\|_2. \end{aligned}$$

This proves the right-hand inequality $W_1(\mu,\nu)\le \sup_x \|x-T(x)\|_2$.

As we know that

$$d_{\mathrm{BL}}(\mu,\nu) \le W_1(\mu,\nu),$$

combining both results completes the proof. □

We now state a version of Taylor's theorem for decoders that is used in the proof of the main theorems. To accommodate an arbitrary bottleneck dimension $b \geq 1$, we formulate this using the Jacobian matrix and a vector-valued bilinear form.

Lemma 3 (Taylor's Theorem for the decoder)**.** *Let $\psi(\alpha; u) \in \mathbb{R}^d$ be twice continuously differentiable with respect to the bottleneck variable $\alpha \in \mathbb{R}^b$. For any base point $\alpha^* \in \mathbb{R}^b$, step $\Delta\alpha \in \mathbb{R}^b$, and scalar $\lambda \in [0, 1]$, the exact second-order expansion is*

$$\psi(\alpha^*+\lambda\Delta\alpha; u) = \psi(\alpha^*; u)+\lambda\, J_\alpha\psi(\alpha^*; u)\,\Delta\alpha+\lambda^2 \int_0^1 (1-s)\,\nabla^2_\alpha\psi\big(\alpha^*+s\lambda\Delta\alpha; u\big)[\Delta\alpha, \Delta\alpha]\, ds.$$

Proof. Let $v = \lambda\Delta\alpha \in \mathbb{R}^b$. Now, we define the function $g : [0, 1] \to \mathbb{R}^d$ as

$$g(s) = \psi(\alpha^* + sv; u).$$

By the multivariate chain rule, the derivatives of $g(s)$ with respect to the scalar s are $g'(s) = J_\alpha\psi(\alpha^* + sv; u)v \in \mathbb{R}^d$ and $g''(s) = \nabla^2_\alpha\psi(\alpha^* + sv; u)[v, v] \in \mathbb{R}^d$.

By the fundamental theorem of calculus we can conclude that

$$g(1) - g(0) = \int_0^1 g'(s)\, ds.$$

We apply integration by parts to the right-hand side. After applying integration by parts, we get

$$\begin{aligned}\int_0^1 g'(s)\, ds &= \Big[- g'(s)(1 - s)\Big]_0^1 - \int_0^1 \big(- (1 - s)\big)g''(s)\, ds \\ &= g'(0) + \int_0^1 (1 - s)g''(s)\, ds.\end{aligned}$$

So,

$$g(1) = g(0) + g'(0) + \int_0^1 (1 - s)g''(s)\, ds.$$

By substituting the definitions of $g(0), g(1), g'(0)$ and $g''(s)$ back into this equation and factoring out λ and λ^2, we exactly yield the stated lemma. □

We reproduce the Azuma-Höeffding inequality (see Lemma A.1 in [43]).

Lemma 4 (Azuma-Höeffding bound)**.** *Let $(X_k)_{k\geq 0}$ be a martingale taking values in $\mathbb{R}^D$ with respect to the filtration $(\mathcal{F}_k)_{k\geq 0}$, with $X_0 = 0$. Assume that the following holds almost surely for all $k \geq 1$:*

$$\mathbb{E}\Big[e^{\langle\lambda, X_k - X_{k-1}\rangle} \,\Big|\, \mathcal{F}_{k-1}\Big] \leq e^{L^2\|\lambda\|_2^2/2}.$$

Then we have

$$\mathbb{P}\left(\max_{k\leq n} \|X_k\|_2 \geq 2L\sqrt{n}(\sqrt{D} + t)\right) \leq e^{-t^2}.$$

Although the d_{BL} and W_p metrics are key parts of the analysis, to rigorously measure the distance between the empirical and population measures, we would like to evaluate the measures strictly through the lens of the network's predictive functions. We formalize this idea of metric induced by such a parametric family below.

Let $\mathcal{F} = \{f_\theta\}_{\theta\in\Theta}$ be a family of test functions from Ω to $\mathbb{R}^p$, which is indexed by a parameter set Θ. For any two probability measures μ and ν supported on Ω, the **parametric distance** $d_\mathcal{F}$ between them is defined as

$$d_\mathcal{F}(\mu, \nu) = \sup_{f\in\mathcal{F}} \left\| \int_\Omega f(\omega)\, d\mu(\omega) - \int_\Omega f(\omega)\, d\nu(\omega) \right\|_2.$$

For scalar-valued function classes ($p = 1$), the Euclidean norm simplifies to the standard absolute value. We now prove a result from empirical process theory that bounds the statistical concentration of the specific parametric functions evaluated by our network.

Lemma 5 (Parametric Concentration of Empirical Measures)**.** *Let μ be a probability measure supported on a compact parameter space $\Omega \subset \mathbb{R}^D$. Let $X_1, \ldots, X_N \sim \mu$ be i.i.d., and define the empirical measure $\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$. Let $\mathcal{F} = \{f_\theta : \Omega \to \mathbb{R}^p\}_{\theta\in\Theta}$ be a family of test functions indexed by a compact set $\Theta \subset \mathbb{R}^d$ (e.g., the data domain) with diameter R. Assume that for all $\theta \in \Theta$, $\sup_{w\in\Omega} \|f_\theta(w)\|_2 \le M$, and the map $\theta \mapsto f_\theta(w)$ is L-Lipschitz for all $w \in \Omega$ (i.e., $\|f_\theta(w) - f_{\theta'}(w)\|_2 \le L\|\theta - \theta'\|_2$).*

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$d_{\mathcal{F}}(\mu, \mu_N) \le \frac{1}{\sqrt{N}}\left(C + M\sqrt{d\log N} + M\sqrt{2\log(1/\delta)}\right),$$

where $C := 2L + M\sqrt{2d\log(3R)} + M\sqrt{2\log 2}$.

Proof. Let us first note that the parametric distance $d_{\mathcal{F}}$ for μ and μ_N reduces to

$$d_{\mathcal{F}}(\mu, \mu_N) = \sup_{\theta\in\Theta}\left\|\mathbb{E}_\mu[f_\theta] - \frac{1}{N}\sum_{i=1}^N f_\theta(X_i)\right\|_2.$$

We aim to bound the uniform deviation over the parametric class $\mathcal{F}$ for which we need to construct a cover of the finite-dimensional index set $\Theta \subset \mathbb{R}^d$. Let Θ_ϵ be an ϵ-cover of Θ in the Euclidean norm. By standard volumetric arguments we can always guarantee that the covering number is bounded by $|\Theta_\epsilon| \le (3R/\epsilon)^d$. For any $\theta \in \Theta$, there exists a $\theta' \in \Theta_\epsilon$ such that $\|\theta - \theta'\|_2 \le \epsilon$. By the L-Lipschitz property of the test functions with respect to the index θ, we can decompose the error using the triangle inequality:

$$\begin{aligned}\left\|\mathbb{E}_\mu[f_\theta] - \frac{1}{N}\sum_{i=1}^N f_\theta(X_i)\right\|_2 &\le \|\mathbb{E}_\mu[f_\theta] - \mathbb{E}_\mu[f_{\theta'}]\|_2 + \left\|\mathbb{E}_\mu[f_{\theta'}] - \frac{1}{N}\sum_{i=1}^N f_{\theta'}(X_i)\right\|_2 \\ &\quad + \left\|\frac{1}{N}\sum_{i=1}^N f_{\theta'}(X_i) - \frac{1}{N}\sum_{i=1}^N f_\theta(X_i)\right\|_2 \\ &\le 2L\epsilon + \left\|\mathbb{E}_\mu[f_{\theta'}] - \frac{1}{N}\sum_{i=1}^N f_{\theta'}(X_i)\right\|_2.\end{aligned}$$

Taking the supremum over all $\theta \in \Theta$, the worst-case continuous error is bounded by the worst-case error on the finite cover:

$$\sup_{\theta\in\Theta}\left\|\mathbb{E}_\mu[f_\theta] - \frac{1}{N}\sum_{i=1}^N f_\theta(X_i)\right\|_2 \le 2L\epsilon + \max_{\theta'\in\Theta_\epsilon}\left\|\mathbb{E}_\mu[f_{\theta'}] - \frac{1}{N}\sum_{i=1}^N f_{\theta'}(X_i)\right\|_2.$$

For a fixed $\theta' \in \Theta_\epsilon$, the random vectors $f_{\theta'}(X_i) \in \mathbb{R}^p$ are iid, and bounded by M. Applying the vector version of Hoeffding's inequality for this specific θ', we get,

$$\mathbb{P}\left(\left\|\mathbb{E}_\mu[f_{\theta'}] - \frac{1}{N}\sum_{i=1}^N f_{\theta'}(X_i)\right\|_2 \ge \tau\right) \le 2\exp\left(-\frac{N\tau^2}{2M^2}\right).$$

We want this to hold for all elements in the cover. So applying a union bound over Θ_ϵ, the probability that the maximum fluctuation exceeds τ is

$$\mathbb{P}\left(\max_{\theta'\in\Theta_\epsilon}\left\|\mathbb{E}_\mu[f_{\theta'}] - \frac{1}{N}\sum_{i=1}^N f_{\theta'}(X_i)\right\|_2 \ge \tau\right) \le 2\left(\frac{3R}{\epsilon}\right)^d \exp\left(-\frac{N\tau^2}{2M^2}\right).$$

Setting the right-hand side equal to δ and solving for τ gives us

$$\tau = M\sqrt{\frac{2}{N}}\sqrt{d\log\left(\frac{3R}{\epsilon}\right) + \log\left(\frac{2}{\delta}\right)}.$$

The total error is bounded by $2L\epsilon + \tau$. If we choose the covering radius $\epsilon = 1/\sqrt{N}$. Substituting this ϵ, then $\log(3R/\epsilon) = \log(3R) + \frac{1}{2}\log N$.

Using the subadditivity $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ repeatedly, we can expand τ:

$$\begin{aligned}\tau &\leq M\sqrt{\frac{2}{N}}\left(\sqrt{\frac{d}{2}\log N} + \sqrt{d\log(3R)} + \sqrt{\log 2} + \sqrt{\log(1/\delta)}\right) \\ &= \frac{1}{\sqrt{N}}\left(M\sqrt{d\log N} + M\sqrt{2d\log(3R)} + M\sqrt{2\log 2} + M\sqrt{2\log(1/\delta)}\right).\end{aligned}$$

Adding the $2L\epsilon = \frac{2L}{\sqrt{N}}$ term and letting $C := 2L + M\sqrt{2d\log(3R)} + M\sqrt{2\log 2}$, we get

$$d_{\mathcal{F}}(\mu, \mu_N) \leq \frac{1}{\sqrt{N}}\left(C + M\sqrt{d\log N} + M\sqrt{2\log(1/\delta)}\right).$$

□

Lemma 6 (Relationship between Parametric and Bounded Lipschitz Distances)**.** *Let $\mathcal{F}$ be a family of test functions defined on a domain Ω. Let every $f \in \mathcal{F}$ be uniformly bounded by M and is M-Lipschitz continuous. Then, for any two probability measures μ and ν on Ω, $d_{\mathcal{F}}(\mu, \nu) \leq M d_{\mathrm{BL}}(\mu, \nu)$.*

Proof. Lets evaluate the integral difference for a single, fixed function $f \in \mathcal{F}$. As the norm of any vector can be expressed as the supremum of its inner product with all unit vectors, we get

$$\begin{aligned}\left\|\int f d\mu - \int f d\nu\right\|_2 &= \sup_{\|v\|_2 \leq 1}\left\langle v, \int f d\mu - \int f d\nu\right\rangle \\ &= \sup_{\|v\|_2 \leq 1}\left(\int \langle v, f(x)\rangle d\mu(x) - \int \langle v, f(x)\rangle d\nu(x)\right).\end{aligned}$$

Now fix an arbitrary unit vector $v \in \mathbb{R}^p$ with $\|v\|_2 \leq 1$. Let $h(x) = \langle v, f(x)\rangle$ be the projection of $f(x)$ on to v. We now show that this is boundedness and Lipschitz. By the Cauchy-Schwarz inequality and our assumptions on $\mathcal{F}$, the function h is bounded,

$$|h(x)| \leq \|v\|_2 \|f(x)\|_2 \leq 1 \cdot M = M.$$

Similarly, h inherits the Lipschitz continuity of f:

$$|h(x) - h(y)| = |\langle v, f(x) - f(y)\rangle| \leq \|v\|_2 \|f(x) - f(y)\|_2 \leq M\|x - y\|_2.$$

Now $h(x)/M$ is bounded by 1 and is 1-Lipschitz and hence by definition,

$$\frac{1}{M}\int h(x) d\mu(x) - \int h(x) d\nu(x) \leq d_{\mathrm{BL}}(\mu, \nu).$$

which implies

$$\int \langle v, f(x)\rangle d\mu(x) - \int \langle v, f(x)\rangle d\nu(x) \leq M d_{\mathrm{BL}}(\mu, \nu).$$

Because this inequality holds for every unit vector v, it must hold for the supremum over all unit vectors, which implies that $\left\|\int f d\mu - \int f d\nu\right\|_2 \leq M d_{\mathrm{BL}}(\mu, \nu)$. Finally, taking the supremum over all functions $f \in \mathcal{F}$ yields the target result

$$\sup_{f \in \mathcal{F}}\left\|\int f d\mu - \int f d\nu\right\|_2 \leq M d_{\mathrm{BL}}(\mu, \nu).$$

□

F BOUNDEDNESS OF THE COUPLED PDE AND SGD

One of our assumptions was that the data distribution is supported on some bounded ball. For our setting to be consistent, the SDG updates need to be within the ball. In this section, we show exactly this, we prove that the regularized SDG updates and the corresponding mean-field dynamics both remain uniformly bounded. We first bound all the quantities in the forward pass and in the descent fields, and the corresponding population bounds. We then use these to construct a fixed ball which is invariant under both the SDG iterates and the characteristic dynamics associated with the coupled PDE.

It is important to stress here that the sixth-order regularizer plays a crucial role. As we shall see in the proof, the radial component of the unregualrized versions of the encoder and decoder fields grows at most as a fifth order of the radius. So for a sufficiently large radius, the sixth-order regularization is needed. One can choose a lower order regularizer, in which case one either looses the ability to find a uniform radius or the computations become messy.

F.1 STANDING ASSUMPTIONS

We recall the assumptions here for reference.

Assumption 1. *There exist finite constants M_x, M_ξ, L_ξ, R_0 such that:*

(i) the data distribution has bounded support, $\mathrm{supp}(P) \subseteq B_d(0, M_x)$*;*

(ii) the activation satisfies $\sigma \in C_b^2(\mathbb{R})$*, with* $S_0 := \|\sigma\|_\infty$*,* $S_1 := \|\sigma'\|_\infty$*,* $S_2 := \|\sigma''\|_\infty$*;*

(iii) the learning-rate schedule $\xi : [0,\infty) \to [0,\infty)$ *is* L_ξ*-Lipschitz and* $\|\xi\|_\infty \le M_\xi$*;*

(iv) the initial population measures have compact support, $\mathrm{supp}(\rho_{U,0}) \cup \mathrm{supp}(\rho_{W,0}) \subseteq B_D(0, R_0)$*, and the finite particles are initialized i.i.d. from them, so that* $\max_j \|u_j^0\|_2 \le R_0$ *and* $\max_i \|w_i^0\|_2 \le R_0$ *for every realization.*

F.2 BOUNDS FOR THE SAMPLE AND POPULATION FIELDS

We first derive the estimates needed for the finite-particle SGD dynamics. Fix $R \ge 1$, and suppose that at some iteration k,

$$\max_{1\le j\le N} \|u_j^k\|_2 \le R, \qquad \max_{1\le i\le N} \|w_i^k\|_2 \le R. \tag{21}$$

Equivalently,

$$\mathrm{supp}\, \hat\rho_{U,k}^N \subseteq B_D(0,R), \qquad \mathrm{supp}\, \hat\rho_{W,k}^N \subseteq B_D(0,R).$$

For notational convenience, we define

$$E_R := M_x + S_0 R. \tag{22}$$

Lemma 7 (Bounds for the single-sample quantities)**.** *Suppose that Assumption 1 holds and that the particles at iteration k satisfy* (21) *for some $R \ge 1$. Then, for every $x_{k+1} \in \mathrm{supp}(P)$, the quantities appearing in the forward and backward passes satisfy: $\|\alpha_k\|_2, \|\hat x_k\|_2, \|r_k\|_2 = \mathcal{O}(R)$ and $\|\overline{J}_k\|, \|g_k\|_2 = \mathcal{O}(R^2)$*

Furthermore, for every $u = (u_1, u_2) \in B_D(0,R)$ and $w = (w_1, w_2) \in B_D(0,R)$, the single-sample descent fields satisfy

$$\begin{aligned} &\left\|F_U\big(u; \hat\rho_{U,k}^N, \hat\rho_{W,k}^N, x_{k+1}\big)\right\|_2 = \mathcal{O}(R^3), \quad \left\|F_W\big(w; \hat\rho_{U,k}^N, \hat\rho_{W,k}^N, x_{k+1}\big)\right\|_2 = \mathcal{O}(R^4), \\ &\sup \max\left\{\langle u, F_U(u; \rho_U, \rho_W, x)\rangle, \langle w, F_W(w; \rho_U, \rho_W, x)\rangle\right\} = \mathcal{O}(R^5), \end{aligned} \tag{23}$$

where the supremum is taken over $x \in \mathrm{supp}(P), u, w \in B_D(0,R)$, and over probability measures ρ_U, ρ_W supported in $B_D(0,R)$.

Proof. Recall that the encoder neuron is

$$\phi(x; w) = w_2\, \sigma(\langle w_1, x\rangle), \qquad w = (w_1, w_2).$$

Writing $w_i^k = (w_{i,1}^k, w_{i,2}^k)$, the assumption $\|w_i^k\|_2 \le R$ implies $\|w_{i,2}^k\|_2 \le R$. Hence, using $|\sigma| \le S_0$,

$$\|\phi(x_{k+1}; w_i^k)\|_2 \le S_0 R.$$

Therefore,

$$\|\alpha_k\|_2 \le \frac{1}{N}\sum_{i=1}^{N} \|\phi(x_{k+1}; w_i^k)\|_2 \le S_0 R.$$

Similarly, the decoder neuron is

$$\psi(\alpha; u) = u_1\, \sigma(\langle u_2, \alpha\rangle), \qquad u = (u_1, u_2).$$

Writing $u_j^k = (u_{j,1}^k, u_{j,2}^k)$, the bound $\|u_j^k\|_2 \le R$ gives $\|u_{j,1}^k\|_2 \le R$, and consequently

$$\|\psi(\alpha_k; u_j^k)\|_2 \le S_0 R.$$

Thus,

$$\|\hat{x}_k\|_2 \le \frac{1}{N}\sum_{j=1}^{N} \|\psi(\alpha_k; u_j^k)\|_2 \le S_0 R.$$

Since $r_k = \hat{x}_k - x_{k+1}$ and $\|x_{k+1}\|_2 \le M_x$,

$$\|r_k\|_2 \le \|\hat{x}_k\|_2 + \|x_{k+1}\|_2 \le S_0 R + M_x (=: E_R).$$

Next,

$$J_\alpha \psi(\alpha_k; u_j^k) = u_{j,1}^k (u_{j,2}^k)^\top \sigma'\big(\langle u_{j,2}^k, \alpha_k\rangle\big).$$

Since $\|u_{j,1}^k\|_2, \|u_{j,2}^k\|_2 \le R$ and $|\sigma'| \le S_1$,

$$\|J_\alpha \psi(\alpha_k; u_j^k)\| \le S_1 R^2.$$

Hence,

$$\|\overline{J}_k\| \le \frac{1}{N}\sum_{j=1}^{N} \|J_\alpha \psi(\alpha_k; u_j^k)\| \le S_1 R^2.$$

Since $g_k = \overline{J}_k^\top r_k$, we further obtain

$$\|g_k\|_2 \le \|\overline{J}_k\|\, \|r_k\|_2 \le S_1 R^2 E_R.$$

We next estimate the decoder sample field. For $u = (u_1, u_2)$,

$$F_U = -\begin{pmatrix} \sigma(\langle u_2, \alpha_k\rangle) r_k \\ \alpha_k \sigma'(\langle u_2, \alpha_k\rangle)\langle u_1, r_k\rangle \end{pmatrix}.$$

Since $\|u_1\|_2 \le R$,

$$|\langle u_1, r_k\rangle| \le R\|r_k\|_2 \le R E_R.$$

Using also $\|\alpha_k\|_2 \le S_0 R$, $\|r_k\|_2 \le E_R$, $|\sigma| \le S_0$, and $|\sigma'| \le S_1$, we obtain

$$\begin{aligned} \|F_U\|_2^2 &= |\sigma(\langle u_2, \alpha_k\rangle)|^2 \|r_k\|_2^2 + \|\alpha_k\|_2^2 |\sigma'(\langle u_2, \alpha_k\rangle)|^2 |\langle u_1, r_k\rangle|^2 \\ &\le S_0^2 E_R^2 + S_0^2 S_1^2 R^4 E_R^2 = S_0^2 E_R^2 (1 + S_1^2 R^4). \end{aligned}$$

Therefore,

$$\|F_U\|_2 \le S_0 E_R \sqrt{1 + S_1^2 R^4} (=: B_U(R)).$$

For the encoder sample field, with $w = (w_1, w_2)$,

$$F_W = -\begin{pmatrix} x_{k+1} \sigma'(\langle w_1, x_{k+1}\rangle)\langle w_2, g_k\rangle \\ \sigma(\langle w_1, x_{k+1}\rangle) g_k \end{pmatrix}.$$

Since $\|w_2\|_2 \le R$,

$$|\langle w_2, g_k\rangle| \le R\|g_k\|_2.$$

Thus, using $\|x_{k+1}\|_2 \le M_x$,

$$\begin{aligned}\|F_W\|_2^2 &= \|x_{k+1}\|_2^2|\sigma'(\langle w_1, x_{k+1}\rangle)|^2|\langle w_2, g_k\rangle|^2 + |\sigma(\langle w_1, x_{k+1}\rangle)|^2\|g_k\|_2^2 \\ &\le \left(S_1^2M_x^2R^2 + S_0^2\right)\|g_k\|_2^2 \le \left(S_1^2M_x^2R^2 + S_0^2\right)S_1^2R^4E_R^2.\end{aligned}$$

Consequently,

$$\|F_W\|_2 \le S_1R^2E_R\sqrt{S_0^2 + S_1^2M_x^2R^2}(=: B_W(R)).$$

It remains to obtain a common polynomial bound. Since $R \ge 1$,

$$E_R = M_x + S_0R \le (M_x + S_0)R,$$

while

$$\sqrt{1 + S_1^2R^4} \le (1 + S_1)R^2, \qquad \sqrt{S_0^2 + S_1^2M_x^2R^2} \le (S_0 + S_1M_x)R.$$

Let $C_{\mathrm{grad}} := (M_x + S_0)\max\{S_0(1 + S_1), S_1(S_0 + S_1M_x)\}$, then

$$B_U(R) \le (M_x + S_0)S_0(1 + S_1)R^3 \le C_{\mathrm{grad}}R^4,$$

and

$$B_W(R) \le (M_x + S_0)S_1(S_0 + S_1M_x)R^4 \le C_{\mathrm{grad}}R^4.$$

So, for any $u, w \in B_D(0, R)$, Cauchy–Schwarz yields

$$\langle u, F_U\rangle \le \|u\|_2\|F_U\|_2 \le R \cdot C_{\mathrm{grad}}R^4, \qquad \langle w, F_W\rangle \le \|w\|_2\|F_W\|_2 \le R \cdot C_{\mathrm{grad}}R^4.$$

Taking the supremum gives us the last bound. □

Note that the estimate (23) shows that the possible outward contribution of the unregularized dynamics is at most of order R^5. On the other hand, the regularization field satisfies

$$\left\langle v, -\lambda\|v\|_2^4v\right\rangle = -\lambda\|v\|_2^6.$$

Thus, the sixth-order regularization produces a strictly higher-order inward radial contribution.

We now state the population counterpart of Lemma 7.

Lemma 8 (Bounds for the population quantities)**.** *Suppose that Assumption 1 holds and that, for some $R \ge 1$, $\mathrm{supp}(\rho_U) \subseteq B_D(0, R)$ and $\mathrm{supp}(\rho_W) \subseteq B_D(0, R)$. Then, uniformly over $x \in \mathrm{supp}(P)$ we get the following bounds: $\|\alpha(x; \rho_W)\|_2, \|\hat{x}(x; \rho_U, \rho_W)\|_2, \|r(x; \rho_U, \rho_W)\|_2 = \mathcal{O}(R)$ and $\|\overline{J}(x; \rho_U, \rho_W)\| = \mathcal{O}(R^2)$ $\left\|\overline{J}(x; \rho_U, \rho_W)^\top r(x; \rho_U, \rho_W)\right\|_2 = \mathcal{O}(R^3)$ Moreover, the unregularized population fields satisfy*

$$\max\{\|G_U(u; \rho_U, \rho_W)\|_2, \|G_W(w; \rho_U, \rho_W)\|_2\} = \mathcal{O}(R^4), \text{ and}$$
$$\max\{\langle u, G_U(u; \rho_U, \rho_W)\rangle, \langle w, G_W(w; \rho_U, \rho_W)\rangle\} = \mathcal{O}(R^5).$$

Proof. The proof follows from exactly the same calculations as in Lemma 7. The only difference is that the empirical averages defining α_k, $\hat{x}_k$, and $\overline{J}_k$ are replaced by integrals with respect to the probability measures ρ_W and ρ_U.

For example,

$$\alpha(x; \rho_W) = \int \phi(x; w)\rho_W(dw),$$

and, since ρ_W is supported in $B_D(0, R)$,

$$\|\alpha(x; \rho_W)\|_2 \le \int \|\phi(x; w)\|_2\rho_W(dw) \le S_0R.$$

The remaining forward estimates follow in exactly the same manner.

Furthermore,

$$G_U(u; \rho_U, \rho_W) = \mathbb{E}_{x\sim P}\left[F_U(u; \rho_U, \rho_W, x)\right]$$

and

$$G_W(w; \rho_U, \rho_W) = \mathbb{E}_{x\sim P}\left[F_W(w; \rho_U, \rho_W, x)\right].$$

Therefore, Jensen's inequality and Lemma 7 give

$$\|G_U\|_2 \le B_U(R), \qquad \|G_W\|_2 \le B_W(R).$$

The radial estimates follow similarly by taking expectations in (23). □

F.3 Uniform boundedness of the SGD iterates

We now construct a fixed ball that contains every SGD iterate. Define

$$R_* := \max\left\{R_0, 1, \frac{128 C_{\mathrm{grad}}}{\lambda}\right\}. \tag{24}$$

By the definition of R_*,

$$C_{\mathrm{grad}} \le \frac{\lambda R_*}{128}.$$

To save space let us denote, $A(R) := \sup \max\{\langle u, F_U(u;\rho_U,\rho_W,x)\rangle, \langle w, F_W(w;\rho_U,\rho_W,x)\rangle\}$. Using (23),

$$A(R_*) \le C_{\mathrm{grad}} R_*^5 \le \frac{\lambda R_*^6}{128}. \tag{25}$$

If we let $B(R) := \max\{B_U(R), B_W(R)\}$, and define $M_* := B(R_*) + \lambda R_*^5$, then M_* is an upper bound for the norm of the full regularized sample field on $B_D(0, R_*)$, since

$$\left\|F_U - \lambda\|u\|_2^4 u\right\|_2 \le B(R_*) + \lambda R_*^5 = M_*, \tag{26}$$

$$\left\|F_W - \lambda\|w\|_2^4 w\right\|_2 \le B(R_*) + \lambda R_*^5 = M_*. \tag{27}$$

Finally, define

$$s_* := \min\left\{\frac{R_*}{2M_*}, \frac{\lambda R_*^6}{64 M_*^2}\right\}.$$

Theorem 6 (Uniform boundedness of the regularized SGD iterates)**.** *Suppose that Assumption 1 holds and let $\lambda > 0$. Suppose further that the SGD step size $s_k = \epsilon\xi(k\epsilon)$ satisfies*

$$\epsilon M_\xi \le s_*. \tag{28}$$

Then, for every width $N \ge 1$, every realization of the bounded initialization, and every data sequence $x_1, x_2, \ldots \in \mathrm{supp}(P)$, the regularized SGD iterates satisfy

$$\max_{1\le j\le N} \|u_j^k\|_2 \le R_*, \qquad \max_{1\le i\le N} \|w_i^k\|_2 \le R_*, \qquad k \ge 0. \tag{29}$$

Equivalently,

$$\mathrm{supp}\,\hat\rho_{U,k}^N \cup \mathrm{supp}\,\hat\rho_{W,k}^N \subseteq B_D(0, R_*), \qquad k \ge 0. \tag{30}$$

The radius R_* is independent of N, k, and the realization of the data sequence.

Proof. We prove the theorem by induction on k. At $k = 0$, Assumption 1 and the fact that $R_* \ge R_0$ give

$$\|u_j^0\|_2 \le R_*, \qquad \|w_i^0\|_2 \le R_* \qquad \text{for every} \quad i, j.$$

Suppose now that the result holds at iteration k,

$$\max_j \|u_j^k\|_2 \le R_*, \qquad \max_i \|w_i^k\|_2 \le R_*.$$

We prove that the same is true at iteration $k+1$.

The decoder and encoder updates have the common form

$$v^+ = v + s_k H, \tag{31}$$

where v denotes either u_j^k or w_i^k, and $H = F - \lambda\|v\|_2^4 v$. Here F denotes the corresponding unregularized single-sample field F_U or F_W.

By (26)–(27), we can conclude that

$$\|H\|_2 \le M_*. \tag{32}$$

We divide the argument into two cases. The cases are case 1: $\|v\|_2 \le \frac{R_*}{2}$ and case 2: $\frac{R_*}{2} \le \|v\|_2 \le R_*$.

We first consider *case 1:* $\|v\|_2 \le \frac{R_*}{2}$. Using (31) and (32), we can conclude that

$$\|v^+\|_2 \le \|v\|_2 + s_k\|H\|_2 \le \frac{R_*}{2} + s_k M_*.$$

Since $s_k = \epsilon\xi(k\epsilon) \le \epsilon M_\xi \le s_*$, and $s_* \le \frac{R_*}{2M_*}$, we obtain $s_k M_* \le \frac{R_*}{2}$. Hence

$$\|v^+\|_2 \le R_*.$$

We now consider *case 2:* $\frac{R_*}{2} < \|v\|_2 \le R_*$. We now use the radial effect of the regularization. Squaring (31) gives

$$\begin{aligned}\|v^+\|_2^2 &= \|v\|_2^2 + 2s_k\langle v, H\rangle + s_k^2\|H\|_2^2 \\ &= \|v\|_2^2 + 2s_k\left(\langle v, F\rangle - \lambda\|v\|_2^6\right) + s_k^2\|H\|_2^2. \end{aligned} \tag{33}$$

By Lemma 7, we obtain that $\langle v, F\rangle \le A(R_*)$. Since $\|v\|_2 > \frac{R_*}{2}$, we have

$$\|v\|_2^6 \ge \frac{R_*^6}{64}.$$

So, we can conclude that,

$$\langle v, F\rangle - \lambda\|v\|_2^6 \le A(R_*) - \frac{\lambda R_*^6}{64} \le \frac{\lambda R_*^6}{128} - \frac{\lambda R_*^6}{64} = -\frac{\lambda R_*^6}{128}. \tag{34}$$

Substituting (34) and (32) into (33) yields

$$\|v^+\|_2^2 \le \|v\|_2^2 - \frac{s_k\lambda R_*^6}{64} + s_k^2 M_*^2. \tag{35}$$

By the definition of s_*,

$$s_k \le s_* \le \frac{\lambda R_*^6}{64M_*^2}.$$

Therefore,

$$s_k^2 M_*^2 \le \frac{s_k\lambda R_*^6}{64}.$$

Using this in (35), we obtain

$$\|v^+\|_2^2 \le \|v\|_2^2 \le R_*^2.$$

Thus,

$$\|v^+\|_2 \le R_*.$$

Both cases therefore show that a particle lying in $B_D(0, R_*)$ at iteration k remains in $B_D(0, R_*)$ at iteration $k+1$. Since the same argument applies to every encoder and decoder particle, we conclude that

$$\max_j \|u_j^{k+1}\|_2 \le R_*, \qquad \max_i \|w_i^{k+1}\|_2 \le R_*.$$

Induction over k proves (29).

Finally, by the definitions

$$\hat\rho_{U,k}^N = \frac{1}{N}\sum_{j=1}^N \delta_{u_j^k}, \qquad \hat\rho_{W,k}^N = \frac{1}{N}\sum_{i=1}^N \delta_{w_i^k},$$

the statement (30) follows immediately. □

F.4 UNIFORM BOUNDEDNESS OF THE COUPLED PDE

We now prove the corresponding boundedness result for the coupled PDE. Recall that the regularized population descent fields are

$$V_U(u;\rho_U,\rho_W) = G_U(u;\rho_U,\rho_W) - \lambda\|u\|_2^4 u,$$
$$V_W(w;\rho_U,\rho_W) = G_W(w;\rho_U,\rho_W) - \lambda\|w\|_2^4 w.$$

The coupled Vlasov–McKean system is

$$\partial_t \rho_{U,t} + \nabla_u \cdot (\rho_{U,t}\, \xi(t) V_U(\cdot;\rho_{U,t},\rho_{W,t})) = 0, \tag{36}$$
$$\partial_t \rho_{W,t} + \nabla_w \cdot (\rho_{W,t}\, \xi(t) V_W(\cdot;\rho_{U,t},\rho_{W,t})) = 0. \tag{37}$$

The associated characteristic equations are

$$\dot{u}_t = \xi(t) V_U(u_t;\rho_{U,t},\rho_{W,t}),$$
$$\dot{w}_t = \xi(t) V_W(w_t;\rho_{U,t},\rho_{W,t}).$$

Theorem 7 (Uniform boundedness of the coupled mean-field dynamics)**.** *Suppose that Assumption 1 holds and let $\lambda > 0$. Let $(\rho_{U,t},\rho_{W,t})_{t\geq 0}$ be a characteristic solution of* (36)–(37) *with initial data $\rho_{U,0}$ and $\rho_{W,0}$. Then, for the same radius R_* defined in* (24)*,*

$$\operatorname{supp}(\rho_{U,t}) \cup \operatorname{supp}(\rho_{W,t}) \subseteq B_D(0,R_*), \qquad t \geq 0. \tag{38}$$

Equivalently, every decoder and encoder characteristic starting in the support of the corresponding initial measure satisfies

$$\|u_t\|_2 \leq R_*, \qquad \|w_t\|_2 \leq R_*, \qquad t \geq 0. \tag{39}$$

Proof. We prove that the closed ball $B_D(0,R_*)$ is invariant under both characteristic flows. The only point requiring some care is that the population vector fields depend on the evolving measures. To avoid any circularity, we first work on a slightly larger ball.

By (25), $A(R_*) \leq \lambda R_*^6/128$. Since $A(R) \leq RB(R) \leq C_{\mathrm{grad}} R^5 \leq \lambda R_* R^5/128$ and $\lambda R_* R^5/128$ is a continuous function in R, we may choose $\varepsilon > 0$ sufficiently small so that

$$A(R_*+\varepsilon) \leq \lambda R_*(R_*+\varepsilon)^5/128 < \lambda R_*^6.$$

Set $c_* := \lambda R_*^6 - A(R_*+\varepsilon) > 0$.

Suppose that at some time t the two population measures satisfy $\operatorname{supp}(\rho_{U,t}) \cup \operatorname{supp}(\rho_{W,t}) \subseteq B_D(0,R_*+\varepsilon)$. Lemma 8, applied with $R = R_* + \varepsilon$, then gives, for every decoder particle u with $\|u\|_2 = R_*$,

$$\langle u, G_U(u;\rho_{U,t},\rho_{W,t})\rangle \leq A(R_*+\varepsilon),$$

and the analogous estimate holds for the encoder field. Hence, using the regularized vector fields, for every $\|u\|_2 = R_*$ and $\|w\|_2 = R_*$,

$$\langle u, V_U(u;\rho_{U,t},\rho_{W,t})\rangle \leq -c_* < 0, \qquad \langle w, V_W(w;\rho_{U,t},\rho_{W,t})\rangle \leq -c_* < 0. \tag{40}$$

Thus, as long as both population measures remain in the slightly larger ball $B_D(0,R_*+\varepsilon)$, both vector fields point strictly inward on the boundary $\partial B_D(0,R_*)$.

Let v_t denote either a decoder characteristic u_t or an encoder characteristic w_t, and let V denote the corresponding field V_U or V_W. In either case, $\dot{v}_t = \xi(t) V(v_t;\rho_{U,t},\rho_{W,t})$.

Let $P_* : \mathbb{R}^D \to B_D(0,R_*)$ denote the metric projection onto the closed ball, namely

$$P_*(v) = \begin{cases} v, & \|v\|_2 \leq R_*, \\ R_* \dfrac{v}{\|v\|_2}, & \|v\|_2 > R_*, \end{cases}$$

and define $d_*(v) := \|v - P_*(v)\|_2 = \max\{\|v\|_2 - R_*, 0\}$, which is exactly the Euclidean distance from v to $B_D(0,R_*)$.

By Assumption 1, the condition $\sigma \in C_b^2(\mathbb{R})$, and the definitions of the population vector fields, V_U and V_W are uniformly locally Lipschitz in their particle variables on bounded sets. Consequently, there exists $L_\varepsilon < \infty$ such that, whenever the population measures are supported in $B_D(0, R_* + \varepsilon)$,

$$\|V(v; \rho_{U,t}, \rho_{W,t}) - V(v'; \rho_{U,t}, \rho_{W,t})\|_2 \le L_\varepsilon \|v - v'\|_2 \tag{41}$$

for $v, v' \in B_D(0, R_* + \varepsilon)$, and for either choice $V = V_U$ or $V = V_W$.

Consider now any time interval on which both population measures are supported in $B_D(0, R_* + \varepsilon)$. Since the squared distance to a closed convex set is continuously differentiable, along a characteristic we have

$$\frac{d}{dt}\frac{1}{2} d_*(v_t)^2 = \langle v_t - P_*(v_t), \dot{v}_t \rangle .$$

If $\|v_t\|_2 \le R_*$, then $v_t = P_*(v_t)$ and the right-hand side is zero. If $\|v_t\|_2 > R_*$, then $\|P_*(v_t)\|_2 = R_*$, and therefore

$$\begin{aligned}\frac{d}{dt}\frac{1}{2} d_*(v_t)^2 = \xi(t) \langle v_t - P_*(v_t), V(P_*(v_t); \rho_{U,t}, \rho_{W,t}) \rangle \\ + \xi(t) \langle v_t - P_*(v_t), V(v_t; \rho_{U,t}, \rho_{W,t}) - V(P_*(v_t); \rho_{U,t}, \rho_{W,t}) \rangle .\end{aligned}$$

The vector $v_t - P_*(v_t)$ is a nonnegative scalar multiple of $P_*(v_t)$. Since $\xi(t) \ge 0$, the first term on the right-hand side is nonpositive by (40). For the second term, (41) and $\xi(t) \le M_\xi$ give

$$|\xi(t) \langle v_t - P_*(v_t), V(v_t; \rho_{U,t}, \rho_{W,t}) - V(P_*(v_t); \rho_{U,t}, \rho_{W,t}) \rangle| \le M_\xi L_\varepsilon d_*(v_t)^2.$$

Hence, on every such interval,

$$\frac{d}{dt} d_*(v_t)^2 \le 2 M_\xi L_\varepsilon d_*(v_t)^2. \tag{42}$$

We now close the argument by continuation. Since $R_* \ge R_0$, the initial supports satisfy $\operatorname{supp}(\rho_{U,0}) \cup \operatorname{supp}(\rho_{W,0}) \subseteq B_D(0, R_*)$. In particular, every initial characteristic satisfies $d_*(v_0) = 0$. By continuity of the characteristic flow, there is a nontrivial initial time interval on which both population measures remain supported in $B_D(0, R_* + \varepsilon)$. On this interval, Gronwall's inequality applied to (42) yields $d_*(v_t) = 0$ for every decoder and encoder characteristic. Therefore all characteristics in fact remain in the smaller ball $B_D(0, R_*)$ throughout this interval.

We can continue for as long as the characteristic solution exists. If at time t, the support goes outside the ball $B_D(0, R_* + \varepsilon)$, the preceding estimate would show that all characteristics remain in the strictly smaller ball $B_D(0, R_*)$ up to that time. By continuity of the flow would then imply that the solution remains inside $B_D(0, R_* + \varepsilon)$ for a further time interval $t' > t$ contradicting the fact that the follow escapes at t. Consequently,

$$\|u_t\|_2 \le R_*, \qquad \|w_t\|_2 \le R_*, \qquad t \ge 0.$$

Finally, $\rho_{U,t}$ and $\rho_{W,t}$ are the pushforwards of $\rho_{U,0}$ and $\rho_{W,0}$ under their corresponding characteristic flows. Since every characteristic starting in the initial supports remains in $B_D(0, R_*)$, we obtain

$$\operatorname{supp}(\rho_{U,t}) \subseteq B_D(0, R_*), \qquad \operatorname{supp}(\rho_{W,t}) \subseteq B_D(0, R_*), \qquad t \ge 0.$$

This proves (38), and equivalently (39). □

Remark 2 (Common trapping ball)**.** *Theorems 6 and 7 show that the same deterministic ball*

$$B_D(0, R_*)$$

can be used for both the finite-particle SGD dynamics and the limiting mean-field dynamics. More precisely,

$$\operatorname{supp} \hat{\rho}_{U,k}^N \cup \operatorname{supp} \hat{\rho}_{W,k}^N \subseteq B_D(0, R_*), \qquad k \ge 0,$$

provided the SGD step size satisfies (28)*, whereas*

$$\operatorname{supp}(\rho_{U,t}) \cup \operatorname{supp}(\rho_{W,t}) \subseteq B_D(0, R_*), \qquad t \ge 0,$$

for the characteristic mean-field dynamics. The SGD dynamics require the additional small-step condition because the explicit discrete update can otherwise overshoot the boundary of the invariant ball.

G Existence and Uniqueness of the Coupled PDE

In this section, we prove the global existence and uniqueness of the coupled Vlasov–McKean system:

$$\partial_t \rho_{U,t} + \nabla_u \cdot (\rho_{U,t}\, \xi(t) V_U(\cdot; \rho_{U,t}, \rho_{W,t})) = 0, \tag{43}$$
$$\partial_t \rho_{W,t} + \nabla_w \cdot (\rho_{W,t}\, \xi(t) V_W(\cdot; \rho_{U,t}, \rho_{W,t})) = 0. \tag{44}$$

Recall that the regularized descent vector fields are given by

$$V_U(u; \rho_U, \rho_W) = G_U(u; \rho_U, \rho_W) - \lambda \|u\|_2^4 u,$$
$$V_W(w; \rho_U, \rho_W) = G_W(w; \rho_U, \rho_W) - \lambda \|w\|_2^4 w,$$

where the unregularized population fields are

$$G_U(u; \rho_U, \rho_W) = -\mathbb{E}_{x\sim P}\left[J_u \psi(\alpha(x; \rho_W); u)^\top r(x; \rho_U, \rho_W)\right],$$
$$G_W(w; \rho_U, \rho_W) = -\mathbb{E}_{x\sim P}\left[J_w \phi(x; w)^\top \overline{J}(x; \rho_U, \rho_W)^\top r(x; \rho_U, \rho_W)\right].$$

We use the trapping radius R_* obtained in the previous section along with the 1. So the initial measures satisfies

$$\operatorname{supp}(\rho_{U,0}) \cup \operatorname{supp}(\rho_{W,0}) \subseteq B_D(0, R_*),$$

and by previous section, the characteristic trajectories associated with the dynamics cannot leave this ball.

For $R < \infty$, let

$$\mathcal{P}_R := \left\{\rho \in \mathcal{P}(\mathbb{R}^D) : \operatorname{supp}(\rho) \subseteq B_D(0, R)\right\}.$$

For two pairs of measures $(\mu, \nu), (\mu', \nu') \in \mathcal{P}_R^2$, we use the product Wasserstein distance

$$\mathcal{W}_1\left((\mu, \nu), (\mu', \nu')\right) := W_1(\mu, \mu') + W_1(\nu, \nu').$$

Our proof proceeds in three steps. First, we establish uniform Lipschitz estimates for all quantities entering the vector fields on a fixed bounded ball. Second, we solve the nonlinear characteristic equations by a fixed-point argument. Finally, we identify the characteristic solution with the unique distributional solution of (43)–(44).

G.1 Lipschitz estimates on the trapping ball

We first record the regularity properties of the basic encoder and decoder maps. Recall that $w = (w_1, w_2)$ and $u = (u_1, u_2)$, with

$$\phi(x; w) = w_2\, \sigma(\langle w_1, x\rangle), \qquad \psi(\alpha; u) = u_1\, \sigma(\langle u_2, \alpha\rangle).$$

The respective Jacobians are

$$J_\alpha \psi(\alpha; u) = u_1 u_2^\top \sigma'(\langle u_2, \alpha\rangle),$$
$$J_u \psi(\alpha; u)^\top z = \begin{pmatrix} \sigma(\langle u_2, \alpha\rangle) z \\ \alpha \sigma'(\langle u_2, \alpha\rangle)\langle u_1, z\rangle \end{pmatrix}, \qquad z \in \mathbb{R}^d,$$
$$J_w \phi(x; w)^\top h = \begin{pmatrix} x\sigma'(\langle w_1, x\rangle)\langle w_2, h\rangle \\ \sigma(\langle w_1, x\rangle) h \end{pmatrix}, \qquad h \in \mathbb{R}^b.$$

Before we state the next lemma, let us define

$$C_R := 2\,(1 + S_0)\,(S_0 + 2S_1 + S_2)\,(1 + M_x)^2\,(1 + R)^3. \tag{45}$$

Lemma 9 (Uniform local regularity of the building blocks)**.** *For every $R < \infty$, whenever*

$$\|x\|_2 \le M_x, \qquad \|u\|_2 \le R, \qquad \|w\|_2 \le R, \qquad \|\alpha\|_2 \le S_0 R,$$

the maps

$$\phi(x; w), \qquad \psi(\alpha; u), \qquad J_w \phi(x; w), \qquad J_u \psi(\alpha; u), \qquad J_\alpha \psi(\alpha; u)$$

are bounded by C_R in respective norms and C_R-Lipschitz.

Proof. Recall that for $w = (w_1, w_2)$, $u = (u_1, u_2)$,

$$\phi(x; w) = w_2\,\sigma(\langle w_1, x\rangle),\ \psi(\alpha; u) = u_1\,\sigma(\langle u_2, \alpha\rangle),\ J_\alpha\psi(\alpha; u) = \sigma'(\langle u_2, \alpha\rangle)\,u_1 u_2^\top,$$

$$J_w\phi(x; w) = \big[\,\sigma'(\langle w_1, x\rangle)\,w_2 x^\top \,\big|\, \sigma(\langle w_1, x\rangle)\,I_b\,\big], \qquad J_u\psi(\alpha; u) = \big[\,\sigma(\langle u_2, \alpha\rangle)\,I_d \,\big|\, \sigma'(\langle u_2, \alpha\rangle)\,u_1\alpha^\top\,\big],$$

where the two blocks are the derivatives with respect to the first and second component of w, respectively u. As $\|w\|_2 \le R$, so $\|w_1\|_2, \|w_2\|_2 \le R$, and similar thing for u. Hence we have

$$\|\phi\|_2 \le S_0 R,\ \|\psi\|_2 \le S_0 R,\ \|J_\alpha\psi\|_{\mathrm{op}} \le S_1 R^2,$$

As $\|ab^\top\|_{\mathrm{op}} = \|a\|_2\|b\|_2$, we further get

$$\|J_\alpha\psi\|_{\mathrm{op}} \le S_1 R^2,\ \|J_w\phi\|_{\mathrm{op}} \le S_1 R M_x + S_0$$
$$\|J_u\psi\|_{\mathrm{op}} \le S_0 + S_1 R \cdot S_0 R = S_0(1 + S_1 R^2)$$

where for a block matrix $[X \,|\, Y]$ we used $\|[X \,|\, Y]\|_{\mathrm{op}} \le \|X\|_{\mathrm{op}} + \|Y\|_{\mathrm{op}}$. Each of these is at most C_R.

To get the Lipschitz bounds, let (x, w, α, u) and (x', w', α', u') be two sets of arguments in the stated domain. We write Δ for the difference. For notational convenience let

$$a := \langle w_1, x\rangle, \qquad a' := \langle w_1', x'\rangle, \qquad b := \langle u_2, \alpha\rangle, \qquad b' := \langle u_2', \alpha'\rangle.$$

Then the differences are bounded as follows

$$|\Delta a| = |\langle w_1, x\rangle - \langle w_1', x'\rangle| \le M_x\|\Delta w_1\|_2 + R\|\Delta x\|_2,$$
$$|\Delta b| = |\langle u_2, \alpha\rangle - \langle u_2', \alpha'\rangle| \le S_0 R\,\|\Delta u_2\|_2 + R\|\Delta\alpha\|_2,$$
$$|\sigma(a) - \sigma(a')| \le S_1|\Delta a|,\ |\sigma'(a) - \sigma'(a')| \le S_2|\Delta a|,$$
$$|\sigma(b) - \sigma(b')| \le S_1|\Delta b|,\ |\sigma'(b) - \sigma'(b')| \le S_2|\Delta b|.$$

To get all the Lipschitz property, we make repeated use triangles inequality applied to the following for each quantity.

$$FG - F'G' = (F - F')G + F'(G - G'),$$
$$FGH - F'G'H' = (F - F')GH + F'(G - G')H + F'G'(H - H'),$$

Using the above bounds, and triangles inequality we get,

$$\begin{aligned}
\|\Delta\phi\|_2 &= \|w_2\sigma(a) - w_2'\sigma(a')\|_2 \\
&\le S_0\|\Delta w_2\|_2 + RS_1|\Delta a| \\
&\le S_0\|\Delta w_2\|_2 + S_1 R M_x\|\Delta w_1\|_2 + S_1 R^2\|\Delta x\|_2, \\
\|\Delta\psi\|_2 &= \|u_1\sigma(b) - u_1'\sigma(b')\|_2 \\
&\le S_0\|\Delta u_1\|_2 + RS_1|\Delta b| \\
&\le S_0\|\Delta u_1\|_2 + S_0 S_1 R^2\|\Delta u_2\|_2 + S_1 R^2\|\Delta\alpha\|_2, \\
\|\Delta J_\alpha\psi\|_{\mathrm{op}} &= \big\|\sigma'(b)u_1u_2^\top - \sigma'(b')u_1'(u_2')^\top\big\|_{\mathrm{op}} \\
&\le S_2|\Delta b|\,R^2 + S_1 R\|\Delta u_1\|_2 + S_1 R\|\Delta u_2\|_2 \\
&\le S_1 R\big(\|\Delta u_1\|_2 + \|\Delta u_2\|_2\big) + S_2 R^3\big(S_0\|\Delta u_2\|_2 + \|\Delta\alpha\|_2\big), \\
\|\Delta J_w\phi\|_{\mathrm{op}} &\le \underbrace{S_2|\Delta a|\,RM_x + S_1 M_x\|\Delta w_2\|_2 + S_1 R\|\Delta x\|_2}_{\text{first block}} + \underbrace{S_1|\Delta a|}_{\text{second block}} \\
&\le (S_2 R M_x^2 + S_1 M_x)\|\Delta w_1\|_2 + S_1 M_x\|\Delta w_2\|_2 \\
&\quad + (S_2 R^2 M_x + 2S_1 R)\|\Delta x\|_2, \\
\|\Delta J_u\psi\|_{\mathrm{op}} &\le \underbrace{S_1|\Delta b|}_{\text{first block}} + \underbrace{S_2|\Delta b|\,R\cdot S_0 R + S_1 S_0 R\|\Delta u_1\|_2 + S_1 R\|\Delta\alpha\|_2}_{\text{second block}} \\
&\le S_0 S_1 R\|\Delta u_1\|_2 + \big(S_0 S_1 R + S_0^2 S_2 R^3\big)\|\Delta u_2\|_2 \\
&\quad + \big(S_1 R + S_0 S_2 R^3\big)\|\Delta\alpha\|_2.
\end{aligned}$$

All coefficients on the right-hand sides are finite constants depending only on S_0, S_1, S_2, M_x and R. Chooing a larger C_R if necessary, and using $\|\Delta w_1\|_2 + \|\Delta w_2\|_2 \le 2\|\Delta w\|_2$ (similar for (u_1, u_2)), we get Lipschitz bounds. □

We next use Lemma 9 to control the dependence of the mean-field quantities on the probability measures. For two pairs $(\mu,\nu),(\mu',\nu')\in\mathcal{P}_R^2$, let $\mathcal{W}_1\big((\mu,\nu),(\mu',\nu')\big) := W_1(\mu,\mu')+W_1(\nu,\nu')$.

Lemma 10 (Lipschitz dependence of the mean-field vector fields)**.** *For every $R<\infty$, there exists a constant L_R, M_R such that, for all $u,u',w,w'\in B_D(0,R)$ and $(\mu,\nu),(\mu',\nu')\in\mathcal{P}_R^2$, we have*

$$\|V_U(u;\mu,\nu)-V_U(u';\mu',\nu')\|_2 \le L_R\left[\|u-u'\|_2+\mathcal{W}_1\left((\mu,\nu),(\mu',\nu')\right)\right], \tag{46}$$

$$\|V_W(w;\mu,\nu)-V_W(w';\mu',\nu')\|_2 \le L_R\left[\|w-w'\|_2+\mathcal{W}_1\left((\mu,\nu),(\mu',\nu')\right)\right], \tag{47}$$

$$\sup_{\substack{u,w\in B_D(0,R)\\(\mu,\nu)\in\mathcal{P}_R^2}} \left(\|V_U(u;\mu,\nu)\|_2+\|V_W(w;\mu,\nu)\|_2\right)\le M_R.$$

Proof. We will use Lemma 9 to all the quantities in this proof. Let π_W be any coupling of ν and ν', then

$$\begin{aligned}\|\alpha(x;\nu)-\alpha(x;\nu')\|_2 &= \left\|\int\big(\phi(x;w)-\phi(x;w')\big)\pi_W(dw,dw')\right\|_2\\ &\le \int\|\phi(x;w)-\phi(x;w')\|_2\pi_W(dw,dw')\\ &\le C_R\int\|w-w'\|_2\pi_W(dw,dw').\end{aligned}$$

After taking the infimum over all couplings, we get

$$\sup_{x\in\mathrm{supp}(P)}\|\alpha(x;\nu)-\alpha(x;\nu')\|_2\le C_R W_1(\nu,\nu'). \tag{48}$$

Let π_U be any coupling of μ and μ', then

$$\begin{aligned}\|\hat{x}(x;\mu,\nu)-\hat{x}(x;\mu',\nu')\|_2 &\le \int\|\psi(\alpha(x;\nu);u)-\psi(\alpha(x;\nu');u')\|_2\,\pi_U(du,du')\\ &\le C_R\int\|u-u'\|_2\pi_U(du,du')+C_R\|\alpha(x;\nu)-\alpha(x;\nu')\|_2,\end{aligned}$$

which after using (48) and taking the infimum over all couplings π_U gives us

$$\sup_{x\in\mathrm{supp}(P)}\|\hat{x}(x;\mu,\nu)-\hat{x}(x;\mu',\nu')\|_2\le C_R W_1(\mu,\mu')+C_R^2W_1(\nu,\nu').$$

As $r(x;\mu,\nu)=\hat{x}(x;\mu,\nu)-x$,, the same estimate gives

$$\sup_{x\in\mathrm{supp}(P)}\|r(x;\mu,\nu)-r(x;\mu',\nu')\|_2\le C_R W_1(\mu,\mu')+C_R^2W_1(\nu,\nu'). \tag{49}$$

We also have from previous lemma, $\|\hat{x}(x;\mu,\nu)\|_2\le\int\|\psi(\alpha(x;\nu);u)\|_2\,\mu(du)\le C_R$, and hence,

$$\sup_{\substack{x\in\mathrm{supp}(P)\\(\mu,\nu)\in\mathcal{P}_R^2}}\|r(x;\mu,\nu)\|_2\le C_R+M_x. \tag{50}$$

Now let us consider the averaged decoder Jacobian $\overline{J}(x;\mu,\nu)$. We use the same coupling argument and the Lipschitz continuity of $J_\alpha\psi$ to get

$$\sup_{x\in\mathrm{supp}(P)}\|\overline{J}(x;\mu,\nu)-\overline{J}(x;\mu',\nu')\|\le C_R W_1(\mu,\mu')+C_R^2W_1(\nu,\nu'). \tag{51}$$

Also,

$$\sup_{\substack{x\in\mathrm{supp}(P)\\(\mu,\nu)\in\mathcal{P}_R^2}}\|\overline{J}(x;\mu,\nu)\|\le C_R. \tag{52}$$

We now estimate the decoder population field. Let us define $A(x):=J_u\psi(\alpha(x;\nu);u)$, $A'(x):=J_u\psi(\alpha(x;\nu');u)$, $r(x):=r(x;\mu,\nu)$ and $r'(x):=r(x;\mu',\nu')$. Then

$$A(x)^\top r(x)-(A'(x))^\top r'(x)=A(x)^\top(r(x)-r'(x))+(A(x)-A'(x))^\top r'(x).$$

Hence, by Lemma 9, (48), (49), and (50),

$$\begin{aligned}
&\|G_U(u;\mu,\nu) - G_U(u;\mu',\nu')\|_2 \\
&\quad \le \mathbb{E}_x\left[\|A(x)\|\,\|r(x)-r'(x)\|_2\right] + \mathbb{E}_x\left[\|A(x)-A'(x)\|\,\|r'(x)\|_2\right] \\
&\quad \le C_R\left[C_R W_1(\mu,\mu') + C_R^2 W_1(\nu,\nu')\right] + (C_R+M_x)C_R^2 W_1(\nu,\nu') \\
&\quad = C_R^2 W_1(\mu,\mu') + \left(M_x C_R^2 + 2C_R^3\right) W_1(\nu,\nu'). \qquad (53)
\end{aligned}$$

Now consider the encoder field. Since $J_w\phi(x;w)$ does not depend on the measures, it is enough to estimate $\overline{J}(x;\mu,\nu)^\top r(x;\mu,\nu) - \overline{J}(x;\mu',\nu')^\top r(x;\mu',\nu')$. Writing $\overline{J} := \overline{J}(x;\mu,\nu)$ and $\overline{J}' := \overline{J}(x;\mu',\nu')$, we have

$$\overline{J}^\top r - (\overline{J}')^\top r' = \overline{J}^\top (r-r') + (\overline{J}-\overline{J}')^\top r'.$$

Using (49), (50), (51), and (52),

$$\begin{aligned}
&\left\|\overline{J}(x;\mu,\nu)^\top r(x;\mu,\nu) - \overline{J}(x;\mu',\nu')^\top r(x;\mu',\nu')\right\|_2 \\
&\quad \le C_R\left[C_R W_1(\mu,\mu') + C_R^2 W_1(\nu,\nu')\right] + (C_R+M_x)\left[C_R W_1(\mu,\mu') + C_R^2 W_1(\nu,\nu')\right] \\
&\quad = \left(2C_R^2 + M_x C_R\right) W_1(\mu,\mu') + \left(2C_R^3 + M_x C_R^2\right) W_1(\nu,\nu').
\end{aligned}$$

Since $\|J_w\phi(x;w)\| \le C_R$, we obtain

$$\begin{aligned}
&\|G_W(w;\mu,\nu) - G_W(w;\mu',\nu')\|_2 \\
&\quad \le \left(2C_R^3 + M_x C_R^2\right) W_1(\mu,\mu') + \left(2C_R^4 + M_x C_R^3\right) W_1(\nu,\nu'). \qquad (54)
\end{aligned}$$

We next consider the dependence on the particle variables. If the measures are fixed, the functions $r(x;\mu,\nu)$, $\alpha(x;\nu)$, and $\overline{J}(x;\mu,\nu)$ are fixed functions of x. For the decoder field, Lemma 9 and (50) give

$$\|G_U(u;\mu,\nu) - G_U(u';\mu,\nu)\|_2 \le C_R(C_R+M_x)\|u-u'\|_2. \qquad (55)$$

For the encoder field, Lemma 9, (52), and (50) give

$$\|G_W(w;\mu,\nu) - G_W(w';\mu,\nu)\|_2 \le C_R^2(C_R+M_x)\|w-w'\|_2. \qquad (56)$$

It remains to include the regularization term. Let us denote $H(v) := \|v\|_2^4 v$, then $DH(v) = \|v\|_2^4 I_D + 4\|v\|_2^2 vv^\top$. Therefore, for $\|v\|_2 \le R$,

$$\|DH(v)\|_{\mathrm{op}} \le 5R^4.$$

By the mean-value theorem,

$$\|H(v) - H(v')\|_2 \le 5R^4\|v-v'\|_2, \quad v,v' \in B_D(0,R). \qquad (57)$$

Since $V_U(u;\mu,\nu) = G_U(u;\mu,\nu) - \lambda H(u)$ and $V_W(w;\mu,\nu) = G_W(w;\mu,\nu) - \lambda H(w)$, combining (53), (54), (55), (56), and (57) gives (46)–(47).

Now we prove the uniform bound. From Lemma 9, (50), and (52),

$$\|G_U(u;\mu,\nu)\|_2 \le C_R(C_R+M_x), \quad \|G_W(w;\mu,\nu)\|_2 \le C_R^2(C_R+M_x),$$

and for $\|v\|_2 \le R$, $\lambda\|H(v)\|_2 = \lambda\|v\|_2^5 \le \lambda R^5$. Therefore,

$$\|V_U(u;\mu,\nu)\|_2 + \|V_W(w;\mu,\nu)\|_2 \le (C_R + C_R^2)(C_R+M_x) + 2\lambda R^5.$$

□

G.2 Characteristic fixed-point construction

We now construct a solution of the coupled system by solving the characteristic equations. Fix $T>0$ and define $\mathbb{X}_T$ to be the set og continuous curves with values in $\mathcal{P}^2_{R_*}$,

$$\mathbb{X}_T := C\left([0,T];\mathcal{P}^2_{R_*}\right).$$

We equip this space with the metric $d_T(m,m') := \sup_{0\le t\le T}\mathcal{W}_1(m_t,m_t')$. Because $B_D(0,R_*)$ is compact, the metric space $(\mathcal{P}_{R_*},W_1)$ is complete. It follows that $(\mathbb{X}_T,d_T)$ is also complete.

Given a prescribed curve $m = (\mu_t, \nu_t) \in \mathbb{X}_T$, consider the characteristic equations,

$$\frac{d}{dt} X_U^m(t, u_0) = \xi(t) V_U \left(X_U^m(t, u_0); \mu_t, \nu_t \right), \qquad X_U^m(0, u_0) = u_0,$$
$$\frac{d}{dt} X_W^m(t, w_0) = \xi(t) V_W \left(X_W^m(t, w_0); \mu_t, \nu_t \right), \qquad X_W^m(0, w_0) = w_0.$$

By Lemma 10, the right-hand sides are continuous in time and uniformly Lipschitz in the spatial variables on the trapping ball. Therefore the Picard–Lindelöf theorem gives a unique local solution for every initial point. Moreover, because the prescribed measures μ_t, ν_t are supported in $B_D(0, R_*)$, the invariant-ball estimate applies to these characteristic equations. Consequently,

$$X_U^m(t, u_0),\ X_W^m(t, w_0) \in B_D(0, R_*) \qquad \text{whenever } u_0, w_0 \in B_D(0, R_*).$$

The trajectories remain in a compact region on which the vector fields are uniformly bounded and Lipschitz, so the local solutions extend to the entire interval $[0, T]$.

We define a map $\mathcal{T} : \mathbb{X}_T \to \mathbb{X}_T$ by

$$(\mathcal{T} m)_t := \left((X_U^m(t, \cdot))_{\#} \rho_{U,0}, (X_W^m(t, \cdot))_{\#} \rho_{W,0} \right).$$

The invariant-ball property shows that both push-forward measures remain in $\mathcal{P}_{R_*}$. As the vector fields are uniformly bounded, the characteristic maps are uniformly continuous in time i.e. for $0 \le s, t \le T$,

$$\sup_{u_0 \in B_D(0,R_*)} \|X_U^m(t, u_0) - X_U^m(s, u_0)\|_2 + \sup_{w_0 \in B_D(0,R_*)} \|X_W^m(t, w_0) - X_W^m(s, w_0)\|_2 \le C_R |t - s|.$$

So, $t \mapsto (\mathcal{T} m)_t$ is continuous in W_1. Hence $\mathcal{T}$ maps $\mathbb{X}_T$ into itself.

We now show that $\mathcal{T}$ is a contraction on a sufficiently short time interval. Let $m = (\mu, \nu)$ and $m' = (\mu', \nu')$ be two elements of $\mathbb{X}_T$. We couple the two characteristic flows by using same initializing. Let us define $E(t)$ as

$$E(t) := \sup_{u_0 \in B_D(0,R_*)} \|X_U^m(t, u_0) - X_U^{m'}(t, u_0)\|_2 + \sup_{w_0 \in B_D(0,R_*)} \|X_W^m(t, w_0) - X_W^{m'}(t, w_0)\|_2.$$

Using Lemma 10 and $\xi(t) \le M_\xi$, for all $t \le T$ we get,

$$E(t) \le M_\xi L_{R_*} \int_0^t \left(E(s) + 2\, \mathcal{W}_1(m_s, m'_s) \right) ds.$$

Since $\mathcal{W}_1(m_s, m'_s) \le d_T(m, m')$, applying Gronwall's inequality gives us

$$E(t) \le 2 M_\xi L_{R_*}\, t\, e^{M_\xi L_{R_*} t}\, d_T(m, m').$$

On the other hand,

$$\mathcal{W}_1\big((\mathcal{T} m)_t, (\mathcal{T} m')_t\big) \le E(t).$$

Therefore

$$d_T(\mathcal{T} m, \mathcal{T} m') \le 2 M_\xi L_{R_*}\, T\, e^{M_\xi L_{R_*} T}\, d_T(m, m').$$

If we choose $T_0 > 0$ such that $2 M_\xi L_{R_*}\, T_0\, e^{M_\xi L_{R_*} T_0} < 1$, then $\mathcal{T} : \mathbb{X}_{T_0} \to \mathbb{X}_{T_0}$ is a strict contraction.

Proposition 1 (Global characteristic solution)**.** *There exists a unique pair $(\rho_{U,t}, \rho_{W,t})_{t \ge 0}$ generated by the characteristic equations*

$$\dot{X}_U(t, u_0) = \xi(t) V_U \left(X_U(t, u_0); \rho_{U,t}, \rho_{W,t} \right),$$
$$\dot{X}_W(t, w_0) = \xi(t) V_W \left(X_W(t, w_0); \rho_{U,t}, \rho_{W,t} \right),$$

with $X_U(0, u_0) = u_0$, $X_W(0, w_0) = w_0$, and $\rho_{U,t} = (X_U(t, \cdot))_{\#} \rho_{U,0}$, $\rho_{W,t} = (X_W(t, \cdot))_{\#} \rho_{W,0}$. Moreover, $\operatorname{supp}(\rho_{U,t}) \cup \operatorname{supp}(\rho_{W,t}) \subseteq B_D(0, R_*)$ for all $t \ge 0$.

Proof. On $[0, T_0]$, the map $\mathcal{T}$ is a strict contraction on the complete metric space $\mathbb{X}_{T_0}$. By using Banach's fixed-point theorem, we get a unique fixed point $(\rho_{U,t}, \rho_{W,t})_{0\le t\le T_0}$. By definition of the fixed point, these measures are generated by the characteristic flows

$$\rho_{U,t} = (X_U(t,\cdot))_{\#}\rho_{U,0},\ \rho_{W,t} = (X_W(t,\cdot))_{\#}\rho_{W,0}.$$

Our earlier invariant-ball estimate gives

$$\operatorname{supp}(\rho_{U,t}) \cup \operatorname{supp}(\rho_{W,t}) \subseteq B_D(0, R_*) \qquad \text{for } 0 \le t \le T_0.$$

We now extend the solution globally by getting solution on patches and pasting them together. At time T_0, let us use ρ_{U,T_0} and ρ_{W,T_0} as the new initial measures. As the constants in the contraction proof depend only on M_ξ, L_{R_*}, and the trapping ball $B_D(0, R_*)$, and independent on initial measures inside that ball, we can repeat the same fixed-point argument on $[T_0, 2T_0]$. In fact, by same idea, we repeat this construction on

$$[kT_0, (k+1)T_0], \qquad k = 0, 1, 2, \dots,$$

to get patch solutions on for all $t \ge 0$. The pieces agree at the endpoints (by construction) and by uniqueness, these can be patched together to get the global characteristic solution which is again unique. The invariant-ball estimate gives

$$\operatorname{supp}(\rho_{U,t}) \cup \operatorname{supp}(\rho_{W,t}) \subseteq B_D(0, R_*) \qquad \text{for all } t \ge 0.$$

□

G.3 Distributional formulation and uniqueness

Definition 1 (Distributional solution). *Let $T > 0$. A pair $(\rho_{U,t}, \rho_{W,t}) \in C([0,T]; \mathcal{P}^2_{R_*})$ is called a distributional solution of* (43)–(44) *with initial data $(\rho_{U,0}, \rho_{W,0})$ if, for every $\zeta_U, \zeta_W \in C^1_c([0,T) \times \mathbb{R}^D)$, we have*

$$\begin{aligned}
&\int_0^T \int_{\mathbb{R}^D} \Big[\partial_t \zeta_U(t,u) + \xi(t) \nabla_u \zeta_U(t,u) \cdot V_U(u; \rho_{U,t}, \rho_{W,t})\Big] \rho_{U,t}(du)\, dt \\
&\qquad + \int_{\mathbb{R}^D} \zeta_U(0,u) \rho_{U,0}(du) = 0, \\
&\int_0^T \int_{\mathbb{R}^D} \Big[\partial_t \zeta_W(t,w) + \xi(t) \nabla_w \zeta_W(t,w) \cdot V_W(w; \rho_{U,t}, \rho_{W,t})\Big] \rho_{W,t}(dw)\, dt \\
&\qquad + \int_{\mathbb{R}^D} \zeta_W(0,w) \rho_{W,0}(dw) = 0.
\end{aligned}$$

By standard chain-rule evaluations along characteristics, the unique continuous distributional solution of a linear continuity equation with uniformly Lipschitz velocity fields is precisely its characteristic pushforward. Consequently, we obtain the following equivalence.

Proposition 2 (Existence and uniqueness of the distributional coupled PDE). *For every $T > 0$, the characteristic solution constructed in Proposition 1 is the unique distributional solution of* (43)–(44) *in $C([0,T]; \mathcal{P}^2_{R_*})$. Thus, the coupled PDE possesses a unique global continuous distributional solution satisfying $\operatorname{supp}(\rho_{U,t}) \cup \operatorname{supp}(\rho_{W,t}) \subseteq B_D(0, R_*)$ for all $t \ge 0$.*

H Closeness between Finite and Mean-Field Risks

H.1 Technical Lemmas

Lemma 11 (Attainment of the Regularized Mean-Field Minimum). *Let Assumption 1 hold, then the regularized risk functional $\mathcal{R}_\lambda$ attains its minimum on $\mathcal{P}_6(\mathbb{R}^D) \times \mathcal{P}_6(\mathbb{R}^D)$. If (ρ^*_U, ρ^*_W) is a global minimizer then 6th-moment is bounded,*

$$\int \|u\|_2^6 \rho^*_U(du) + \int \|w\|_2^6 \rho^*_W(dw) \le \frac{3M_x^2}{\lambda}.$$

Proof. For the trivial pair (δ_0, δ_0) the network output is identically zero, and $\mathcal{R}_\lambda(\delta_0, \delta_0) = \frac{1}{2}\mathbb{E}\|x\|_2^2 \leq \frac{1}{2}M_x^2$. We can always choose a sequence (ρ_U^n, ρ_W^n) such that $\mathcal{R}_\lambda(\rho_U^n, \rho_W^n)$ monotonically converges to its infrimum value. We need to show that this sequence actually converges to a pair of measures in the domain of the regularized risk which realizes the infimum. Maybe after discarding first few, we can assume that

$$\mathcal{R}_\lambda(\rho_U^n, \rho_W^n) \leq \mathcal{R}_\lambda(\delta_0, \delta_0) + 1 \leq \frac{1}{2}M_x^2 + 1.$$

Since the reconstruction term is nonnegative, this gives the uniform sixth-moment bound

$$\frac{\lambda}{6}\left(\int \|u\|_2^6 \rho_U^n(du) + \int \|w\|_2^6 \rho_W^n(dw)\right) \leq \frac{1}{2}M_x^2 + 1$$

and hence,

$$\sup_n \left(\int \|u\|_2^6 \rho_U^n(du) + \int \|w\|_2^6 \rho_W^n(dw)\right) < \infty.$$

We use this uniform bound to show that the sequence is tight. Applying Markov's inequality, for any $A > 0$ we have

$$\rho_U^n(\{\|u\|_2 > A\}) \leq A^{-6}\int \|u\|_2^6 \rho_U^n(du), \quad \rho_W^n(\{\|w\|_2 > A\}) \leq A^{-6}\int \|w\|_2^6 \rho_W^n(dw).$$

The right-hand sides tend to zero uniformly in n as $A \to \infty$. Hence both sequences $(\rho_U^n)_n$ and $(\rho_W^n)_n$ are tight. By Prokhorov's theorem, there exists a weakly convergent subsequence. Taking the subsequence to be our new subsequence (by abuse of notataion),

$$\rho_U^n \rightharpoonup \rho_U^*, \quad \rho_W^n \rightharpoonup \rho_W^*$$

weakly to some probability measures ρ_U^*, ρ_W^*. By using lower semicontinuity of the map $z \mapsto \|z\|_2^6$ under weak convergence, we have

$$\int \|u\|_2^6 \rho_U^*(du) \leq \liminf_{n\to\infty} \int \|u\|_2^6 \rho_U^n(du),$$
$$\int \|w\|_2^6 \rho_W^*(dw) \leq \liminf_{n\to\infty} \int \|w\|_2^6 \rho_W^n(dw).$$

i.e. they have finite 6th-moment. In particular, $(\rho_U^*, \rho_W^*) \in \mathcal{P}_6(\mathbb{R}^D) \times \mathcal{P}_6(\mathbb{R}^D)$.

We need to check that the reconstruction term is continuous along this subsequence. For fixed x, let

$$\alpha_n(x) := \int \phi(x; w)\rho_W^n(dw), \quad \alpha_*(x) := \int \phi(x; w)\rho_W^*(dw).$$

$\phi(x; w)$ is continuous in w and satisfies $\|\phi(x; w)\|_2 \leq S_0\|w\|_2$. Hencem the uniform 6th-moment bound gives uniform integrability of $\|w\|_2$. Thus weak convergence implies

$$\alpha_n(x) \to \alpha_*(x) \qquad \text{for each } x \in \operatorname{supp}(P).$$

Similarly, using the fact that $\|\psi(\alpha; u)\|_2 \leq S_0\|u\|_2$ and the fact that ψ is continuous in both variables, we get,

$$\int \psi(\alpha_n(x); u)\rho_U^n(du) \longrightarrow \int \psi(\alpha_*(x); u)\rho_U^*(du) \qquad \text{for each } x \in \operatorname{supp}(P).$$

Since $\|x\|_2 \leq M_x$, dominated convergence gives the convergence of the reconstruction risks,

$$\frac{1}{2}\mathbb{E}_x\left[\left\|x - \int \psi(\alpha_n(x); u)\rho_U^n(du)\right\|_2^2\right] \longrightarrow \frac{1}{2}\mathbb{E}_x\left[\left\|x - \int \psi(\alpha_*(x); u)\rho_U^*(du)\right\|_2^2\right].$$

Now if we combine the continuity of the reconstruction term with the lower semicontinuity of the regularization term, we get

$$\mathcal{R}_\lambda(\rho_U^*, \rho_W^*) \leq \liminf_{n\to\infty} \mathcal{R}_\lambda(\rho_U^n, \rho_W^n),$$

i.e. (ρ_U^*, ρ_W^*) attains the infimum of $\mathcal{R}_\lambda$. As it's the global minima, $\mathcal{R}_\lambda(\rho_U^*, \rho_W^*) \leq \mathcal{R}_\lambda(\delta_0, \delta_0) \leq \frac{1}{2}M_x^2$. As the reconstruction term is nonnegative,

$$\frac{\lambda}{6}\left(\int \|u\|_2^6 \rho_U^*(du) + \int \|w\|_2^6 \rho_W^*(dw)\right) \leq \frac{1}{2}M_x^2.$$

giving us our bound. □

For any $\rho_U \in \mathcal{P}_6(\mathbb{R}^D)$, let us define ($\beta \in \mathbb{R}^b$)

$$f_x(\beta) := \frac{1}{2}\left\| \int \psi(\beta;u)\rho_U(du) - x \right\|_2^2, \quad g(\beta) := \int \|\psi(\beta;u)\|_2^2 \rho_U(du).$$

Lemma 12 (Uniform Hessian bounds in the bottleneck variable)**.** *Let $\rho_U \in \mathcal{P}_6(\mathbb{R}^D)$ satify $\int \|u\|_2^6 \, \rho_U(du) \le K_\lambda$, where $K_\lambda := \max\left\{1, \frac{3M_x^2}{\lambda}\right\}$. Then, for every $x \in \mathrm{supp}(P)$ and every $\beta \in \mathbb{R}^b$,*

$$\|D^2 f_x(\beta)\|_{\mathrm{op}} \le H_f, \qquad \|D^2 g(\beta)\|_{\mathrm{op}} \le H_g,$$

for some constants H_f and H_g.

Proof. The moment hypothesis implies, by Lyapunov's inequality, that for every $1 \le p \le 6$,

$$\int \|u\|_2^p \, \rho_U(du) \le \left(\int \|u\|_2^6 \, \rho_U(du) \right)^{p/6} \le K_\lambda^{p/6}.$$

As the $\sigma, \sigma', \sigma''$ and the moments are bounded, we can differentiate under the integral. Let $m(\beta) := \int \psi(\beta;u)\rho_U(du)$. We first bound m and its first two derivatives.

$$\|m(\beta)\|_2 \le S_0 \int \|u_1\|_2 \, \rho_U(du) \le S_0 \int \|u\|_2 \, \rho_U(du) \le S_0 K_\lambda^{1/6},$$

$$Dm(\beta) = \int u_1 u_2^\top \sigma'(\langle u_2, \beta\rangle) \, \rho_U(du).$$

$$\|Dm(\beta)\|_{\mathrm{op}} \le S_1 \int \|u_1\|_2 \|u_2\|_2 \, \rho_U(du) \le S_1 \int \|u\|_2^2 \, \rho_U(du) \le S_1 K_\lambda^{1/3}$$

Similarly, the second derivative of m is the bilinear map

$$D^2 m(\beta)[h,k] = \int u_1 \, \sigma''(\langle u_2, \beta\rangle)\langle u_2, h\rangle\langle u_2, k\rangle \, \rho_U(du).$$

$$\|D^2 m(\beta)\|_{\mathrm{op}} \le S_2 \int \|u_1\|_2 \|u_2\|_2^2 \, \rho_U(du) \le S_2 \int \|u\|_2^3 \, \rho_U(du) \le S_2 K_\lambda^{1/2}.$$

We now bound the Hessian of $f_x(\beta) = \frac{1}{2}\|m(\beta) - x\|_2^2$. For any direction $h \in \mathbb{R}^b$,

$$D^2 f_x(\beta)[h,h] = \|Dm(\beta)h\|_2^2 + \left\langle m(\beta) - x, \; D^2 m(\beta)[h,h] \right\rangle.$$

Using $\|x\|_2 \le M_x$, the bounds above give

$$|D^2 f_x(\beta)[h,h]| \le (S_1 K_\lambda^{1/3})^2 \|h\|_2^2 + (S_0 K_\lambda^{1/6} + M_x)(S_2 K_\lambda^{1/2}) \|h\|_2^2.$$

Thus

$$\|D^2 f_x(\beta)\|_{\mathrm{op}} \le (S_1 K_\lambda^{1/3})^2 + (M_x + S_0 K_\lambda^{1/6})(S_2 K_\lambda^{1/2}) = H_f.$$

It remains to bound the Hessian of g. Since

$$\|\psi(\beta;u)\|_2^2 = \|u_1\|_2^2 \sigma(\langle u_2, \beta\rangle)^2,$$

differentiating twice gives

$$D^2 g(\beta) = 2 \int \|u_1\|_2^2 \left[\left(\sigma'(\langle u_2, \beta\rangle)\right)^2 + \sigma(\langle u_2, \beta\rangle)\sigma''(\langle u_2, \beta\rangle) \right] u_2 u_2^\top \, \rho_U(du).$$

Therefore

$$\|D^2 g(\beta)\|_{\mathrm{op}} \le 2(S_1^2 + S_0 S_2) \int \|u_1\|_2^2 \|u_2\|_2^2 \, \rho_U(du).$$

Since

$$\|u_1\|_2^2 \|u_2\|_2^2 \le \|u\|_2^4,$$

we get

$$\|D^2 g(\beta)\|_{\mathrm{op}} \le 2(S_1^2 + S_0 S_2) \int \|u\|_2^4 \, \rho_U(du) \le 2(S_1^2 + S_0 S_2) K_\lambda^{2/3} = H_g.$$

This proves the claim. □

H.2 Proof of Theorem 3

For the convenience of the reader, we restate Theorem 3, which establishes the closeness between the empirical finite-N risk and the continuous regularized mean-field risk.

Theorem 3 (Closeness between the finite-width and mean-field optimal risks)**.** *Let Assumption 1 hold and let $\lambda > 0$. Then $\mathcal{R}_\lambda$ attains its minimum on $\mathcal{P}_6(\mathbb{R}^D) \times \mathcal{P}_6(\mathbb{R}^D)$, and every minimizer (ρ_U^*, ρ_W^*) satisfies $\int \|u\|_2^6 \rho_U^*(du) + \int \|w\|_2^6 \rho_W^*(dw) \le 3M_x^2/\lambda$. There are constants $A_\lambda, B_\lambda > 0$, depending on M_x, S_0, S_1, S_2 and λ, such that for every $N \ge 1$,*

$$0 \;\le\; \inf_{U,W} \mathcal{R}_\lambda^N(U,W) \;-\; \min_{\rho_U,\rho_W \in \mathcal{P}_6(\mathbb{R}^D)} \mathcal{R}_\lambda(\rho_U,\rho_W) \;\le\; \frac{A_\lambda}{N} + \frac{B_\lambda}{N^2}. \tag{11}$$

Proof. The existence of a minimizer and the sixth-moment bound follow from Lemma 11. It remains to compare the optimal finite-width risk with the optimal mean-field risk.

We first prove the lower bound in eq. (11). Let $\hat{\rho}_U^N := \frac{1}{N}\sum_{j=1}^N \delta_{u_j}$ and $\hat{\rho}_W^N := \frac{1}{N}\sum_{i=1}^N \delta_{w_i}$ be the empirical measures associated with the $U = (u_1, \dots, u_N)$ and $W = (w_1, \dots, w_N)$. So, $\mathcal{R}_\lambda^N(U,W) = \mathcal{R}_\lambda(\hat{\rho}_U^N, \hat{\rho}_W^N)$. Since the mean-field minimum is taken over all probability measures in $\mathcal{P}_6(\mathbb{R}^D)^2$, while empirical measures are only a special subclass, we easily get the lower bound,

$$\inf_{U,W} \mathcal{R}_\lambda^N(U,W) \ge \min_{\rho_U,\rho_W \in \mathcal{P}_6(\mathbb{R}^D)} \mathcal{R}_\lambda(\rho_U,\rho_W).$$

Let $K_\lambda := \max\left\{1, \frac{3M_x^2}{\lambda}\right\}$ and let (ρ_U^*, ρ_W^*) be a minimizer of $\mathcal{R}_\lambda$. By Lemma 11,

$$\int \|u\|_2^6 \rho_U^*(du) + \int \|w\|_2^6 \rho_W^*(dw) \le \frac{3M_x^2}{\lambda} \le K_\lambda.$$

We approximate this optimal mean-field pair by random empirical measures. Let

$$u_1, \dots, u_N \overset{\text{iid}}{\sim} \rho_U^*, \qquad w_1, \dots, w_N \overset{\text{iid}}{\sim} \rho_W^*.$$

For a fixed data point x, define the population and empirical bottlenecks by

$$\alpha(x) := \int \phi(x;w)\rho_W^*(dw), \quad \alpha_N(x) := \frac{1}{N}\sum_{i=1}^N \phi(x;w_i).$$

Since $\mathbb{E}_W[\alpha_N(x)] = \alpha(x)$, independence gives

$$\mathbb{E}_W\|\alpha_N(x) - \alpha(x)\|_2^2 = \frac{1}{N}\mathbb{E}_{w\sim\rho_W^*}\|\phi(x;w) - \alpha(x)\|_2^2.$$

Using $\mathbb{E}\|Z - \mathbb{E}Z\|_2^2 \le \mathbb{E}\|Z\|_2^2$ and the bound $\|\phi(x;w)\|_2 \le S_0\|w\|_2$, we obtain

$$\mathbb{E}_W\|\alpha_N(x) - \alpha(x)\|_2^2 \le \frac{1}{N}\int S_0^2\|w\|_2^2\rho_W^*(dw) \le \frac{S_0^2 K_\lambda^{1/3}}{N},$$

Now we condition on the encoder sample W and average only over the decoder particles. For fixed x and W, let $Y_j := \psi(\alpha_N(x); u_j)$. Let's recall some notations we used earlier lemmas:

$$m(\beta) := \int \psi(\beta;u)\rho_U^*(du),\; g(\beta) := \int \|\psi(\beta;u)\|_2^2\rho_U^*(du),\; f_x(\beta) := \frac{1}{2}\|m(\beta) - x\|_2^2.$$

Then the random variables $Y_1, \dots, Y_N$ are iid conditionally on W, with conditional mean $m(\alpha_N(x))$. Calculating the variance of the average, we get

$$\begin{aligned}\mathbb{E}_U\left[\frac{1}{2}\left\|\frac{1}{N}\sum_{j=1}^N Y_j - x\right\|_2^2 \middle| W\right] &= f_x(\alpha_N(x)) + \frac{1}{2N}\left[g(\alpha_N(x)) - \|m(\alpha_N(x))\|_2^2\right] \\ &\le f_x(\alpha_N(x)) + \frac{1}{2N}g(\alpha_N(x)).\end{aligned}$$

We now average over the encoder sample. The random variable $\alpha_N(x)$ is an unbiased approximation of $\alpha(x)$, so the first-order Taylor term vanishes in expectation. For any C^2 function $h : \mathbb{R}^b \to \mathbb{R}$, the Taylor's formula gives

$$h(\alpha_N) = h(\alpha) + \nabla h(\alpha)^\top(\alpha_N - \alpha) + \int_0^1 (1-s)(\alpha_N - \alpha)^\top D^2 h(\alpha + s(\alpha_N - \alpha))(\alpha_N - \alpha)\, ds.$$

Taking expectation over W and using $\mathbb{E}_W[\alpha_N - \alpha] = 0$, we get

$$\mathbb{E}_W[h(\alpha_N)] \le h(\alpha) + \frac{1}{2} \sup_{\beta \in \mathbb{R}^b} \|D^2 h(\beta)\|_{\mathrm{op}}\, \mathbb{E}_W\|\alpha_N - \alpha\|_2^2.$$

We apply this to $h = f_x$ and then to $h = g$, and use Lemma 12 to get

$$\mathbb{E}_W[f_x(\alpha_N(x))] \le f_x(\alpha(x)) + \frac{H_f S_0^2 K_\lambda^{1/3}}{2N},$$

$$\mathbb{E}_W[g(\alpha_N(x))] \le g(\alpha(x)) + \frac{H_g S_0^2 K_\lambda^{1/3}}{2N}.$$

Moreover,

$$g(\alpha(x)) = \int \|\psi(\alpha(x); u)\|_2^2 \rho_U^*(du) \le S_0^2 \int \|u\|_2^2 \rho_U^*(du) \le S_0^2 K_\lambda^{1/3}.$$

Combining the last three estimates gives, for every fixed x,

$$\mathbb{E}_{U,W}\left[\frac{1}{2}\left\|\frac{1}{N}\sum_{j=1}^N \psi(\alpha_N(x); u_j) - x\right\|_2^2\right] \le f_x(\alpha(x)) + \frac{1}{2N}\left(H_f S_0^2 K_\lambda^{1/3} + S_0^2 K_\lambda^{1/3}\right) + \frac{H_g S_0^2 K_\lambda^{1/3}}{4N^2}.$$

Integrating this over $x \sim P$ we get

$$\mathbb{E}_{U,W}\left[\frac{1}{2}\mathbb{E}_x\left\|\hat{x}_N(x) - x\right\|_2^2\right] \le \mathcal{R}^{\mathrm{rec}}(\rho_U^*, \rho_W^*) + \frac{A_\lambda}{N} + \frac{B_\lambda}{N^2},$$

where

$$A_\lambda := \frac{1}{2}\left(H_f S_0^2 K_\lambda^{1/3} + S_0^2 K_\lambda^{1/3}\right), \qquad B_\lambda := \frac{1}{4} H_g S_0^2 K_\lambda^{1/3}.$$

Finally, the expected empirical regularization terms agree exactly with the mean-field regularization terms:

$$\mathbb{E}_U\left[\frac{\lambda}{6N}\sum_{j=1}^N \|u_j\|_2^6\right] = \frac{\lambda}{6}\int \|u\|_2^6 \rho_U^*(du),$$

$$\mathbb{E}_W\left[\frac{\lambda}{6N}\sum_{i=1}^N \|w_i\|_2^6\right] = \frac{\lambda}{6}\int \|w\|_2^6 \rho_W^*(dw).$$

Adding the reconstruction and regularization bounds, we obtain

$$\mathbb{E}_{U,W}\left[\mathcal{R}_\lambda^N(U,W)\right] \le \mathcal{R}_\lambda(\rho_U^*, \rho_W^*) + \frac{A_\lambda}{N} + \frac{B_\lambda}{N^2}.$$

Since the expectation of a random variable is at least its infimum over all realizations, there exists at least one realization of the sampled particles such that

$$\mathcal{R}_\lambda^N(U,W) \le \mathcal{R}_\lambda(\rho_U^*, \rho_W^*) + \frac{A_\lambda}{N} + \frac{B_\lambda}{N^2}.$$

Therefore

$$\inf_{U,W} \mathcal{R}_\lambda^N(U,W) \le \mathcal{R}_\lambda(\rho_U^*, \rho_W^*) + \frac{A_\lambda}{N} + \frac{B_\lambda}{N^2}.$$

Since (ρ_U^*, ρ_W^*) is a minimizer of the mean-field risk, this is exactly the desired upper bound. □

I VARIATIONAL CHARACTERIZATION OF LOCAL MINIMA

In this section, we prove the following theorem giving a variational characterization of the local minima of the risk functional $\mathcal{R}_\lambda$. The proof follows closely that of [43].

Theorem 4 (Variational characterization of global minima)**.** *Let Assumption 1 hold, let $\lambda > 0$, and let (ρ_U^*, ρ_W^*) be a global minimizer of $\mathcal{R}_\lambda$ on $\mathcal{P}_6(\mathbb{R}^D)^2$. With Φ_U, Φ_W the first variations* (4)–(5)*,*

$$\phi_U^* := \inf_u \Phi_U(u; \rho_U^*, \rho_W^*) > -\infty \text{ and } \operatorname{supp}(\rho_U^*) \subseteq \arg\min_u \Phi_U(u; \rho_U^*, \rho_W^*);$$

$$\phi_W^* := \inf_w \Phi_W(w; \rho_U^*, \rho_W^*) > -\infty \text{ and } \operatorname{supp}(\rho_W^*) \subseteq \arg\min_w \Phi_W(w; \rho_U^*, \rho_W^*).$$

Proof. We evaluate the first variations at the fixed minimizer (ρ_U^*, ρ_W^*) and write

$$\Phi_U(u) := \Phi_U(u; \rho_U^*, \rho_W^*), \qquad \Phi_W(w) := \Phi_W(w; \rho_U^*, \rho_W^*).$$

First note that the unregularized parts of Φ_U and Φ_W grow at most polynomially of lower order than $\|u\|_2^6$ and $\|w\|_2^6$, while the regularization contributes $\frac{\lambda}{6}\|u\|_2^6$ and $\frac{\lambda}{6}\|w\|_2^6$. Hence Φ_U and Φ_W are continuous and coercive. In particular, Φ_U and Φ_W are bounded below. Let us denote their infimum by

$$\phi_U^* := \inf_u \Phi_U(u) > -\infty, \qquad \phi_W^* := \inf_w \Phi_W(w) > -\infty.$$

We first consider the decoder measure ρ_U^*. Let

$$S_U := \arg\min_u \Phi_U(u) = \{u \in \mathbb{R}^D : \Phi_U(u) = \phi_U^*\}.$$

Assume to the contrary that there exists some$u_0 \in \operatorname{supp}(\rho_U^*)$ with $u_0 \notin S_U$; then $\Phi_U(u_0) > \phi_U^*$. That means that there exists a $\Delta > 0$ such that

$$\Phi_U(u_0) \geq \phi_U^* + 2\Delta.$$

As Φ_U is continuous, there is a radius $r_0 > 0$ such that

$$\Phi_U(u) \geq \phi_U^* + \Delta \qquad \text{for all } u \in B(u_0, r_0).$$

Since $u_0 \in \operatorname{supp}(\rho_U^*)$, the ball has positive ρ_U^*-mass. Let $t_0 := \rho_U^*(B(u_0, r_0)) > 0$ and $\nu_U := \frac{1}{t_0}\rho_U^*\big|_{B(u_0,r_0)}$, then ν_U is a probability measure supported on the region where $\Phi_U \geq \phi_U^* + \Delta$. Let us pick any point $\bar{u} \in S_U$ and then define the signed measure $\Delta\rho_U := \delta_{\bar{u}} - \nu_U$ and for any $0 \leq t \leq t_0$, $\rho_U^t := \rho_U^* + t\Delta\rho_U = \rho_U^* - t\nu_U + t\delta_{\bar{u}}$.

For a fixed ρ_W^*, the regularized risk is quadratic in ρ_U in the reconstruction part and linear in ρ_U in the regularization part. So there exists a finite constant C_U such that t small enough

$$\mathcal{R}_\lambda(\rho_U^t, \rho_W^*) - \mathcal{R}_\lambda(\rho_U^*, \rho_W^*) \leq t \int \Phi_U(u)\, d\Delta\rho_U(u) + C_U t^2.$$

The linear term in the above expression is strictly negative as

$$\int \Phi_U(u)\, d\Delta\rho_U(u) = \Phi_U(\bar{u}) - \int \Phi_U(u)\, \nu_U(du) \leq \phi_U^* - (\phi_U^* + \Delta) = -\Delta.$$

Therefore

$$\mathcal{R}_\lambda(\rho_U^t, \rho_W^*) - \mathcal{R}_\lambda(\rho_U^*, \rho_W^*) \leq -t\Delta + C_U t^2.$$

Now choosing $t > 0$ small enough so that $t \leq t_0$ and $C_U t < \Delta$ gives

$$\mathcal{R}_\lambda(\rho_U^t, \rho_W^*) < \mathcal{R}_\lambda(\rho_U^*, \rho_W^*),$$

which contradicts the fact that (ρ_U^*, ρ_W^*) is a global minima. Hence

$$\operatorname{supp}(\rho_U^*) \subseteq \arg\min_u \Phi_U(u; \rho_U^*, \rho_W^*).$$

We now focus on the encoder measure ρ_W^*. Although the argument is similar, but the expansion is no longer exactly quadratic, because ρ_W enters the risk through the nonlinear bottleneck variable. Let

$$S_W := \arg\min_w \Phi_W(w) = \{w \in \mathbb{R}^D : \Phi_W(w) = \phi_W^*\},$$

assume that there exists $w_0 \in \mathrm{supp}(\rho_W^*)$ with $w_0 \notin S_W$. Again, by continuity of Φ_W, there exist $\Delta > 0$ and $r_0 > 0$ such that $\Phi_W(w) \geq \phi_W^* + \Delta$ for all $w \in B(w_0, r_0)$. Since $w_0 \in \mathrm{supp}(\rho_W^*)$, let $t_0 := \rho_W^*(B(w_0, r_0)) > 0$ and $\nu_W := \frac{1}{t_0}\rho_W^*\big|_{B(w_0,r_0)}$. Now let us pick $\bar{w} \in S_W$ and set $\Delta\rho_W := \delta_{\bar{w}} - \nu_W$ and for $0 \leq t \leq t_0$, define $\rho_W^t := \rho_W^* + t\Delta\rho_W = \rho_W^* - t\nu_W + t\delta_{\bar{w}}$. The bottleneck map is linear in ρ_W, so if $\alpha^*(x) := \alpha(x; \rho_W^*)$, then $\alpha(x; \rho_W^t) = \alpha^*(x) + t\,\Delta\alpha(x)$, where $\Delta\alpha(x) := \int \phi(x; w)\, d\Delta\rho_W(w)$. The signed measure $\Delta\rho_W$ is supported on the compact set $B(w_0, r_0) \cup \{\bar{w}\}$. Hence $\Delta\alpha(x)$ is uniformly bounded for $x \in \mathrm{supp}(P)$.

We now expand the reconstruction map in the bottleneck variable. Since $\psi(\alpha; u)$ is C^2 in α and the first and second derivatives are integrable under ρ_U^*, applying Taylor's theorem we get

$$\int \psi(\alpha^*(x) + t\Delta\alpha(x); u)\, \rho_U^*(du) = \hat{x}^*(x) + t\,\overline{J}^*(x)\Delta\alpha(x) + t^2 R_t(x),$$

where

$$\hat{x}^*(x) := \int \psi(\alpha^*(x); u)\rho_U^*(du), \qquad \overline{J}^*(x) := \int J_\alpha \psi(\alpha^*(x); u)\rho_U^*(du),$$

and the remainder satisfies

$$\sup_{0\leq t\leq t_0} \sup_{x\in\mathrm{supp}(P)} \|R_t(x)\|_2 < \infty.$$

The uniform bound follows from the boundedness of $\Delta\alpha(x)$ and the uniform second-derivative bounds for ψ in the bottleneck variable. Let $r^*(x) := \hat{x}^*(x) - x$, and $\tau_1(x) := \overline{J}^*(x)\Delta\alpha(x)$, then the perturbed reconstruction has the form

$$\hat{x}^t(x) - x = r^*(x) + t\tau_1(x) + t^2 R_t(x).$$

Expanding the squared norm and using the uniform bounds on r^*, τ_1, and R_t, we obtain

$$\frac{1}{2}\mathbb{E}_x\|\hat{x}^t(x) - x\|_2^2 - \frac{1}{2}\mathbb{E}_x\|r^*(x)\|_2^2 = t\,\mathbb{E}_x\langle r^*(x), \tau_1(x)\rangle + O(t^2).$$

The $O(t^2)$ term is bounded by $C_W t^2$ for some finite constant C_W. The regularization term in ρ_W is linear, so its first-order term is $t\int \frac{\lambda}{6}\|w\|_2^6\, d\Delta\rho_W(w)$. Combining this with the first-order reconstruction term gives

$$\begin{aligned}
&\mathbb{E}_x\langle r^*(x), \tau_1(x)\rangle + \int \frac{\lambda}{6}\|w\|_2^6\, d\Delta\rho_W(w) \\
&\quad = \int \left[\mathbb{E}_x\left\langle \overline{J}^*(x)^\top r^*(x), \phi(x; w)\right\rangle + \frac{\lambda}{6}\|w\|_2^6\right] d\Delta\rho_W(w) \\
&\quad = \int \Phi_W(w)\, d\Delta\rho_W(w).
\end{aligned}$$

$$\mathcal{R}_\lambda(\rho_U^*, \rho_W^t) - \mathcal{R}_\lambda(\rho_U^*, \rho_W^*) \leq t\int \Phi_W(w)\, d\Delta\rho_W(w) + C_W t^2.$$

Similar to the decoder case, we have

$$\int \Phi_W(w)\, d\Delta\rho_W(w) = \Phi_W(\bar{w}) - \int \Phi_W(w)\, \nu_W(dw) \leq \phi_W^* - (\phi_W^* + \Delta) = -\Delta.$$

$$\mathcal{R}_\lambda(\rho_U^*, \rho_W^t) - \mathcal{R}_\lambda(\rho_U^*, \rho_W^*) \leq -t\Delta + C_W t^2.$$

Now choosing $t > 0$ small enough so that $t \leq t_0$ and $C_W t < \Delta$, we get

$$\mathcal{R}_\lambda(\rho_U^*, \rho_W^t) < \mathcal{R}_\lambda(\rho_U^*, \rho_W^*),$$

again contradicting the global minimality of (ρ_U^*, ρ_W^*). Hence

$$\mathrm{supp}(\rho_W^*) \subseteq \arg\min_w \Phi_W(w; \rho_U^*, \rho_W^*).$$

□

J Main Result: Mean-Field Dynamics and Propagation of Chaos

Before giving the proof of the main theorem, we recall the key quantities governing the mean-field dynamics (see Section C and Section D). The Vlasov–McKean dynamics are driven by the spatial gradients of the regularized first variations of the risk functional $\Phi_U = \frac{\delta \mathcal{R}_\lambda}{\delta \rho_U}$ and $\Phi_W = \frac{\delta \mathcal{R}_\lambda}{\delta \rho_W}$, given by

$$\Phi_U(u;\rho_U,\rho_W) = \mathbb{E}_x\big[\langle r(x;\rho_U,\rho_W), \psi(\alpha(x;\rho_W);u)\rangle\big] + \frac{\lambda}{6}\|u\|_2^6,$$
$$\Phi_W(w;\rho_U,\rho_W) = \mathbb{E}_x\big[\langle \overline{J}(x;\rho_U,\rho_W)^\top r(x;\rho_U,\rho_W),\, \phi(x;w)\rangle\big] + \frac{\lambda}{6}\|w\|_2^6,$$

where $r = \hat{x} - x$ is the residual and $\overline{J}(x;\rho_U,\rho_W) = \int J_\alpha \psi(\alpha(x;\rho_W);u)\rho_U(du) \in \mathbb{R}^{d\times b}$ is the averaged decoder Jacobian. The particle dynamics are governed by the negative spatial gradients of these potentials. For the decoder, differentiating Φ_U with respect to u yields the regularized continuous vector field V_U:

$$V_U(u;\rho_U,\rho_W) := -\nabla_u \Phi_U(u;\rho_U,\rho_W) = \underbrace{-\mathbb{E}_x\left[J_u\psi(\alpha(x;\rho_W);u)^\top r(x;\rho_U,\rho_W)\right]}_{=:G_U(u;\rho_U,\rho_W)} - \lambda\|u\|_2^4 u. \tag{58}$$

For the encoder, we differentiate Φ_W with respect to w. Observe that r and $\overline{J}$ depend only on the measures and hence are constant with respect to the microscopic variable w. Thus, the regularized continuous vector field V_W is

$$V_W(w;\rho_U,\rho_W) := -\nabla_w \Phi_W(w;\rho_U,\rho_W) = \underbrace{-\mathbb{E}_x\left[J_w\phi(x;w)^\top \overline{J}(x;\rho_U,\rho_W)^\top r(x;\rho_U,\rho_W)\right]}_{=:G_W(w;\rho_U,\rho_W)} - \lambda\|w\|_2^4 w. \tag{59}$$

These regularized vector fields drive the coupled Vlasov–McKean system (6). Let us also recall the gradients driving SGD from Section D. At iteration k a single sample $x_{k+1} \sim P$ is drawn, and with α_k, r_k and $g_k = \overline{J}_k^\top r_k$ as in (15), (16), (18), the single-sample unregularized descent fields are

$$F_U\big(u;\hat{\rho}_{U,k}^N,\hat{\rho}_{W,k}^N,x_{k+1}\big) := -J_u\psi(\alpha_k;u)^\top r_k,$$
$$F_W\big(w;\hat{\rho}_{U,k}^N,\hat{\rho}_{W,k}^N,x_{k+1}\big) := -J_w\phi(x_{k+1};w)^\top g_k,$$

see (19) and (20), and the SGD updates are (7). We distinguish between the continuous vector fields (G_U, G_W) and the gradients used by SGD (F_U, F_W). While G represents the ideal, noise-free population gradient averaged over the entire data distribution, SGD computes a noisy gradient using a single random data sample. Since x_{k+1} is drawn independently of the current configuration, the conditional expectations of the single-sample fields are exactly the population fields evaluated at the empirical measures,

$$\begin{aligned}\mathbb{E}\big[F_U(u;\hat{\rho}_{U,k}^N,\hat{\rho}_{W,k}^N,x_{k+1}) \,\big|\, U^k, W^k\big] &= G_U(u;\hat{\rho}_{U,k}^N,\hat{\rho}_{W,k}^N)\\ \mathbb{E}\big[F_W(w;\cdot) \,\big|\, U^k, W^k\big] &= G_W(w;\hat{\rho}_{U,k}^N,\hat{\rho}_{W,k}^N).\end{aligned} \tag{60}$$

For the convenience of the reader we restate the main theorem.

Theorem 2 (Mean-field dynamics and propagation of chaos)**.** *Let Assumption 1 hold, let $\lambda > 0$, and let $(\rho_{U,t},\rho_{W,t})_{t\geq 0}$ be the solution of* (6) *given by Theorem 1. Consider the SGD iterates* (7) *initialized i.i.d. from $(\rho_{U,0},\rho_{W,0})$ with step size $s_k = \epsilon\,\xi(k\epsilon)$, $\epsilon M_\xi \leq s_*$. Fix $T > 0$ and define* $\mathrm{err}_{N,\epsilon}(\delta) := \sqrt{N^{-1}\vee\epsilon}\,\left(\sqrt{D + \log\big(N\,(\frac{T}{\epsilon}\vee 1)\big)} + \delta\right)$*. There exist constants $C_1, C_2 > 0$, depending only on $b, d, M_x, S_0, S_1, S_2, \lambda, R_*, L_\xi, M_\xi$, such that for every $\delta > 0$ and every fixed test function f with $\|f\|_\infty, \|f\|_{\mathrm{Lip}} \leq 1$, with probability at least $1 - 7e^{-\delta^2}$,*

$$\max_{k\in[0,T/\epsilon]\cap\mathbb{N}} \left|\frac{1}{N}\sum_{j=1}^N f(u_j^k) - \int f\, d\rho_{U,k\epsilon}\right| \leq C_1(\sqrt{T}\vee T\vee 1)\, e^{C_2T^2}\,\mathrm{err}_{N,\epsilon}(\delta), \tag{9}$$

$$\max_{k\in[0,T/\epsilon]\cap\mathbb{N}} \left|\frac{1}{N}\sum_{i=1}^N f(w_i^k) - \int f\, d\rho_{W,k\epsilon}\right| \leq C_1(\sqrt{T}\vee T\vee 1)\, e^{C_2T^2}\,\mathrm{err}_{N,\epsilon}(\delta),$$

$$\max_{k\in[0,T/\epsilon]\cap\mathbb{N}} \left|\mathcal{R}_\lambda^N(U^k,W^k) - \mathcal{R}_\lambda(\rho_{U,k\epsilon},\rho_{W,k\epsilon})\right| \leq C_1(\sqrt{T}\vee T\vee 1)\, e^{C_2T^2}\,\mathrm{err}_{N,\epsilon}(\delta). \tag{10}$$

Consequently, for every fixed $t \geq 0$, along any sequence $(N, \epsilon = \epsilon_N)$ with $N \to \infty$, $\epsilon_N \to 0$, $N/\log(N/\epsilon_N) \to \infty$ and $\epsilon_N \log(N/\epsilon_N) \to 0$, the empirical measures converge weakly, almost surely,

$$\hat{\rho}^N_{U,\lfloor t/\epsilon_N \rfloor} \Rightarrow \rho_{U,t}, \qquad \hat{\rho}^N_{W,\lfloor t/\epsilon_N \rfloor} \Rightarrow \rho_{W,t}.$$

J.1 Uniform constants on the trapping ball

By Theorem 6, under the step condition $\epsilon M_\xi \leq s_*$, every SGD particle stays in $B_D(0, R_*)$ for every $k \geq 0$, and by Theorem 7 every mean-field characteristic stays in the same ball. Consequently, all the quantities appearing in this section are evaluated on the fixed compact set $B_D(0, R_*)$, and we now collect the constants that we shall use. Recall from (22) that $E_* := M_x + S_0 R_*$ and from (45) that

$$C_{R_*} = 2(1+S_0)(S_0+2S_1+S_2)(1+M_x)^2(1+R_*)^3, \quad C_* := \max\{1, C_{R_*}\}.$$

By Lemma 8 (with $R = R_*$), for all $x \in \mathrm{supp}(P)$ and $(\rho_U, \rho_W) \in \mathcal{P}^2_{R_*}$,

$$\|\alpha(x;\rho_W)\|_2 \leq S_0 R_*,\ \|\hat{x}(x;\rho_U,\rho_W)\|_2 \leq S_0 R_*,\ \|r(x;\rho_U,\rho_W)\|_2 \leq E_*,\ \|\overline{J}(x;\rho_U,\rho_W)\|_{\mathrm{op}} \leq S_1 R_*^2, \tag{61}$$

and the same bounds hold for the empirical quantities $\alpha_k, \hat{x}_k, r_k, \overline{J}_k$ by Lemma 7. By Lemma 9, whenever $\|x\|_2 \leq M_x$, $\|u\|_2, \|w\|_2 \leq R_*$ and $\|\alpha\|_2 \leq S_0 R_*$, the maps

$$\phi(x;w), \quad \psi(\alpha;u), \quad J_w\phi(x;w), \quad J_u\psi(\alpha;u), \quad J_\alpha\psi(\alpha;u)$$

are bounded by C_* and are C_*-Lipschitz jointly in all of their arguments (x, w, α, u). We shall use in particular that ψ, $J_u\psi$ and $J_\alpha\psi$ are C_*-Lipschitz in the bottleneck variable α, that ϕ and $J_w\phi$ are C_*-Lipschitz in x, and that ψ, $J_\alpha\psi$ (resp. ϕ) are C_*-Lipschitz in the particle u (resp. w). Since $J_\alpha\psi(\alpha;u) = \sigma'(\langle u_2, \alpha\rangle) u_1 u_2^\top$ has rank one, its operator and Frobenius norms coincide, so all statements about $J_\alpha\psi$ hold in both norms.

By Lemmas 7 and 8, on $B_D(0, R_*)$ we have $\|F_U\|_2,\ \|G_U\|_2 \leq B_U(R_*)$, $\|F_W\|_2,\ \|G_W\|_2 \leq B_W(R_*)$, and $B_* := \max\{B_U(R_*), B_W(R_*)\} \leq C_{\mathrm{grad}} R_*^4$, where $B_U(R) = S_0 E_R \sqrt{1 + S_1^2 R^4}$ and $B_W(R) = S_1 R^2 E_R \sqrt{S_0^2 + S_1^2 M_x^2 R^2}$, and hence, by (26)–(27),

$$\|V_U(u;\rho_U,\rho_W)\|_2 \leq M_*, \quad \|V_W(w;\rho_U,\rho_W)\|_2 \leq M_*, \quad M_* = B_* + \lambda R_*^5. \tag{62}$$

In particular, every mean-field characteristic moves with speed at most $M_\xi M_*$. Finally, we define the two Lipschitz constants that govern the whole tracking argument,

$$L_{\mathrm{sp}} := C_* E_* \big(1 + S_1 R_*^2\big) + 5\lambda R_*^4, \quad L_{\mathrm{me}} := C_*^2 \,(E_* + C_*), \tag{63}$$

which will bound the Lipschitz dependence of V_U, V_W on the particle variable and on the pair of measures, respectively (Lemma 13).

J.2 Lipschitz Continuity via Parametric Distances

We now establish the Lipschitz continuity of the regularized vector fields on the trapping ball. Lemma 10 already gives such a statement with respect to the Wasserstein distance. For the propagation of chaos argument, we need a sharper statement: the dependence on the measures must be measured through the functions that the network actually evaluates, since these are the only functions for which the empirical concentration of Lemma 5 is dimension-free in N. We consider the following parametric test function classes. For the decoder, we define two separate families:

$$\mathcal{F}_{U,1} = \{u' \mapsto \psi(\alpha;u')\}_{\alpha\in\mathcal{A}} \quad \text{and} \quad \mathcal{F}_{U,2} = \{u' \mapsto J_\alpha\psi(\alpha;u')\}_{\alpha\in\mathcal{A}}, \quad \mathcal{A} := B_b(0, S_0 R_*),$$

where $\mathcal{A} \subset \mathbb{R}^b$ contains every bottleneck value that can occur, population or empirical (by (61)). The functions in $\mathcal{F}_{U,2}$ are $\mathbb{R}^{d\times b}$-valued; we identify $\mathbb{R}^{d\times b} \cong \mathbb{R}^{db}$ and measure them in the Frobenius norm, which dominates the operator norm. For the encoder, we define

$$\mathcal{F}_W = \{w' \mapsto \phi(x;w')\}_{x\in\mathrm{supp}(P)}.$$

These are the families of functions that our network *sees*: the encoder measure enters the fields only through $\alpha = \int \phi \, d\rho_W$, and the decoder measure only through $\hat{x} = \int \psi \, d\rho_U$ and $\overline{J} = \int J_\alpha\psi \, d\rho_U$. Their parametric distances, in the sense of Section E, are

$$d_{\mathcal{F}_{U,1}}(\mu,\nu) = \sup_{\alpha\in\mathcal{A}} \left\| \int \psi(\alpha;u') d(\mu-\nu)(u') \right\|_2,$$

$$d_{\mathcal{F}_{U,2}}(\mu,\nu) = \sup_{\alpha\in\mathcal{A}} \left\| \int J_\alpha \psi(\alpha;u')\, d(\mu-\nu)(u') \right\|_F,$$
$$d_{\mathcal{F}_W}(\mu,\nu) = \sup_{x\in\mathrm{supp}(P)} \left\| \int \phi(x;w')\, d(\mu-\nu)(w') \right\|_2.$$

We combine the decoder families into a single metric $d_{\mathcal{F}_U}(\mu,\nu) = \max\big(d_{\mathcal{F}_{U,1}}(\mu,\nu), d_{\mathcal{F}_{U,2}}(\mu,\nu)\big)$. The regularity of these classes follows directly from Lemma 9 and (61). Every $f \in \mathcal{F}_{U,1}\cup\mathcal{F}_{U,2}\cup\mathcal{F}_W$ on $B_D(0,R_*)$ has the following regularity: $\sup\|f\| \le C_*$, f is C_*-Lipschitz in the particle variable and $\theta \mapsto f_\theta$ is C_*-Lipschitz in the index θ, where the index is $\theta = \alpha \in \mathcal{A} \subset \mathbb{R}^b$ (diameter $2S_0R_*$) for the decoder classes and $\theta = x \in \mathrm{supp}(P) \subset \mathbb{R}^d$ (diameter at most $2M_x$) for the encoder class. Two consequences will be used repeatedly. First, by Lemma 6 (applied to $\mathbb{R}^p$-valued classes, with $p = d$, $p = db$, and $p = b$ respectively) and $d_{\mathrm{BL}} \le W_1$,

$$d_{\mathcal{F}_U}(\mu,\nu) \le C_*\, W_1(\mu,\nu), \quad d_{\mathcal{F}_W}(\mu,\nu) \le C_*\, W_1(\mu,\nu), \quad \mu,\nu \in \mathcal{P}_{R_*}. \tag{64}$$

Second, applying Lemma 5 with $M = L = C_*$, index dimension b or d (both at most D), and index diameter at most $R_{\max} := \max\{1, 2S_0R_*, 2M_x\}$ (recall that the covering bound in that lemma only increases with the diameter, so we may enlarge it to be at least one), we obtain: if $\mu \in \mathcal{P}_{R_*}$ and μ_N is the empirical measure of N i.i.d. samples from μ, then for each of the three classes $\mathcal{F} \in \{\mathcal{F}_{U,1}, \mathcal{F}_{U,2}, \mathcal{F}_W\}$ and every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$d_{\mathcal{F}}(\mu,\mu_N) \le \frac{1}{\sqrt{N}}\Big(C_{\mathcal{F}} + C_*\sqrt{D\log N} + C_*\sqrt{2\log(1/\delta)}\Big), \tag{65}$$

where $C_{\mathcal{F}} := 2C_* + C_*\sqrt{2D\log(3R_{\max})} + C_*\sqrt{2\log 2}$.

Lemma 13 (Lipschitz Continuity of Regularized Vector Fields via Parametric Distances)**.** *Let Assumption 1 hold and let L_{sp}, L_{me} be as in* (63)*.*

(a) (Spatial Lipschitz continuity.) For all $(\rho_U,\rho_W) \in \mathcal{P}^2_{R_*}$ and $u,u',w,w' \in B_D(0,R_*)$,

$$\|V_U(u;\rho_U,\rho_W) - V_U(u';\rho_U,\rho_W)\|_2 \le L_{\mathrm{sp}}\|u-u'\|_2,$$
$$\|V_W(w;\rho_U,\rho_W) - V_W(w';\rho_U,\rho_W)\|_2 \le L_{\mathrm{sp}}\|w-w'\|_2.$$

(b) (Lipschitz continuity in the measures.) For all $(\rho_U,\rho_W),(\hat\rho_U,\hat\rho_W) \in \mathcal{P}^2_{R_*}$ and $u,w \in B_D(0,R_*)$,

$$\|V_U(u;\hat\rho_U,\hat\rho_W) - V_U(u;\rho_U,\rho_W)\|_2 \le L_{\mathrm{me}}\Big(d_{\mathcal{F}_U}(\hat\rho_U,\rho_U) + d_{\mathcal{F}_W}(\hat\rho_W,\rho_W)\Big),$$
$$\|V_W(w;\hat\rho_U,\hat\rho_W) - V_W(w;\rho_U,\rho_W)\|_2 \le L_{\mathrm{me}}\Big(d_{\mathcal{F}_U}(\hat\rho_U,\rho_U) + d_{\mathcal{F}_W}(\hat\rho_W,\rho_W)\Big).$$

Proof. We first prove (a). With the measures fixed, $\alpha(x;\rho_W)$, $r(x;\rho_U,\rho_W)$ and $\overline{J}(x;\rho_U,\rho_W)$ are fixed functions of x. For the decoder, by (58), the C_*-Lipschitz continuity of $u \mapsto J_u\psi(\alpha;u)$ from Lemma 9, and $\|r\|_2 \le E_*$ from (61),

$$\begin{aligned}
&\|G_U(u;\rho_U,\rho_W) - G_U(u';\rho_U,\rho_W)\|_2 \\
&\le \mathbb{E}_x\big[\|J_u\psi(\alpha(x);u) - J_u\psi(\alpha(x);u')\|_{\mathrm{op}}\, \|r(x)\|_2\big] \\
&\le C_*E_*\|u-u'\|_2.
\end{aligned}$$

For the encoder, by (59), the C_*-Lipschitz continuity of $w \mapsto J_w\phi(x;w)$, and $\|\overline{J}^\top r\|_2 \le S_1R_*^2E_*$, we have

$$\begin{aligned}
&\|G_W(w;\rho_U,\rho_W) - G_W(w';\rho_U,\rho_W)\|_2 \\
&\le \mathbb{E}_x\big[\|J_w\phi(x;w) - J_w\phi(x;w')\|_{\mathrm{op}}\, \|\overline{J}(x)^\top r(x)\|_2\big] \\
&\le C_*S_1R_*^2E_*\|w-w'\|_2.
\end{aligned}$$

It remains to include the regularizer. By (57), the map $v \mapsto \|v\|_2^4 v$ is $5R_*^4$-Lipschitz on $B_D(0,R_*)$, so $v \mapsto \lambda\|v\|_2^4 v$ is $5\lambda R_*^4$-Lipschitz. Adding the two contributions, both fields are Lipschitz in the particle variable with constant at most $C_*E_*\max\{1,S_1R_*^2\} + 5\lambda R_*^4 \le L_{\mathrm{sp}}$.

We now prove (b). Because the regularizer terms $-\lambda\|u\|_2^4 u$ and $-\lambda\|w\|_2^4 w$ depend only on the individual particle and not on the measures, they cancel in the measure-wise variation:

$$\|V_U(u;\hat\rho_U,\hat\rho_W) - V_U(u;\rho_U,\rho_W)\|_2 = \|G_U(u;\hat\rho_U,\hat\rho_W) - G_U(u;\rho_U,\rho_W)\|_2,$$

and likewise for V_W. For both unregularized fields, we decouple the simultaneous variation of the two measures by adding and subtracting the mixed term $G(\rho_U,\hat\rho_W)$,

$$\|G(\hat\rho_U,\hat\rho_W) - G(\rho_U,\rho_W)\|_2 \le \underbrace{\|G(\hat\rho_U,\hat\rho_W) - G(\rho_U,\hat\rho_W)\|_2}_{\text{varying } \rho_U \text{ only}} + \underbrace{\|G(\rho_U,\hat\rho_W) - G(\rho_U,\rho_W)\|_2}_{\text{varying } \rho_W \text{ only}}.$$

Throughout, we write $\alpha = \alpha(x;\rho_W)$ and $\hat\alpha = \alpha(x;\hat\rho_W)$, and we indicate by $\hat x(\rho_U)$, $\hat x(\hat\alpha)$, $\overline{J}(\rho_U)$, $\overline{J}(\hat\alpha)$, etc. which argument is being varied. Since $w \mapsto \phi(x;w)$ belongs to $\mathcal{F}_W$ for every x, the bottleneck shift is bounded by

$$\|\hat\alpha - \alpha\|_2 = \Big\| \int \phi(x;w')\, d(\hat\rho_W - \rho_W)(w') \Big\|_2 \le d_{\mathcal{F}_W}(\hat\rho_W,\rho_W), \tag{66}$$

uniformly in x. Similarly, since $u' \mapsto \psi(\alpha;u')$ and $u' \mapsto J_\alpha\psi(\alpha;u')$ belong to $\mathcal{F}_{U,1}$ and $\mathcal{F}_{U,2}$ for every $\alpha \in \mathcal{A}$,

$$\begin{aligned} &\|\hat x(\hat\rho_U) - \hat x(\rho_U)\|_2 \le d_{\mathcal{F}_U}(\hat\rho_U,\rho_U) \\ &\|\overline{J}(\hat\rho_U) - \overline{J}(\rho_U)\|_{\mathrm{op}} \le \|\overline{J}(\hat\rho_U) - \overline{J}(\rho_U)\|_F \le d_{\mathcal{F}_U}(\hat\rho_U,\rho_U), \end{aligned} \tag{67}$$

with the encoder measure (hence $\hat\alpha$) held fixed. Finally, when the decoder measure is fixed and only the bottleneck moves, the C_*-Lipschitz continuity of ψ and $J_\alpha\psi$ in α gives

$$\begin{aligned} &\|\hat x(\hat\alpha) - \hat x(\alpha)\|_2 \le \int \|\psi(\hat\alpha;u') - \psi(\alpha;u')\|_2\, \rho_U(du') \le C_*\|\hat\alpha - \alpha\|_2, \\ &\|\overline{J}(\hat\alpha) - \overline{J}(\alpha)\|_{\mathrm{op}} \le C_*\|\hat\alpha - \alpha\|_2. \end{aligned} \tag{68}$$

Decoder field, varying ρ_U. Because the encoder measure $\hat\rho_W$ is fixed, $\hat\alpha$ is fixed and $J_u\psi(\hat\alpha;u)$ is constant. Using $\|J_u\psi\|_{\mathrm{op}} \le C_*$ and (67),

$$\begin{aligned} \|G_U(u;\hat\rho_U,\hat\rho_W) - G_U(u;\rho_U,\hat\rho_W)\|_2 &= \Big\|\mathbb{E}_x\Big[J_u\psi(\hat\alpha;u)^\top\big(\hat x(\hat\rho_U) - \hat x(\rho_U)\big)\Big]\Big\|_2 \\ &\le \mathbb{E}_x\Big[\|J_u\psi(\hat\alpha;u)\|_{\mathrm{op}}\|\hat x(\hat\rho_U) - \hat x(\rho_U)\|_2\Big] \le C_*\, d_{\mathcal{F}_U}(\hat\rho_U,\rho_U). \end{aligned}$$

Decoder field, varying ρ_W. Keeping ρ_U fixed, the bottleneck shifts from α to $\hat\alpha$. Adding and subtracting the mixed product term $J_u\psi(\alpha;u)^\top(\hat x(\hat\alpha) - x)$ gives

$$\begin{aligned} &\|G_U(u;\rho_U,\hat\rho_W) - G_U(u;\rho_U,\rho_W)\|_2 \\ &\le \mathbb{E}_x\Big[\|J_u\psi(\hat\alpha;u) - J_u\psi(\alpha;u)\|_{\mathrm{op}}\|\hat x(\hat\alpha) - x\|_2 + \|J_u\psi(\alpha;u)\|_{\mathrm{op}}\|\hat x(\hat\alpha) - \hat x(\alpha)\|_2\Big] \\ &\le \mathbb{E}_x\Big[(C_*\|\hat\alpha - \alpha\|_2)\, E_* + C_*\,(C_*\|\hat\alpha - \alpha\|_2)\Big] \\ &= C_*(E_* + C_*)\,\|\hat\alpha - \alpha\|_2 \le C_*(E_* + C_*)\, d_{\mathcal{F}_W}(\hat\rho_W,\rho_W), \end{aligned}$$

where we used the C_*-Lipschitz continuity of $J_u\psi$ in α, $\|r\|_2 \le E_*$, (68) and (66). Combining the two decoder estimates and using $C_* \ge 1$,

$$\|G_U(u;\hat\rho_U,\hat\rho_W) - G_U(u;\rho_U,\rho_W)\|_2 \le C_*(E_* + C_*)\Big(d_{\mathcal{F}_U}(\hat\rho_U,\rho_U) + d_{\mathcal{F}_W}(\hat\rho_W,\rho_W)\Big).$$

Encoder field, varying ρ_U. Keeping $\hat\rho_W$ fixed, the bottleneck $\hat\alpha$ is fixed and $J_w\phi(x;w)$ is constant with respect to ρ_U. Hence

$$\begin{aligned} &\|G_W(w;\hat\rho_U,\hat\rho_W) - G_W(w;\rho_U,\hat\rho_W)\|_2 \\ &\le \mathbb{E}_x\Big[\|J_w\phi(x;w)\|_{\mathrm{op}}\big\|\overline{J}(\hat\rho_U)^\top(\hat x(\hat\rho_U) - x) - \overline{J}(\rho_U)^\top(\hat x(\rho_U) - x)\big\|_2\Big]. \end{aligned}$$

We add and subtract $\overline{J}(\rho_U)^\top(\hat{x}(\hat{\rho}_U) - x)$ to isolate the differences:

$$\begin{aligned}\big\|\overline{J}(\hat{\rho}_U)^\top(\hat{x}(\hat{\rho}_U) - x) - \overline{J}(\rho_U)^\top(\hat{x}(\rho_U) - x)\big\|_2 &\le \|\overline{J}(\hat{\rho}_U) - \overline{J}(\rho_U)\|_{\mathrm{op}}\|\hat{x}(\hat{\rho}_U) - x\|_2 \\ &\quad + \|\overline{J}(\rho_U)\|_{\mathrm{op}}\|\hat{x}(\hat{\rho}_U) - \hat{x}(\rho_U)\|_2 \\ &\le E_*\, d_{\mathcal{F}_U}(\hat{\rho}_U, \rho_U) + S_1 R_*^2\, d_{\mathcal{F}_U}(\hat{\rho}_U, \rho_U),\end{aligned}$$

by (67) and (61). Since $\|J_w\phi\|_{\mathrm{op}} \le C_*$ and $S_1 R_*^2 \le C_*$, we get

$$\|G_W(w; \hat{\rho}_U, \hat{\rho}_W) - G_W(w; \rho_U, \hat{\rho}_W)\|_2 \le C_*(E_* + C_*)\, d_{\mathcal{F}_U}(\hat{\rho}_U, \rho_U).$$

Encoder field, varying ρ_W. The decoder measure ρ_U is fixed, but the bottleneck shifts from α to $\hat{\alpha}$, which affects both $\hat{x}$ and $\overline{J}$. Adding and subtracting the mixed term $\overline{J}(\alpha)^\top(\hat{x}(\hat{\alpha}) - x)$,

$$\begin{aligned}&\big\|\overline{J}(\hat{\alpha})^\top(\hat{x}(\hat{\alpha}) - x) - \overline{J}(\alpha)^\top(\hat{x}(\alpha) - x)\big\|_2 \\ &\le \|\overline{J}(\hat{\alpha}) - \overline{J}(\alpha)\|_{\mathrm{op}}\|\hat{x}(\hat{\alpha}) - x\|_2 + \|\overline{J}(\alpha)\|_{\mathrm{op}}\|\hat{x}(\hat{\alpha}) - \hat{x}(\alpha)\|_2 \\ &\le C_*\|\hat{\alpha} - \alpha\|_2\, E_* + S_1 R_*^2\, C_*\|\hat{\alpha} - \alpha\|_2 \le C_*(E_* + C_*)\, \|\hat{\alpha} - \alpha\|_2,\end{aligned}$$

by (68). Multiplying by $\|J_w\phi\|_{\mathrm{op}} \le C_*$ and using (66), we get

$$\|G_W(w; \rho_U, \hat{\rho}_W) - G_W(w; \rho_U, \rho_W)\|_2 \le C_*^2(E_* + C_*)\, d_{\mathcal{F}_W}(\hat{\rho}_W, \rho_W).$$

Combining the decoupled bounds, we can conclude that both unregularized vector fields are Lipschitz in the pair of measures with respect to the parametric distances, with constant $\max\{C_*(E_* + C_*),\, C_*^2(E_* + C_*)\} = C_*^2(E_* + C_*) = L_{\mathrm{me}}$, since $C_* \ge 1$. This proves (b). □

J.3 Coupled Trajectory Bounds

We define two coupled sets of trajectories, all sharing the exact same initializations $u_j^0 \sim \rho_{U,0}$ and $w_i^0 \sim \rho_{W,0}$, drawn i.i.d.

- **Continuous Ideal Trajectories** $(\tilde{u}_j^t, \tilde{w}_i^t)$: Governed by the nonlinear Vlasov–McKean characteristics of Proposition 1, driven by the true infinite-width population measures $\rho_{U,s}$ and $\rho_{W,s}$:

$$\begin{aligned}\tilde{u}_j^t &= u_j^0 + \int_0^t \xi(s) V_U(\tilde{u}_j^s; \rho_{U,s}, \rho_{W,s})\, ds, \\ \tilde{w}_i^t &= w_i^0 + \int_0^t \xi(s) V_W(\tilde{w}_i^s; \rho_{U,s}, \rho_{W,s})\, ds.\end{aligned}$$

  In the notation of Proposition 1, $\tilde{u}_j^t = X_U(t, u_j^0)$ and $\tilde{w}_i^t = X_W(t, w_i^0)$. Since the characteristic flow is a deterministic map applied to independent initial conditions, for every fixed t the ideal particles $\{\tilde{u}_j^t\}_{j \le N}$ are i.i.d. with law $(X_U(t, \cdot))_{\#}\rho_{U,0} = \rho_{U,t}$, and similarly $\{\tilde{w}_i^t\}_{i \le N}$ are i.i.d. with law $\rho_{W,t}$. By Theorem 7, all ideal particles remain in $B_D(0, R_*)$.

- **Discrete SGD Trajectories** (u_j^k, w_i^k): Driven by the regularized single-sample gradients as in (7), with $s_k = \epsilon\xi(k\epsilon)$:

$$\begin{aligned}u_j^{k+1} &= u_j^k + \epsilon\xi(k\epsilon)\Big(F_U(u_j^k; \hat{\rho}_{U,k}^N, \hat{\rho}_{W,k}^N, x_{k+1}) - \lambda\|u_j^k\|_2^4 u_j^k\Big), \\ w_i^{k+1} &= w_i^k + \epsilon\xi(k\epsilon)\Big(F_W(w_i^k; \hat{\rho}_{U,k}^N, \hat{\rho}_{W,k}^N, x_{k+1}) - \lambda\|w_i^k\|_2^4 w_i^k\Big).\end{aligned}$$

  By Theorem 6, under $\epsilon M_\xi \le s_*$ all SGD particles remain in $B_D(0, R_*)$ for every realization.

The total continuous time interval is $[0, T]$ and SGD updates the particles at the grid times $t_k = k\epsilon$, so the number of update steps within $[0, T]$ is $K_T = \lfloor T/\epsilon \rfloor$. We use the delayed time index $[s] := \epsilon\lfloor s/\epsilon \rfloor$, and we write $u_j^{[s]} := u_j^{\lfloor s/\epsilon \rfloor}$, $\hat{\rho}_{U,[s]}^N := \hat{\rho}_{U,\lfloor s/\epsilon \rfloor}^N$, and similarly for the encoder. Because the network weights are frozen between updates, integrating a constant field over an interval of width

ϵ multiplies it by ϵ, and the SGD recursion can be rewritten exactly as a continuous-time integral evaluated at grid times:

$$u_j^{[t]} = u_j^0 + \int_0^{[t]} \xi([s])\Big(F_U\big(u_j^{[s]}; \hat{\rho}^N_{U,[s]}, \hat{\rho}^N_{W,[s]}, x_{[s]/\epsilon+1}\big) - \lambda\|u_j^{[s]}\|_2^4 u_j^{[s]}\Big)ds, \tag{69}$$

and likewise for $w_i^{[t]}$. We compare the two trajectories at grid times only, which is all that Theorem 2 requires. Accordingly, we define

$$\Delta_U(t) := \max_{j\le N} \max_{0\le k\le \lfloor t/\epsilon\rfloor} \|\tilde{u}_j^{k\epsilon} - u_j^k\|_2^2,\ \Delta_W(t) := \max_{i\le N} \max_{0\le k\le \lfloor t/\epsilon\rfloor} \|\tilde{w}_i^{k\epsilon} - w_i^k\|_2^2,\ \Delta(t) := \Delta_U(t) + \Delta_W(t). \tag{70}$$

Both are nondecreasing step functions of t, bounded by $4R_*^2$, and $\Delta_U(t) = \max_j \sup_{s\le t} \|\tilde{u}_j^{[s]} - u_j^{[s]}\|_2^2$. Because the discrete trajectories are constant between updates, any concentration bound that must hold uniformly over $[0,T]$ needs to be enforced only on the $K_T + 1 \le 2X_T$ grid points, where $X_T := T/\epsilon \vee 1$.

We now define the constants entering the trajectory bound:

$$C_{\text{time}} := L_\xi M_* + M_\xi^2 M_* \big(L_{\text{sp}} + 2C_* L_{\text{me}}\big), \quad C_{\text{mart}} := 32 M_\xi^2 B_*^2, \tag{71}$$

$$C_{\text{exp}} := 4M_\xi^2\big(L_{\text{sp}}^2 + 8C_*^2 L_{\text{me}}^2\big), \tag{72}$$

$$C_{\text{poly}} := \max\Big\{\, 8\bar{\epsilon}\, C_{\text{time}}^2,\ \ 192\, M_\xi^2 L_{\text{me}}^2 \big(C_{\mathcal{F}}^2 + 8C_*^2\big)\, D,\ \ 8\, C_{\text{mart}} \Big\}, \quad \bar{\epsilon} := \frac{s_*}{M_\xi}. \tag{73}$$

All of these depend only on $M_x, S_0, S_1, S_2, \lambda, R_*, L_\xi, M_\xi$ and, polynomially, on the particle dimension D (through the factor D and through $C_{\mathcal{F}}$).

Lemma 14 (Coupled Trajectory Bound)**.** *Let Assumption 1 hold, let $\lambda > 0$ and $\epsilon M_\xi \le s_*$. With probability at least $1 - 4e^{-\delta^2}$, for all $T \ge 0$,*

$$\Delta(T) \le C_{\text{poly}}(T \vee T^2)\left(\epsilon + \frac{1}{N}\right)\Big[D + \log\big(NX_T\big) + \delta^2\Big]\exp\Big(C_{\text{exp}} T^2\Big), \quad X_T = T/\epsilon \vee 1,$$

where C_{poly} and C_{exp} are given by (73) *and* (72)*.*

Proof. Fix a decoder particle j and a time $t \in [0,T]$. The ideal particle at the grid time $[t]$ is

$$\tilde{u}_j^{[t]} = u_j^0 + \int_0^{[t]} \xi(s) V_U(\tilde{u}_j^s; \rho_{U,s}, \rho_{W,s})\, ds,$$

while the SGD particle is given by (69). Subtracting the two, the shared initial positions u_j^0 cancel. To break the integral difference into manageable parts, we add and subtract the following three intermediate evaluations of the regularized field:

1. $\xi([s]) V_U(\tilde{u}_j^{[s]}; \rho_{U,[s]}, \rho_{W,[s]})$ (time discretization),
2. $\xi([s]) V_U(u_j^{[s]}; \rho_{U,[s]}, \rho_{W,[s]})$ (spatial tracking),
3. $\xi([s]) V_U(u_j^{[s]}; \hat{\rho}^N_{U,[s]}, \hat{\rho}^N_{W,[s]})$ (measure gap).

Using $\|a+b+c+d\|_2^2 \le 4(\|a\|_2^2 + \|b\|_2^2 + \|c\|_2^2 + \|d\|_2^2)$, we split the error into four distinct integrals:

$$\|\tilde{u}_j^{[t]} - u_j^{[t]}\|_2^2 \le 4\left(E_{\text{time}}(t) + E_{\text{track}}(t) + E_{\text{meas}}(t) + E_{\text{mart}}(t)\right), \tag{74}$$

where

$$E_{\text{time}}(t) := \left\| \int_0^{[t]} \Big(\xi(s) V_U(\tilde{u}_j^s; \rho_{U,s}, \rho_{W,s}) - \xi([s]) V_U(\tilde{u}_j^{[s]}; \rho_{U,[s]}, \rho_{W,[s]}) \Big)\, ds \right\|_2^2,$$

$$E_{\text{track}}(t) := \left\| \int_0^{[t]} \xi([s]) \Big(V_U(\tilde{u}_j^{[s]}; \rho_{U,[s]}, \rho_{W,[s]}) - V_U(u_j^{[s]}; \rho_{U,[s]}, \rho_{W,[s]}) \Big)\, ds \right\|_2^2,$$

$$E_{\text{meas}}(t) := \left\| \int_0^{[t]} \xi([s]) \Big(V_U(u_j^{[s]}; \rho_{U,[s]}, \rho_{W,[s]}) - V_U(u_j^{[s]}; \hat{\rho}_{U,[s]}^N, \hat{\rho}_{W,[s]}^N) \Big)\, ds \right\|_2^2,$$

$$E_{\text{mart}}(t) := \left\| \int_0^{[t]} \xi([s]) \Big(V_U(u_j^{[s]}; \hat{\rho}_{U,[s]}^N, \hat{\rho}_{W,[s]}^N) - \big(F_U(u_j^{[s]}; \hat{\rho}_{U,[s]}^N, \hat{\rho}_{W,[s]}^N, x_{[s]/\epsilon+1}) - \lambda \|u_j^{[s]}\|_2^4 u_j^{[s]} \big) \Big) ds \right\|_2^2.$$

All particles and measures appearing above stay in $B_D(0, R_*)$ and $\mathcal{P}_{R_*}$ respectively, so every estimate of Section J.1 and Lemma 13 applies.

1. Time discretization term (E_{time}). We must control the difference between the field evaluated at the continuous time s and at the floored time $[s]$. By the Cauchy–Schwarz inequality $\| \int_0^{[t]} f \|_2^2 \le [t] \int_0^{[t]} \|f\|_2^2 \le t \int_0^t \|f\|_2^2$,

$$E_{\text{time}}(t) \le t \int_0^t \|\xi(s)V(s) - \xi([s])V([s])\|_2^2 \, ds, \quad V(s) := V_U(\tilde{u}_j^s; \rho_{U,s}, \rho_{W,s}), \quad V([s]) := V_U(\tilde{u}_j^{[s]}; \rho_{U,[s]}, \rho_{W,[s]}).$$

Adding and subtracting $\xi([s])V(s)$ and applying the triangle inequality,

$$\|\xi(s)V(s) - \xi([s])V([s])\|_2 \le |\xi(s) - \xi([s])| \|V(s)\|_2 + |\xi([s])| \|V(s) - V([s])\|_2.$$

By (62), $\|V(s)\|_2 \le M_*$; since ξ is L_ξ-Lipschitz and $|s - [s]| \le \epsilon$, the first term is bounded by $L_\xi M_* \epsilon$. For the second term, $|\xi([s])| \le M_\xi$, and we add and subtract $V_U(\tilde{u}_j^{[s]}; \rho_{U,s}, \rho_{W,s})$:

$$\begin{aligned} \|V(s) - V([s])\|_2 \le & \|V_U(\tilde{u}_j^s; \rho_{U,s}, \rho_{W,s}) - V_U(\tilde{u}_j^{[s]}; \rho_{U,s}, \rho_{W,s})\|_2 \\ & + \|V_U(\tilde{u}_j^{[s]}; \rho_{U,s}, \rho_{W,s}) - V_U(\tilde{u}_j^{[s]}; \rho_{U,[s]}, \rho_{W,[s]})\|_2. \end{aligned} \tag{75}$$

For the first term in (75), Lemma 13(a) gives the bound $L_{\text{sp}} \|\tilde{u}_j^s - \tilde{u}_j^{[s]}\|_2$. Since the ideal particles move with speed at most $M_\xi M_*$ by (62), the displacement over a time interval of length at most ϵ satisfies

$$\|\tilde{u}_j^s - \tilde{u}_j^{[s]}\|_2 \le M_\xi M_* \epsilon, \quad \|\tilde{w}_i^s - \tilde{w}_i^{[s]}\|_2 \le M_\xi M_* \epsilon. \tag{76}$$

For the second term in (75), Lemma 13(b) gives the bound $L_{\text{me}}\big(d_{\mathcal{F}_U}(\rho_{U,s}, \rho_{U,[s]}) + d_{\mathcal{F}_W}(\rho_{W,s}, \rho_{W,[s]})\big)$. To bound these parametric distances, recall from Proposition 1 that $\rho_{U,s}$ is the push-forward of $\rho_{U,[s]}$ under the characteristic flow map $T_U := X_U(s, \cdot) \circ X_U([s], \cdot)^{-1}$ from time $[s]$ to time s (the flow is invertible since the fields are Lipschitz on the invariant ball). By the same speed bound as in (76), $\sup_{u \in \text{supp}(\rho_{U,[s]})} \|T_U(u) - u\|_2 \le M_\xi M_* \epsilon$, and Lemma 2 together with (64) yields

$$d_{\mathcal{F}_U}(\rho_{U,s}, \rho_{U,[s]}) \le C_* \, W_1(\rho_{U,s}, \rho_{U,[s]}) \le C_* M_\xi M_* \epsilon, \quad d_{\mathcal{F}_W}(\rho_{W,s}, \rho_{W,[s]}) \le C_* M_\xi M_* \epsilon. \tag{77}$$

Substituting these into (75), we obtain $\|V(s) - V([s])\|_2 \le M_\xi M_* \big(L_{\text{sp}} + 2C_* L_{\text{me}}\big)\epsilon$, and therefore, with C_{time} defined in (71),

$$\|\xi(s)V(s) - \xi([s])V([s])\|_2 \le L_\xi M_* \epsilon + M_\xi^2 M_* \big(L_{\text{sp}} + 2C_* L_{\text{me}}\big)\epsilon = C_{\text{time}}\, \epsilon.$$

Substituting this uniform bound into the integral gives

$$E_{\text{time}}(t) \le t \int_0^t (C_{\text{time}} \epsilon)^2 \, ds = t^2 C_{\text{time}}^2 \epsilon^2. \tag{78}$$

2. Spatial tracking term (E_{track}). This term isolates the error between the ideal trajectory $\tilde{u}_j$ and the SGD trajectory u_j evaluated at the same grid time $[s]$ and with the same population measures. Applying Cauchy–Schwarz, $|\xi([s])| \le M_\xi$, and Lemma 13(a),

$$\begin{aligned} E_{\mathrm{track}}(t) &\le t \int_0^t \|\xi([s])\big(V_U(\tilde{u}_j^{[s]}; \rho_{U,[s]}, \rho_{W,[s]}) - V_U(u_j^{[s]}; \rho_{U,[s]}, \rho_{W,[s]})\big)\|_2^2 \, ds \\ &\le t M_\xi^2 L_{\mathrm{sp}}^2 \int_0^t \|\tilde{u}_j^{[s]} - u_j^{[s]}\|_2^2 \, ds \le t M_\xi^2 L_{\mathrm{sp}}^2 \int_0^t \Delta_U(s) \, ds, \end{aligned} \tag{79}$$

where in the last step we used $\|\tilde{u}_j^{[s]} - u_j^{[s]}\|_2^2 \le \Delta_U(s)$, see (70).

3. Measure gap term (E_{meas}). This term captures the macroscopic error introduced because the discrete network evaluates its gradients on the empirical measures $\hat{\rho}_U^N, \hat{\rho}_W^N$, whereas the ideal flow evaluates on the population measures ρ_U, ρ_W. Applying Cauchy–Schwarz and Lemma 13(b),

$$E_{\mathrm{meas}}(t) \le t M_\xi^2 L_{\mathrm{me}}^2 \int_0^t \Big(d_{\mathcal{F}_U}(\rho_{U,[s]}, \hat{\rho}_{U,[s]}^N) + d_{\mathcal{F}_W}(\rho_{W,[s]}, \hat{\rho}_{W,[s]}^N) \Big)^2 \, ds.$$

To bound these parametric distances, we introduce the *bridge measures*

$$\bar{\rho}_{U,[s]}^N := \frac{1}{N} \sum_{m=1}^N \delta_{\tilde{u}_m^{[s]}}, \quad \bar{\rho}_{W,[s]}^N := \frac{1}{N} \sum_{m=1}^N \delta_{\tilde{w}_m^{[s]}},$$

i.e. the empirical measures of the ideal particles at time $[s]$. By the triangle inequality, we decompose each distance into a statistical concentration term and a particle shift term:

$$d_{\mathcal{F}_U}(\rho_{U,[s]}, \hat{\rho}_{U,[s]}^N) \le \underbrace{d_{\mathcal{F}_U}(\rho_{U,[s]}, \bar{\rho}_{U,[s]}^N)}_{\text{Concentration } Q_{U,[s]}} + \underbrace{d_{\mathcal{F}_U}(\bar{\rho}_{U,[s]}^N, \hat{\rho}_{U,[s]}^N)}_{\text{Particle Shift } P_{U,[s]}}, \quad d_{\mathcal{F}_W}(\rho_{W,[s]}, \hat{\rho}_{W,[s]}^N) \le Q_{W,[s]} + P_{W,[s]},$$

with $Q_{W,[s]}, P_{W,[s]}$ defined analogously.

Particle shift. The bridge and the SGD empirical measures are coupled through the index m, so the diagonal coupling $\frac{1}{N}\sum_m \delta_{(\tilde{u}_m^{[s]}, u_m^{[s]})}$ is admissible in the definition of W_1. Together with (64) and Jensen's inequality $(\frac{1}{N}\sum_m a_m)^2 \le \frac{1}{N}\sum_m a_m^2$,

$$P_{U,[s]}^2 \le C_*^2 \, W_1(\bar{\rho}_{U,[s]}^N, \hat{\rho}_{U,[s]}^N)^2 \le C_*^2 \Big(\frac{1}{N} \sum_{m=1}^N \|\tilde{u}_m^{[s]} - u_m^{[s]}\|_2 \Big)^2 \le C_*^2 \, \Delta_U(s), \quad P_{W,[s]}^2 \le C_*^2 \, \Delta_W(s). \tag{80}$$

Concentration. As observed above, for every fixed grid time $[s]$ the ideal particles $\{\tilde{u}_m^{[s]}\}_{m \le N}$ are i.i.d. samples from $\rho_{U,[s]} \in \mathcal{P}_{R_*}$, so $\bar{\rho}_{U,[s]}^N$ is a genuine empirical measure of i.i.d. samples, and likewise for the encoder. We may therefore apply (65) to each of the three classes $\mathcal{F}_{U,1}, \mathcal{F}_{U,2}, \mathcal{F}_W$ at each of the $K_T + 1 \le 2X_T$ grid times. Allocating the failure probability $\delta' := e^{-\delta_1^2}/(6X_T)$ to each of the at most $3(K_T+1) \le 6X_T$ events, the union bound guarantees that with probability at least $1 - e^{-\delta_1^2}$ all of them hold simultaneously. Since $\log(1/\delta') = \log(6X_T) + \delta_1^2$, squaring (65) and using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we conclude that with probability at least $1 - e^{-\delta_1^2}$, simultaneously for all grid times $[s] \in [0, T]$,

$$Q_{U,[s]}^2 \le \frac{3}{N} \Big(C_{\mathcal{F}}^2 + C_*^2 D \log N + 2C_*^2 \big(\log(6X_T) + \delta_1^2 \big) \Big) =: \frac{3}{N} \, \Xi_T(\delta_1), \quad Q_{W,[s]}^2 \le \frac{3}{N} \, \Xi_T(\delta_1), \tag{81}$$

where we used that $d_{\mathcal{F}_U}$ is the maximum of $d_{\mathcal{F}_{U,1}}$ and $d_{\mathcal{F}_{U,2}}$.

Substituting the decompositions into the integral and applying $(a+b+c+d)^2 \le 4(a^2+b^2+c^2+d^2)$ together with (80), we obtain, on the event of (81),

$$E_{\mathrm{meas}}(t) \le 4t M_\xi^2 L_{\mathrm{me}}^2 \int_0^t \Big(Q_{U,[s]}^2 + Q_{W,[s]}^2 + C_*^2 \Delta(s) \Big) ds \le 4t M_\xi^2 L_{\mathrm{me}}^2 \Big(\frac{6t}{N} \Xi_T(\delta_1) + C_*^2 \int_0^t \Delta(s) \, ds \Big), \tag{82}$$

simultaneously for all $j \le N$ and $t \le T$.

4. Martingale noise term (E_{mart}). This term captures the accumulated error from using the noisy single-sample gradient F_U instead of the population field G_U evaluated on the empirical measures. Notice that the regularizer $-\lambda\|u_j^{[s]}\|_2^4 u_j^{[s]}$ is deterministic given the current configuration and is independent of the data sample $x_{[s]/\epsilon+1}$; it therefore cancels exactly between the regularized population field and the regularized stochastic gradient. Replacing the integral over $[0,[t]]$ by the discrete sum,

$$\begin{aligned} E_{\mathrm{mart}}(t) &= \Big\| \epsilon \sum_{k=1}^{\lfloor t/\epsilon \rfloor} \xi\big((k-1)\epsilon\big)\Big(G_U(u_j^{k-1};\hat\rho^N_{U,k-1},\hat\rho^N_{W,k-1}) - F_U\big(u_j^{k-1};\hat\rho^N_{U,k-1},\hat\rho^N_{W,k-1},x_k\big)\Big)\Big\|_2^2 \\ &= \big\|\epsilon X^{(j)}_{\lfloor t/\epsilon\rfloor}\big\|_2^2, \end{aligned}$$

where we define, with respect to the filtration $\mathcal{F}_k := \sigma\big(\{u_j^0\}_j, \{w_i^0\}_i, x_1,\dots,x_k\big)$, the $\mathbb{R}^D$-valued increments and partial sums

$$\Delta X_k^{(j)} := \xi\big((k-1)\epsilon\big)\Big(G_U(u_j^{k-1};\hat\rho^N_{U,k-1},\hat\rho^N_{W,k-1}) - F_U\big(u_j^{k-1};\hat\rho^N_{U,k-1},\hat\rho^N_{W,k-1},x_k\big)\Big),$$

$$X_n^{(j)} := \sum_{k=1}^{n} \Delta X_k^{(j)}, \quad X_0^{(j)} := 0.$$

We verify the two conditions of Lemma 4.

1. *Martingale property.* The configuration (U^{k-1}, W^{k-1}) is $\mathcal{F}_{k-1}$-measurable, and x_k is independent of $\mathcal{F}_{k-1}$. By (60), $\mathbb{E}[F_U(u_j^{k-1};\cdot,x_k)\,|\,\mathcal{F}_{k-1}] = G_U(u_j^{k-1};\hat\rho^N_{U,k-1},\hat\rho^N_{W,k-1})$, so $\mathbb{E}[\Delta X_k^{(j)}\,|\,\mathcal{F}_{k-1}] = 0$ and $(X_n^{(j)})_n$ is a martingale.
2. *Sub-Gaussian increments.* Because all particles are confined to $B_D(0,R_*)$, so $\|F_U\|_2 \le B_U(R_*) \le B_*$ and $\|G_U\|_2 \le B_*$, hence $\|\Delta X_k^{(j)}\|_2 \le M_\xi \cdot 2B_* =: c$ almost surely. Crucially, the confinement radius R_* enters this bound only through B_*, and the sixth-order regularizer, which is what forces the large radius, does not appear in the noise at all. For any fixed $\gamma \in \mathbb{R}^D$, the scalar variable $Y = \langle \gamma, \Delta X_k^{(j)}\rangle$ has conditional mean zero and satisfies $|Y| \le c\|\gamma\|_2$ by Cauchy–Schwarz. Hoeffding's lemma for bounded zero-mean variables gives

$$\mathbb{E}\big[e^Y \,\big|\, \mathcal{F}_{k-1}\big] \le \exp\Big(\frac{(2c\|\gamma\|_2)^2}{8}\Big) = \exp\Big(\frac{c^2\|\gamma\|_2^2}{2}\Big),$$

which is exactly the sub-Gaussian condition of Lemma 4 with $L = c = 2M_\xi B_*$.

Applying Lemma 4 to particle j with $n = K_T \le T/\epsilon$ steps, for every $y > 0$,

$$\mathbb{P}\Big(\max_{k\le K_T} \|X_k^{(j)}\|_2 \ge 2c\sqrt{T/\epsilon}\,(\sqrt{D}+y)\Big) \le e^{-y^2}.$$

We require this to hold uniformly over all N decoder particles. By a union bound the failure probability is at most Ne^{-y^2}; choosing $y = \sqrt{\log N + \delta_2^2}$ makes it at most $e^{-\delta_2^2}$. Since $E_{\mathrm{mart}}(t) = \|\epsilon X^{(j)}_{\lfloor t/\epsilon\rfloor}\|_2^2$, multiplying by ϵ and squaring, and using $(a+b)^2 \le 2(a^2+b^2)$, we conclude that with probability at least $1-e^{-\delta_2^2}$, simultaneously for all $j \le N$ and $t \le T$,

$$E_{\mathrm{mart}}(t) \le \epsilon^2\Big[2c\sqrt{T/\epsilon}\big(\sqrt{D}+\sqrt{\log N+\delta_2^2}\big)\Big]^2 \le 8c^2\,T\epsilon\,\big(D+\log N+\delta_2^2\big) = C_{\mathrm{mart}}\,T\epsilon\,\big(D+\log N+\delta_2^2\big), \tag{83}$$

with $C_{\mathrm{mart}} = 8c^2 = 32M_\xi^2 B_*^2$ as in (71).

Aggregation for the decoder. We set $\delta_1 = \delta_2 = \delta$. By a union bound over the two probabilistic events (81) and (83), with probability at least $1-2e^{-\delta^2}$ the bounds (78), (79), (82) and (83) hold simultaneously for all $j \le N$ and all $t \le T$. Substituting them into (74), using $t \le T$ in the

prefactors, and taking the maximum over j and over the grid times up to t (the right-hand side is nondecreasing in t), we obtain

$$\Delta_U(t) \le 4\Big(T^2 C_{\mathrm{time}}^2 \epsilon^2 + T M_\xi^2 L_{\mathrm{sp}}^2 \int_0^t \Delta_U(s)\,ds + 4T M_\xi^2 L_{\mathrm{me}}^2 \Big(\frac{6T}{N}\Xi_T(\delta) + C_*^2 \int_0^t \Delta(s)\,ds\Big) + C_{\mathrm{mart}} T\epsilon\big(D + \log N + \delta^2\big)\Big). \tag{84}$$

Aggregation for the encoder. The identical decomposition applies to the encoder particles, with V_W, G_W, F_W in place of V_U, G_U, F_U. All the constants used above were defined symmetrically: the speed bound $M_\xi M_*$ and (77) are the same, Lemma 13 gives the same L_{sp} and L_{me}, the measure gap is controlled by the very same concentration terms $Q_{U,[s]}, Q_{W,[s]}$ and particle shifts (80), and the martingale increments are bounded by $2M_\xi B_W(R_*) \le 2M_\xi B_* = c$. Hence, on an event of probability at least $1 - 2e^{-\delta^2}$ (a fresh union bound for the encoder concentration and the encoder martingale),

$$\Delta_W(t) \le 4\Big(T^2 C_{\mathrm{time}}^2 \epsilon^2 + T M_\xi^2 L_{\mathrm{sp}}^2 \int_0^t \Delta_W(s)\,ds + 4T M_\xi^2 L_{\mathrm{me}}^2 \Big(\frac{6T}{N}\Xi_T(\delta) + C_*^2 \int_0^t \Delta(s)\,ds\Big) + C_{\mathrm{mart}} T\epsilon\big(D + \log N + \delta^2\big)\Big). \tag{85}$$

Gronwall. Adding (84) and (85), and using $\int_0^t \Delta_U + \int_0^t \Delta_W = \int_0^t \Delta$, we obtain, with probability at least $1 - 4e^{-\delta^2}$, for all $t \le T$,

$$\Delta(t) \le A_T + K_T' \int_0^t \Delta(s)\,ds, \tag{86}$$

where

$$A_T := 8T^2 C_{\mathrm{time}}^2 \epsilon^2 + \frac{192\,T^2 M_\xi^2 L_{\mathrm{me}}^2}{N}\,\Xi_T(\delta) + 8C_{\mathrm{mart}} T\epsilon\big(D + \log N + \delta^2\big),$$
$$K_T' := 4T M_\xi^2 \big(L_{\mathrm{sp}}^2 + 8C_*^2 L_{\mathrm{me}}^2\big) = C_{\exp}\,T.$$

Since Δ is bounded and nondecreasing, Gronwall's inequality applied to (86) yields $\Delta(t) \le A_T \exp(K_T' t)$ for all $t \le T$; evaluating at $t = T$ gives

$$\Delta(T) \le A_T \exp\big(C_{\exp} T^2\big).$$

Simplification of the constants. It remains to bound A_T in the form claimed. Recall $X_T = T/\epsilon \vee 1$ and $N \ge 1$, so $\log N \le \log(NX_T)$ and $\log(6X_T) \le \log 6 + \log(NX_T)$. Using $2\log 6 \le 4$ and $D \ge 2$, the bracket in (81) satisfies

$$\Xi_T(\delta) = C_{\mathcal{F}}^2 + C_*^2 D \log N + 2C_*^2 \log(6X_T) + 2C_*^2\delta^2 \le \big(C_{\mathcal{F}}^2 + 8C_*^2\big)\,D\,\Big[D + \log(NX_T) + \delta^2\Big].$$

Here the factor D is unavoidable: it comes from the covering number of the b- and d-dimensional index sets in (65), and it is the only place where the dimension enters other than through the additive D inside the bracket. For the time-discretization term we use the step condition $\epsilon M_\xi \le s_*$, i.e. $\epsilon \le \bar\epsilon$, to write $\epsilon^2 \le \bar\epsilon\,\epsilon$. Finally, the bracket $[D + \log(NX_T) + \delta^2]$ is at least 1, and $T^2 \le T \vee T^2$, $T \le T \vee T^2$. Therefore

$$\begin{aligned} A_T &\le 8\bar\epsilon\, C_{\mathrm{time}}^2\, T^2\epsilon\,\big[D + \log(NX_T) + \delta^2\big] + 192 M_\xi^2 L_{\mathrm{me}}^2 \big(C_{\mathcal{F}}^2 + 8C_*^2\big) D\,\frac{T^2}{N}\big[D + \log(NX_T) + \delta^2\big] \\ &\quad + 8C_{\mathrm{mart}}\,T\epsilon\,\big[D + \log(NX_T) + \delta^2\big] \\ &\le C_{\mathrm{poly}}\,(T \vee T^2)\Big(\epsilon + \frac{1}{N}\Big)\Big[D + \log(NX_T) + \delta^2\Big], \end{aligned}$$

with C_{poly} as in (73). This completes the proof. □

J.4 Risk Tracking

We now bound the deviation between the finite-width regularized risk $\mathcal{R}_\lambda^N(U^k, W^k)$ along SGD and the mean-field regularized risk $\mathcal{R}_\lambda(\rho_{U,k\epsilon}, \rho_{W,k\epsilon})$. Let $\tilde{U}^{k\epsilon} = \{\tilde{u}_j^{k\epsilon}\}_{j=1}^N$ and $\tilde{W}^{k\epsilon} = \{\tilde{w}_i^{k\epsilon}\}_{i=1}^N$ denote the ideal continuous particle configurations at the grid time $k\epsilon$. Let

$$K_R := E_* C_* (1 + C_*) + \lambda R_*^5. \tag{87}$$

Lemma 15 (Finite-N Risk Lipschitzness)**.** *Let Assumption 1 hold. For any two configurations (U, W) and (U', W') of N decoder and N encoder particles in $B_D(0, R_*)$,*

$$\big|\mathcal{R}_\lambda^N(U, W) - \mathcal{R}_\lambda^N(U', W')\big| \le K_R \Big(\max_j \|u_j - u'_j\|_2 + \max_i \|w_i - w'_i\|_2 \Big).$$

In particular,

$$\max_{k \in [0, T/\epsilon] \cap \mathbb{N}} \big|\mathcal{R}_\lambda^N(\tilde{U}^{k\epsilon}, \tilde{W}^{k\epsilon}) - \mathcal{R}_\lambda^N(U^k, W^k)\big| \le K_R \max_{k \in [0, T/\epsilon] \cap \mathbb{N}} \Big(\max_j \|\tilde{u}_j^{k\epsilon} - u_j^k\|_2 + \max_i \|\tilde{w}_i^{k\epsilon} - w_i^k\|_2 \Big).$$

Proof. Fix $x \in \mathrm{supp}(P)$ and let $\alpha_N = \frac{1}{N}\sum_i \phi(x; w_i)$ and $\alpha'_N = \frac{1}{N}\sum_i \phi(x; w'_i)$ be the two empirical bottlenecks. By the C_*-Lipschitz continuity of ϕ in w (Lemma 9),

$$\|\alpha_N - \alpha'_N\|_2 \le \frac{1}{N}\sum_{i=1}^N \|\phi(x; w_i) - \phi(x; w'_i)\|_2 \le C_* \max_i \|w_i - w'_i\|_2.$$

Next, with $\hat{x}_N = \frac{1}{N}\sum_j \psi(\alpha_N; u_j)$ and $\hat{x}'_N = \frac{1}{N}\sum_j \psi(\alpha'_N; u'_j)$, the C_*-Lipschitz continuity of ψ in both α and u gives

$$\begin{aligned}\|\hat{x}_N - \hat{x}'_N\|_2 &\le \frac{1}{N}\sum_{j=1}^N \|\psi(\alpha_N; u_j) - \psi(\alpha'_N; u'_j)\|_2 \le C_* \|\alpha_N - \alpha'_N\|_2 + C_* \max_j \|u_j - u'_j\|_2 \\ &\le C_*(1 + C_*)\Big(\max_i \|w_i - w'_i\|_2 + \max_j \|u_j - u'_j\|_2 \Big). \end{aligned} \tag{88}$$

For the reconstruction part of the risk, expanding the difference of squares,

$$\Big| \tfrac{1}{2}\|\hat{x}_N - x\|_2^2 - \tfrac{1}{2}\|\hat{x}'_N - x\|_2^2 \Big| = \tfrac{1}{2}\big|\langle (\hat{x}_N - x) + (\hat{x}'_N - x),\, \hat{x}_N - \hat{x}'_N \rangle\big| \le E_* \|\hat{x}_N - \hat{x}'_N\|_2,$$

because both residuals are bounded by E_* by (61). Taking the expectation over x and using (88), the reconstruction risks differ by at most $E_* C_*(1 + C_*)(\max_i \|w_i - w'_i\|_2 + \max_j \|u_j - u'_j\|_2)$. For the regularization part, the map $v \mapsto \frac{\lambda}{6}\|v\|_2^6$ has gradient $\lambda \|v\|_2^4 v$, whose norm is at most λR_*^5 on $B_D(0, R_*)$; by the mean value theorem along the segment (which stays in the ball by convexity),

$$\Big| \frac{\lambda}{6N}\sum_{j=1}^N \|u_j\|_2^6 - \frac{\lambda}{6N}\sum_{j=1}^N \|u'_j\|_2^6 \Big| \le \lambda R_*^5 \max_j \|u_j - u'_j\|_2,$$

and likewise for the encoder. Adding the three contributions proves the first claim with K_R as in (87). The second claim follows by applying the first one to $(U, W) = (\tilde{U}^{k\epsilon}, \tilde{W}^{k\epsilon})$ and $(U', W') = (U^k, W^k)$ for each k; both configurations lie in $B_D(0, R_*)$ by Theorems 6 and 7. □

For the concentration of the risk of the ideal particles we need the following constants:

$$\begin{aligned} H_f^* &:= S_1^2 R_*^4 + E_* S_2 R_*^3, \quad K_{\mathrm{bias}} := \tfrac{1}{2} S_0^2 R_*^2 (1 + H_f^*), \\ c_{\mathrm{MD}} &:= 2E_* C_* S_0 R_* + \tfrac{\lambda}{6} R_*^6, \quad K_C := \max\{c_{\mathrm{MD}}, K_{\mathrm{bias}}\}. \end{aligned} \tag{89}$$

Lemma 16 (Risk Concentration)**.** *Let Assumption 1 hold. With probability at least $1 - e^{-\delta^2}$,*

$$\max_{k \in [0, T/\epsilon] \cap \mathbb{N}} \big|\mathcal{R}_\lambda^N(\tilde{U}^{k\epsilon}, \tilde{W}^{k\epsilon}) - \mathcal{R}_\lambda(\rho_{U,k\epsilon}, \rho_{W,k\epsilon})\big| \le \frac{K_C}{\sqrt{N}} \Big[\sqrt{\log(4X_T)} + 1 + \delta \Big], \quad X_T = T/\epsilon \vee 1.$$

Proof. Fix a grid time $k\epsilon$ and abbreviate $\rho_U = \rho_{U,k\epsilon}$, $\rho_W = \rho_{W,k\epsilon}$, $\mathcal{R}^N = \mathcal{R}^N_\lambda(\tilde{U}^{k\epsilon}, \tilde{W}^{k\epsilon})$. Recall that $\tilde{U}^{k\epsilon}$ and $\tilde{W}^{k\epsilon}$ consist of $2N$ independent random variables, the decoder particles being i.i.d. from ρ_U and the encoder particles i.i.d. from ρ_W, all supported in $B_D(0, R_*)$. We decompose the error into a deterministic bias term and a stochastic fluctuation term:

$$\big|\mathcal{R}^N - \mathcal{R}_\lambda(\rho_U, \rho_W)\big| \le \underbrace{\big|\mathcal{R}^N - \mathbb{E}[\mathcal{R}^N]\big|}_{\text{Fluctuation}} + \underbrace{\big|\mathbb{E}[\mathcal{R}^N] - \mathcal{R}_\lambda(\rho_U, \rho_W)\big|}_{\text{Bias}}.$$

Bias. The expected empirical regularizers agree exactly with their population counterparts, $\mathbb{E}[\frac{\lambda}{6N}\sum_j \|\tilde{u}_j\|_2^6] = \frac{\lambda}{6}\int \|u\|_2^6 \rho_U(du)$ and similarly for the encoder, so only the reconstruction term contributes to the bias. We follow the argument of the proof of Theorem 3, now on the ball. Fix x and write, as there, $m(\beta) := \int \psi(\beta; u)\rho_U(du)$, $g(\beta) := \int \|\psi(\beta; u)\|_2^2 \rho_U(du)$, $f_x(\beta) := \frac{1}{2}\|m(\beta) - x\|_2^2$, and let $\alpha(x) = \alpha(x; \rho_W)$ and $\alpha_N(x) = \frac{1}{N}\sum_i \phi(x; \tilde{w}_i)$. Conditionally on the encoder particles, the decoder outputs $\psi(\alpha_N(x); \tilde{u}_j)$ are i.i.d. with mean $m(\alpha_N(x))$, and the exact variance identity gives

$$\mathbb{E}_U\Big[\tfrac{1}{2}\|\hat{x}_N(x) - x\|_2^2 \,\Big|\, \tilde{W}\Big] = f_x(\alpha_N(x)) + \frac{1}{2N}\Big[g(\alpha_N(x)) - \|m(\alpha_N(x))\|_2^2\Big],$$

$$0 \le g(\alpha_N(x)) - \|m(\alpha_N(x))\|_2^2 \le S_0^2 R_*^2,$$

where we used $\|\psi(\beta; u)\|_2 \le S_0\|u\|_2 \le S_0 R_*$. Next, $\alpha_N(x)$ is an unbiased estimator of $\alpha(x)$ with $\mathbb{E}_W\|\alpha_N(x) - \alpha(x)\|_2^2 \le \frac{1}{N}\int \|\phi(x; w)\|_2^2 \rho_W(dw) \le \frac{S_0^2 R_*^2}{N}$. Taylor's formula with integral remainder for the C^2 function f_x around $\alpha(x)$ shows that the first-order term vanishes in expectation, so

$$\big|\mathbb{E}_W[f_x(\alpha_N(x))] - f_x(\alpha(x))\big| \le \tfrac{1}{2}\sup_{\beta \in \mathcal{A}} \|D^2 f_x(\beta)\|_{\mathrm{op}}\, \mathbb{E}_W\|\alpha_N(x) - \alpha(x)\|_2^2 \le \frac{H_f^* S_0^2 R_*^2}{2N},$$

where the Hessian bound $H_f^* = S_1^2 R_*^4 + E_* S_2 R_*^3$ follows from the computation in the proof of Lemma 12, with the moment bounds $K_\lambda^{p/6}$ replaced by R_*^p (which is legitimate since ρ_U is supported in $B_D(0, R_*)$): indeed $\|Dm\|_{\mathrm{op}} \le S_1 R_*^2$, $\|D^2 m\|_{\mathrm{op}} \le S_2 R_*^3$ and $\|m(\beta) - x\|_2 \le E_*$. Combining the two displays and integrating over $x \sim P$, the bias is two-sided bounded by

$$\big|\mathbb{E}[\mathcal{R}^N] - \mathcal{R}_\lambda(\rho_U, \rho_W)\big| \le \frac{S_0^2 R_*^2 (1 + H_f^*)}{2N} = \frac{K_{\mathrm{bias}}}{N}.$$

Fluctuation. We apply McDiarmid's inequality to $\mathcal{R}^N$ as a function of the $2N$ independent particles. Modifying a single decoder particle $\tilde{u}_j$ changes $\hat{x}_N(x)$ by at most $\frac{2S_0 R_*}{N}$, since each decoder output is bounded by $S_0 R_*$. Modifying a single encoder particle $\tilde{w}_i$ changes $\alpha_N(x)$ by at most $\frac{2S_0 R_*}{N}$, which by the C_*-Lipschitz continuity of ψ in α changes $\hat{x}_N(x)$ by at most $\frac{2C_* S_0 R_*}{N}$. As in the proof of Lemma 15, a change $\|\Delta\hat{x}_N\|_2$ in the prediction changes the reconstruction risk by at most $E_*\|\Delta\hat{x}_N\|_2$, and a change of one particle changes the corresponding regularizer by at most $\frac{\lambda}{6N} R_*^6$. Using $C_* \ge 1$, the bounded-difference constant for every one of the $2N$ coordinates is therefore

$$c_i \le \frac{1}{N}\Big(2E_* C_* S_0 R_* + \frac{\lambda}{6} R_*^6\Big) = \frac{c_{\mathrm{MD}}}{N}, \quad \sum_{i=1}^{2N} c_i^2 \le \frac{2c_{\mathrm{MD}}^2}{N}.$$

McDiarmid's inequality gives, for every $\tau > 0$,

$$\mathbb{P}\big(|\mathcal{R}^N - \mathbb{E}[\mathcal{R}^N]| \ge \tau\big) \le 2\exp\Big(-\frac{2\tau^2}{2c_{\mathrm{MD}}^2/N}\Big) = 2\exp\Big(-\frac{N\tau^2}{c_{\mathrm{MD}}^2}\Big).$$

We require this to hold uniformly over the $K_T + 1 \le 2X_T$ grid times. After a union bound, setting $4X_T \exp(-N\tau^2/c_{\mathrm{MD}}^2) = e^{-\delta^2}$ and solving for τ,

$$\tau = \frac{c_{\mathrm{MD}}}{\sqrt{N}}\sqrt{\log(4X_T) + \delta^2} \le \frac{c_{\mathrm{MD}}}{\sqrt{N}}\Big(\sqrt{\log(4X_T)} + \delta\Big),$$

by subadditivity of the square root. Adding the fluctuation and the bias, and using $1/N \le 1/\sqrt{N}$, we obtain the claim with $K_C = \max\{c_{\mathrm{MD}}, K_{\mathrm{bias}}\}$. □

J.5 Proof of Theorem 2

We are now ready to prove the main theorem. Throughout, $X_T = T/\epsilon \vee 1$, and we set

$$C_1 := \max\Big\{\sqrt{2C_{\mathrm{poly}}} + \sqrt{2},\ \ 2K_R\sqrt{C_{\mathrm{poly}}} + 2K_C\Big\}, \quad C_2 := \frac{C_{\exp}}{2}, \tag{90}$$

with $C_{\mathrm{poly}}, C_{\exp}, K_R, K_C$ from (73), (72), (87), (89).

Step 1: the square root of the trajectory bound. By Lemma 14, with probability at least $1-4e^{-\delta^2}$,

$$\Delta(T) \le C_{\mathrm{poly}}(T \vee T^2)\Big(\epsilon + \frac{1}{N}\Big)\Big[D + \log(NX_T) + \delta^2\Big]e^{C_{\exp}T^2}.$$

Using $\epsilon + 1/N \le 2(\epsilon \vee N^{-1})$, $\sqrt{T \vee T^2} = \sqrt{T} \vee T$, and the subadditivity $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we obtain on the same event

$$\begin{aligned}\sqrt{\Delta(T)} &\le \sqrt{2C_{\mathrm{poly}}}\,(\sqrt{T} \vee T)\,e^{C_2T^2}\,\sqrt{\epsilon \vee N^{-1}}\Big(\sqrt{D + \log(NX_T)} + \delta\Big) \\ &= \sqrt{2C_{\mathrm{poly}}}\,(\sqrt{T} \vee T)\,e^{C_2T^2}\,\mathrm{err}_{N,\epsilon}(\delta),\end{aligned} \tag{91}$$

where $\mathrm{err}_{N,\epsilon}(\delta) = \sqrt{N^{-1} \vee \epsilon}\,(\sqrt{D + \log(NX_T)} + \delta)$ is the error term of Theorem 2. Both $\sqrt{\Delta_U(T)}$ and $\sqrt{\Delta_W(T)}$ are bounded by the right-hand side of (91).

Step 2: test functions on the decoder particles. Let f be a fixed test function with $\|f\|_\infty \le 1$ and $\|f\|_{\mathrm{Lip}} \le 1$. Adding and subtracting the evaluation of f on the ideal particles and applying the triangle inequality,

$$\Big|\frac{1}{N}\sum_{j=1}^N f(u_j^k) - \int f\,d\rho_{U,k\epsilon}\Big| \le \underbrace{\Big|\frac{1}{N}\sum_{j=1}^N f(u_j^k) - \frac{1}{N}\sum_{j=1}^N f(\tilde{u}_j^{k\epsilon})\Big|}_{\text{Term 1 (Trajectory Error)}} + \underbrace{\Big|\frac{1}{N}\sum_{j=1}^N f(\tilde{u}_j^{k\epsilon}) - \int f\,d\rho_{U,k\epsilon}\Big|}_{\text{Term 2 (Measure Concentration)}}.$$

Term 1. Since f is 1-Lipschitz, and by Cauchy–Schwarz and (70),

$$\text{Term 1} \le \frac{1}{N}\sum_{j=1}^N \|u_j^k - \tilde{u}_j^{k\epsilon}\|_2 \le \sqrt{\frac{1}{N}\sum_{j=1}^N \|u_j^k - \tilde{u}_j^{k\epsilon}\|_2^2} \le \sqrt{\Delta_U(T)} \quad \text{for all } k \le K_T.$$

Term 2. For fixed k, the ideal particles $\tilde{u}_j^{k\epsilon}$ are i.i.d. from $\rho_{U,k\epsilon}$ and $f(\tilde{u}_j^{k\epsilon}) \in [-1,1]$. Hoeffding's inequality gives $\mathbb{P}(\text{Term 2} \ge \tau) \le 2\exp(-N\tau^2/2)$, and a union bound over the $K_T + 1 \le 2X_T$ grid times gives $\mathbb{P}(\max_k \text{Term 2} \ge \tau) \le 4X_T \exp(-N\tau^2/2)$. Setting the right-hand side equal to $e^{-\delta^2}$ and solving, with probability at least $1 - e^{-\delta^2}$,

$$\max_{k \le K_T} \text{Term 2} \le \sqrt{\frac{2}{N}}\sqrt{\log(4X_T) + \delta^2} \le \sqrt{\frac{2}{N}}\Big(\sqrt{\log(4X_T)} + \delta\Big) \le \sqrt{2}\,\mathrm{err}_{N,\epsilon}(\delta), \tag{92}$$

where in the last step we used $N^{-1} \le \epsilon \vee N^{-1}$ and $\log(4X_T) \le \log 4 + \log(NX_T) \le D + \log(NX_T)$, since $D \ge 2 \ge \log 4$.

Combining Term 1 with (91) and Term 2 with (92), on the intersection of the two events (probability at least $1 - 5e^{-\delta^2}$),

$$\begin{aligned}\max_{k \in [0,T/\epsilon] \cap \mathbb{N}} \Big|\frac{1}{N}\sum_{j=1}^N f(u_j^k) - \int f\,d\rho_{U,k\epsilon}\Big| &\le \Big(\sqrt{2C_{\mathrm{poly}}}\,(\sqrt{T} \vee T)\,e^{C_2T^2} + \sqrt{2}\Big)\mathrm{err}_{N,\epsilon}(\delta) \\ &\le C_1\,(\sqrt{T} \vee T \vee 1)\,e^{C_2T^2}\,\mathrm{err}_{N,\epsilon}(\delta).\end{aligned} \tag{93}$$

Step 3: test functions on the encoder particles. The identical argument for the encoder particles, with $\Delta_W(T) \le \Delta(T)$ in Term 1 and a separate Hoeffding union bound (costing another $e^{-\delta^2}$) in Term 2, gives on an event of probability at least $1 - 6e^{-\delta^2}$ both (93) and

$$\max_{k \in [0,T/\epsilon] \cap \mathbb{N}} \Big|\frac{1}{N}\sum_{i=1}^N f(w_i^k) - \int f\,d\rho_{W,k\epsilon}\Big| \le C_1\,(\sqrt{T} \vee T \vee 1)\,e^{C_2T^2}\,\mathrm{err}_{N,\epsilon}(\delta). \tag{94}$$

Step 4: the risk. By the triangle inequality, we decompose the risk error into the finite-particle tracking gap and the measure concentration gap:

$$\max_k \left|\mathcal{R}_\lambda^N(U^k, W^k) - \mathcal{R}_\lambda(\rho_{U,k\epsilon}, \rho_{W,k\epsilon})\right| \\ \leq \underbrace{\max_k \left|\mathcal{R}_\lambda^N(U^k, W^k) - \mathcal{R}_\lambda^N(\tilde{U}^{k\epsilon}, \tilde{W}^{k\epsilon})\right|}_{\text{Trajectory Gap}} + \underbrace{\max_k \left|\mathcal{R}_\lambda^N(\tilde{U}^{k\epsilon}, \tilde{W}^{k\epsilon}) - \mathcal{R}_\lambda(\rho_{U,k\epsilon}, \rho_{W,k\epsilon})\right|}_{\text{Measure Gap}}.$$

By Lemma 15, $\sqrt{a}+\sqrt{b} \leq \sqrt{2(a+b)}$, and (91), on the event of Lemma 14,

$$\text{Trajectory Gap} \leq K_R\Big(\sqrt{\Delta_U(T)} + \sqrt{\Delta_W(T)}\Big) \\ \leq K_R\sqrt{2\Delta(T)} \leq 2K_R\sqrt{C_{\text{poly}}}\,(\sqrt{T}\vee T)\,e^{C_2T^2}\,\text{err}_{N,\epsilon}(\delta).$$

By Lemma 16, with probability at least $1-e^{-\delta^2}$, using $\sqrt{\log(4X_T)} \leq \sqrt{D+\log(NX_T)}$ and $1 \leq \sqrt{D+\log(NX_T)}$,

$$\text{Measure Gap} \leq \frac{K_C}{\sqrt{N}}\Big[\sqrt{\log(4X_T)}+1+\delta\Big] \leq \frac{2K_C}{\sqrt{N}}\Big(\sqrt{D+\log(NX_T)}+\delta\Big) \leq 2K_C\,\text{err}_{N,\epsilon}(\delta).$$

Adding the two gaps, on an event of probability at least $1-7e^{-\delta^2}$ on which (93) and (94) also hold,

$$\max_{k\in[0,T/\epsilon]\cap\mathbb{N}} \left|\mathcal{R}_\lambda^N(U^k, W^k) - \mathcal{R}_\lambda(\rho_{U,k\epsilon}, \rho_{W,k\epsilon})\right| \leq C_1\,(\sqrt{T}\vee T\vee 1)\,e^{C_2T^2}\,\text{err}_{N,\epsilon}(\delta).$$

This proves (9)–(10) with the constants C_1, C_2 of (90).

Step 5: almost sure weak convergence. Finally, we show the almost sure weak convergence of the empirical SGD measures to the Vlasov–McKean solution at a fixed time $t \geq 0$ along a sequence (N, ϵ_N) as in the statement. We assume, as usual, that all networks are realized on a common probability space carrying the initializations and the data samples for every N, and we fix $T \geq t$. Let $k_N := \lfloor t/\epsilon_N \rfloor$ be the grid index corresponding to time t, so that $k_N\epsilon_N \leq t \leq T$.

Borel–Cantelli. For each N, set the confidence parameter to $\delta_N := \sqrt{2\log N}$ and let E_N be the event on which the bound of Lemma 14 (for width N, step ϵ_N, horizon T, and level δ_N) fails. Then $\mathbb{P}(E_N) \leq 4e^{-\delta_N^2} = 4/N^2$, so $\sum_N \mathbb{P}(E_N) < \infty$ and the first Borel–Cantelli lemma shows that almost surely E_N occurs only finitely often. Hence, almost surely, for all sufficiently large N,

$$\Delta^{(N)}(T) \leq C_{\text{poly}}(T\vee T^2)\Big(\epsilon_N + \frac{1}{N}\Big)\Big[D + \log\big(N(T/\epsilon_N \vee 1)\big) + 2\log N\Big]e^{C_{\exp}T^2},$$

where $\Delta^{(N)}$ denotes the deviation (70) for the N-th network. Since $\log(N(T/\epsilon_N \vee 1)) + 2\log N \leq 3\log(N/\epsilon_N) + \log(T\vee 1)$, the right-hand side is bounded by a constant multiple of $(\epsilon_N + N^{-1})\log(N/\epsilon_N)$ for large N, which tends to zero under the assumptions $\epsilon_N \log(N/\epsilon_N) \to 0$ and $N/\log(N/\epsilon_N) \to \infty$. Therefore

$$\Delta^{(N)}(T) \longrightarrow 0 \quad \text{almost surely.} \tag{95}$$

Wasserstein decomposition. Because all measures involved are supported in the compact ball $B_D(0, R_*)$, weak convergence is equivalent to convergence in W_1, so it suffices to show $W_1(\hat{\rho}^N_{U,k_N}, \rho_{U,t}) \to 0$ almost surely. By the triangle inequality through the bridge measure $\bar{\rho}^N_{U,k_N\epsilon_N} = \frac{1}{N}\sum_j \delta_{u_j^{k_N\epsilon_N}}$,

$$W_1(\hat{\rho}^N_{U,k_N}, \rho_{U,t}) \leq W_1(\hat{\rho}^N_{U,k_N}, \bar{\rho}^N_{U,k_N\epsilon_N}) + W_1(\bar{\rho}^N_{U,k_N\epsilon_N}, \rho_{U,k_N\epsilon_N}) + W_1(\rho_{U,k_N\epsilon_N}, \rho_{U,t}). \tag{96}$$

For the first term, the diagonal coupling of the two empirical measures through the particle index gives, exactly as in (80),

$$W_1(\hat{\rho}^N_{U,k_N}, \bar{\rho}^N_{U,k_N\epsilon_N}) \leq \frac{1}{N}\sum_{j=1}^N \|u_j^{k_N} - \tilde{u}_j^{k_N\epsilon_N}\|_2 \leq \sqrt{\Delta_U^{(N)}(T)} \longrightarrow 0$$

almost surely by (95). For the third term, $\rho_{U,t}$ is the push-forward of $\rho_{U,k_N\epsilon_N}$ under the characteristic flow from time $k_N\epsilon_N$ to time t, which displaces every point by at most $M_\xi M_*|t - k_N\epsilon_N| \le M_\xi M_*\epsilon_N$ by (62); Lemma 2 gives

$$W_1(\rho_{U,k_N\epsilon_N}, \rho_{U,t}) \le M_\xi M_*\epsilon_N \longrightarrow 0.$$

For the middle term we use that both measures are push-forwards under the same map: $\bar\rho^N_{U,s} = (X_U(s,\cdot))_\#\bar\rho^N_{U,0}$ and $\rho_{U,s} = (X_U(s,\cdot))_\#\rho_{U,0}$, where $\bar\rho^N_{U,0} = \frac{1}{N}\sum_j \delta_{u_j^0}$ is the empirical measure of the i.i.d. initializations. The flow map $X_U(s,\cdot)$ is Lipschitz on $B_D(0,R_*)$: for two initial points a, a', Lemma 13(a) and $\xi \le M_\xi$ give $\frac{d}{ds}\|X_U(s,a) - X_U(s,a')\|_2 \le M_\xi L_{\mathrm{sp}}\|X_U(s,a) - X_U(s,a')\|_2$, hence $\|X_U(s,a) - X_U(s,a')\|_2 \le e^{M_\xi L_{\mathrm{sp}} s}\|a - a'\|_2$ by Gronwall. Pushing forward an optimal coupling of $(\bar\rho^N_{U,0}, \rho_{U,0})$ through $(X_U(s,\cdot), X_U(s,\cdot))$ therefore shows

$$W_1(\bar\rho^N_{U,k_N\epsilon_N}, \rho_{U,k_N\epsilon_N}) \le e^{M_\xi L_{\mathrm{sp}} T}\, W_1(\bar\rho^N_{U,0}, \rho_{U,0}).$$

By Varadarajan's theorem, the empirical measure of i.i.d. samples from the fixed law $\rho_{U,0}$ converges weakly to $\rho_{U,0}$ almost surely, and since $\rho_{U,0}$ has compact support this is equivalent to $W_1(\bar\rho^N_{U,0}, \rho_{U,0}) \to 0$ almost surely. Thus the middle term of (96) tends to zero almost surely as well.

Combining the three terms, $W_1(\hat\rho^N_{U,k_N}, \rho_{U,t}) \to 0$ almost surely, which on the compact ball $B_D(0,R_*)$ is exactly the weak convergence

$$\hat\rho^N_{U,\lfloor t/\epsilon_N\rfloor} \Rightarrow \rho_{U,t} \quad \text{almost surely.}$$

By a perfectly symmetric argument for the encoder measures, using $\Delta^{(N)}_W(T) \le \Delta^{(N)}(T)$, the same time-shift bound, and the Lipschitz flow $X_W(s,\cdot)$, we concurrently obtain $\hat\rho^N_{W,\lfloor t/\epsilon_N\rfloor} \Rightarrow \rho_{W,t}$ almost surely. This completes the proof of Theorem 2. □

Remark 3. *The proof differs from the additive two-layer mean-field setting of [43] in the measure-gap term only. In the additive case the field seen by a particle depends on the other particles through a single average, and the corresponding concentration term is controlled by one function class. Here each decoder particle sees the encoder population through the composition $\alpha \mapsto \psi(\alpha;\cdot)$, and each encoder particle sees the decoder population through $\hat x$ and the backpropagated Jacobian $\overline{J}$; this is why three parametric classes $\mathcal{F}_{U,1}, \mathcal{F}_{U,2}, \mathcal{F}_W$ are needed in Lemma 13, and why the encoder and decoder deviations Δ_U, Δ_W cannot be closed separately but only jointly through Δ in the Gronwall step. The trapping ball of Theorem 1 is what allows every constant to be uniform despite the absence of a compact parameter domain.*

K Numerical Validation and Experimental Details

This appendix documents and extends the main experiments on finite-width agreement, the transition from a PCA-like to a nonlinear representation, and empirical self-consistency. We report variability across initializations, sensitivity to regularization and kernel bandwidth, and replications with Adam and biases on four datasets.

K.1 Experimental Setup and Diagnostics

Synthetic data and model. The main four-panel experiment and training-stage illustration use one fixed sample of $n = 500$ observations, $x_i = (s_i, \tanh s_i) + \varepsilon_i, \qquad s_i \in [-2.5, 2.5], \qquad \varepsilon_i \sim \mathcal{N}(0, 0.08^2 I_2)$, where the s_i are equally spaced and the data-generation seed is 7. We use the paper's bias-free, $1/N$-normalized encoder and decoder with $\tanh$ activation, input dimension $d = 2$, scalar bottleneck $b = 1$, and $\lambda = 10^{-6}$ unless stated otherwise. For weight particles $w_i = (w_{1,i}, w_{2,i})$ and $u_i = (u_{1,i}, u_{2,i})$, the empirical objective is

$$\widehat{\mathcal{R}}^N_{\lambda,n} = \frac{1}{2n}\sum_{j=1}^{n}\|\widehat{x}_j - x_j\|_2^2 + \frac{\lambda}{6N}\sum_{i=1}^{N}\big(\|u_i\|_2^6 + \|w_i\|_2^6\big). \tag{97}$$

All plotted risks evaluate this objective on the full fixed sample; the figures suppress the hat. SGD uses particle step size 0.1, equivalently learning rate $\eta_N = 0.1N$ for ordinary gradients of (97). Width comparisons use coupled initializations and, for mini-batch training, a common sample stream within each repetition.

Table 1: Settings for the noisy-$\tanh$ experiments.

Experiment	Specification
Finite width	$N \in \{8, 16, 32, 64, 128, 256, 512\}$, $N_{\mathrm{ref}} = 1024$; 10 coupled initializations (seeds $10000, \ldots, 10009$); batch size 128, sampled with replacement (stream seeds $20000, \ldots, 20009$); 4000 SGD updates, evaluated every 40 updates including initialization. Thus $t_m = 0.1m$ and $T = 400$.
Training geometry	$N \in \{8, 64, 1024\}$; 3 coupled initializations (seeds $30000, 30001, 30002$); full-batch SGD, evaluated every 100 epochs; 20000–30000 epochs under the stopping rule below.
Bandwidth	The same $N = 64$ trajectories, evaluated at $k \in \{10^{-4}, 0.005, 0.02, 0.05, 0.15, 0.3, 0.5, 2\}$.
Training snapshots	$N = 64$, seed 30000, $\eta_{64} = 6.4$; snapshots at epochs 2000 and 20000.

Diagnostics. Let h_i be the scalar encoder output, $\widehat{x}_i$ its reconstruction, c the sample mean, and v_1 the leading principal direction of the data. We measure discrepancy from a wider reference network on the saved iteration grid $\mathcal{I}$, and distance from the affine rank-one PCA line, by $e_N = \max_{m\in\mathcal{I}} \left|\widehat{\mathcal{R}}^N_{\lambda,n}(m) - \widehat{\mathcal{R}}^{N_{\mathrm{ref}}}_{\lambda,n}(m)\right|$, and $d_{\mathrm{PCA}} = \left[\frac{1}{n}\sum_{i=1}^n \left\|(I - v_1 v_1^\top)(\widehat{x}_i - c)\right\|_2^2\right]^{1/2}$. For Gaussian bandwidth $k > 0$, the kernel self-consistency diagnostic is

$$\widehat{m}_k(h_i) = \frac{\sum_{j=1}^n \exp[-(h_j - h_i)^2/(2k^2)]x_j}{\sum_{j=1}^n \exp[-(h_j - h_i)^2/(2k^2)]}, \qquad \bar{\Delta}_k = \frac{1}{n}\sum_{i=1}^n \|\widehat{x}_i - \widehat{m}_k(h_i)\|_2.$$

The default bandwidth is $k = 0.15$; we write $\bar{\Delta} = \bar{\Delta}_{0.15}$. This is a mean norm evaluated at learned codes, whereas the theoretical self-consistency residual uses a squared norm and an exact conditional mean. It does not test whether the encoder is a nearest-point projection. All diagnostics are in-sample; shaded bands and error bars denote one sample standard deviation across initializations, not confidence intervals.

Stopping and visualization. After 20000 epochs, full-batch training stops when all three runs satisfy $d_{\mathrm{PCA}} > 0.05$ and both the relative risk change and relative PCA-distance change over the preceding 1000 epochs are below 5×10^{-3}; otherwise it stops at 30000 epochs. The recorded stopping epochs are 20000, 23000, and 20000 for $N = 8, 64, 1024$, respectively. This is an operational stopping rule, not an optimality certificate. The individual run in Figure 3 satisfies it at epoch 20000. Its curves decode 500 uniformly spaced codes between the minimum and maximum observed codes; the PCA line uses 500 scores spanning the observed range with a 5% margin at each end.

Implementation. Computations use single-precision arithmetic on the CPU, with at most four PyTorch threads for the mean-field experiments. The reported environment is Python 3.12.13, NumPy 2.0.1, PyTorch 2.5.1, scikit-learn 1.9.0, Seaborn 0.13.2, and Matplotlib 3.11.0. Code is provided in the supplementary material.

K.2 Validation on Noisy Nonlinear Data

Finite width and training geometry. Figure 2 shows the finite width and training geometry. The mean discrepancy decreases with width, on a scale broadly consistent with the $N^{-1/2}$ reference. Because the comparison uses a finite reference network, a fixed step size, and a discrete evaluation grid, it is a numerical check of agreement between widths rather than a direct measurement of error to the continuous mean-field limit. Across the tested widths, reconstructed samples first approach the PCA line and subsequently move away while kernel self-consistency improves overall. Figure 3 makes this transition explicit: the epoch-2000 decoder is approximately linear, whereas the final curve follows the nonlinear data geometry. Thus the snapshots support the interpretation of d_{PCA}, whose magnitude alone would not establish nonlinearity.

Bandwidth sensitivity. Figure 2**(d)** shows that the magnitude and evolution of $\bar{\Delta}_k$ depend on smoothing. The estimator includes the observation itself: for distinct codes, $\widehat{m}_k(h_i) \to x_i$ as $k \to 0$, whereas large bandwidths average over progressively more of the sample. The diagnostic

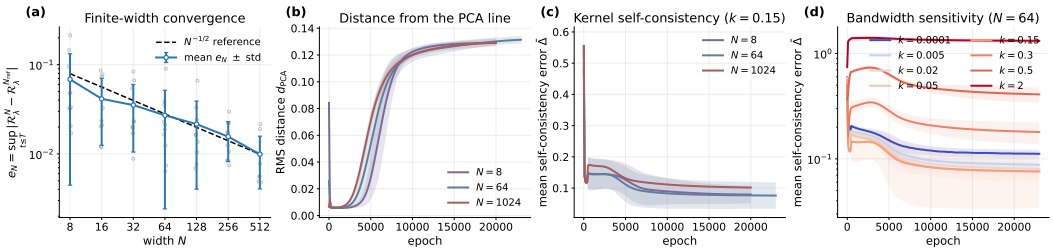


Figure 2: **Finite width and training geometry, including variability.** **(a)** e_N for 10 coupled initializations, with mean and standard deviation; the $N^{-1/2}$ guide is anchored at $N = 512$. **(b)** d_{PCA} and **(c)** $\bar{\Delta}$ for $N = 8, 64, 1024$. **(d)** Bandwidth sensitivity at $N = 64$. Curves and bands in **(b)–(d)** show means and standard deviations over three runs on the same fixed sample.

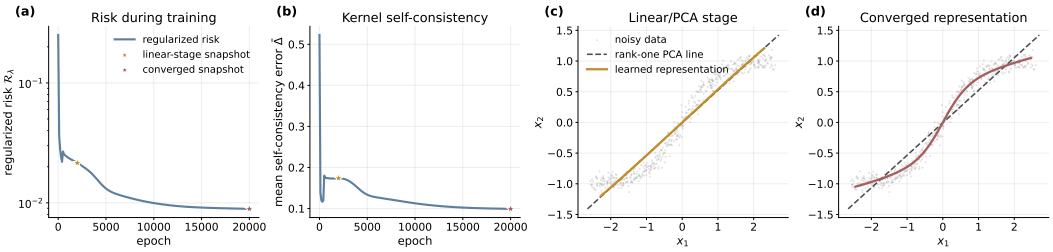


Figure 3: **From a PCA-like representation to a nonlinear curve.** **(a)** Regularized empirical risk and **(b)** kernel self-consistency for one $N = 64$ full-batch SGD run. Stars mark the states shown in **(c)** at epoch 2000 and **(d)** at epoch 20000. The decoder initially follows the PCA line and later follows the nonlinear structure of the data.

therefore interpolates between reconstruction error and comparison with a global mean; a low value at one bandwidth alone is insufficient evidence for population self-consistency.

Regularization sensitivity. Figure 4 compares $\lambda \in \{10^{-8}, 10^{-6}, 10^{-4}\}$ using the same observations, coupled initializations, mini-batch streams, evaluation grids, and horizon across regularization strengths. The strongest penalty suppresses departure from the PCA line over the displayed horizon, while the two weaker penalties allow nonlinear representations. Among these three settings, $\lambda = 10^{-6}$ gives the smallest final $\bar{\Delta}$ at $k = 0.15$. The objective decomposition also shows that lower reconstruction error need not imply lower kernel self-consistency error.

K.3 Additional Datasets and Adam Optimization

Architecture and protocol. We repeat the diagnostics on half moons, Iris, Breast Cancer Wisconsin, and Digits using Adam and trainable biases. All biases start at zero and are trained jointly with the weights; the sixth-order penalty in (97) remains weight-only. Adam uses $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$, without weight decay or a learning-rate schedule. The learning rate is fixed across widths within each dataset. Finite-width comparisons use $N \in \{8, 16, 32, 64, 128\}$, $N_{\mathrm{ref}} = 256$, and ten coupled mini-batch runs. Training trajectories use $N \in \{8, 64, 128\}$, three initializations, and 12000 full-batch updates. Bandwidth comparisons reuse the $N = 64$ trajectories and the bandwidths in Table 1. The regularization sweep uses $\lambda \in \{10^{-8}, 10^{-6}, 10^{-3}\}$ and three runs.

Parameter selection. For each dataset, one $N = 64$ initialization is screened for 12000 full-batch updates over $\lambda \in \{10^{-8}, 10^{-6}, 10^{-4}, 10^{-3}\}$, and $\eta \in \{3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$. Eligible settings reduce $\bar{\Delta}$ by at least 5%, have relative range at most 8% over the final fifth of training, and attain reconstruction loss at most 1.05 times the rank-one PCA loss. They are ranked by a score combining final self-consistency error, late-training slope, and variability; PCA departure is not a selection criterion. Table 2 reports the selected settings and the three-run means.

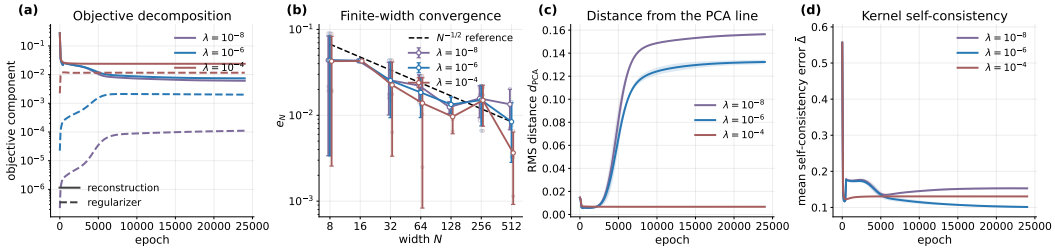


Figure 4: **Regularization sensitivity on noisy** $\tanh$ **data.** Colors identify $\lambda \in \{10^{-8}, 10^{-6}, 10^{-4}\}$. **(a)** Reconstruction (solid) and penalty (dashed) terms. **(b)** e_N relative to $N_{\text{ref}} = 1024$, with individual runs and mean $\pm$ standard deviation; the $N^{-1/2}$ guide is anchored at the $\lambda = 10^{-6}$ result for $N = 512$. **(c)** d_{PCA} and **(d)** $\bar{\Delta}$. Trajectories in **(a), (c), (d)** use $N = 64$ and average three full-batch runs; bands in **(c), (d)** show standard deviations.

Table 2: Selected Adam settings and mean self-consistency errors at $N = 64$ over three initializations.

Dataset	λ	η	$\bar{\Delta}(0)$	$\bar{\Delta}(12000)$
Half moons	10^{-6}	3×10^{-3}	0.249	0.067
Iris	10^{-8}	10^{-2}	0.376	0.089
Breast Cancer Wisconsin	10^{-6}	3×10^{-3}	0.215	0.138
Digits	10^{-8}	10^{-2}	0.294	0.158

Since selection uses self-consistency, these comparisons demonstrate the behavior of the selected configurations rather than improvement for arbitrary hyperparameters.

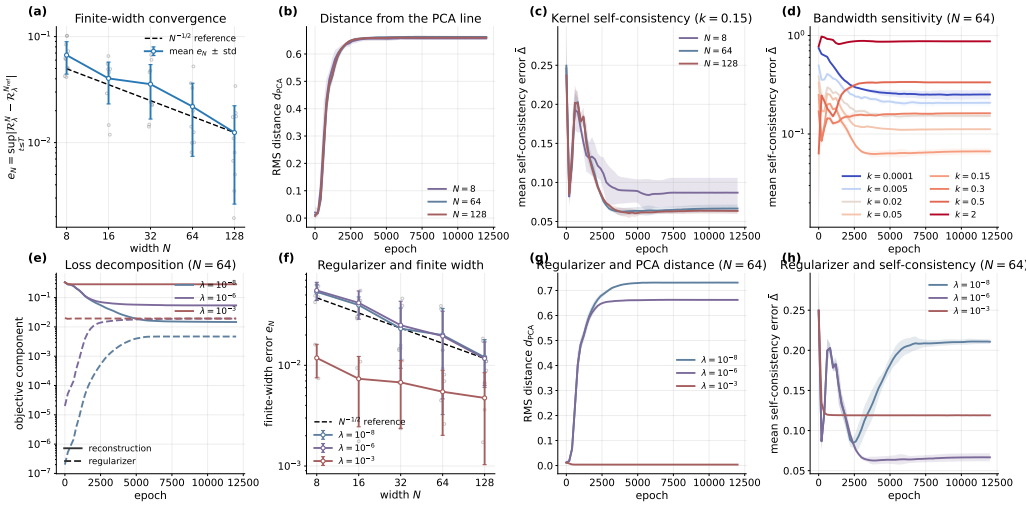
Results. Figure 5 shows decreasing mean finite-width discrepancy for the selected configurations on all four datasets. The $N = 64$ self-consistency errors fall below their initial values and approach plateaus, although the trajectories can be nonmonotone and increasing width does not uniformly improve the final error. Reconstructions depart from the rank-one PCA line; a distinct PCA-like intermediate stage is less evident than in the noisy-$\tanh$ experiment. Strong regularization generally limits this departure, while its effect on self-consistency is dataset-dependent. These replications support the relevance of the numerical diagnostics beyond the bias-free SGD setting. Adam dynamics and unpenalized biases are outside the stated mean-field convergence theorem.

K.4 Illustration of Latent Dimension

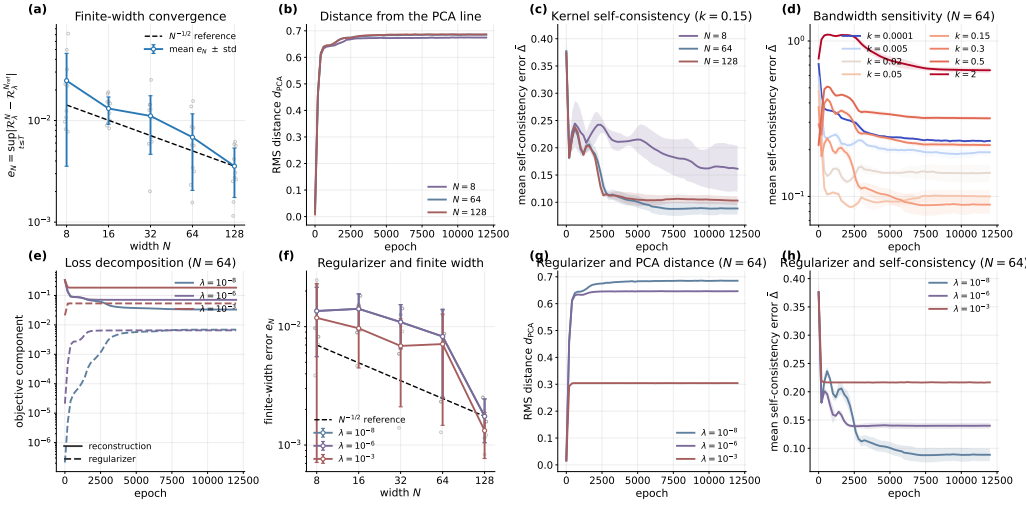
Figure 6 illustrates the difference between PCA and nonlinear representations on 1000 observations generated by `sklearn.datasets.make_moons` with noise standard deviation 0.05. The one-dimensional AE separates the two components more clearly than the leading PCA coordinate; the three-dimensional representation illustrates the additional geometric freedom of a larger latent space. Since $b = 3 > d = 2$, the latter is an embedding illustration, not a compression experiment. The architectures are $2 \to 128 \to 1 \to 128 \to 2$ and $2 \to 64 \to 3 \to 64 \to 2$, with hidden $\tanh$ activations, linear latent and output layers, and biases in every affine map. They minimize unregularized full-batch mean squared error using Adam with learning rate 0.01, the coefficients stated above, and no weight decay, for 1000 and 300 updates, respectively. The plots use the final encoder outputs; class labels are used only for visualization.

L Self-Consistency and Principal Manifolds

Just as linear bottleneck autoencoders are fundamentally linked to Principal Component Analysis, their nonlinear counterparts yield a geometric structure known as a *principal manifold*. Such a manifold is a low-dimensional representation formed by the decoder, defined by a self-consistency property where each point on the surface is the conditional expected value of the data mapped to it



(i) Half moons.



(ii) Iris.

Figure 5: **Dataset replications with Adam and biases.** **(a)** Finite-width discrepancy relative to $N_{\mathrm{ref}} = 256$; open points are individual runs, bars show mean $\pm$ standard deviation, and the dashed line is an $N^{-1/2}$ guide. **(b)** PCA distance, **(c)** self-consistency at $k = 0.15$, and **(d)** bandwidth sensitivity at $N = 64$. **(e)–(h)** Regularization sensitivity: reconstruction and penalty terms, finite-width discrepancy, PCA distance, and self-consistency. Solid and dashed curves in **(e)** denote reconstruction and penalty terms; shaded bands denote one standard deviation.

via nearest-point projection. To make this geometric statement mathematically precise, we evaluate the macroscopic maps generated by the mean-field limits of the encoder and decoder, confining ourtheo analysis to the unregularized reconstruction risk.

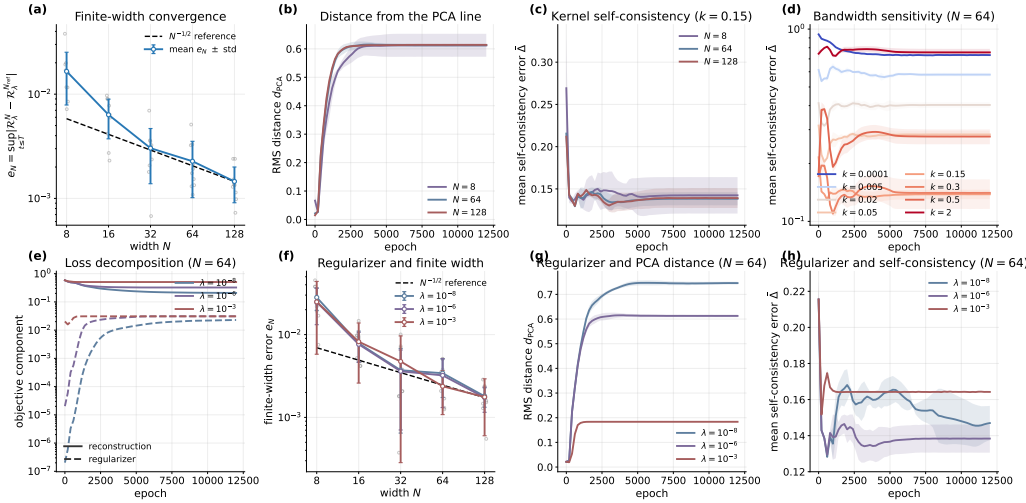
Let $X \sim P$. The bounded-support assumption from the preceding sections implies

$$\mathbb{E}\|X\|_2^2 < \infty.$$

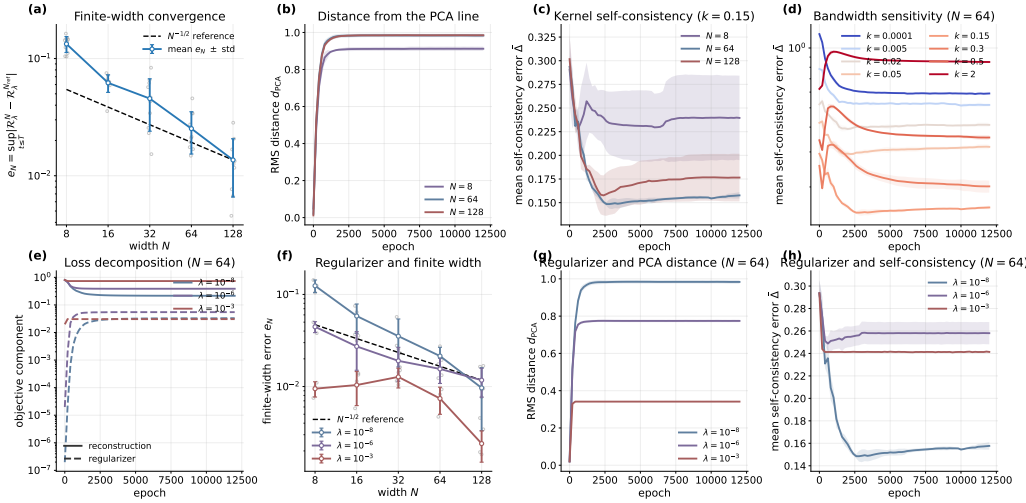
For a mean-field pair (ρ_U, ρ_W), define the induced encoder and decoder maps by

$$f_{\rho_W}(x) := \alpha(x; \rho_W) = \int \phi(x; w)\, \rho_W(dw), \qquad g_{\rho_U}(z) := \int \psi(z; u)\, \rho_U(du).$$

Thus f_{ρ_W} is the encoder and g_{ρ_U} is the decoder function. We change the notation here to focus take the focus away from the measure ρ_U and ρ_W. With this change in perspective the unregularized



(iii) Breast Cancer Wisconsin.



(iv) Digits.

Figure 5: Dataset replications with Adam and biases (continued). Panel definitions and uncertainty summaries are as above.

mean-field risk can be rewritten as,

$$\mathcal{R}(\rho_U, \rho_W) = \mathcal{L}(f_{\rho_W}, g_{\rho_U}), \qquad \mathcal{L}(f, g) := \frac{1}{2}\,\mathbb{E}\big[\|X - g(f(X))\|_2^2\big]\,.$$

Fix a nonempty compact latent set $\mathcal{K} \subset \mathbb{R}^b$. We write

$$\mathfrak{F}_W(\mathcal{K}) \subseteq \{f : \mathbb{R}^d \to \mathcal{K} : f \text{ is Borel}\}, \qquad \mathfrak{G}_U(\mathcal{K}) \subseteq C(\mathcal{K}; \mathbb{R}^d)$$

for the encoder and decoder map classes. Because of the fixed ball argument one can always choose such a compact set.

Definition 2 (Blockwise optimality)**.** *A pair* $(f^\star, g^\star) \in \mathfrak{F}_W(\mathcal{K}) \times \mathfrak{G}_U(\mathcal{K})$ *is* blockwise optimal *if*

$$\mathcal{L}(f^\star, g^\star) \le \mathcal{L}(f, g^\star) \qquad \text{for every } f \in \mathfrak{F}_W(\mathcal{K}),$$
$$\mathcal{L}(f^\star, g^\star) \le \mathcal{L}(f^\star, g) \qquad \text{for every } g \in \mathfrak{G}_U(\mathcal{K}).$$

Thus each map is a global minimizer when the other map is held fixed. Every joint global minimizer is blockwise optimal, but the converse need not hold.

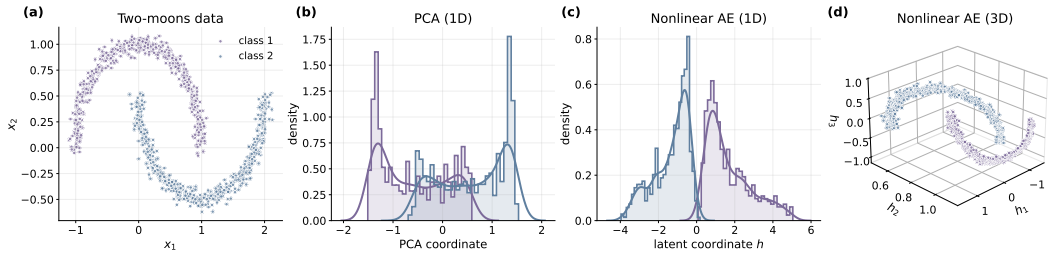


Figure 6: **Linear and nonlinear representations of two moons. (a)** Noisy observations, colored by component. **(b)** Class-conditional densities of the leading PCA coordinate. **(c)** Densities of the one-dimensional nonlinear AE code. **(d)** Three-dimensional nonlinear AE representation of the same observations.

For a Borel encoder $f : \mathbb{R}^d \to \mathcal{K}$, let the push forward probabaility measure be $Q_f := f_{\#}P$. Since $\mathcal{K}$ is a standard Borel space, the conditional mean has a Borel version $m_f : \mathcal{K} \to \mathbb{R}^d$ satisfying

$$m_f(f(X)) = \mathbb{E}[X \mid \sigma(f(X))] \qquad \text{almost surely.} \tag{98}$$

It is unique Q_f-almost everywhere. For a continuous decoder $g : \mathcal{K} \to \mathbb{R}^d$, define the nearest-code correspondence

$$A_g(x) := \arg\min_{z \in \mathcal{K}} \|x - g(z)\|_2^2.$$

As $\mathcal{K}$ is compact and g is continuous, $A_g(x)$ is nonempty and compact for every $x \in \mathbb{R}^d$. We measure the failures of the conditional-mean and nearest-point identities by

$$\mathsf{SC}(f, g) := \int_{\mathcal{K}} \|g(z) - m_f(z)\|_2^2 \, Q_f(dz),$$
$$\mathsf{Proj}(f, g) := \mathbb{E}\left[\|X - g(f(X))\|_2^2 - \min_{z \in \mathcal{K}} \|X - g(z)\|_2^2\right].$$

Both are nonnegative, and $\mathsf{SC}(f, g)$ does not depend on the chosen Q_f-almost-everywhere version of m_f.

Lemma 17 (Conditional-mean decomposition)**.** *Let $f : \mathbb{R}^d \to \mathcal{K}$ be Borel. There is a Borel version m_f satisfying* (98)*, and $m_f \in L^2(Q_f; \mathbb{R}^d)$. For every Borel map $a : \mathcal{K} \to \mathbb{R}^d$ such that $a(f(X)) \in L^2$, one has*

$$\mathbb{E}\|X - a(f(X))\|_2^2 = \mathbb{E}\|X - m_f(f(X))\|_2^2 + \int_{\mathcal{K}} \|m_f(z) - a(z)\|_2^2 \, Q_f(dz). \tag{99}$$

Consequently,

$$\mathcal{L}(f, g) = \mathcal{L}(f, m_f) + \frac{1}{2}\mathsf{SC}(f, g) \tag{100}$$

for every continuous decoder $g : \mathcal{K} \to \mathbb{R}^d$.

Proof. Because $\mathcal{K}$ and $\mathbb{R}^d$ are standard Borel spaces, there is a regular conditional distribution $\kappa(z, dx)$ of X given $f(X) = z$. Defining

$$m_f(z) := \int_{\mathbb{R}^d} x \, \kappa(z, dx)$$

whenever the integral is finite, and setting $m_f(z) = 0$ on the remaining Q_f-null set, gives a Borel map satisfying (98). Conditional Jensen's inequality yields

$$\int_{\mathcal{K}} \|m_f(z)\|_2^2 \, Q_f(dz) \le \mathbb{E}\|X\|_2^2 < \infty.$$

Set $Z = f(X)$. Expanding

$$X - a(Z) = (X - m_f(Z)) + (m_f(Z) - a(Z))$$

gives the two squared terms in (99) and the cross term

$$2\,\mathbb{E}[\langle X - m_f(Z), m_f(Z) - a(Z)\rangle]\,.$$

Since $m_f(Z) - a(Z)$ is $\sigma(Z)$-measurable and $\mathbb{E}[X - m_f(Z) \mid \sigma(Z)] = 0$, the cross term is zero. This proves (99); multiplying by $1/2$ and taking $a = g$ gives (100). □

We next record the measurable-selection result used to construct a nearest-point encoder from the nearest-code correspondence A_g.

Theorem 8 (Kuratowski–Ryll-Nardzewski measurable selection)**.** *Let E be a standard Borel space and let K be a compact metric space. Suppose that a set-valued map $A : E \rightrightarrows K$ has nonempty closed values and is weakly measurable, meaning that*

$$\{x \in E : A(x) \cap O \neq \varnothing\}$$

is Borel for every relatively open set $O \subseteq K$. Then there exists a Borel map $s : E \to K$ such that $s(x) \in A(x)$ for every $x \in E$.

Lemma 18 (Borel nearest-point encoder)**.** *Let $g \in C(\mathcal{K}; \mathbb{R}^d)$. There exists a Borel map $\pi_g^0 : \mathbb{R}^d \to \mathcal{K}$ such that*

$$\pi_g^0(x) \in A_g(x) \qquad \text{for every } x \in \mathbb{R}^d. \tag{101}$$

Moreover, every Borel encoder $f : \mathbb{R}^d \to \mathcal{K}$ satisfies

$$\mathcal{L}(f, g) - \mathcal{L}(\pi_g^0, g) = \frac{1}{2}\mathsf{Proj}(f, g) \geq 0, \tag{102}$$

and equality holds if and only if $f(x) \in A_g(x)$ for P-almost every x.

Proof. Define

$$c(x, z) := \|x - g(z)\|_2^2, \qquad v(x) := \min_{z \in \mathcal{K}} c(x, z).$$

Since $g(\mathcal{K})$ is compact,

$$v(x) = \operatorname{dist}(x, g(\mathcal{K}))^2,$$

so v is continuous. For every x,

$$A_g(x) = \{z \in \mathcal{K} : c(x, z) = v(x)\}$$

is nonempty and compact.

It remains to verify weak measurability. Let $O \subseteq \mathcal{K}$ be relatively open. The case $O = \mathcal{K}$ is immediate. Otherwise, for $m \geq 1$, define

$$F_m := \left\{ z \in \mathcal{K} : \operatorname{dist}(z, \mathcal{K} \setminus O) \geq \frac{1}{m} \right\}.$$

Each F_m is compact and $O = \bigcup_{m \geq 1} F_m$. If $F \subseteq \mathcal{K}$ is nonempty and compact, then

$$x \longmapsto \min_{z \in F} c(x, z) = \operatorname{dist}(x, g(F))^2$$

is continuous. Therefore

$$\{x : A_g(x) \cap O \neq \varnothing\} = \bigcup_{m : F_m \neq \varnothing} \left\{ x : \min_{z \in F_m} c(x, z) = v(x) \right\},$$

which is Borel. Theorem 8 now gives a Borel map π_g^0 satisfying (101).

For every $x \in \mathbb{R}^d$,

$$\|x - g(f(x))\|_2^2 \geq \min_{z \in \mathcal{K}} \|x - g(z)\|_2^2 = \|x - g(\pi_g^0(x))\|_2^2.$$

The difference is a nonnegative Borel function. Integrating it and multiplying by $1/2$ proves (102). Its expectation is zero if and only if it is zero P-almost surely, which is equivalent to $f(x) \in A_g(x)$ for P-almost every x. □

Definition 3 (Self-consistent principal manifold)**.** *Let $\mathcal{K}$ be a compact b-dimensional C^1 manifold with boundary, let $g : \mathcal{K} \to \mathbb{R}^d$ be a C^1 embedding, and put $M := g(\mathcal{K})$. Let π_g^0 be an everywhere-defined Borel nearest-point selector from Lemma 18, and define*

$$\Pi_M(x) := g(\pi_g^0(x)) \in M.$$

Assume that the nearest point in M is unique for P-almost every x. We call M a self-consistent principal b-manifold *for P if*

$$\Pi_M(X) = \mathbb{E}[X \mid \sigma(\Pi_M(X))] \qquad \textit{almost surely}. \tag{103}$$

The almost-sure uniqueness assumption makes this definition independent of the chosen Borel selector.

Theorem 5. *Let $(f^\star, g^\star) \in \mathfrak{F}_W(\mathcal{K}) \times \mathfrak{G}_U(\mathcal{K})$ be blockwise optimal and assume (i) some Borel encoder $\pi_{g^\star} \in \mathfrak{F}_W(\mathcal{K})$ satisfies $\pi_{g^\star}(x) \in A_{g^\star}(x)$ for P-a.e. x, and (ii) $m_{f^\star}$ has a continuous representative in $\mathfrak{G}_U(\mathcal{K})$. Then*

$$f^\star(x) \in A_{g^\star}(x) \text{ for } P\text{-a.e. } x, \qquad g^\star(z) = m_{f^\star}(z) \text{ for } Q_{f^\star}\text{-a.e. } z, \tag{12}$$

equivalently $\mathsf{Proj}(f^\star, g^\star) = \mathsf{SC}(f^\star, g^\star) = 0$. If, in addition, $g^\star$ is injective and the nearest point in $g^\star(\mathcal{K})$ is unique P-almost surely, then $f^\star = \pi_{g^\star}$ P-almost surely and

$$g^\star(z) = m_{\pi_{g^\star}}(z) \qquad Q_{f^\star}\textit{-almost everywhere}. \tag{13}$$

If, furthermore, $\mathcal{K}$ is a compact b-dimensional C^1 manifold with boundary and $g^\star$ is a C^1 embedding, then $g^\star(\mathcal{K})$ is a self-consistent principal b-manifold in the sense of Definition 3. For $b = 1$, this is the classical principal-curve self-consistency condition of Hastie and Stuetzle [49].

Proof. Let $\pi_{g^\star}^0$ be the everywhere-defined selector from Lemma 18. Assumption (i) and the equality characterization in that lemma give

$$\mathcal{L}(\pi_{g^\star}, g^\star) = \mathcal{L}(\pi_{g^\star}^0, g^\star) \le \mathcal{L}(f^\star, g^\star).$$

Because $\pi_{g^\star}$ is an admissible encoder competitor, blockwise optimality gives the reverse inequality. Hence equality holds. Equation (102) then gives $\mathsf{Proj}(f^\star, g^\star) = 0$, and the equality characterization in Lemma 18 completes the proof of the first part of 12.

Let $\overline{m} \in \mathfrak{G}_U(\mathcal{K})$ be the continuous representative from assumption (ii). Since $\overline{m} = m_{f^\star}$ $Q_{f^\star}$-almost everywhere, Lemma 17 gives

$$\mathcal{L}(f^\star, g^\star) = \mathcal{L}(f^\star, \overline{m}) + \frac{1}{2}\mathsf{SC}(f^\star, g^\star). \tag{104}$$

Blockwise optimality in the decoder block implies $\mathcal{L}(f^\star, g^\star) \le \mathcal{L}(f^\star, \overline{m})$. Since the residual in (104) is nonnegative, it must vanish. This proves in particula second part of (12).

Now assume injectivity and almost-sure uniqueness of the nearest image point. First part of the equation (12) shows that $g^\star(f^\star(X))$ is a nearest point in $g^\star(\mathcal{K})$. Uniqueness of that image point and injectivity of $g^\star$ imply $f^\star(X) = \pi_{g^\star}(X) = \pi_{g^\star}^0(X)$ almost surely. Hence these latent variables have the same law and generate the same completed sigma-field, so second part of (12) is equivalent to (13).

Finally, let $M = g^\star(\mathcal{K})$ and $\Pi_M(X) = g^\star(\pi_{g^\star}^0(X))$. Since a continuous injection from compact $\mathcal{K}$ into $\mathbb{R}^d$ is a homeomorphism onto its image, $(g^\star)^{-1} : M \to \mathcal{K}$ is Borel. Therefore

$$\sigma(\pi_{g^\star}^0(X)) = \sigma(\Pi_M(X))$$

up to completion. Combining this with (13) gives

$$\Pi_M(X) = \mathbb{E}[X \mid \sigma(\Pi_M(X))] \qquad \text{almost surely},$$

which is precisely (103). □

M EARLY-TIME PCA-FIRST DYNAMICS FOR AN ARBITRARY BOTTLENECK DIMENSION

Appendix L describes the geometry of an optimum. Here we look at the other end of training and ask what the model learns first when it starts close to the zero map. The answer is a PCA-first mechanism, and we establish it in two steps.

In Section M.1 we take the linear activation $\sigma(z) = z$. Then the mean-field dynamics reduce exactly to a system of ODEs for two moment matrices, and we show that these matrices align with the top eigenvector of the data covariance before anything else happens. In Section M.2 we return to the bounded nonlinear activation of the paper. When the network is initialized close to the zero map, the encoder output, and hence the bottleneck representation, is small. Thus both the encoder and the decoder operate near the origin, where their activations can be approximated by linear functions, and during this initial regime the encoder features change only slightly. As a result, the same two moment matrices satisfy a linear system of the same form as in the linear model, except that the data covariance is replaced by a covariance matrix weighted by the initial feature kernel. When this kernel is in its linear regime, the weighted covariance has the same leading eigenvector as the data covariance, so the nonlinear model also starts by learning the first principal component.

Throughout Sections M.1–M.2 we study $\lambda = 0$. This is what gives a closed system of moments and a clean PCA-first picture.

We assume that the data are centered, $\mathbb{E}[x] = 0$, and denote the covariance matrix of the data by $\Sigma := \mathbb{E}[xx^\top] \in \mathbb{R}^{d\times d}$, with eigenvalues $\mu_1 > \mu_2 \geq \cdots \geq \mu_d \geq 0$ and orthonormal eigenvectors $v_1, \ldots, v_d$. We use the intrinsic time $s(t) := \int_0^t \xi(r)\, dr$ and write a dot for the derivative with respect to s; this removes the learning-rate schedule from all equations. The encoder and decoder maps are $\phi(x; w) = w_2\sigma(\langle w_1, x\rangle)$ and $\psi(a; u) = u_1\sigma(\langle u_2, a\rangle)$, and we write

$$\alpha_s(x) := \int \phi(x; w)\, \rho_{W,s}(dw) \in \mathbb{R}^b, \qquad \Psi_s(a) := \int \psi(a; u)\, \rho_{U,s}(du) \in \mathbb{R}^d,$$

so that $\widehat{x}_s(x) = \Psi_s(\alpha_s(x))$. The second moments of the particle laws are

$$Q_{w_1,s} := \int w_1 w_1^\top \rho_{W,s}(dw) \in \mathbb{R}^{d\times d}, \qquad Q_{w_2,s} := \int w_2 w_2^\top \rho_{W,s}(dw) \in \mathbb{R}^{b\times b},$$
$$Q_{u_1,s} := \int u_1 u_1^\top \rho_{U,s}(du) \in \mathbb{R}^{d\times d}, \qquad Q_{u_2,s} := \int u_2 u_2^\top \rho_{U,s}(du) \in \mathbb{R}^{b\times b}.$$

For a nonzero matrix $M \in \mathbb{R}^{d\times b}$ and a unit vector $v \in \mathbb{R}^d$ we measure how much of M lies outside the direction v by

$$\sin\angle(M, v) := \frac{\|(I_d - vv^\top)M\|_{\mathrm{F}}}{\|M\|_{\mathrm{F}}}.$$

This is zero exactly when every column of M is a multiple of v. This $\sin\angle$ is a notation to indicate that alignment between M and v. This has nothing to do with actual $\sin$ angle between two vectors.

M.1 THE LINEAR AUTOENCODER

We begin with the linear activation $\sigma(z) = z$, which is the classical setting in which autoencoders are related to PCA. The goal of this subsection is Theorem 9 below: if the linear model is started from a small initialization, then the moment matrices that describe the encoder and the decoder grow exponentially, and by the time they reach a fixed macroscopic size they are aligned with the first principal direction of the data. Note that the linear activation is unbounded and does not satisfy Assumption 1, so nothing in this subsection is a global statement. All results are finite-time comparisons that identify the direction in which the moments move before higher-order terms take over.

With a linear activation the encoder and decoder are matrix multiplications. Define

$$A_s := \int w_1 w_2^\top \rho_{W,s}(dw) \in \mathbb{R}^{d\times b}, \qquad B_s := \int u_1 u_2^\top \rho_{U,s}(du) \in \mathbb{R}^{d\times b}. \tag{105}$$

Then $\alpha_s(x) = A_s^\top x$ and $\widehat{x}_s(x) = B_s A_s^\top x$, so the reconstruction risk depends on the particle laws only through (A, B),

$$\mathcal{R}_{\text{lin}}(A, B) := \frac{1}{2}\mathbb{E}_x\|x - BA^\top x\|_2^2,$$

and its matrix gradients are

$$F(A, B) := \nabla_A \mathcal{R}_{\text{lin}}(A, B) = \Sigma A B^\top B - \Sigma B \in \mathbb{R}^{d\times b}, \tag{106}$$

$$H(A, B) := \nabla_B \mathcal{R}_{\text{lin}}(A, B) = BA^\top \Sigma A - \Sigma A \in \mathbb{R}^{d\times b}. \tag{107}$$

We write $Z_s := (A_s, B_s)$ with $\|Z\|_2^2 := \|A\|_{\text{F}}^2 + \|B\|_{\text{F}}^2$. The number $\|Z_0\|_2$ is the size of the initialization and will appear throughout.

We assume that the initial seeds are compactly supported and isotropic:

$$Q_{w_1,0} = Q_{u_1,0} = I_d, \qquad Q_{w_2,0} = Q_{u_2,0} = I_b. \tag{108}$$

Such an initialization can be constructed for any prescribed matrices A_0 and B_0 with operator norm at most one. One may choose w_1 and w_2 along eigenvectors of suitable block matrices and multiply them by independent Rademacher signs. The random signs ensure that the individual second moments are identity matrices, while the joint construction can be chosen so that the cross moment satisfies $\int w_1 w_2^\top = A_0$. An analogous construction gives the cross moment B_0 for u_1 and u_2. For details of above construction, see lemma 20.

To see what drives the dynamics near the origin, note that for $Z \approx 0$ and $Q \approx I$ the gradients (106)–(107) are, to first order in Z, $F \approx -\Sigma B$ and $H \approx -\Sigma A$, and we will see in Proposition 3 that the moments move by $\dot{A} = -Q_{w_1} F - F Q_{w_2}$ and $\dot{B} = -Q_{u_1} H - H Q_{u_2}$. So near the origin $\dot{A} \approx 2\Sigma B$ and $\dot{B} \approx 2\Sigma A$, and the linear part of the moment system is the block operator

$$\mathcal{M}(A, B) := (\dot{A}, \dot{B}) = (2\Sigma B,\, 2\Sigma A).$$

The eigenvectors of $\mathcal{M}$ are $(v_i, \pm v_i)/\sqrt{2}$ with eigenvalues $\pm 2\mu_i$, so the linearized flow $e^{\mathcal{M}s} Z_0$ amplifies the v_1-component of $A + B$ fastest. This is the main idea behind the theorem.

M.1.1 Main result

Now we describe the first result for linear activation. The result basically says that if one starts the linear model from a small initialization of size $\|Z_0\|_2$ with isotropic seeds, then both moment matrices grow exponentially, and by the time they have reached a fixed macroscopic size δ almost all of their column energy (by column energy we mean vv^T for a column v) lies along the first principal direction v_1. The bound has three parts: a term $(\|Z_0\|_2/\delta)^{1-\mu_2/\mu_1}$ coming from the second principal direction, which grows too but more slowly; a term $\|Z_0\|_2/\delta$ from the modes that decay and a term δ^2 from the nonlinear corrections that become visible at size δ. The gap $\mu_1 > \mu_2$ is what makes the first term small.

Theorem 9 (Linear PCA-first)**.** *Let $\sigma(z) = z$ and $\lambda = 0$, and assume $d \geq 2$, $\mu_1 > \mu_2$ and assume* (108)*. Assume $\|Z_0\|_2 > 0$ and that the initialization has some overlap with the top eigenvector of Σ,*

$$\kappa_1 := \Big\| v_1^\top \frac{A_0 + B_0}{\sqrt{2}} \Big\|_2 > 0. \tag{109}$$

Then there are constants $c_\star, C_\star > 0$, depending only on μ_1, such that the following holds. Let δ be such that $\|Z_0\|_2 < \delta < c_\star$ and

$$\frac{\|Z_0\|_2}{\kappa_1}\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^{1-\mu_2/\mu_1} + \frac{\|Z_0\|_2}{\delta} + \delta^2\Big] \leq c_\star. \tag{110}$$

Then, at the time

$$s_\delta := \frac{1}{2\mu_1}\log\frac{\delta}{\|Z_0\|_2}$$

at which the linearized flow reaches size δ, the matrices A_{s_δ} and B_{s_δ} are nonzero and

$$\sin\angle(A_{s_\delta}, v_1) \vee \sin\angle(B_{s_\delta}, v_1) \leq C_\star \frac{\|Z_0\|_2}{\kappa_1}\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^{1-\mu_2/\mu_1} + \frac{\|Z_0\|_2}{\delta} + \delta^2\Big]. \tag{111}$$

Remark 4. *The theorem has four requirements on the initialization: the isotropic moments* (108)*, a nonzero overlap $\kappa_1 > 0$, a size $\|Z_0\|_2$ below the scale δ, and eq.* (110)*. All four are satisfied in the following way.*

First we pick a matrix $M \in \mathbb{R}^{d\times b}$ with $\|M\|_{\mathrm{F}} = 1$ and $\bar{\kappa} := \|v_1^\top M\|_2 > 0$, and a scale $\tau \in (0, 1]$. Now we set $A_0 = B_0 = \tau M$ and $Q_{w_1,0}, Q_{w_2,0}, Q_{u_1,0}, Q_{u_2,0}$ by the Rademacher construction described in lemma 20 which we can do because $\|\tau M\|_{\mathrm{op}} \le 1$. Then

$$\|Z_0\|_2 \le \sqrt{2}\,\tau, \qquad \kappa_1 = \left\| v_1^\top \frac{2\tau M}{\sqrt{2}} \right\|_2 = \sqrt{2}\,\tau\,\bar{\kappa}, \qquad \frac{\|Z_0\|_2}{\kappa_1} \le \frac{1}{\bar{\kappa}}.$$

Therefore, to argue that a valid δ exists under the constraints given in theorem 9, it is enough to choose δ and τ such that

$$\frac{1}{\bar{\kappa}}\left[\left(\frac{\sqrt{2}\,\tau}{\delta}\right)^{1-\mu_2/\mu_1} + \frac{\sqrt{2}\,\tau}{\delta} + \delta^2\right] \le c_\star, \qquad \sqrt{2}\,\tau < \delta < c^*.$$

To satisfy the above conditions, we first choose $\delta > 0$ small enough so that $\delta < c^*$ and $\delta^2 \le \frac{\bar{\kappa}c_\star}{3}$. Once δ is fixed, we choose $\tau > 0$ sufficiently small so that $\sqrt{2}\,\tau < \delta$ and $\left(\frac{\sqrt{2}\,\tau}{\delta}\right)^{1-\mu_2/\mu_1} \le \frac{\bar{\kappa}c_\star}{3}$. Since $\sqrt{2}\,\tau/\delta < 1$ and $1 - \mu_2/\mu_1 \in (0, 1)$, this implies

$$\frac{\sqrt{2}\,\tau}{\delta} \le \left(\frac{\sqrt{2}\,\tau}{\delta}\right)^{1-\mu_2/\mu_1} \le \frac{\bar{\kappa}c_\star}{3}.$$

Therefore,

$$\frac{1}{\bar{\kappa}}\left[\left(\frac{\sqrt{2}\,\tau}{\delta}\right)^{1-\mu_2/\mu_1} + \frac{\sqrt{2}\,\tau}{\delta} + \delta^2\right] \le c_\star.$$

Hence a valid choice of δ and τ exists. In particular, we may first fix δ sufficiently small and then choose the initialization scale τ sufficiently small relative to δ.

Since v_1 is unknown, we cannot directly choose M so that $\bar{\kappa} = \|v_1^\top M\|_2 > 0$. Instead, we choose M at random. For example, let $M = OM_0$, where M_0 is fixed and $O \in O(d)$ is Haar-distributed. Then, for every $\eta \in (0, 1)$, we have

$$\mathbb{P}\left(\bar{\kappa} \ge \frac{c\,\eta}{\sqrt{bd}}\right) \ge 1 - \eta.$$

As, $v_1^\top O$ is uniformly distributed on the unit sphere in $\mathbb{R}^d$, so $v_1^\top O M_0$ behaves like the projection of a random direction onto the columns of M_0. Thus, with high probability, $\bar{\kappa}$ is of order $1/\sqrt{bd}$. So, with high probability, we can apply the above construction with $\bar{\kappa} \gtrsim 1/\sqrt{bd}$. Hence, we may take $\bar{\kappa} = \Theta(1/\sqrt{bd})$ in the above argument, choose δ accordingly, and then choose τ sufficiently small relative to δ.

M.1.2 Proof of Theorem 9

In this section, we give the proof of Theorem 9. The main idea of the proof is first to write mean-field PDE for the linear model into an equivalent a closed system of ODEs for (A, B) and the second moments (Proposition 3). And, then this system of odes differs from the linear system $\dot{Z} = \mathcal{M}Z$ only by terms of higher order in Z (Lemma 19). The proof is just then a comparison between the exact flow and the explicit linearized flow.

Proposition 3 (ODEs for the linear model)**.** *Let $\sigma(z) = z$ and $\lambda = 0$. Along intrinsic time s, the particles move by*

$$\begin{aligned} \dot{w}_1 &= -F(A, B)w_2, & \dot{w}_2 &= -F(A, B)^\top w_1, \\ \dot{u}_1 &= -H(A, B)u_2, & \dot{u}_2 &= -H(A, B)^\top u_1, \end{aligned} \tag{112}$$

and the moment matrices evolve by

$$\dot{A} = -Q_{w_1} F(A,B) - F(A,B) Q_{w_2}, \qquad \dot{B} = -Q_{u_1} H(A,B) - H(A,B) Q_{u_2}. \tag{113}$$

The second moments satisfy $\dot{Q}_{w_1} = -FA^\top - AF^\top$, $\dot{Q}_{w_2} = -F^\top A - A^\top F$, and likewise for the decoder with (H, B). In particular $(A, B, Q_{w_1}, Q_{w_2}, Q_{u_1}, Q_{u_2})$ satisfies a closed system of ODEs.

Proof. We give explicit proof for the encoder and then the decoder is identical with (w, A, F) replaced by (u, B, H).

The risk depends on the encoder particles only through $A = \int w_1 w_2^\top \rho_W(dw)$. If a single particle $w = (w_1, w_2)$ is moved by $(\delta w_1, \delta w_2)$, then A changes by $\delta A = \delta w_1 w_2^\top + w_1 \delta w_2^\top$, and by the chain rule we can say that

$$\delta \mathcal{R}_{\text{lin}} = \langle F, \delta A \rangle_F = \langle F, \delta w_1 w_2^\top \rangle_F + \langle F, w_1 \delta w_2^\top \rangle_F = \langle F w_2, \delta w_1 \rangle + \langle F^\top w_1, \delta w_2 \rangle,$$

where we used $\langle F, ab^\top \rangle_F = a^\top F b$. So $\nabla_{w_1} \mathcal{R}_{\text{lin}} = F w_2$ and $\nabla_{w_2} \mathcal{R}_{\text{lin}} = F^\top w_1$. So, this completes the proof for (112) because characteristics are equal to minus of gradients.

Now we want to differentiate A_s with respect to the intrinsic time s. Recall that $\rho_{W,s}$ is the push-forward of $\rho_{W,0}$ along the characteristic flow which means if $w_s = (w_{1,s}, w_{2,s})$ denotes the position of the particle at time s which starts at w_0, then for every test function g,

$$\int g(w)\, \rho_{W,s}(dw) = \int g(w_s)\, \rho_{W,0}(dw_0).$$

Applying this with $g(w) = w_1 w_2^\top$ gives $A_s = \int w_{1,s} w_{2,s}^\top \rho_{W,0}(dw_0)$, where now only the integrand depends on s. Now, differentiating under the integral (this is allowed because the particles remain bounded during the time interval we consider, and the characteristic fields are continuous) and using the product rule, we get

$$\dot{A}_s = \int \big(\dot{w}_{1,s} w_{2,s}^\top + w_{1,s} \dot{w}_{2,s}^\top\big) \rho_{W,0}(dw_0) = \int \big(\dot{w}_1 w_2^\top + w_1 \dot{w}_2^\top\big) \rho_{W,s}(dw),$$

where in the second equality holds by the push-forward identity. Now after substituting the characteristic equations (112) above, we get

$$\begin{aligned} \dot{A}_s &= \int \big(- F w_2 w_2^\top - w_1 w_1^\top F\big) \rho_{W,s}(dw) = -F \int w_2 w_2^\top \rho_{W,s}(dw) - \Big(\int w_1 w_1^\top \rho_{W,s}(dw)\Big) F \\ &= -F Q_{w_2,s} - Q_{w_1,s} F, \end{aligned}$$

where in the last step, $F = F(A_s, B_s)$ depends only on the moments, so it is constant with respect to w and can be taken outside the integral. The same computation with $g(w) = w_1 w_1^\top$ gives

$$\begin{aligned} \dot{Q}_{w_1,s} &= \int \big(\dot{w}_1 w_1^\top + w_1 \dot{w}_1^\top\big) \rho_{W,s}(dw) = -F \int w_2 w_1^\top \rho_{W,s}(dw) - \int w_1 w_2^\top \rho_{W,s}(dw)\, F^\top \\ &= -F A_s^\top - A_s F^\top, \end{aligned}$$

and with $g(w) = w_2 w_2^\top$ it gives $\dot{Q}_{w_2,s} = -F^\top A_s - A_s^\top F$. Since F and H are functions of (A_s, B_s) only, the right-hand sides of all these equations are expressed through $(A_s, B_s, Q_{w_1,s}, Q_{w_2,s}, Q_{u_1,s}, Q_{u_2,s})$, so the system closes. □

The next lemma bounds the difference between the right-hand side of (113) and $\mathcal{M}Z$, and it bounds how far the second moments drift from the identity. The lemma also shows that both errors are of higher order in the size of Z.

Lemma 19. *Let $\gamma_s := \|Q_{w_1,s} - I_d\|_{\text{op}} + \|Q_{u_1,s} - I_d\|_{\text{op}} + \|Q_{w_2,s} - I_b\|_{\text{op}} + \|Q_{u_2,s} - I_b\|_{\text{op}}$ measure how far the second moments have moved away from the isotropic initialization* (108)*. Whenever $\gamma_s \le 1$,*

$$\|\dot{Z}_s - \mathcal{M} Z_s\|_{\text{F}} \le 8\mu_1 \big(\gamma_s \|Z_s\|_2 + \|Z_s\|_2^3\big),$$
$$\gamma_s \le 8\mu_1 \int_0^s \big(\|Z_r\|_2^2 + \|Z_r\|_2^4\big)\, dr.$$

Proof. First, we prove Lemma 19 and the we go for Lemma 19. We use the following facts about matrix norms throughout. For matrices X, Y of compatible sizes, $\|XY\|_{\mathrm{F}} \le \|X\|_{\mathrm{op}}\|Y\|_{\mathrm{F}}$, $\|XY\|_{\mathrm{F}} \le \|X\|_{\mathrm{F}}\|Y\|_{\mathrm{op}}$, and $\|X\|_{\mathrm{op}} \le \|X\|_{\mathrm{F}}$. Note that $\|\Sigma\|_{\mathrm{op}} = \mu_1$. For the moment matrices, $\|A\|_{\mathrm{F}} \le \|Z\|_2$ and $\|B\|_{\mathrm{F}} \le \|Z\|_2$ by the definition $\|Z\|_2^2 = \|A\|_{\mathrm{F}}^2 + \|B\|_{\mathrm{F}}^2$, and $\|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}} \le \frac{1}{2}\|Z\|_2^2$. For the second moments, the assumption $\gamma_s \le 1$ gives $\|Q\|_{\mathrm{op}} \le \|I\|_{\mathrm{op}} + \|Q - I\|_{\mathrm{op}} \le 2$ for each of the four matrices Q. Finally, for the Gram matrix $B^\top B$ we have $\|B^\top B\|_{\mathrm{op}} = \|B\|_{\mathrm{op}}^2 \le \|B\|_{\mathrm{F}}^2$, and likewise $\|A^\top \Sigma A\|_{\mathrm{op}} \le \|\Sigma\|_{\mathrm{op}}\|A\|_{\mathrm{op}}^2 \le \mu_1\|A\|_{\mathrm{F}}^2$.

First, we handle the encoder momnet matrix. Now, by using Proposition 3, we know that $\dot{A} = -Q_{w_1}F - FQ_{w_2}$ where $F = \Sigma AB^\top B - \Sigma B$. Now, by expanding $\dot{A}$, we get

$$\dot{A} = Q_{w_1}\Sigma B + \Sigma B Q_{w_2} - Q_{w_1}\Sigma AB^\top B - \Sigma AB^\top B Q_{w_2}.$$

From Section M.1, the linear part $\mathcal{M}$ gives an approximation of $\dot{A}$ to be $2\Sigma B$. Now, after subtracting $2\Sigma B$ from $\dot{A}$, we get

$$\dot{A} - 2\Sigma B = \underbrace{(Q_{w_1} - I_d)\Sigma B}_{(1)} + \underbrace{\Sigma B(Q_{w_2} - I_b)}_{(2)} \underbrace{- Q_{w_1}\Sigma AB^\top B}_{(3)} \underbrace{- \Sigma AB^\top B Q_{w_2}}_{(4)}.$$

Terms (1) and (2) are there because the second moments are not exactly the identity whereas terms (3) and (4) are the cubic part of the gradient. Now, we bound each in Frobenius norm:

$$\begin{aligned}
&\|(1)\|_{\mathrm{F}} \le \|Q_{w_1} - I_d\|_{\mathrm{op}}\,\|\Sigma\|_{\mathrm{op}}\,\|B\|_{\mathrm{F}} \le \gamma_s\,\mu_1\,\|Z_s\|_2,\\
&\|(2)\|_{\mathrm{F}} \le \|\Sigma\|_{\mathrm{op}}\,\|B\|_{\mathrm{F}}\,\|Q_{w_2} - I_b\|_{\mathrm{op}} \le \mu_1\,\gamma_s\,\|Z_s\|_2,\\
&\|(3)\|_{\mathrm{F}} \le \|Q_{w_1}\|_{\mathrm{op}}\,\|\Sigma\|_{\mathrm{op}}\,\|A\|_{\mathrm{F}}\,\|B^\top B\|_{\mathrm{op}} \le 2\mu_1\|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}}^2 \le 2\mu_1\|Z_s\|_2^3,\\
&\|(4)\|_{\mathrm{F}} \le \|\Sigma\|_{\mathrm{op}}\,\|A\|_{\mathrm{F}}\,\|B^\top B\|_{\mathrm{op}}\,\|Q_{w_2}\|_{\mathrm{op}} \le 2\mu_1\|Z_s\|_2^3.
\end{aligned}$$

Now, after adding up the above bounds we get

$$\|\dot{A} - 2\Sigma B\|_{\mathrm{F}} \le 2\mu_1\gamma_s\|Z_s\|_2 + 4\mu_1\|Z_s\|_2^3.$$

So, this completes the calculation for the encoder moment matrix. Now, we for for the decoder moment matrix. The calculation for the decoder moment matrix is analogous to encoder moment matrix. We give a short exposition of calculation for decoder moment matrix below. Again, by using Proposition 3, we know that $\dot{B} = -Q_{u_1}H - HQ_{u_2}$ where $H = BA^\top\Sigma A - \Sigma A$, and the same manipulation as above gives

$$\dot{B} - 2\Sigma A = (Q_{u_1} - I_d)\Sigma A + \Sigma A(Q_{u_2} - I_b) - Q_{u_1}BA^\top\Sigma A - BA^\top\Sigma A Q_{u_2}.$$

The first two terms are bounded by $\mu_1\gamma_s\|Z_s\|_2$ each, exactly as before. For the cubic terms we use $\|Q_{u_1}BA^\top\Sigma A\|_{\mathrm{F}} \le \|Q_{u_1}\|_{\mathrm{op}}\|BA^\top\Sigma A\|_{\mathrm{F}} \le 2\|B\|_{\mathrm{F}}\|A^\top\Sigma A\|_{\mathrm{op}} \le 2\mu_1\|B\|_{\mathrm{F}}\|A\|_{\mathrm{F}}^2 \le \mu_1\|Z_s\|_2^3$, where we have used the fact that $\|Q\|_{\mathrm{op}} \le 2$, so each cubic term is bounded by $2\mu_1\|Z_s\|_2^3$. Hence

$$\|\dot{B} - 2\Sigma A\|_{\mathrm{F}} \le 2\mu_1\gamma_s\|Z_s\|_2 + 4\mu_1\|Z_s\|_2^3.$$

Since $\dot{Z} - \mathcal{M}Z = (\dot{A} - 2\Sigma B,\ \dot{B} - 2\Sigma A)$, its Frobenius norm is at most the sum of the Frobenius norms of the two blocks. So after combining the above bounds we get

$$\|\dot{Z}_s - \mathcal{M}Z_s\|_{\mathrm{F}} \le 4\mu_1\gamma_s\|Z_s\|_2 + 8\mu_1\|Z_s\|_2^3 \le 8\mu_1\big(\gamma_s\|Z_s\|_2 + \|Z_s\|_2^3\big).$$

So, this completes the proof for Lemma 19.

Now we go for Lemma 19 By, using Proposition 3, we know that

$$\dot{Q}_{w_1} = -FA^\top - AF^\top, \quad \dot{Q}_{w_2} = -F^\top A - A^\top F, \quad \dot{Q}_{u_1} = -HB^\top - BH^\top, \quad \dot{Q}_{u_2} = -H^\top B - B^\top H.$$

We first bound $\|F\|_{\mathrm{F}}$ and $\|H\|_{\mathrm{F}}$. From Equation (106) and the triangle inequality, we can say that

$$\|F\|_{\mathrm{F}} \le \|\Sigma AB^\top B\|_{\mathrm{F}} + \|\Sigma B\|_{\mathrm{F}} \le \mu_1\|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}}^2 + \mu_1\|B\|_{\mathrm{F}},$$

and in the same way, from Equation (107), and the triangle inequality, we can say that $\|H\|_{\mathrm{F}} \le \mu_1\|B\|_{\mathrm{F}}\|A\|_{\mathrm{F}}^2 + \mu_1\|A\|_{\mathrm{F}}$. Then, we can conclude that

$$\|\dot{Q}_{w_1}\|_{\mathrm{op}} \le \|\dot{Q}_{w_1}\|_{\mathrm{F}} \le 2\|F\|_{\mathrm{F}}\|A\|_{\mathrm{F}} \le 2\mu_1\big(\|A\|_{\mathrm{F}}^2\|B\|_{\mathrm{F}}^2 + \|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}}\big) \le 2\mu_1\big(\|Z_s\|_2^4 + \|Z_s\|_2^2\big),$$

where last inequality follows from $\|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}} \le \|Z_s\|_2^2$ and $\|A\|_{\mathrm{F}}^2\|B\|_{\mathrm{F}}^2 \le \|Z_s\|_2^4$. The same bound holds for $\dot{Q}_{w_2}$ and for $\dot{Q}_{u_1}, \dot{Q}_{u_2}$ with H and B in place of F and A. So the sum of the four operator norms of the derivatives is at most $8\mu_1(\|Z_s\|_2^2 + \|Z_s\|_2^4)$.

By Equation (108), each Q starts at the identity, so we can write $Q_{w_1,s} - I_d = \int_0^s \dot{Q}_{w_1,r}\, dr$ (by using fundamental theorem of calculus) and

$$\|Q_{w_1,s} - I_d\|_{\mathrm{op}} \le \int_0^s \|\dot{Q}_{w_1,r}\|_{\mathrm{op}}\, dr,$$

and likewise for the other three. Summing the four inequalities gives

$$\gamma_s \le \int_0^s 8\mu_1\big(\|Z_r\|_2^2 + \|Z_r\|_2^4\big)dr,$$

which proves Lemma 19. Note that the hypothesis $\gamma_s \le 1$ was used only to bound $\|Q\|_{\mathrm{op}} \le 2$ in the cubic terms of the first inequality. □

Now, we give the proof of Theorem 9.

Proof of Theorem 9. We use $C := 8$ throughout the proof. We define $q_s : [0,\infty) \longrightarrow \mathbb{R}$ by $q_s = \|Z_0\|_2\, e^{2\mu_1 s}$ for the size of the linearized flow at time s. We also set $s_\star := \frac{1}{2\mu_1}\log\frac{c_\star}{\|Z_0\|_2}$, the time at which the linearized flow reaches size $c_\star$. So, we can conclude that $q_s \le c_\star$ if and only if $s \le s_\star$, and $s_\delta < s_\star$ because $\delta < c_\star$ (by assumption). Note that $q_{s_\delta} = \delta$ by the definition of s_δ. Note that the block operator

$$\mathcal{M} := \begin{pmatrix} 0 & 2\Sigma \\ 2\Sigma & 0 \end{pmatrix},$$

acting on $Z = \begin{pmatrix} A \\ B \end{pmatrix}$, so that $\mathcal{M}Z = \begin{pmatrix} 2\Sigma B \\ 2\Sigma A \end{pmatrix}$. We define

$$Z_s^{\mathrm{lin}} := \begin{pmatrix} A_s^{\mathrm{lin}} \\ B_s^{\mathrm{lin}} \end{pmatrix} = e^{\mathcal{M}s} Z_0$$

as the solution of

$$\dot{Z}_s^{\mathrm{lin}} = \mathcal{M} Z_s^{\mathrm{lin}}, \qquad Z_0^{\mathrm{lin}} = Z_0.$$

So, we can say that

$$\dot{A}_s^{\mathrm{lin}} = 2\Sigma B_s^{\mathrm{lin}}, \qquad \dot{B}_s^{\mathrm{lin}} = 2\Sigma A_s^{\mathrm{lin}}.$$

We also define

$$Z_{+,0} := \frac{A_0 + B_0}{\sqrt{2}}, \qquad Z_{-,0} := \frac{A_0 - B_0}{\sqrt{2}},$$

for the sum and difference of the initial moments. With this notation, $\kappa_1 = \|v_1^\top Z_{+,0}\|_2$. Note that $P_1 := v_1 v_1^\top$ is the orthogonal projection onto the top eigenvector, so that $\|(I_d - P_1)M\|_{\mathrm{F}} = \sin\angle(M, v_1)\,\|M\|_{\mathrm{F}}$ for any nonzero M.

First we give the outline for this proof. We divide the proof into four steps. First step solves the linear system explicitly. The second step shows that the exact moments stay close to the linear solution as long as q_s is small. The third step measures the size of each part of A_{s_δ}. And, the last step divides the part orthogonal to v_1 by the total to get the angle bound.

First, we try to solve linear systems explicitely. Since $\mathcal{M}(A,B) = (2\Sigma B, 2\Sigma A)$ swaps the two blocks, it is convenient to look at the sum and the difference: $\frac{d}{ds}(A+B) = 2\Sigma(A+B)$ and $\frac{d}{ds}(A-B) = -2\Sigma(A-B)$. Solving these two decoupled equations gives

$$A_s^{\mathrm{lin}} = \tfrac{1}{\sqrt{2}}\big(e^{2\Sigma s}Z_{+,0} + e^{-2\Sigma s}Z_{-,0}\big), \qquad B_s^{\mathrm{lin}} = \tfrac{1}{\sqrt{2}}\big(e^{2\Sigma s}Z_{+,0} - e^{-2\Sigma s}Z_{-,0}\big). \tag{114}$$

So the sum grows and the difference decays. In the eigenbasis of Σ, the component of $Z_{+,0}$ along v_i is multiplied by $e^{2\mu_i s}$, so the component along v_1 grows fastest. As, $\|e^{\mathcal{M}s}\|_{\mathrm{op}} = e^{2\mu_1 s}$, so we can conclude that $\|Z_s^{\mathrm{lin}}\|_2 \le q_s$.

Now, we try to show that the exact flow stays close to the linear one. In particular, we show that there are constants $c_\star, C_\star > 0$, depending only on μ_1, such that

$$\|Z_s - Z_s^{\mathrm{lin}}\|_{\mathrm{F}} \le C_\star q_s^3 \qquad \text{and} \qquad \gamma_s \le C_\star q_s^2 \qquad \text{whenever } q_s \le c_\star. \tag{115}$$

Let S be the largest time in $[0, s_\star]$ such that

$$\|Z_r\|_2 \le 2q_r \quad \text{and} \quad \gamma_r \le 1 \qquad \text{for all } r \le S.$$

Note that the both of the above conditions hold at $r = 0$ ($\|Z_0\|_2 = q_0$, $\gamma_0 = 0$), so this implies $S > 0$. As, $S \le s_\star$ and q is increasing, so we can conclude that for every $r \le s \le S$ we have $q_r \le q_s \le q_{s_\star} = c_\star < 1$, . This is the only property of $s_\star$ that we use.

Since on the time interval $[0, S]$, $\gamma_s \le 1$ for all $s \le S$, we can apply Lemma 19 (particularly Lemma 19) to conclude that

$$\gamma_s \le C\mu_1 \int_0^s \big(\|Z_r\|_2^2 + \|Z_r\|_2^4\big) dr \le C\mu_1 \int_0^s \big(4q_r^2 + 16q_r^4\big) dr \le 20C\mu_1 \int_0^s q_r^2\, dr \le 5C\, q_s^2,$$

where the second inequality follows from the fact that, on the time interval $[0, S]$, $\|Z_r\|_2 \le 2q_r$. The third inequality follows from $q_r \le 1$. For the last inequality, we use $\int_0^s q_r^2\, dr = \|Z_0\|_2^2 \int_0^s e^{4\mu_1 r}\, dr \le \frac{q_s^2}{4\mu_1}$.

Now we write the exact equation as $\dot{Z} = \mathcal{M}Z + (\dot{Z} - \mathcal{M}Z)$ and then use Duhamel's formula, which is given by

$$Z_s - Z_s^{\mathrm{lin}} = \int_0^s e^{\mathcal{M}(s-r)} \big(\dot{Z}_r - \mathcal{M}Z_r\big) dr.$$

Note that the the integrand has norm at most $e^{2\mu_1(s-r)} \cdot \big\|\dot{Z}_r - \mathcal{M}Z_r\big\|_F$. Since on the time interval $[0, S]$, $\gamma_s \le 1$ for all $s \le S$, we can apply Lemma 19 (particularly, Lemma 19) to conclude that $\big\|\dot{Z}_r - \mathcal{M}Z_r\big\|_F \le C\mu_1(\gamma_r \|Z_r\|_2 + \|Z_r\|_2^3)$. Now, with $\gamma_r \le 5Cq_r^2$ and $\|Z_r\|_2 \le 2q_r$, the integrand has norm at most $e^{2\mu_1(s-r)} \cdot C\mu_1(10C+8)q_r^3$, so

$$\|Z_s - Z_s^{\mathrm{lin}}\|_{\mathrm{F}} \le C\mu_1(10C+8)\, \|Z_0\|_2^3\, e^{2\mu_1 s} \int_0^s e^{6\mu_1 r} dr \le \frac{C(10C+8)}{6}\, q_s^3.$$

This proves (115) on $[0, S]$ with $C_\star := \max\{5C,\, C(10C+8)/6\}$.

Now we have to check that $S = s_\star$. At this point we fix the constant $c_\star$. Before this, $c_\star$ is a free constant, and we only use the property of $c_\star$ that $c_\star \le 1$. Recall that $C_\star = \max\{5C,\, C(10C+8)/6\} = 352/3$, so in particular $C_\star > 1$. We choose

$$c_\star := \frac{1}{2\sqrt{2}\, C_\star}.$$

Two properties of this choice are used below. First property is, $c_\star < 1$, because $C_\star > 1$ and second property is, $C_\star c_\star^2 = \frac{C_\star}{8C_\star^2} = \frac{1}{8C_\star} \le \frac{1}{2}$. Now, on the time interval $[0, S]$ we have already shown that $\|Z_s - Z_s^{\mathrm{lin}}\|_{\mathrm{F}} \le C_\star q_s^3$ and $\gamma_s \le C_\star q_s^2$. Using $\|Z_s^{\mathrm{lin}}\|_2 \le q_s$ from the first part of the proof, the triangle inequality gives

$$\|Z_s\|_2 \le \|Z_s^{\mathrm{lin}}\|_2 + \|Z_s - Z_s^{\mathrm{lin}}\|_{\mathrm{F}} \le q_s + C_\star q_s^3 = q_s\big(1 + C_\star q_s^2\big) \le q_s\big(1 + C_\star c_\star^2\big) \le \tfrac{3}{2}\, q_s,$$

where we use $q_s \le c_\star$ on $[0, S]$ and then $C_\star c_\star^2 \le \frac{1}{2}$. In the same way,

$$\gamma_s \le C_\star q_s^2 \le C_\star c_\star^2 \le \tfrac{1}{2}.$$

So at the time S we have $\|Z_S\|_2 \le \frac{3}{2} q_S < 2q_S$ and $\gamma_S \le \frac{1}{2} < 1$. So, this implies that the both conditions defining S hold with strict inequality and since $s \mapsto Z_s$ and $s \mapsto \gamma_s$ are continuous, both conditions then continue to hold on a small interval $[S, S+\eta]$ for some $\eta > 0$. If $S < s_\star$, this contradicts the fact that S is the largest time in $[0, s_\star]$ with these two properties. Hence $S = s_\star$, and we can conclude that (115) holds on the whole interval $[0, s_\star]$, that is, for every s with $q_s \le c_\star$. In particular it holds at $s = s_\delta$, because $s_\delta < s_\star$.

Now we measure the size of each part of A_{s_δ}. Recall that $q_{s_\delta} = \delta$, so $e^{2\mu_1 s_\delta} = \delta/\|Z_0\|_2$. We write A_{s_δ} as the linear solution plus the error, and then split the linear solution part using (114) and $I_d = P_1 + (I_d - P_1)$. So, after writing A_{s_δ} as above, we get

$$A_{s_\delta} = \underbrace{\tfrac{1}{\sqrt{2}} P_1 e^{2\Sigma s_\delta} Z_{+,0}}_{\text{(i)}} + \underbrace{\tfrac{1}{\sqrt{2}} (I_d - P_1) e^{2\Sigma s_\delta} Z_{+,0}}_{\text{(ii)}} + \underbrace{\tfrac{1}{\sqrt{2}} e^{-2\Sigma s_\delta} Z_{-,0}}_{\text{(iii)}} + \underbrace{A_{s_\delta} - A^{\text{lin}}_{s_\delta}}_{\text{(iv)}}.$$

We now bound each of the four parts.

The first part (i) is the part of the growing mode along v_1. Since v_1 is an eigenvector of Σ with eigenvalue μ_1, we have $P_1 e^{2\Sigma s} = v_1 v_1^\top e^{2\Sigma s} = e^{2\mu_1 s} v_1 v_1^\top$, so

$$\text{(i)} = \tfrac{1}{\sqrt{2}} e^{2\mu_1 s_\delta}\, v_1\, (v_1^\top Z_{+,0}), \qquad \|\text{(i)}\|_{\text{F}} = \tfrac{1}{\sqrt{2}} e^{2\mu_1 s_\delta} \|v_1^\top Z_{+,0}\|_2 = \tfrac{1}{\sqrt{2}} \frac{\kappa_1}{\|Z_0\|_2}\, \delta,$$

where we use $\kappa_1 = \|v_1^\top Z_{+,0}\|_2$ and $e^{2\mu_1 s_\delta} = \delta/\|Z_0\|_2$. Note that every column of (i) is a multiple of v_1.

The second part (ii) collects the growing modes along $v_2, \dots, v_d$. On the orthogonal complement of v_1 the matrix $e^{2\Sigma s}$ has eigenvalues $e^{2\mu_2 s}, \dots, e^{2\mu_d s}$, the largest of which is $e^{2\mu_2 s}$. So $\|(I_d - P_1) e^{2\Sigma s_\delta}\|_{\text{op}} = e^{2\mu_2 s_\delta}$, and since $\|Z_{+,0}\|_{\text{F}} \le \|Z_0\|_2$,

$$\|\text{(ii)}\|_{\text{F}} \le \tfrac{1}{\sqrt{2}} e^{2\mu_2 s_\delta} \|Z_0\|_2.$$

Now $e^{2\mu_2 s_\delta} = \big(e^{2\mu_1 s_\delta}\big)^{\mu_2/\mu_1} = (\delta/\|Z_0\|_2)^{\mu_2/\mu_1}$, so

$$\|\text{(ii)}\|_{\text{F}} \le \tfrac{1}{\sqrt{2}}\, \|Z_0\|_2 \Big(\frac{\delta}{\|Z_0\|_2}\Big)^{\mu_2/\mu_1} = \tfrac{1}{\sqrt{2}}\, \delta \Big(\frac{\|Z_0\|_2}{\delta}\Big)^{1-\mu_2/\mu_1}.$$

The third part (iii) collects the decaying modes. Since $\Sigma \succeq 0$, all eigenvalues of $e^{-2\Sigma s}$ lie in $(0, 1]$, so $\|e^{-2\Sigma s_\delta}\|_{\text{op}} \le 1$, and $\|Z_{-,0}\|_{\text{F}} \le \|Z_0\|_2$. Hence

$$\|\text{(iii)}\|_{\text{F}} \le \tfrac{1}{\sqrt{2}} \|Z_0\|_2.$$

The last part (iv) is the nonlinear error. Now, from Equation (115) we can conclude that at $s = s_\delta$,

$$\|\text{(iv)}\|_{\text{F}} \le C_\star q_{s_\delta}^3 = C_\star \delta^3.$$

Note that only the first part (i) lies along v_1 while the other three parts may point in any direction. Let's define

$$E := \delta \Big(\frac{\|Z_0\|_2}{\delta}\Big)^{1-\mu_2/\mu_1} + \|Z_0\|_2 + \sqrt{2}\, C_\star \delta^3,$$

so that $\|\text{(ii)}\|_{\text{F}} + \|\text{(iii)}\|_{\text{F}} + \|\text{(iv)}\|_{\text{F}} \le \frac{1}{\sqrt{2}} E$. Since $(I_d - P_1)$ maps (i) to zero and $\|I_d - P_1\|_{\text{op}} = 1$, we can conclude that

$$\|(I_d - P_1) A_{s_\delta}\|_{\text{F}} \le \|\text{(ii)}\|_{\text{F}} + \|\text{(iii)}\|_{\text{F}} + \|\text{(iv)}\|_{\text{F}} \le \tfrac{1}{\sqrt{2}} E.$$

On the other hand, by the triangle inequality in the form $\|a + b\| \ge \|a\| - \|b\|$ with $a = \text{(i)}$ and $b = \text{(ii)} + \text{(iii)} + \text{(iv)}$, we can conlcude that

$$\|A_{s_\delta}\|_{\text{F}} \ge \|\text{(i)}\|_{\text{F}} - \tfrac{1}{\sqrt{2}} E = \tfrac{1}{\sqrt{2}} \Big(\frac{\kappa_1}{\|Z_0\|_2}\, \delta - E\Big).$$

Finally, we prove the angle bound. Dividing the two inequalities just obtained, and recalling that $\sin \angle(A_{s_\delta}, v_1) = \|(I_d - P_1) A_{s_\delta}\|_{\text{F}} / \|A_{s_\delta}\|_{\text{F}}$, we get

$$\sin \angle(A_{s_\delta}, v_1) \le \frac{E}{\frac{\kappa_1}{\|Z_0\|_2}\delta - E} = \frac{\rho}{1 - \rho}, \qquad \text{where} \qquad \rho := \frac{\|Z_0\|_2}{\kappa_1} \cdot \frac{E}{\delta},$$

provided $\rho < 1$, which we check now. By dividing the E by δ, we get

$$\frac{E}{\delta} = \Big(\frac{\|Z_0\|_2}{\delta}\Big)^{1-\mu_2/\mu_1} + \frac{\|Z_0\|_2}{\delta} + \sqrt{2}\, C_\star \delta^2 \le \sqrt{2}\, C_\star \Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^{1-\mu_2/\mu_1} + \frac{\|Z_0\|_2}{\delta} + \delta^2\Big],$$

where we use $\sqrt{2}C_\star \geq 1$. Now, from the hypothesis (110) of the theorem, we know that $\frac{\|Z_0\|_2}{\kappa_1}\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^{1-\mu_2/\mu_1} + \frac{\|Z_0\|_2}{\delta} + \delta^2\Big] \leq c_\star$, so by using this we can conclude that

$$\rho \leq \sqrt{2}\, C_\star \cdot \frac{\|Z_0\|_2}{\kappa_1}\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^{1-\mu_2/\mu_1} + \frac{\|Z_0\|_2}{\delta} + \delta^2\Big] \leq \sqrt{2}\, C_\star\, c_\star = \frac{1}{2},$$

where we use $c_\star = 1/(2\sqrt{2}C_\star)$. So $1-\rho \geq \frac{1}{2}$. Since $1-\rho \geq \frac{1}{2}$, we can conclude that $\|A_{s_\delta}\|_{\mathrm{F}} \geq \frac{1}{\sqrt{2}}\frac{\kappa_1}{\|Z_0\|_2}\delta\,(1-\rho) > 0$, so A_{s_δ} is nonzero and

$$\sin\angle(A_{s_\delta}, v_1) \leq \frac{\rho}{1-\rho} \leq 2\rho \leq 2\sqrt{2}\, C_\star\, \frac{\|Z_0\|_2}{\kappa_1}\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^{1-\mu_2/\mu_1} + \frac{\|Z_0\|_2}{\delta} + \delta^2\Big].$$

So, this proves the Equation (111) for A_{s_δ}, with the constant in the theorem equal to $2\sqrt{2}C_\star$.

For B_{s_δ} we repeat the same argument. From (114), we can say that the only difference between B_s^{lin} and A_s^{lin} is the sign in front of $e^{-2\Sigma s}Z_{-,0}$, so the decomposition of B_{s_δ} has the same four parts with (iii) replaced by $-$(iii). All four bounds only involve the norms of the pieces, and $\|-(\mathrm{iii})\|_{\mathrm{F}} = \|(\mathrm{iii})\|_{\mathrm{F}}$, so same bound holds for B_{s_δ}. This finishes the proof. □

Lemma 20 (Existence of isotropic seeds with prescribed cross moments)**.** *Let $A_0 \in \mathbb{R}^{d\times b}$ with $\|A_0\|_{\mathrm{op}} \leq 1$. Then there is a compactly supported probability measure on $\mathbb{R}^d \times \mathbb{R}^b$ whose second moments are*

$$\int w_1 w_1^\top = I_d, \qquad \int w_2 w_2^\top = I_b, \qquad \int w_1 w_2^\top = A_0.$$

The same holds for the decoder with A_0 replaced by any B_0 with $\|B_0\|_{\mathrm{op}} \leq 1$.

Proof. Write $v = (w_1, w_2) \in \mathbb{R}^{d+b}$ and consider the block matrix

$$G := \begin{pmatrix} I_d & A_0 \\ A_0^\top & I_b \end{pmatrix} \in \mathbb{R}^{(d+b)\times(d+b)}.$$

The three requirements of the lemma say exactly that $\mathbb{E}[vv^\top] = G$. Now, we know that a matrix is a second-moment matrix of some probability measure if and only if it is symmetric positive semidefinite, so we first check that $G \succeq 0$. As, $I_d \succ 0$, so by using the Schur complement criterion, we can conclude that

$$G \succeq 0 \quad\Longleftrightarrow\quad I_b - A_0^\top I_d^{-1} A_0 = I_b - A_0^\top A_0 \succeq 0 \quad\Longleftrightarrow\quad \|A_0\|_{\mathrm{op}} \leq 1,$$

which is the hypothesis.

Since $G \succeq 0$, it has a spectral decomposition $G = V\Lambda V^\top$ with V orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{d+b})$, $\lambda_k \geq 0$. Put $L := V\Lambda^{1/2}$, so that $LL^\top = G$. Let $\xi = (\xi_1, \ldots, \xi_{d+b})$ be a vector of independent Rademacher signs, $\mathbb{P}(\xi_k = \pm 1) = \frac{1}{2}$, and define

$$v := L\xi, \qquad \text{i.e.} \qquad v = \sum_{k=1}^{d+b} \xi_k \sqrt{\lambda_k}\, V_{\cdot k},$$

where $V_{\cdot k}$ is the k-th eigenvector of G. Then

$$\mathbb{E}[vv^\top] = L\,\mathbb{E}[\xi\xi^\top]\,L^\top = L\, I_{d+b}\, L^\top = G,$$

because the Rademacher variables are independent with $\mathbb{E}[\xi_k] = 0$ and $\mathbb{E}[\xi_k^2] = 1$. The law of v is supported on the 2^{d+b} points $\{L\xi : \xi \in \{\pm 1\}^{d+b}\}$, all of norm at most $\|L\|_{\mathrm{op}}\sqrt{d+b} = \|G\|_{\mathrm{op}}^{1/2}\sqrt{d+b} \leq \sqrt{2(d+b)}$, so it is compactly supported. Splitting v into its first d and last b coordinates gives (w_1, w_2) and this satisfies all three moment identities.

For the decoder, we apply the same construction with G replaced by $\begin{pmatrix} I_d & B_0 \\ B_0^\top & I_b \end{pmatrix}$ and $v = (u_1, u_2)$. □

M.2 THE BOUNDED NONLINEAR ACTIVATION

We now return to the bounded activation function with $\lambda = 0$. The goal of this subsection is to show the same PCA-first kind of behavior for non-linear activation (Theorem 10). Note that the network is nonlinear, so there is no exact moment system that defines PCA-first kind of behavior as it was there for linear case (Section M.1). What we show instead is that, as long as the encoder and decoder outputs are small, two moment matrices satisfy a linear system of the same form as in the linear model, up to an error of second order in the output size. The catch is that now covariance Σ is replaced by a covariance $\mathcal{C}$ in which pairs of data points are weighted by the kernel of the first-layer features at initialization. Next, we show in Section M.2.3 that when the features are in the linear range of σ, the top eigenvector of $\mathcal{C}$ is the top eigenvector of Σ, which brings us back to ordinary PCA.

We define

$$S_k := \|\sigma^{(k)}\|_\infty, \qquad k = 0, 1, 2, 3,$$

for the bounds on σ and its first three derivatives. We also assume that σ is smooth and bounded with $\sigma(0) = 0$ and $\sigma'(0) = 1$. One of possible example of σ is $\tanh$.

For non-linear case, the matrices we track are

$$C_s := \mathbb{E}_x\big[x\,\alpha_s(x)^\top\big] \in \mathbb{R}^{d\times b}, \qquad B_s := D\Psi_s(0) = \int u_1 u_2^\top \rho_{U,s}(du) \in \mathbb{R}^{d\times b}.$$

The first is the correlation between the data and the bottleneck. The second is the Jacobian of the decoder at the origin because $D\Psi_s(a) = \int \sigma'(\langle u_2, a\rangle) u_1 u_2^\top \rho_{U,s}(du)$ and $\sigma'(0) = 1$. For a linear activation $C_s = \Sigma A_s$ and B_s is the same matrix in (105), so these are the natural replacements of (A_s, B_s). As in the linear case we write $Z_s := (C_s, B_s)$ with $\|Z\|_2^2 := \|C\|_{\mathrm{F}}^2 + \|B\|_{\mathrm{F}}^2$, and $\|Z_0\|_2$ is the size of the initialization. Note that $\|C_s\|_{\mathrm{F}} \le M_x \|\alpha_s\|_{L^\infty}$, because $\|\mathbb{E}_x[x\alpha_s(x)^\top]\|_{\mathrm{F}} \le \mathbb{E}_x[\|x\|_2 \|\alpha_s(x)\|_2]$.

We use an initialization in which the first-layer directions w_1 and u_2 are of order one, while the output coefficients w_2 and u_1 are small. The following assumption makes this precise.

Assumption 2 (Small-output initialization)**.** *The initial laws $\rho_{W,0}, \rho_{U,0}$ are compactly supported and satisfy the following two conditions.*

(i) The first-layer coordinates are bounded: there is a constant $R_1 \ge 1$ such that $\|w_1\|_2 \le R_1$ for $\rho_{W,0}$-almost every w and $\|u_2\|_2 \le R_1$ for $\rho_{U,0}$-almost every u.

(ii) The output coefficients are of the order of the initial moments: $\|Z_0\|_2 > 0$, and there is a constant $\kappa_0 \ge 1$ such that

$$\operatorname*{ess\,sup}_{\rho_{W,0}} \|w_2\|_2 + \operatorname*{ess\,sup}_{\rho_{U,0}} \|u_1\|_2 \le \kappa_0 \|Z_0\|_2,$$

where for a probability measure ρ on $\mathbb{R}^d \times \mathbb{R}^b$ and a map $w = (w_1, w_2) \mapsto \|w_2\|_2$, the essential supremum ($\operatorname{ess\,sup}$*) is*

$$\operatorname*{ess\,sup}_{\rho} \|w_2\|_2 := \inf\Big\{M \ge 0 :\ \rho\big(\{w : \|w_2\|_2 > M\}\big) = 0\Big\},$$

that is, the smallest number M such that $\|w_2\|_2 \le M$ for ρ-almost every w. The $\operatorname{ess\,sup}_{\rho_{U,0}} \|u_1\|_2$ is defined in the same way with the map $u = (u_1, u_2) \mapsto \|u_1\|_2$.

Two objects stay frozen to leading order during the small-output phase, because they depend only on the first-layer marginals, which we will show barely move. They are the feature kernels at initialization,

$$k_W(x, x') := \int \sigma(\langle w_1, x\rangle)\sigma(\langle w_1, x'\rangle)\, \rho_{W,0}(dw), \tag{116}$$

$$k_U(a, a') := \int \sigma(\langle u_2, a\rangle)\sigma(\langle u_2, a'\rangle)\, \rho_{U,0}(du).$$

We define the kernel covariance as follows

$$\mathcal{C} := \int h(w_1)h(w_1)^\top \rho_{W,0}(dw) = \mathbb{E}_{x,x'}\big[k_W(x, x')\, x x'^\top\big] \succeq 0., \tag{117}$$

where

$$h(w_1) := \mathbb{E}_x\big[\sigma(\langle w_1, x\rangle)x\big] \in \mathbb{R}^d.$$

The second expression for $\mathcal{C}$ follows from the first by writing $h(w_1)h(w_1)^\top = \mathbb{E}_{x,x'}[\sigma(\langle w_1, x\rangle)\sigma(\langle w_1, x'\rangle)xx'^\top]$ and exchanging the integral over w with the expectation. So $\mathcal{C}$ is a covariance of the data in which each pair (x, x') is weighted by $k_W(x, x')$. For a linear activation $h(w_1) = \Sigma w_1$ and $\mathcal{C} = \Sigma Q_{w_1,0}\Sigma$, so $\mathcal{C}$ plays the role of Σ^2 because $Q_{w_1,0} = \mathbb{I}_d$. We write $\mathcal{C}\bar{v}_i = \nu_i\bar{v}_i$ with $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_d \geq 0$ and orthonormal $\bar{v}_1, \ldots, \bar{v}_d$. Again, we define

$$m(a) := \int \sigma(\langle u_2, a\rangle)u_2\,\rho_{U,0}(du) \in \mathbb{R}^b, \qquad Dm(0) = \int \sigma'(0)\,u_2u_2^\top\rho_{U,0}(du) = Q_{u_2,0}, \quad (118)$$

where the last identity again uses $\sigma'(0) = 1$. We also assume that the decoder directions are isotropic in the latent space,

$$Q_{u_2,0} = qI_b, \qquad q > 0, \quad (119)$$

and set

$$\rho_i := \sqrt{q\nu_i}\ (i = 1, \ldots, d), \qquad \theta := 1 - \frac{\rho_2}{\rho_1} = 1 - \sqrt{\frac{\nu_2}{\nu_1}}, \qquad \gamma := \sqrt{\frac{\nu_1}{q}}.$$

The number ρ_1 will be the growth rate of the top mode, θ the exponent in the alignment bound, and γ the factor that converts the decoder moment B_0 into the same units as C_0.

For the ease of calculation, we need a single number that measures how small the outputs are. Let $\mathcal{A} := \{a \in \mathbb{R}^b : \|a\|_2 \leq 1\}$ be the closed unit ball, and for a C^2 map $f : \mathcal{A} \to \mathbb{R}^d$ write $\|f\|_{C^2(\mathcal{A})} := \sup_{\mathcal{A}}\|f\|_2 + \sup_{\mathcal{A}}\|Df\|_{\mathrm{op}} + \sup_{\mathcal{A}}\|D^2f\|_{\mathrm{op}}$, and similarly for C^0, C^1. Set

$$\mathfrak{a}_s := \|\alpha_s\|_{L^\infty(\operatorname{supp}P)} + \|\Psi_s\|_{C^2(\mathcal{A})} + \operatorname*{ess\,sup}_{\rho_{W,s}}\|w_2\|_2 + \operatorname*{ess\,sup}_{\rho_{U,s}}\|u_1\|_2.$$

This controls the encoder output, the decoder map with two derivatives, and the output coefficients of every particle. Note that $\|B_s\|_{\mathrm{op}} = \|D\Psi_s(0)\|_{\mathrm{op}} \leq \mathfrak{a}_s$ and $\|C_s\|_{\mathrm{F}} \leq M_x\mathfrak{a}_s$, so $\|Z_s\|_2 \leq (1 + M_x)\mathfrak{a}_s$. At time zero we can bound $\mathfrak{a}_0$ by the initial moments. Indeed, $\|\alpha_0(x)\|_2 \leq S_0\operatorname{ess\,sup}\|w_2\|_2$, and since $\Psi_0(a) = \int u_1\sigma(\langle u_2, a\rangle)\rho_{U,0}(du)$ with $\|u_2\|_2 \leq R_1$, differentiating under the integral gives $\sup_{\mathcal{A}}\|\Psi_0\|_2 \leq S_0\operatorname{ess\,sup}\|u_1\|_2$, $\sup_{\mathcal{A}}\|D\Psi_0\|_{\mathrm{op}} \leq S_1R_1\operatorname{ess\,sup}\|u_1\|_2$ and $\sup_{\mathcal{A}}\|D^2\Psi_0\|_{\mathrm{op}} \leq S_2R_1^2\operatorname{ess\,sup}\|u_1\|_2$. Adding these and using condition (ii) of Assumption 2, we get

$$\mathfrak{a}_0 \leq C_0\|Z_0\|_2, \qquad C_0 := \kappa_0\big(1 + S_0 + S_1R_1 + S_2R_1^2\big). \quad (120)$$

As in the linear case we define the exit time through the leading linear system so we do it here also. Since the top mode grows at rate ρ_1, it reaches size δ at

$$s_\delta := \frac{1}{\rho_1}\log\frac{\delta}{\|Z_0\|_2}, \qquad \text{so that} \qquad \|Z_0\|_2\,e^{\rho_1 s} \leq \delta \ \text{ for all } s \leq s_\delta, \quad \text{with equality at } s = s_\delta. \quad (121)$$

Finally we need the analogue of κ_1 in (109). Set

$$\kappa_1 := \frac{1}{2}\Big\|\bar{v}_1^\top C_0 + \gamma\,\bar{v}_1^\top B_0\Big\|_2, \qquad \gamma = \sqrt{\nu_1/q}. \quad (122)$$

Later in this subsection, we will see that the growing mode of the leading linear system along $\bar{v}_1$ has coefficient exactly κ_1, in the same way as $\|v_1^\top(A_0 + B_0)/\sqrt{2}\|_2$ is the coefficient of the growing mode along v_1 in the linear case. By the triangle inequality, $\kappa_1 \leq \frac{1}{2}(1 + \gamma)\|Z_0\|_2$, so the ratio $\|Z_0\|_2/\kappa_1$ is always at least $2/(1 + \gamma)$, so it is small when the initialization lies mostly along the growing top mode.

M.2.1 Main result

Now we describe the result for the bounded activation. The result says the following. If one starts the nonlinear model from a small-output initialization of size $\|Z_0\|_2$, then the correlation matrix C_s between the data and the bottleneck, and the decoder Jacobian B_s at the origin, both grow exponentially at rate ρ_1, and by the time the leading mode reaches a macroscopic size δ they are aligned with the top eigenvector $\bar{v}_1$ of $\mathcal{C}$. The bound has the same three parts as in Theorem 9:

a term $(\|Z_0\|_2/\delta)^\theta$ from the second eigenvector of $\mathcal{C}$, which grows too but more slowly; a term $\|Z_0\|_2/\delta$ from the decaying modes; and a term δ from the nonlinear corrections that become visible at size δ. The exponent is $\theta = 1 - \sqrt{\nu_2/\nu_1}$ rather than $1 - \mu_2/\mu_1$ because the growth rates are square roots of the eigenvalues. Then, in Section M.2.3, we show that in the linear-feature regime $\mathcal{C} \propto \Sigma^2$, $\theta = 1 - \mu_2/\mu_1$. The last error term is δ rather than δ^2 (linear case) because the error is only accurate to second order, whereas the linear model is exact to third order.

Theorem 10 (Early kernel-PCA alignment)**.** *Let $\lambda = 0$. Assume Assumption 2, latent isotropy* (119)*, and a strict gap $\nu_1 > \nu_2 > 0$ between the two largest eigenvalues of $\mathcal{C}$. Assume that the initialization has some overlap with the growing top mode, that is, $\kappa_1 > 0$ in* (122)*. Then there are constants $c_\star, C_\star > 0$, which do not depend on $\|Z_0\|_2$, such that the following holds. Let δ be such that $\|Z_0\|_2 < \delta < c_\star$ and*

$$\frac{\|Z_0\|_2}{\kappa_1}\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^\theta + \frac{\|Z_0\|_2}{\delta} + \delta\Big] \le c_\star. \tag{123}$$

Then, at the time s_δ defined in (121)*, the matrices C_{s_δ} and B_{s_δ} are nonzero and*

$$\sin\angle(C_{s_\delta}, \bar v_1) \ \vee\ \sin\angle(B_{s_\delta}, \bar v_1) \le C_\star\, \frac{\|Z_0\|_2}{\kappa_1}\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^\theta + \frac{\|Z_0\|_2}{\delta} + \delta\Big]. \tag{124}$$

Moreover, at time s_δ the decoder map takes values along the line $\mathbb{R}\bar v_1$: there is a vector $z_{\Psi,\delta} \in \mathbb{R}^b$ with $\|z_{\Psi,\delta}\|_2 \le C_\star\delta$ such that

$$\big\|\Psi_{s_\delta} - \bar v_1\, z_{\Psi,\delta}^\top m(\cdot)\big\|_{C^1(\mathcal{A})} \le C_\star\,\delta\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^\theta + \frac{\|Z_0\|_2}{\delta} + \delta\Big]. \tag{125}$$

The condition (123) has the same meaning as (110): the initialization must lie mostly along the growing top mode, and it must be small compared with the exit scale δ, so that the leading mode has sufficient time to separate from the second mode before the trajectory reaches the exit scale. The last statement says that at the first exit from the small-initialization regime the decoder image is approximately a line in the direction $\bar v_1$, even when $b > 1$.

Remark 5. *The theorem has four requirements on the initialization: Assumption 2, latent isotropy* (119)*, the gap $\nu_1 > \nu_2 > 0$, and $\kappa_1 > 0$ together with* (123)*. All four can be satisfied simultaneously in the following way.*

First we choose the first layer. Take $w_1 = \eta z$ with z compactly supported and $\mathbb{E}[zz^\top] = \kappa I_d$, and take u_2 compactly supported with $\mathbb{E}[u_2 u_2^\top] = qI_b$ (for instance we can sample z and u_2 independently uniform on a sphere of the appropriate radius). This gives condition (i) of Assumption 2 and (119)*. For the eigen-gap part, first we show in Section M.2.3 that $\mathcal{C} = \kappa\eta^2\Sigma^2 + E_\eta$ with $\|E_\eta\|_{\mathrm{op}} \le C\eta^4$. Then, by Weyl's inequality the we can conclude that ith eigenvalues of $\mathcal{C}$ lie $[\kappa\eta^2\mu_i^2 - C\eta^4, \kappa\eta^2\mu_i^2 + C\eta^4]$ for all $i = 1, 2, \dots, d$, where μ_i is the ith eigenvalue of Σ matrix. So if $\mu_1 > \mu_2 > 0$ and η is small enough such that $2C\eta^4 < \kappa\eta^2 \min\{\mu_1^2 - \mu_2^2, \mu_2^2\}$, then $\nu_1 > \nu_2 > 0$. So the gap condition on $\mathcal{C}$ is inherited from the gap condition on Σ once the η is small.*

Now we choose the output coefficients. Take $u_1 = t\,Mu_2$ and $w_2 = t\,N^\top w_1$ for fixed matrices $M, N \in \mathbb{R}^{d\times b}$ and a scale $t > 0$. Then

$$B_0 = t\int Mu_2u_2^\top \rho_{U,0}(du) = t\,q\,M, \qquad C_0 = t\,\mathbb{E}_x\Big[x\int \sigma(\langle w_1, x\rangle) w_1^\top \rho_{W,0}(dw)\Big]N = t\,\Gamma N,$$

where $\Gamma := \int h(w_1)w_1^\top\rho_{W,0}(dw)$ depends only on the encoder. So $\|Z_0\|_2 = t\,\|(\Gamma N, qM)\|_2$ and $\kappa_1 = \frac{t}{2}\|\bar v_1^\top\Gamma N + \gamma q\,\bar v_1^\top M\|_2$ both scale linearly in t, and the ratio $\|Z_0\|_2/\kappa_1$ does not depend on t. Condition (ii) of Assumption 2 holds with $\kappa_0 = (\|M\|_{\mathrm{op}} + \|N\|_{\mathrm{op}})R_1/\|(\Gamma N, qM)\|_2$, which also does not depend on t. The overlap $\kappa_1 > 0$ holds as soon as $\bar v_1^\top\Gamma N + \gamma q\,\bar v_1^\top M \neq 0$. The simplest choice is $N = 0$: then $\kappa_1 = \frac{t}{2}\gamma q\,\|\bar v_1^\top M\|_2$, and the only requirement is $\bar v_1^\top M \neq 0$. Since $\bar v_1$ is unknown, we choose M at random, exactly as in Remark 4.

Finally we choose the scales. With M, N and the first-layer laws fixed, the ratio $r := \|Z_0\|_2/\kappa_1$ and the constants $c_\star, C_\star$ are fixed numbers. We can rewrite Equation (123) *as*

$$r\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^\theta + \frac{\|Z_0\|_2}{\delta} + \delta\Big] \le c_\star.$$

So like the linear case, we first choose δ so small that $\delta < c_\star$ and $\delta \le c_\star/(3r)$. Then choose t, and hence $\|Z_0\|_2 = t\|(\Gamma N, qM)\|_2$, so small that $\|Z_0\|_2 < \delta$ and $(\|Z_0\|_2/\delta)^\theta \le c_\star/(3r)$. Since $\|Z_0\|_2/\delta < 1$ and $\theta \in (0,1)$, this also gives $\|Z_0\|_2/\delta \le (\|Z_0\|_2/\delta)^\theta \le c_\star/(3r)$, and the three terms add up to at most $c_\star/r$. So, choice of parameters is exactly same as in the linear case as we first fix the exit scale δ, then fix the initialization scale t small relative to δ.

Remark 6. *One might expect that a bottleneck of dimension b learns the top b principal directions at the same time. Theorem 10 says that this is not what happens at the beginning. Near the zero map, all b latent coordinates are driven by the same fastest-growing direction $\bar{v}_1$ like each column of C_{s_δ} and of B_{s_δ} points along $\bar{v}_1$, and the decoder image is close to the line $\mathbb{R}\bar{v}_1$. In other words, at the first exit from the small-initialization regime the network has learned one data direction, and it has stored this one direction b times in the latent space. The second principal direction only becomes visible later, once the first one has stopped growing. The same remark applies to the linear model in Theorem 9.*

M.2.2 Proof of Theorem 10

In this section we give the proof of Theorem 10. The main idea is the same as for Theorem 9. First, we show that as long as the outputs are small, the moments $Z_s = (C_s, B_s)$ satisfy a linear system up to errors of second order in $\mathfrak{a}_s$ (Lemma 21). And then, we show that linear system can be solved explicitly in the eigenbasis of $\mathcal{C}$, which identifies the growing and decaying modes (Lemma 22). And then at last, we show that the exact moments stay close to the linear solution up to time s_δ (Lemma 23). The proof of the theorem is then the same comparison of modes as in the linear case.

Throughout the proofs, we use the following two properties of σ:

$$|\sigma(t)| \le S_1|t|, \qquad |\sigma(t) - t| \le \tfrac{1}{2}S_2 t^2 \qquad \text{for all } t \in \mathbb{R}, \tag{126}$$

both of which follow from Taylor's theorem together with $\sigma(0) = 0$ and $\sigma'(0) = 1$.

The following lemma is the replacement of Proposition 3 for the nonlinear activation function. It says that as long as $\mathfrak{a}_s$ is small, the first-layer coordinates move only at second order, and the outputs α_s, Ψ_s and hence the moments (C_s, B_s) satisfy linear equations up to second-order errors.

Lemma 21 (Small-amplitude moment closure)**.** *Let Δ_s be the largest distance travelled by a first-layer coordinate on $[0, s]$,*

$$\Delta_s := \max\Big\{ \operatorname*{ess\,sup}_{\rho_{W,0}} \|w_{1,s} - w_{1,0}\|_2,\ \operatorname*{ess\,sup}_{\rho_{U,0}} \|u_{2,s} - u_{2,0}\|_2 \Big\},$$

where w_s, u_s denote the positions at time s of the particles starting at w_0, u_0. Suppose that, for all $0 \le r \le s$, $\mathfrak{a}_r \le 1$ and $\Delta_r \le R_1$. Then, there exists explicit constants $C_\Delta, C_\alpha, C_1, R_2$ and C_Ψ such that the followings hold,

(a) (First-layer drift.) $\Delta_s \le C_\Delta \int_0^s \mathfrak{a}_r^2\, dr$.

(b) (Encoder output.)

$$\big\|\dot{\alpha}_s - k_W\big[B_s^\top(\cdot)\big]\big\|_{L^\infty} \le C_\alpha\big(\mathfrak{a}_s^2 + \mathfrak{a}_s\Delta_s\big), \tag{127}$$

where $k_W[f](x) = \mathbb{E}_{x'}[k_W(x,x')f(x')]$ for vector-valued function f.

(c) (Decoder map.)

$$\big\|\dot{\Psi}_s(\cdot) - C_s\, m(\cdot)\big\|_{C^2(\mathcal{A})} \le C_\Psi\big(\mathfrak{a}_s^2 + \mathfrak{a}_s\Delta_s\big). \tag{128}$$

(d) (Moments.)

$$\|\dot{C}_s - \mathcal{C}B_s\|_{\mathrm{F}} + \|\dot{B}_s - C_s Q_{u_2,0}\|_{\mathrm{F}} \le C_1\big(\mathfrak{a}_s^2 + \mathfrak{a}_s\Delta_s\big), \tag{129}$$

where

$$C_\Delta := S_1(M_x+1)^2, \qquad C_\alpha := S_0^2(M_x+2) + S_1^2 M_x^2(M_x+1) + 2S_0S_1M_x^2,$$

$$C_\Psi := 16(1+R_2)^3\big(S_0+S_1+S_2+S_3\big)^2\big(M_x+2\big)^2, \quad C_1 := M_xC_\alpha + \sqrt{b}C_\Psi \quad R_2 = 2R_1.$$

Proof. We first bound some necessary terms that will be useful in proving the four statements above, and then we prove the main statements. Throughout the proof s is fixed and we write $\mathfrak{a}$ for $\mathfrak{a}_s$ and Δ for Δ_s. By the hypothesis $\Delta_r \le R_1$ and condition (i) of Assumption 2, every first-layer coordinate satisfies $\|w_1\|_2 \le R_1 + R_1 = R_2$ and $\|u_2\|_2 \le R_2$ at every time $r \le s$. By the hypothesis $\mathfrak{a} \le 1$ and the definition of $\mathfrak{a}$, we have $\|\alpha_s(x)\|_2 \le \mathfrak{a} \le 1$, so $\alpha_s(x) \in \mathcal{A}$ for every x. The term $\|\Psi_s\|_{C^2(\mathcal{A})}$ in $\mathfrak{a}$ controls Ψ_s, $D\Psi_s$ and $D^2\Psi_s$ on $\mathcal{A}$ and since $\alpha_s(x) \in \mathcal{A}$, we can say that $\|\Psi_s(\alpha_s(x))\|_2 \le \mathfrak{a}$, $\|D\Psi_s(\alpha_s(x))\|_{\mathrm{op}} \le \mathfrak{a}$ and $\|D^2\Psi_s(a)\|_{\mathrm{op}} \le \mathfrak{a}$ for all a on the segment from 0 to $\alpha_s(x)$.

We first collect three bounds that are used repeatedly. Let $r_s(x) := \Psi_s(\alpha_s(x)) - x$ be the residual. Since $\|\Psi_s\|_{C^0(\mathcal{A})} \le \mathfrak{a}$,

$$\|r_s(x) + x\|_2 \le \mathfrak{a}, \qquad \text{hence} \qquad \|r_s(x)\|_2 \le M_x + 1. \tag{130}$$

Next, $B_s = D\Psi_s(0)$ and, by the mean value theorem applied to $D\Psi_s$ on the segment from 0 to $\alpha_s(x)$,

$$\|D\Psi_s(\alpha_s(x)) - B_s\|_{\mathrm{op}} \le \sup_{\mathcal{A}} \|D^2\Psi_s\|_{\mathrm{op}}\, \|\alpha_s(x)\|_2 \le \mathfrak{a} \cdot \mathfrak{a} = \mathfrak{a}^2, \qquad \|B_s\|_{\mathrm{op}} \le \mathfrak{a}. \tag{131}$$

Finally, let $g_s(x) := D\Psi_s(\alpha_s(x))^\top r_s(x)$ be the signal that the decoder sends back to the bottleneck. So, writing $g_s(x) + B_s^\top x = (D\Psi_s(\alpha_s(x)) - B_s)^\top r_s(x) + B_s^\top (r_s(x) + x)$ and using (130)–(131), we get

$$\|g_s(x) + B_s^\top x\|_2 \le \mathfrak{a}^2(M_x + 1) + \mathfrak{a} \cdot \mathfrak{a} = (M_x + 2)\mathfrak{a}^2, \qquad \|g_s(x)\|_2 \le \mathfrak{a}(M_x + 1). \tag{132}$$

So, up to an error of second order, the backpropagated signal is $-B_s^\top x$.

First, we deal with encoder particles. For $w = (w_1, w_2)$ we have $\phi(x; w) = w_2\sigma(\langle w_1, x\rangle)$, so $\partial_{w_2}\phi = \sigma(\langle w_1, x\rangle) I_{b}$ and $\partial_{w_1}\phi = \sigma'(\langle w_1, x\rangle) w_2 x^\top$. With $\lambda = 0$ the encoder particles move by $\dot w = -\mathbb{E}_x[J_w\phi(x; w)^\top g_s(x)]$, that is,

$$\dot w_2 = -\mathbb{E}_x\big[\sigma(\langle w_1, x\rangle)\, g_s(x)\big], \qquad \dot w_1 = -\mathbb{E}_x\big[\sigma'(\langle w_1, x\rangle)\, \langle w_2, g_s(x)\rangle\, x\big].$$

Now, by (132),

$$\Big\|\dot w_2 - \mathbb{E}_x\big[\sigma(\langle w_1, x\rangle) B_s^\top x\big]\Big\|_2 = \Big\|\mathbb{E}_x\big[\sigma(\langle w_1, x\rangle)\big(g_s(x) + B_s^\top x\big)\big]\Big\|_2 \le S_0(M_x + 2)\mathfrak{a}^2, \tag{133}$$

and, since $\|w_2\|_2 \le \mathfrak{a}$, $\|g_s\|_2 \le \mathfrak{a}(M_x + 1)$ and $\|x\|_2 \le M_x$,

$$\|\dot w_1\|_2 \le S_1\, \mathfrak{a} \cdot \mathfrak{a}(M_x + 1) \cdot M_x = S_1 M_x(M_x + 1)\mathfrak{a}^2. \tag{134}$$

So the output coefficient w_2 moves at first order, driven by B_s, while the feature direction w_1 moves only at second order.

Next, we deal with decoder particles. For $u = (u_1, u_2)$ we have $\psi(a; u) = u_1\sigma(\langle u_2, a\rangle)$, so $\partial_{u_1}\psi = \sigma(\langle u_2, a\rangle) I_d$ and $\partial_{u_2}\psi = \sigma'(\langle u_2, a\rangle) u_1 a^\top$. Let's define $t_x := \langle u_2, \alpha_s(x)\rangle$. The decoder particles move by $\dot u = -\mathbb{E}_x[J_u\psi(\alpha_s(x); u)^\top r_s(x)]$, that is,

$$\dot u_1 = -\mathbb{E}_x\big[\sigma(t_x)\, r_s(x)\big], \qquad \dot u_2 = -\mathbb{E}_x\big[\sigma'(t_x)\, \langle u_1, r_s(x)\rangle\, \alpha_s(x)\big].$$

Note that $|t_x| \le \|u_2\|_2 \|\alpha_s(x)\|_2 \le R_2\mathfrak{a}$. We compare $\dot u_1$ with $C_s u_2 = \mathbb{E}_x[x\, \alpha_s(x)^\top] u_2 = \mathbb{E}_x[t_x\, x]$. So,

$$\dot u_1 - C_s u_2 = -\mathbb{E}_x\big[\sigma(t_x)\, (r_s(x) + x)\big] + \mathbb{E}_x\big[(\sigma(t_x) - t_x)\, x\big].$$

By (126) and (130), the first term has norm at most $S_1 R_2 \mathfrak{a} \cdot \mathfrak{a}$ and the second at most $\frac{1}{2} S_2 R_2^2 \mathfrak{a}^2 M_x$. Hence

$$\|\dot u_1 - C_s u_2\|_2 \le C_6\, \mathfrak{a}^2, \qquad C_6 := S_1 R_2 + \tfrac{1}{2} S_2 R_2^2 M_x. \tag{135}$$

Also, since $\|u_1\|_2 \le \mathfrak{a}$, $\|r_s\|_2 \le M_x + 1$ and $\|\alpha_s\|_2 \le \mathfrak{a}$,

$$\|\dot u_2\|_2 \le S_1\, \mathfrak{a}(M_x + 1)\, \mathfrak{a} = S_1(M_x + 1)\mathfrak{a}^2. \tag{136}$$

So u_1 moves at first order, driven by C_s, and u_2 only at second order.

We now give the proof of the main statements above. We prove them one after another using the above bounds.

We first give proof of statement (a). So, by integrating (134) and (136) over $[0, s]$ along each particle, and using $M_x \le M_x + 1$, we get

$$\|w_{1,s} - w_{1,0}\|_2 \le \int_0^s \|\dot{w}_{1,r}\|_2 \, dr \le S_1(M_x+1)^2 \int_0^s \mathfrak{a}_r^2 \, dr,$$

and the same bound holds for u_2. Taking the essential supremum over particles gives (a) with $C_\Delta = S_1(M_x+1)^2$.

We now give the proof of statement (b). By the push-forward identity, exactly as in the proof of Proposition 3, we can differentiate $\alpha_s(x) = \int w_2 \sigma(\langle w_1, x\rangle) \rho_{W,s}(dw)$ along the particles and we get

$$\dot{\alpha}_s(x) = \int \Big(\dot{w}_2 \, \sigma(\langle w_1, x\rangle) + w_2 \, \sigma'(\langle w_1, x\rangle) \langle \dot{w}_1, x\rangle \Big) \rho_{W,s}(dw) =: T_1(x) + T_2(x).$$

So, by using $\|w_2\|_2 \le \mathfrak{a}$, Equation (134), $\|x\|_2 \le M_x$, and $\mathfrak{a} \le 1$ we get that

$$\|T_2(x)\|_2 \le \mathfrak{a} \cdot S_1 \cdot S_1 M_x (M_x+1) \mathfrak{a}^2 \cdot M_x \le S_1^2 M_x^2 (M_x+1) \mathfrak{a}^2.$$

For bounding T_1 we use (133). Note that

$$T_1(x) - \int \sigma(\langle w_1, x\rangle) \, \mathbb{E}_{x'} \big[\sigma(\langle w_1, x'\rangle) B_s^\top x' \big] \rho_{W,s}(dw) = \int \sigma(\langle w_1, x\rangle) (\dot{w}_2 - \mathbb{E}_x \big[\sigma(\langle w_1, x\rangle) B_s^\top x \big]) \rho_{W,s}(dw).$$

So, by using Equation (133) and $\|\sigma\|_\infty \le S_0$, we get that

$$\Big\| T_1(x) - \int \sigma(\langle w_1, x\rangle) \, \mathbb{E}_{x'} \big[\sigma(\langle w_1, x'\rangle) B_s^\top x' \big] \rho_{W,s}(dw) \Big\|_2 \le S_0 \cdot S_0 (M_x + 2) \mathfrak{a}^2.$$

The integral in the last display equals $\mathbb{E}_{x'}[k_{W,s}(x, x') B_s^\top x']$, where $k_{W,s}$ is the kernel (116) computed with the law of $w_{1,s}$ instead of $w_{1,0}$. It remains to replace $k_{W,s}$ by the frozen kernel k_W. By the push-forward identity, $k_{W,s}(x,x') - k_W(x,x') = \int \big[\sigma(\langle w_{1,s}, x\rangle) \sigma(\langle w_{1,s}, x'\rangle) - \sigma(\langle w_{1,0}, x\rangle) \sigma(\langle w_{1,0}, x'\rangle) \big] \rho_{W,0}(dw_0)$, and since $|\sigma(a)\sigma(a') - \sigma(b)\sigma(b')| \le S_0 S_1 (|a-b| + |a'-b'|)$ with $|\langle w_{1,s} - w_{1,0}, x\rangle| \le M_x \Delta$, we can conclude that

$$|k_{W,s}(x,x') - k_W(x,x')| \le 2 S_0 S_1 M_x \Delta.$$

Hence $\|\mathbb{E}_{x'}[(k_{W,s} - k_W)(x,x') B_s^\top x']\|_2 \le 2 S_0 S_1 M_x \Delta \cdot \mathfrak{a} \cdot M_x$. Adding the three errors gives (127) with $C_\alpha = S_0^2 (M_x + 2) + S_1^2 M_x^2 (M_x + 1) + 2 S_0 S_1 M_x^2$.

We now give the proof of statement (c). Throughout, $a \in \mathcal{A}$, so $\|a\|_2 \le 1$, and we recall $\|u_1\|_2 \le \mathfrak{a}$ and $\|u_2\|_2 \le R_2$ for every decoder particle at time s. We also recall that for a map $f : \mathcal{A} \to \mathbb{R}^d$ the norm $\|f\|_{C^2(\mathcal{A})}$ is the sum of the suprema over $\mathcal{A}$ of $\|f(a)\|_2$, $\|Df(a)\|_{\mathrm{op}}$ and $\|D^2 f(a)\|_{\mathrm{op}}$, where $D^2 f(a)$ is regarded as a bilinear map. The derivatives in a that we need are, for a fixed vector $v \in \mathbb{R}^b$ with $\|v\|_2 \le R_2$,

$$D_a \big[\sigma(\langle v, a\rangle) \big] = \sigma'(\langle v, a\rangle) \, v^\top, \qquad D_a^2 \big[\sigma(\langle v, a\rangle) \big] = \sigma''(\langle v, a\rangle) \, v v^\top, \tag{137}$$

with operator norms at most $S_1 R_2$ and $S_2 R_2^2$ respectively.

By using the push-forward identity, we know that $\Psi_s(a) = \int u_{1,s} \, \sigma(\langle u_{2,s}, a\rangle) \, \rho_{U,0}(du_0)$, where only the integrand depends on s. Now, differentiating under the integral (allowed because the particles stay bounded and the characteristic fields are continuous) and using the product rule and the chain rule, we get

$$\dot{\Psi}_s(a) = \int \Big(\dot{u}_1 \, \sigma(\langle u_2, a\rangle) + u_1 \, \sigma'(\langle u_2, a\rangle) \, \langle \dot{u}_2, a\rangle \Big) \rho_{U,s}(du) =: T_3(a) + T_4(a).$$

Now we insert the decoder equations. By (135) we can write $\dot{u}_1 = C_s u_2 + e_1(u)$ with $\|e_1(u)\|_2 \le C_6 \mathfrak{a}^2$, and by (136) we have $\|\dot{u}_2\|_2 \le S_1(M_x+1)\mathfrak{a}^2$. Substituting the first into T_3,

$$T_3(a) = C_s \int \sigma(\langle u_2, a\rangle) \, u_2 \, \rho_{U,s}(du) + \int e_1(u) \, \sigma(\langle u_2, a\rangle) \, \rho_{U,s}(du) = C_s \, m_s(a) + T_5(a),$$

where $m_s(a) := \int \sigma(\langle u_2, a\rangle) u_2 \, \rho_{U,s}(du)$ is the function (118) computed with the law of $u_{2,s}$ instead of $u_{2,0}$, so that $m_0 = m$. Altogether,

$$\dot\Psi_s(a) - C_s m(a) = \underbrace{C_s\big(m_s(a) - m(a)\big)}_{\text{(A)}} + \underbrace{T_5(a)}_{\text{(B)}} + \underbrace{T_4(a)}_{\text{(C)}}. \tag{138}$$

Term (A) is the price of freezing the decoder kernel at its initial value, term (B) comes from the second-order error in $\dot u_1$, and term (C) comes from the motion of the first-layer coordinates u_2. We bound the $C^2(\mathcal{A})$ norm of each term separately.

We first bound term (C). We have $T_4(a) = \int u_1 \, \sigma'(\langle u_2, a\rangle)\langle \dot u_2, a\rangle \, \rho_{U,s}(du)$. For the value, $\|T_4(a)\|_2 \le \mathfrak{a} \cdot S_1 \cdot \|\dot u_2\|_2 \|a\|_2 \le S_1^2(M_x+1)\mathfrak{a}^3$. For the first derivative, by the product rule and (137),

$$D_a\big[\sigma'(\langle u_2, a\rangle)\langle \dot u_2, a\rangle\big] = \sigma''(\langle u_2, a\rangle)\langle \dot u_2, a\rangle \, u_2^\top + \sigma'(\langle u_2, a\rangle)\, \dot u_2^\top,$$

whose norm is at most $(S_2R_2 + S_1)\|\dot u_2\|_2$, so $\|DT_4(a)\|_{\mathrm{op}} \le \mathfrak{a}(S_1 + S_2R_2)S_1(M_x+1)\mathfrak{a}^2$. Differentiating once more gives $\sigma'''(\langle u_2, a\rangle)\langle \dot u_2, a\rangle u_2 u_2^\top + \sigma''(\langle u_2, a\rangle)(u_2 \dot u_2^\top + \dot u_2 u_2^\top)$, whose norm is at most $(S_3R_2^2 + 2S_2R_2)\|\dot u_2\|_2$, so $\|D^2T_4(a)\|_{\mathrm{op}} \le \mathfrak{a}(2S_2R_2 + S_3R_2^2)S_1(M_x+1)\mathfrak{a}^2$. Adding the three bounds and using $\mathfrak{a}^3 \le \mathfrak{a}^2$ (because $\mathfrak{a} \le 1$),

$$\|T_4\|_{C^2(\mathcal{A})} \le S_1(M_x+1)\big(2S_1 + 3S_2R_2 + S_3R_2^2\big)\,\mathfrak{a}^2.$$

We now bound term (B). We have $T_5(a) = \int e_1(u)\, \sigma(\langle u_2, a\rangle)\, \rho_{U,s}(du)$ with $\|e_1(u)\|_2 \le C_6\mathfrak{a}^2$. The vector $e_1(u)$ does not depend on a, so by (137) the value, first and second derivative of the integrand are bounded by $C_6\mathfrak{a}^2$ times S_0, S_1R_2 and $S_2R_2^2$ respectively. Hence

$$\|T_5\|_{C^2(\mathcal{A})} \le C_6\big(S_0 + S_1R_2 + S_2R_2^2\big)\,\mathfrak{a}^2.$$

We now bound term (A). By the push-forward identity applied to both m_s and $m = m_0$,

$$m_s(a) - m(a) = \int \Big(\sigma(\langle u_{2,s}, a\rangle)\, u_{2,s} - \sigma(\langle u_{2,0}, a\rangle)\, u_{2,0}\Big)\rho_{U,0}(du_0).$$

For fixed a, consider the map $F_a(v) := \sigma(\langle v, a\rangle)v$ on $\{v \in \mathbb{R}^b : \|v\|_2 \le R_2\}$. Its derivative in v is $DF_a(v) = \sigma'(\langle v, a\rangle)\, va^\top + \sigma(\langle v, a\rangle)I_b$, of operator norm at most $S_1R_2 + S_0$. By the mean value theorem, $\|F_a(u_{2,s}) - F_a(u_{2,0})\|_2 \le (S_0 + S_1R_2)\|u_{2,s} - u_{2,0}\|_2 \le (S_0 + S_1R_2)\Delta$, and integrating over $\rho_{U,0}$ gives $\sup_{\mathcal{A}} \|m_s - m\|_2 \le (S_0 + S_1R_2)\Delta$. For the first derivative in a we apply the same argument to $D_aF_a(v) = \sigma'(\langle v, a\rangle)\, vv^\top$: its derivative in v has norm at most $S_2R_2^2\|a\|_2 + 2S_1R_2 \le 2S_1R_2 + S_2R_2^2$, so $\sup_{\mathcal{A}} \|D(m_s - m)\|_{\mathrm{op}} \le (2S_1R_2 + S_2R_2^2)\Delta$. For the second derivative we apply it to $D_a^2F_a(v) = \sigma''(\langle v, a\rangle)\, v \otimes vv^\top$, whose derivative in v has norm at most $S_3R_2^3 + 3S_2R_2^2$, so $\sup_{\mathcal{A}} \|D^2(m_s - m)\|_{\mathrm{op}} \le (3S_2R_2^2 + S_3R_2^3)\Delta$. Adding the three bounds,

$$\|m_s - m\|_{C^2(\mathcal{A})} \le \big(S_0 + 3S_1R_2 + 4S_2R_2^2 + S_3R_2^3\big)\Delta.$$

Since C_s is a constant matrix in a with $\|C_s\|_{\mathrm{F}} \le M_x\mathfrak{a}$,

$$\|C_s(m_s - m)\|_{C^2(\mathcal{A})} \le M_x\big(S_0 + 3S_1R_2 + 4S_2R_2^2 + S_3R_2^3\big)\,\mathfrak{a}\Delta.$$

Adding the bounds of Steps 2–4 in (138), we obtain

$$\|\dot\Psi_s - C_s m\|_{C^2(\mathcal{A})} \le C_{\Psi,1}\,\mathfrak{a}^2 + C_{\Psi,2}\,\mathfrak{a}\Delta \le C_\Psi\big(\mathfrak{a}^2 + \mathfrak{a}\Delta\big),$$

with the explicit constants $C_{\Psi,1} := C_6\big(S_0 + S_1R_2 + S_2R_2^2\big) + S_1(M_x+1)\big(2S_1 + 3S_2R_2 + S_3R_2^2\big)$, $C_{\Psi,2} := M_x\big(S_0 + 3S_1R_2 + 4S_2R_2^2 + S_3R_2^3\big)$, and $C_\Psi :\ge \max\{C_{\Psi,1}, C_{\Psi,2}\}$. This proves (128).

We now give the proof of statement (d). There are two moment equations to check, one for C_s and one for B_s, and they are obtained from statements (b) and (c) respectively, by one linear operation: an average against x for C_s, a derivative at $a = 0$ for B_s.

By (b) we can write

$$\dot\alpha_s(x) = k_W\big[B_s^\top(\cdot)\big](x) + e_s(x) = \mathbb{E}_{x'}\big[k_W(x, x')\, B_s^\top x'\big] + e_s(x), \qquad \sup_x \|e_s(x)\|_2 \le C_\alpha\big(\mathfrak{a}^2 + \mathfrak{a}\Delta\big).$$

Recall that $C_s = \mathbb{E}_x[x\,\alpha_s(x)^\top]$. So, by differentiating in s under the expectation (the data are bounded and $\dot\alpha_s$ is bounded, so this is allowed) gives us $\dot C_s = \mathbb{E}_x[x\,\dot\alpha_s(x)^\top]$. For the main term, since $B_s^\top x'$ is a vector and $(B_s^\top x')^\top = x'^\top B_s$, we can write

$$\mathbb{E}_x\Big[x\,\big(\mathbb{E}_{x'}[k_W(x,x')B_s^\top x']\big)^\top\Big] = \mathbb{E}_x\Big[x\,\mathbb{E}_{x'}\big[k_W(x,x')\,x'^\top\big]\Big]B_s = \mathbb{E}_{x,x'}\big[k_W(x,x')\,xx'^\top\big]B_s = \mathcal{C}B_s,$$

where we take the constant matrix B_s out of the expectation and used the second expression for $\mathcal{C}$ in (117). For the error term, $\|xe_s(x)^\top\|_{\mathrm{F}} = \|x\|_2\|e_s(x)\|_2$, so

$$\big\|\mathbb{E}_x[x\,e_s(x)^\top]\big\|_{\mathrm{F}} \le \mathbb{E}_x\big[\|x\|_2\|e_s(x)\|_2\big] \le M_x\,C_\alpha\big(\mathfrak{a}^2+\mathfrak{a}\Delta\big).$$

Putting the two together, we get

$$\|\dot C_s - \mathcal{C}B_s\|_{\mathrm{F}} \le M_xC_\alpha\big(\mathfrak{a}^2+\mathfrak{a}\Delta\big).$$

By (c) we can write

$$\dot\Psi_s(a) = C_s\,m(a) + f_s(a) \quad \text{for all } a\in\mathcal{A}, \qquad \|f_s\|_{C^2(\mathcal{A})} \le C_\Psi\big(\mathfrak{a}^2+\mathfrak{a}\Delta\big).$$

Recall that $B_s = D\Psi_s(0)$. Since Ψ_s is smooth in a and in s, the derivatives in a and in s commute, so $\dot B_s = \frac{d}{ds}D\Psi_s(0) = D\dot\Psi_s(0)$. So, by differentiating the above at $a=0$ and using $Dm(0) = Q_{u_2,0}$ from (118), we get

$$\dot B_s = D\dot\Psi_s(0) = C_s\,Dm(0) + Df_s(0) = C_sQ_{u_2,0} + Df_s(0).$$

It remains to bound $Df_s(0)\in\mathbb{R}^{d\times b}$ in Frobenius norm. By definition of the $C^2(\mathcal{A})$ norm, $\|Df_s(0)\|_{\mathrm{op}} \le \|f_s\|_{C^2(\mathcal{A})}$, and for a $d\times b$ matrix the Frobenius norm is at most $\sqrt{b}$ times the operator norm (the Frobenius norm is the ℓ^2 norm of the b singular values, each of which is at most the operator norm). Hence

$$\|\dot B_s - C_sQ_{u_2,0}\|_{\mathrm{F}} = \|Df_s(0)\|_{\mathrm{F}} \le \sqrt{b}\,\|Df_s(0)\|_{\mathrm{op}} \le \sqrt{b}\,C_\Psi\big(\mathfrak{a}^2+\mathfrak{a}\Delta\big).$$

So, after adding the above two bounds, we get

$$\|\dot C_s - \mathcal{C}B_s\|_{\mathrm{F}} + \|\dot B_s - C_sQ_{u_2,0}\|_{\mathrm{F}} \le \big(M_xC_\alpha + \sqrt{b}\,C_\Psi\big)\big(\mathfrak{a}^2+\mathfrak{a}\Delta\big),$$

which is (129) with $C_1 := M_xC_\alpha + \sqrt{b}\,C_\Psi$. □

To leading order the nonlinear mean-field PDE has therefore been reduced to the linear system

$$\dot C_s = \mathcal{C}B_s, \qquad \dot B_s = C_sQ_{u_2,0} = q\,C_s \tag{139}$$

for two $d\times b$ matrices. This is the precise sense in which the early dynamics are linear. The activation is nonlinear and the nonlinearity survives inside the frozen kernel k_W, that is, inside $\mathcal{C}$.

The following lemma solves (139) explicitly. It plays the role of the display (114) in the linear proof.

Lemma 22 (Solution of the leading linear system)**.** *Assume* (119) *and* $\nu_1 > 0$*. Let* $(\widetilde C_s, \widetilde B_s)$ *be the solution of* (139) *with* $(\widetilde C_0, \widetilde B_0) = (C_0, B_0)$*, and write* $c_i(s) := \bar v_i^\top \widetilde C_s$ *and* $b_i(s) := \bar v_i^\top \widetilde B_s \in \mathbb{R}^{1\times b}$ *for its rows in the eigenbasis of* $\mathcal{C}$*. Then, for every* i*,*

$$c_i(s) = \cosh(\rho_i s)\,c_i(0) + \frac{\nu_i}{\rho_i}\sinh(\rho_i s)\,b_i(0),\ b_i(s) = \cosh(\rho_i s)\,b_i(0) + \frac{q}{\rho_i}\sinh(\rho_i s)\,c_i(0), \tag{140}$$

where, when $\nu_i = 0$ *(and hence* $\rho_i = 0$*), the coefficients are defined by their limiting values:* $\frac{\nu_i}{\rho_i}\sinh(\rho_i s) = 0$ *and* $\frac{q}{\rho_i}\sinh(\rho_i s) = qs$*. Consequently:*

(a) The top row splits into a growing and a decaying part,

$$c_1(s) = \tfrac12 e^{\rho_1 s}\big(c_1(0) + \gamma\,b_1(0)\big) + \tfrac12 e^{-\rho_1 s}\big(c_1(0) - \gamma\,b_1(0)\big),$$

$$b_1(s) = \tfrac{1}{2\gamma} e^{\rho_1 s}\big(c_1(0) + \gamma\,b_1(0)\big) - \tfrac{1}{2\gamma} e^{-\rho_1 s}\big(c_1(0) - \gamma\,b_1(0)\big),$$

and at time= 0, the growing coefficient has norm $\frac12\|c_1(0) + \gamma b_1(0)\|_2 = \kappa_1$.

(b) For $i \geq 2$, $\|c_i(s)\|_2 \leq e^{\rho_2 s}\big(\|c_i(0)\|_2 + \gamma\|b_i(0)\|_2\big)$ and, if $\nu_2 > 0$, $\|b_i(s)\|_2 \leq e^{\rho_2 s}\big(\|b_i(0)\|_2 + \gamma'\|c_i(0)\|_2\big)$ with $\gamma' := \sqrt{q/\nu_2}$.

(c) For all $s \geq 0$, $\|(\widetilde{C}_s, \widetilde{B}_s)\|_2 \leq C_L e^{\rho_1 s}\|Z_0\|_2$ with $C_L := \sqrt{2}\,\big(2 + (\nu_1 + q)/\rho_1\big)$.

Proof. We now solve (139). Fix i and multiply both equations of (139) on the left by the row vector $\bar{v}_i^\top$. Since $\bar{v}_i$ is an eigenvector of $\mathcal{C}$ with eigenvalue ν_i, we have $\bar{v}_i^\top \mathcal{C} = \nu_i \bar{v}_i^\top$, and therefore

$$\dot{c}_i(s) = \bar{v}_i^\top \dot{C}_s = \bar{v}_i^\top \mathcal{C} B_s = \nu_i\, \bar{v}_i^\top B_s = \nu_i\, b_i(s), \qquad \dot{b}_i(s) = \bar{v}_i^\top \dot{B}_s = q\, \bar{v}_i^\top C_s = q\, c_i(s).$$

So the pair (c_i, b_i) satisfies the closed two-dimensional linear system

$$\dot{c}_i = \nu_i\, b_i, \qquad \dot{b}_i = q\, c_i,$$

with initial values $c_i(0)$ and $b_i(0)$. Differentiating the first equation once more and substituting the second, we get

$$\ddot{c}_i = \nu_i\, \dot{b}_i = \nu_i q\, c_i = \rho_i^2\, c_i.$$

This is a second-order linear equation with constant coefficient, whose general solution is a combination of $\cosh(\rho_i s)$ and $\sinh(\rho_i s)$. The initial position is $c_i(0)$, and the initial velocity is $\dot{c}_i(0) = \nu_i b_i(0)$ by the first equation. Matching these two initial conditions gives

$$c_i(s) = c_i(0)\cosh(\rho_i s) + \nu_i\, b_i(0)\,\frac{\sinh(\rho_i s)}{\rho_i},$$

which is the first formula in (140). For b_i we differentiate this formula and use $b_i = \dot{c}_i/\nu_i$ (when $\nu_i \neq 0$):

$$b_i(s) = \frac{1}{\nu_i}\dot{c}_i(s) = \frac{1}{\nu_i}\Big(c_i(0)\,\rho_i \sinh(\rho_i s) + \nu_i\, b_i(0)\cosh(\rho_i s)\Big) = b_i(0)\cosh(\rho_i s) + q\, c_i(0)\,\frac{\sinh(\rho_i s)}{\rho_i},$$

where in the last step we used $\rho_i^2 = \nu_i q$, so $\rho_i/\nu_i = q/\rho_i$. This is the second formula in (140).

It remains to treat the degenerate case $\nu_i = 0$, where $\rho_i = 0$ and the division by ν_i above is not allowed. In this case the system reads $\dot{c}_i = 0$ and $\dot{b}_i = q\, c_i$, so $c_i(s) = c_i(0)$ is constant and $b_i(s) = b_i(0) + q\, c_i(0)\, s$ grows linearly. These are exactly the limits of the two formulas above as $\rho_i \to 0$, since $\cosh(\rho_i s) \to 1$ and $\sinh(\rho_i s)/\rho_i \to s$. Hence (140) holds for every i, with the convention $\sinh(\rho_i s)/\rho_i := s$ when $\rho_i = 0$.

For (a), put $\cosh x = \frac{1}{2}(e^x + e^{-x})$ and $\sinh x = \frac{1}{2}(e^x - e^{-x})$ in (140) for $i = 1$, and use $\nu_1/\rho_1 = \sqrt{\nu_1/q} = \gamma$ and $q/\rho_1 = \sqrt{q/\nu_1} = 1/\gamma$. The norm of the growing coefficient is κ_1 follows from the definition (122), because $c_1(0) = \bar{v}_1^\top C_0$ and $b_1(0) = \bar{v}_1^\top B_0$.

For (b), let $i \geq 2$, so $\rho_i \leq \rho_2$. Then $\cosh(\rho_i s) \leq e^{\rho_i s} \leq e^{\rho_2 s}$. For the second coefficient of c_i, note that $\frac{\nu_i}{\rho_i}\sinh(\rho_i s) = \nu_i s \cdot \frac{\sinh(\rho_i s)}{\rho_i s}$ and that $x \mapsto \sinh(x)/x$ is increasing on $[0, \infty)$, so this implies $\frac{\nu_i}{\rho_i}\sinh(\rho_i s) \leq \nu_i s \cdot \frac{\sinh(\rho_2 s)}{\rho_2 s} = \frac{\nu_i}{\rho_2}\sinh(\rho_2 s) \leq \frac{\nu_2}{\rho_2} e^{\rho_2 s} = \sqrt{\nu_2/q}\, e^{\rho_2 s} \leq \gamma\, e^{\rho_2 s}$ because $\nu_i \leq \nu_2 \leq \nu_1$. In the same way, if $\nu_2 > 0$, $\frac{q}{\rho_i}\sinh(\rho_i s) \leq \frac{q}{\rho_2}\sinh(\rho_2 s) \leq \sqrt{q/\nu_2}\, e^{\rho_2 s} = \gamma' e^{\rho_2 s}$. Now, if we take norms in (140) then it gives (b).

For (c), the same argument with ρ_2 replaced by ρ_1 (valid for every i, since $\rho_i \leq \rho_1$) gives $\frac{\nu_i}{\rho_i}\sinh(\rho_i s) \leq \frac{\nu_1}{\rho_1}\sinh(\rho_1 s)$ and $\frac{q}{\rho_i}\sinh(\rho_i s) \leq \frac{q}{\rho_1}\sinh(\rho_1 s)$, and $\sinh(\rho_1 s), \cosh(\rho_1 s) \leq e^{\rho_1 s}$. So

$$\|c_i(s)\|_2 + \|b_i(s)\|_2 \leq e^{\rho_1 s}\Big(2 + \frac{\nu_1 + q}{\rho_1}\Big)\big(\|c_i(0)\|_2 + \|b_i(0)\|_2\big) \leq e^{\rho_1 s}\Big(2 + \frac{\nu_1 + q}{\rho_1}\Big)\sqrt{2}\,\big\|(c_i(0), b_i(0))\big\|_2,$$

where the last step is the Cauchy–Schwarz inequality. Since $\|(c_i(s), b_i(s))\|_2 \leq \|c_i(s)\|_2 + \|b_i(s)\|_2$, squaring and summing over i gives $\|(\widetilde{C}_s, \widetilde{B}_s)\|_2^2 = \sum_i \|(c_i(s), b_i(s))\|_2^2 \leq C_L^2 e^{2\rho_1 s}\sum_i \|(c_i(0), b_i(0))\|_2^2 = C_L^2 e^{2\rho_1 s}\|Z_0\|_2^2$. □

So the component of Z_s along $\bar{v}_1$ grows like $e^{\rho_1 s}$ with coefficient κ_1, every other component grows at most like $e^{\rho_2 s}$, and $\rho_2 < \rho_1$ when $\nu_1 > \nu_2$. This is the same picture as for $\mathcal{M}$ in Section M.1, with $2\mu_i$ replaced by $\sqrt{q\nu_i}$ and $\|v_1^\top(A_0 + B_0)/\sqrt{2}\|_2$ replaced by κ_1.

The next lemma is about the comparison of actual odes with its linear counterparts on the small-output window. The next lemma replaces (115) of the linear proof, with the explicit function $\|Z_0\|_2 e^{\rho_1 s}$ playing the role of q_s. It says that, as long as the outputs are small, they grow at most at the rate $e^{\rho_1 s}$ of the top mode, the first layer moves by less than R_1 (so the frozen kernels remain valid), and the true moments (C_s, B_s) stay within second order of the solution $(\widetilde{C}_s, \widetilde{B}_s)$ of the leading system (139). The threshold δ^* below is the largest exit level for which all of this holds on the whole window $[0, s_\delta]$; it enters only through the hypothesis $\delta \le \delta^*$.

Lemma 23. *Assume Assumption 2, latent isotropy* (119)*, and $\nu_1 > 0$. Let $(\widetilde{C}_s, \widetilde{B}_s)$ be the solution of* (139) *from Lemma 22, and let s_δ be the exit time of* (121)*. Then there exist explicit constants $K,\ C_4,\ \delta^* > 0$ such that, whenever the initial moments satisfy $0 < \|Z_0\|_2 < \delta \le \delta^*$, the following hold for all $0 \le s \le s_\delta$.*

(a) (Growth of the outputs at the top-mode rate.)

$$\mathfrak{a}_s \le K \, \|Z_0\|_2 \, e^{\rho_1 s} \le K\delta \le 1.$$

(b) (The first layer barely moves.) $\Delta_s \le R_1$.

(c) (Second-order closeness to the linear moments.)

$$\big\|(C_s, B_s) - (\widetilde{C}_s, \widetilde{B}_s)\big\|_{\mathrm{F}} \le C_4 \, \|Z_0\|_2^2 \, e^{2\rho_1 s} \le C_4 \, \delta^2. \tag{141}$$

The constants are defined as follows. We use C_0 from (120)*, C_L from Lemma 22, $R_2 = 2R_1$, and $C_\Delta, C_\alpha, C_\Psi, C_1, C_6$ from Lemma 21. Two auxiliary constants are new here,*

$$C_2 := S_0^2 M_x + R_1\big(S_0 + S_1 R_1 + S_2 R_1^2\big) + S_0 M_x + R_2, \qquad C_3 := C_\alpha + C_\Psi + S_0(M_x + 2) + C_6.$$

In terms of these,

$$K := 2\Big(C_0 + \frac{C_2 C_L}{\rho_1}\Big),\ C_4 := \frac{2\,C_1 C_L K^2}{\rho_1},\ \delta^* := \min\Big\{\frac{1}{K},\ \frac{2\rho_1}{C_\Delta K},\ \sqrt{\frac{2\rho_1 R_1}{C_\Delta K^2}},\ \frac{\rho_1}{8 C_3 K}\Big(1 + \frac{C_2 C_4}{C_3 K^2}\Big)^{-1}\Big\}.$$

The four terms in δ^* are used, in this order, to guarantee $K\delta \le 1$ (so that Lemma 21 applies), to make the drift Δ_s no larger than the amplitude bound $K\|Z_0\|_2 e^{\rho_1 s}$ (which yields the second-order rate in (c)), to keep $\Delta_s \le R_1$ (which is (b)), and to close the bootstrap inequality for $\mathfrak{a}_s$ with a strict improvement (which yields (a)). Now we give the proof for lemma 23.

Proof. We prove the result similarly to the proof of Theorem 9. By using Equation (121), we can say that for every $r \in [0, s_\delta]$,

$$\|Z_0\|_2 \, e^{\rho_1 r} \le \delta \le \delta^*, \qquad \text{with strict inequality } \|Z_0\|_2 \, e^{\rho_1 r} < \delta \ \text{ for } r < s_\delta. \tag{142}$$

This is the only place where the smallness of δ enters the argument; every condition we impose on δ^* is of the form "a constant times δ is small".

We define S as the largest time in $[0, s_\delta]$ such that

$$\mathfrak{a}_r \le K \, \|Z_0\|_2 \, e^{\rho_1 r} \qquad \text{for all } r \in [0, S].$$

At $r = 0$ this holds because by using Equation (120) we know that $\mathfrak{a}_0 \le C_0 \|Z_0\|_2$ and $C_0 < K$. Since $r \mapsto \mathfrak{a}_r$ is continuous, we get $S > 0$. Our goal is to show $S = s_\delta$. We do this in three steps: first we check that we can apply Lemma 21 on $[0, S]$, then we prove (c) on $[0, S]$, and finally we improve the constant K to $\frac{3}{4}K$ on $[0, S]$, which forces $S = s_\delta$.

In order to apply Lemma 21, we need to show things two things: $\mathfrak{a}_r \le 1$ and $\Delta_r \le R_1$ for all $r \le S$. The first follows from the Equation (142), where for $r \le S$,

$$\mathfrak{a}_r \le K\|Z_0\|_2 e^{\rho_1 r} \le K\delta \le K\delta^* \le 1,$$

where the last inequality is the first term in the definition of δ^*.

For the drift we need a different argument. We define S' as the largest time in $[0, S]$ such that $\Delta_r \le R_1$ for all $r \le S'$. As, $\Delta_0 = 0$, we have $S' > 0$. On $[0, S']$ both hypotheses of Lemma 21 hold, so by using part (a) of that lemma and the bound on $\mathfrak{a}$ give us that, for every $r \le S'$,

$$\Delta_r \le C_\Delta \int_0^r \mathfrak{a}_{r'}^2 \, dr' \le C_\Delta K^2 \|Z_0\|_2^2 \int_0^r e^{2\rho_1 r'} \, dr' \le \frac{C_\Delta K^2}{2\rho_1} \, \|Z_0\|_2^2 \, e^{2\rho_1 r}. \tag{143}$$

We now use the third term in the definition of δ^*, namely $\delta^2 \le 2\rho_1 R_1/(C_\Delta K^2)$, together with (142). For $r < s_\delta$ the inequality in (142) is strict, so (143) gives $\Delta_r < \frac{C_\Delta K^2}{2\rho_1}\delta^2 \le R_1$ and at $r = s_\delta$ it gives $\Delta_r \le R_1$. Suppose now that $S' < S$. Then $S' < s_\delta$, so $\Delta_{S'} < R_1$. Now by using continuity, we can say that $\Delta_r \le R_1$ also holds on a small interval to the right of S', which contradicts the maximality of S'. Hence $S' = S$. So, the drift bound $\Delta_r \le R_1$ holds on all of $[0, S]$. So we can apply Lemma 21 on the time interval $[0, S]$ all the four conclusion of the Lemma 21 hold on the time interval $[0, S]$.

We now combine the bound on $\mathfrak{a}_r$ with the drift bound (143) to get

$$\begin{aligned}\mathfrak{a}_r^2 + \mathfrak{a}_r\Delta_r &\le K^2\|Z_0\|_2^2 e^{2\rho_1 r} + K\|Z_0\|_2 e^{\rho_1 r}\cdot\frac{C_\Delta K^2}{2\rho_1}\|Z_0\|_2^2 e^{2\rho_1 r}\\ &= K^2\|Z_0\|_2^2 e^{2\rho_1 r}\Big(1 + \frac{C_\Delta K}{2\rho_1}\,\|Z_0\|_2 e^{\rho_1 r}\Big).\end{aligned}$$

Now, by using Equation (142) and the second term in the definition of δ^*, we can say that the factor $\frac{C_\Delta K}{2\rho_1}\|Z_0\|_2 e^{\rho_1 r}$ is at most $\frac{C_\Delta K}{2\rho_1}\delta^* \le 1$. So, we can conclude that

$$\mathfrak{a}_r^2 + \mathfrak{a}_r\Delta_r \le 2K^2\,\|Z_0\|_2^2\,e^{2\rho_1 r} \qquad \text{for all } r \le S. \tag{144}$$

Now we make comparison of our actual odes with the linear solution, on $[0, S]$. We write L for the linear operator $(C, B) \mapsto (\mathcal{C}B,\, qC)$ of (139), so that $(\widetilde{C}_s, \widetilde{B}_s) = e^{Ls}Z_0$, and we recall from Lemma 22(c) that $\|e^{Ls}\|_{\mathrm{op}} \le C_L e^{\rho_1 s}$. Now, by using Lemma 21(d), we can say that the exact moments $Z_s = (C_s, B_s)$ solve the perturbed equation

$$\dot{Z}_s = LZ_s + \epsilon_s, \qquad \|\epsilon_s\|_2 \le C_1\big(\mathfrak{a}_s^2 + \mathfrak{a}_s\Delta_s\big) \le 2C_1K^2\|Z_0\|_2^2 e^{2\rho_1 s},$$

where the last inequality follows from (144). Duhamel's formula expresses the difference as $Z_s - e^{Ls}Z_0 = \int_0^s e^{L(s-r)}\epsilon_r\,dr$. We take norms and insert the two bounds above to get

$$\begin{aligned}\|Z_s - (\widetilde{C}_s, \widetilde{B}_s)\|_2 &\le \int_0^s C_L e^{\rho_1(s-r)}\cdot 2C_1K^2\|Z_0\|_2^2 e^{2\rho_1 r}\,dr\\ &= 2C_1C_LK^2\|Z_0\|_2^2\,e^{\rho_1 s}\int_0^s e^{\rho_1 r}\,dr \le \frac{2C_1C_LK^2}{\rho_1}\,\|Z_0\|_2^2\,e^{2\rho_1 s}.\end{aligned}$$

The constant on the right is C_4, so this proves (141) on $[0, S]$.

We now show that on $[0, S]$ the amplitude bound holds with the smaller constant $\frac{3}{4}K$. We first bound the moments. Now, by using Lemma 22(c) and above step, we can conclude that

$$\|Z_r\|_2 \le \|(\widetilde{C}_r, \widetilde{B}_r)\|_2 + \|Z_r - (\widetilde{C}_r, \widetilde{B}_r)\|_2 \le C_L\|Z_0\|_2 e^{\rho_1 r} + C_4\|Z_0\|_2^2 e^{2\rho_1 r}.$$

Next we bound the growth of $\mathfrak{a}$. Recall that $\mathfrak{a}_s$ is the sum of the four quantities $\|\alpha_s\|_{L^\infty}$, $\|\Psi_s\|_{C^2(\mathcal{A})}$, $\operatorname{ess\,sup}\|w_{2,s}\|_2$ and $\operatorname{ess\,sup}\|u_{1,s}\|_2$. Each of the four quantities that adds up $\mathfrak{a}_s$ is at most its value at time 0 plus the integral of the norm of its time derivative. We use Lemma 21 to control each derivative by a constant times $\|Z_r\|_2$ plus the second-order error term.

First, we give an upper bound for bottleneck term. By using (127), we know that $\dot{\alpha}_r$ differs from the linear map $k_W[B_r^\top\,\cdot\,]$ by at most $C_\alpha(\mathfrak{a}_r^2 + \mathfrak{a}_r\Delta_r)$ in L^∞. For the linear map itself, $k_W[B_r^\top\,\cdot\,](x) = \mathbb{E}_{x'}[k_W(x, x')B_r^\top x']$, and the frozen kernel is bounded by $|k_W(x, x')| = \big|\int\sigma(\langle w_{1,0}, x\rangle)\sigma(\langle w_{1,0}, x'\rangle)\rho_{W,0}(dw_0)\big| \le S_0^2$, while $\|B_r^\top x'\|_2 \le \|B_r\|_{\mathrm{op}}M_x$. Hence $\|k_W[B_r^\top\,\cdot\,]\|_{L^\infty} \le S_0^2 M_x\|B_r\|_{\mathrm{op}}$, and the triangle inequality gives

$$\|\dot{\alpha}_r\|_{L^\infty} \le S_0^2 M_x\,\|B_r\|_{\mathrm{op}} + C_\alpha\big(\mathfrak{a}_r^2 + \mathfrak{a}_r\Delta_r\big).$$

Now, we give an upper bound for the decoder map. Again, by using (128), we know that $\dot{\Psi}_r$ differs from $C_r\,m(\cdot)$ by at most $C_\Psi(\mathfrak{a}_r^2 + \mathfrak{a}_r\Delta_r)$ in $C^2(\mathcal{A})$. Since C_r is a constant matrix, $\|C_r m\|_{C^2(\mathcal{A})} \le \|C_r\|_{\mathrm{op}}\|m\|_{C^2(\mathcal{A})} \le \|C_r\|_{\mathrm{F}}\|m\|_{C^2(\mathcal{A})}$. To bound $\|m\|_{C^2(\mathcal{A})}$ we use the definition (118), $m(a) = \int\sigma(\langle u_{2,0}, a\rangle)\,u_{2,0}\,\rho_{U,0}(du_0)$, and the initial bound $\|u_{2,0}\|_2 \le R_1$ from Assumption 2.By differentiating under the integral, we get $Dm(a) = \int\sigma'(\langle u_{2,0}, a\rangle)\,u_{2,0}u_{2,0}^\top\rho_{U,0}(du_0)$ and

$D^2 m(a) = \int \sigma''(\langle u_{2,0}, a\rangle)\, u_{2,0} \otimes u_{2,0} u_{2,0}^\top \rho_{U,0}(du_0)$. So, the value, first and second derivative are bounded by $S_0 R_1$, $S_1 R_1^2$ and $S_2 R_1^3$ respectively, and $\|m\|_{C^2(\mathcal{A})} \le R_1(S_0 + S_1 R_1 + S_2 R_1^2)$. Therefore

$$\|\dot\Psi_r\|_{C^2(\mathcal{A})} \le R_1\big(S_0 + S_1 R_1 + S_2 R_1^2\big)\|C_r\|_{\mathrm{F}} + C_\Psi\big(\mathfrak{a}_r^2 + \mathfrak{a}_r\Delta_r\big).$$

Now, we consider output coefficients of the encoder (w_2) and give an upper bound on it. By using (133), we know that $\dot w_2$ differs from $\mathbb{E}_x[\sigma(\langle w_1, x\rangle) B_r^\top x]$ by at most $S_0(M_x+2)\mathfrak{a}_r^2$. The leading term has norm at most $S_0\|B_r\|_{\mathrm{op}} M_x$, so

$$\|\dot w_2\|_2 \le S_0 M_x\, \|B_r\|_{\mathrm{op}} + S_0(M_x+2)\, \mathfrak{a}_r^2.$$

Now, we consider output coefficients of the decoder (u_1) and give an upper bound on it. By using (135), we can say that $\dot u_1$ differs from $C_r u_2$ by at most $C_6\mathfrak{a}_r^2$. On the time interval $[0, S]$, we have shown $\nabla_r \le R_1$, which further implies that $\|u_2\|_2 \le R_2$, so the leading term has norm at most $\|C_r\|_{\mathrm{op}} R_2 \le R_2\|C_r\|_{\mathrm{F}}$, and

$$\|\dot u_1\|_2 \le R_2\, \|C_r\|_{\mathrm{F}} + C_6\, \mathfrak{a}_r^2.$$

We now add the four bounds. We use $\|B_r\|_{\mathrm{op}}, \|C_r\|_{\mathrm{F}} \le \|Z_r\|_2$ for the leading terms and $\mathfrak{a}_r^2 \le \mathfrak{a}_r^2 + \mathfrak{a}_r\Delta_r$ for the errors. The coefficients of $\|Z_r\|_2$ add up to $S_0^2 M_x + R_1(S_0 + S_1R_1 + S_2R_1^2) + S_0M_x + R_2 = C_2$, and the coefficients of $\mathfrak{a}_r^2 + \mathfrak{a}_r\Delta_r$ add up to $C_\alpha + C_\Psi + S_0(M_x+2) + C_6 = C_3$. Integrating in time, we conclude

$$\mathfrak{a}_s \le \mathfrak{a}_0 + \int_0^s \Big(C_2\|Z_r\|_2 + C_3\big(\mathfrak{a}_r^2 + \mathfrak{a}_r\Delta_r\big)\Big)dr.$$

Using Equation (120) to bound $\mathfrak{a}_0$, Section M.2.2 to bound $\|Z_r\|_2$, and (144) to bound the error term, and then integrating the resulting exponential terms, we obtain

$$\mathfrak{a}_s \le C_0\|Z_0\|_2 + \frac{C_2C_L}{\rho_1}\|Z_0\|_2 e^{\rho_1 s} + \frac{C_2C_4 + 2C_3K^2}{2\rho_1}\|Z_0\|_2^2 e^{2\rho_1 s}.$$

We now use $1 \le e^{\rho_1 s}$ in the first term and $\|Z_0\|_2 e^{\rho_1 s} \le \delta$ from (142) in the last term. Therefore,

$$\mathfrak{a}_s \le \left(C_0 + \frac{C_2C_L}{\rho_1}\right)\|Z_0\|_2 e^{\rho_1 s} + \frac{C_2C_4 + 2C_3K^2}{2\rho_1}\,\delta\,\|Z_0\|_2 e^{\rho_1 s}.$$

By the definition of K, the first coefficient is equal to $K/2$. Also, from the fourth term in the definition of δ^*, we have

$$\delta \le \frac{\rho_1}{8C_3K}\left(1 + \frac{C_2C_4}{C_3K^2}\right)^{-1} \le \frac{\rho_1 K}{2(C_2C_4 + 2C_3K^2)}.$$

Hence, the second coefficient is at most $K/4$. Combining these two bounds, we get

$$\mathfrak{a}_s \le \frac{3}{4}K\,\|Z_0\|_2 e^{\rho_1 s}, \qquad \text{for all } s \le S.$$

In particular $\mathfrak{a}_S < K\|Z_0\|_2 e^{\rho_1 S}$ strictly, since $\|Z_0\|_2 > 0$. Suppose that $S < s_\delta$. By continuity of $s \mapsto \mathfrak{a}_s$, the inequality $\mathfrak{a}_r \le K\|Z_0\|_2 e^{\rho_1 r}$ then holds on a small interval to the right of S, which contradicts the maximality of S. Therefore $S = s_\delta$. □

Now we give the proof of Theorem 10.

Proof of Theorem 10. We use the constants C_4 and δ_* of Lemma 23, the constant C_L of Lemma 22, and $\gamma = \sqrt{\nu_1/q}$, $\gamma' = \sqrt{q/\nu_2}$. We define

$$C_5 := \max\{1,\gamma\}\cdot\max\{1+\gamma,\ 1+\gamma',\ \tfrac{1}{2}(1+\gamma)\}$$

and

$$C_\star := 2\max\Big\{C_5,\ \max\{1,\gamma\}C_4,\ \frac{C_L}{\rho_1} + \frac{C_4}{2\rho_1} + C_0 + C_\Psi K^2\Big\}R_1(S_0 + S_1R_1 + S_2R_1^2),\ c_\star := \min\Big\{\delta_*,\ \frac{1}{C_\star}\Big\}.$$

The reason for these choices will become clear in the course of the proof; what matters now is that none of them depends on $\|Z_0\|_2$ or δ, and that $c_\star \le \delta_*$, so that Lemma 23 applies for every δ with $\|Z_0\|_2 < \delta < c_\star$. Recall also that $e^{\rho_1 s_\delta} = \delta/\|Z_0\|_2$ by (121), and that $\bar{P}_1 := \bar{v}_1 \bar{v}_1^\top$ is the orthogonal projection onto $\bar{v}_1$.

Before going into the details we give the outlines of it. The proof follows the last two steps of the proof of Theorem 9. We write C_{s_δ} as a sum of four parts, one along $\bar{v}_1$ and other three small ones using the explicit modes of Lemma 22 and Lemma 23. Then we do the same for B_{s_δ}, and finally we prove the statement about the decoder map.

Now we measure the size of each part of C_{s_δ}. Let $C_{s_\delta} = \widetilde{C}_{s_\delta} + (C_{s_\delta} - \widetilde{C}_{s_\delta})$, and expand $\widetilde{C}_{s_\delta} = \sum_{i=1}^{d} \bar{v}_i c_i(s_\delta)$ in the eigenbasis of $\mathcal{C}$. Now by using Lemma 22(a), we can split the top row of $c_1(s_\delta)$ into its growing and decaying parts giving us

$$C_{s_\delta} = \underbrace{\tfrac{1}{2} e^{\rho_1 s_\delta}\, \bar{v}_1 \big(c_1(0) + \gamma b_1(0)\big)}_{\text{(i)}} + \underbrace{\sum_{i \ge 2} \bar{v}_i c_i(s_\delta)}_{\text{(ii)}} + \underbrace{\tfrac{1}{2} e^{-\rho_1 s_\delta}\, \bar{v}_1 \big(c_1(0) - \gamma b_1(0)\big)}_{\text{(iii)}} + \underbrace{C_{s_\delta} - \widetilde{C}_{s_\delta}}_{\text{(iv)}}.$$

We now bound each of the four parts. The first part (i) is the growing mode along $\bar{v}_1$. Every column of (i) is a multiple of $\bar{v}_1$, and by using Lemma 22(a) and $e^{\rho_1 s_\delta} = \delta/\|Z_0\|_2$, we can conclude that

$$\|\text{(i)}\|_{\mathrm{F}} = \tfrac{1}{2} e^{\rho_1 s_\delta} \big\| c_1(0) + \gamma b_1(0) \big\|_2 = \kappa_1 e^{\rho_1 s_\delta} = \frac{\kappa_1}{\|Z_0\|_2}\, \delta.$$

The second part (ii) collects the rows along $\bar{v}_2, \dots, \bar{v}_d$. Since the $\bar{v}_i$ are orthonormal, $\|\text{(ii)}\|_{\mathrm{F}}^2 = \sum_{i \ge 2} \|c_i(s_\delta)\|_2^2$. Now by using Lemma 22(b) and the Cauchy–Schwarz inequality, we can say that $(\|c\|_2 + \gamma\|b\|_2)^2 \le (1+\gamma^2)\|(c,b)\|_2^2 \le (1+\gamma)^2 \|(c,b)\|_2^2$, which further implies

$$\|\text{(ii)}\|_{\mathrm{F}}^2 \le e^{2\rho_2 s_\delta} (1+\gamma)^2 \sum_{i \ge 2} \|(c_i(0), b_i(0))\|_2^2 \le (1+\gamma)^2 e^{2\rho_2 s_\delta} \|Z_0\|_2^2.$$

By the definition of θ, we have $e^{\rho_2 s_\delta} = (e^{\rho_1 s_\delta})^{\rho_2/\rho_1} = (\delta/\|Z_0\|_2)^{1-\theta}$, so

$$\|\text{(ii)}\|_{\mathrm{F}} \le (1+\gamma)\, \|Z_0\|_2 \Big(\frac{\delta}{\|Z_0\|_2}\Big)^{1-\theta} = (1+\gamma)\, \delta \Big(\frac{\|Z_0\|_2}{\delta}\Big)^{\theta}.$$

The third part (iii) is the decaying mode along $\bar{v}_1$. Since $e^{-\rho_1 s_\delta} \le 1$ and $\|c_1(0)\|_2, \|b_1(0)\|_2 \le \|Z_0\|_2$,

$$\|\text{(iii)}\|_{\mathrm{F}} \le \tfrac{1}{2}(1+\gamma)\|Z_0\|_2.$$

The last part (iv) is the nonlinear error. By using (141) at $s = s_\delta$, we get

$$\|\text{(iv)}\|_{\mathrm{F}} \le C_4 \|Z_0\|_2^2 e^{2\rho_1 s_\delta} = C_4 \delta^2.$$

Note that only the first part (i) lies along $\bar{v}_1$, while the other three parts may point in any direction. Let us define

$$E := (1+\gamma)\, \delta \Big(\frac{\|Z_0\|_2}{\delta}\Big)^{\theta} + \tfrac{1}{2}(1+\gamma)\|Z_0\|_2 + C_4 \delta^2,$$

so that $\|\text{(ii)}\|_{\mathrm{F}} + \|\text{(iii)}\|_{\mathrm{F}} + \|\text{(iv)}\|_{\mathrm{F}} \le E$. Since $I_d - \bar{P}_1$ maps (i) to zero and $\|I_d - \bar{P}_1\|_{\mathrm{op}} = 1$, we can conclude that

$$\|(I_d - \bar{P}_1) C_{s_\delta}\|_{\mathrm{F}} \le \|\text{(ii)}\|_{\mathrm{F}} + \|\text{(iii)}\|_{\mathrm{F}} + \|\text{(iv)}\|_{\mathrm{F}} \le E,$$

and by the triangle inequality in the form $\|a+b\| \ge \|a\| - \|b\|$ with $a = \text{(i)}$ and $b = \text{(ii)}+\text{(iii)}+\text{(iv)}$,

$$\|C_{s_\delta}\|_{\mathrm{F}} \ge \frac{\kappa_1}{\|Z_0\|_2}\, \delta - E.$$

Now we prove the angle bound for C_{s_δ}. Dividing the two inequalities just obtained,

$$\sin \angle(C_{s_\delta}, \bar{v}_1) = \frac{\|(I_d - \bar{P}_1) C_{s_\delta}\|_{\mathrm{F}}}{\|C_{s_\delta}\|_{\mathrm{F}}} \le \frac{E}{\frac{\kappa_1}{\|Z_0\|_2}\delta - E} = \frac{\rho}{1-\rho}, \qquad \text{where} \qquad \rho := \frac{\|Z_0\|_2}{\kappa_1} \cdot \frac{E}{\delta},$$

provided $\rho < 1$, which we check now. Dividing the definition of E by δ,

$$\frac{E}{\delta} = (1+\gamma)\Big(\frac{\|Z_0\|_2}{\delta}\Big)^{\theta} + \tfrac{1}{2}(1+\gamma)\frac{\|Z_0\|_2}{\delta} + C_4\delta \le \max\{1+\gamma, C_4\}\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^{\theta} + \frac{\|Z_0\|_2}{\delta} + \delta\Big].$$

The bracket is exactly one term in (123), and $\max\{1+\gamma, C_4\} \le C_\star/2$ holds by the definition of $C_\star$ (note $C_5 \ge 1+\gamma$ and $R_1(S_0 + S_1R_1 + S_2R_1^2) \ge 1$ because $R_1 \ge 1$ and $S_1 \ge \sigma'(0) = 1$). So, by using (123) and $c_\star \le 1/C_\star$, we get

$$\rho \le \frac{C_\star}{2} \cdot \frac{\|Z_0\|_2}{\kappa_1}\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^{\theta} + \frac{\|Z_0\|_2}{\delta} + \delta\Big] \le \frac{C_\star}{2}\, c_\star \le \frac{1}{2}.$$

So $1 - \rho \ge \frac{1}{2}$. This implies $\|C_{s_\delta}\|_{\mathrm{F}} \ge \frac{\kappa_1}{\|Z_0\|_2}\delta(1-\rho) > 0$, so C_{s_δ} and

$$\sin\angle(C_{s_\delta}, \bar v_1) \le \frac{\rho}{1-\rho} \le 2\rho \le C_\star\, \frac{\|Z_0\|_2}{\kappa_1}\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^{\theta} + \frac{\|Z_0\|_2}{\delta} + \delta\Big].$$

This proves (124) for C_{s_δ}.

Now we do the same for B_{s_δ}. By Lemma 22(a) the top row of $\widetilde{B}_{s_\delta}$ splits as $b_1(s_\delta) = \frac{1}{2\gamma}e^{\rho_1 s_\delta}(c_1(0) + \gamma b_1(0)) - \frac{1}{2\gamma}e^{-\rho_1 s_\delta}(c_1(0) - \gamma b_1(0))$, so the decomposition of B_{s_δ} has the same four parts, with the following changes. The leading part is $\frac{1}{2\gamma}e^{\rho_1 s_\delta}\bar v_1(c_1(0) + \gamma b_1(0))$, whose norm is $\frac{1}{\gamma}\frac{\kappa_1}{\|Z_0\|_2}\delta$. The part along $\bar v_2, \dots, \bar v_d$ is bounded by $(1+\gamma')\delta(\|Z_0\|_2/\delta)^{\theta}$ which follows from Lemma 22(b) (this is where $\nu_2 > 0$ is used) and the same Cauchy–Schwarz step as before. The decaying part has norm at most $\frac{1}{2\gamma}(1+\gamma)\|Z_0\|_2$, and the nonlinear error is again at most $C_4\delta^2$. So the same computation as for C_{s_δ} gives

$$\sin\angle(B_{s_\delta}, \bar v_1) \le \frac{\rho'}{1-\rho'}$$

$$\begin{aligned}\rho' &:= \gamma\,\frac{\|Z_0\|_2}{\kappa_1}\Big[(1+\gamma')\Big(\frac{\|Z_0\|_2}{\delta}\Big)^{\theta} + \tfrac{1+\gamma}{2\gamma}\frac{\|Z_0\|_2}{\delta} + C_4\delta\Big] \\ &\le \max\{1,\gamma\}\max\{1+\gamma', \tfrac{1}{2}(1+\gamma), C_4\} \cdot \frac{\|Z_0\|_2}{\kappa_1}\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^{\theta} + \frac{\|Z_0\|_2}{\delta} + \delta\Big],\end{aligned}$$

where $\max\{1,\gamma\}\max\{1+\gamma', \frac{1}{2}(1+\gamma), C_4\}$ is at most $C_\star/2$ by the definitions of C_5 and $C_\star$. Hence $\rho' \le \frac{1}{2}$, $B_{s_\delta} \ne 0$, and $\sin\angle(B_{s_\delta}, \bar v_1) \le 2\rho'$, which is (124) for B_{s_δ}.

Finally we prove (125). By using Lemma 21(c), we can write $\dot\Psi_r = C_r m + f_r$ on $[0, s_\delta]$, where f_r is the closure error. Since the $C^1(\mathcal{A})$ norm is bounded by the $C^2(\mathcal{A})$ norm, and by using (144), we get

$$\|f_r\|_{C^1(\mathcal{A})} \le \|f_r\|_{C^2(\mathcal{A})} \le C_\Psi\big(\mathfrak{a}_r^2 + \mathfrak{a}_r\Delta_r\big) \le 2C_\Psi K^2\|Z_0\|_2^2 e^{2\rho_1 r} \qquad \text{for all } r \le s_\delta.$$

Now we integrate $\dot\Psi_r = C_r m + f_r$ from 0 to s_δ. As m does not depend on r, we can take it out of the integral and get

$$\Psi_{s_\delta} = \Psi_0 + \Big(\int_0^{s_\delta} C_r\, dr\Big)m + \int_0^{s_\delta} f_r\, dr.$$

So, the decoder map at the exit time is the initial map, plus the time-integrated moment $\int_0^{s_\delta} C_r\, dr$ applied to the fixed function m, plus the integrated closure error. The last term is easy to bound. By using the bound on f_r above, $\int_0^{s_\delta} e^{2\rho_1 r} dr \le e^{2\rho_1 s_\delta}/(2\rho_1)$ and $e^{\rho_1 s_\delta} = \delta/\|Z_0\|_2$ (the definition of the exit time), we can conclude that

$$\Big\|\int_0^{s_\delta} f_r\, dr\Big\|_{C^1(\mathcal{A})} \le \int_0^{s_\delta} \|f_r\|_{C^1(\mathcal{A})}\, dr \le \frac{2C_\Psi K^2}{2\rho_1}\|Z_0\|_2^2 e^{2\rho_1 s_\delta} = \frac{C_\Psi K^2}{\rho_1}\,\delta^2.$$

So, the main work is in the term $\int_0^{s_\delta} C_r\, dr$. We split it exactly as we split C_{s_δ} above, but now integrated in r. Recall that we wrote C_r as a sum of four parts: (i) the growing part of the top mode, which is $\bar v_1 \cdot \frac{1}{2}e^{\rho_1 r}\big(c_1(0) + \gamma b_1(0)\big)^{\top}$; (ii) the contribution of all the other modes; (iii) the

decaying part of the top mode; and (iv) the difference between the true moments and the linear ones. Integrating part (i) in r gives us the rank-one matrix $\bar{v}_1 z_{\Psi,\delta}^\top$, where we define

$$z_{\Psi,\delta}^\top := \int_0^{s_\delta} \tfrac{1}{2} e^{\rho_1 r}\, dr\, \big(c_1(0) + \gamma b_1(0)\big)^\top = \frac{e^{\rho_1 s_\delta} - 1}{\rho_1}\, \kappa_1\, \frac{\big(c_1(0) + \gamma b_1(0)\big)^\top}{\|c_1(0) + \gamma b_1(0)\|_2},$$

and $\kappa_1 = \frac{1}{2}\|c_1(0) + \gamma b_1(0)\|_2$ as before. Note that $z_{\Psi,\delta}$ points in the same direction as $c_1(0) + \gamma b_1(0)$, and only its length depends on δ. By using $e^{\rho_1 s_\delta} - 1 \le e^{\rho_1 s_\delta} = \delta/\|Z_0\|_2$ and $\kappa_1 \le \frac{1}{2}(1+\gamma)\|Z_0\|_2$, we get

$$\|z_{\Psi,\delta}\|_2 \le \frac{\kappa_1}{\rho_1}\, \frac{\delta}{\|Z_0\|_2} \le \frac{1+\gamma}{2\rho_1}\, \delta.$$

Let R denote the integral over $[0, s_\delta]$ of the remaining three parts, so that

$$\int_0^{s_\delta} C_r\, dr = \bar{v}_1 z_{\Psi,\delta}^\top + R, \qquad R := \int_0^{s_\delta} \big[(\mathrm{ii})_r + (\mathrm{iii})_r + (\mathrm{iv})_r\big]\, dr.$$

To bound R we use the three bounds obtained above for C_{s_δ}, with s_δ replaced by r. These bounds say that, for all $0 \le r \le s_\delta$,

$$\|(\mathrm{ii})_r\|_{\mathrm{F}} \le (1+\gamma)\|Z_0\|_2 e^{\rho_2 r}, \qquad \|(\mathrm{iii})_r\|_{\mathrm{F}} \le \frac{1+\gamma}{2}\|Z_0\|_2 e^{-\rho_1 r}, \qquad \|(\mathrm{iv})_r\|_{\mathrm{F}} \le C_4 \|Z_0\|_2^2 e^{2\rho_1 r},$$

where the last one follows from (141). We now integrate each of them. For the first, we use $\int_0^{s_\delta} e^{\rho_2 r} dr \le e^{\rho_2 s_\delta}/\rho_2$ together with $e^{\rho_2 s_\delta} = (e^{\rho_1 s_\delta})^{\rho_2/\rho_1} = (\delta/\|Z_0\|_2)^{\rho_2/\rho_1}$, and get

$$\int_0^{s_\delta} \|(\mathrm{ii})_r\|_{\mathrm{F}}\, dr \le \frac{1+\gamma}{\rho_2}\|Z_0\|_2 \Big(\frac{\delta}{\|Z_0\|_2}\Big)^{\rho_2/\rho_1} = \frac{1+\gamma}{\rho_2}\, \delta \Big(\frac{\|Z_0\|_2}{\delta}\Big)^{\theta}, \qquad \theta := 1 - \frac{\rho_2}{\rho_1}.$$

So, the lower modes contribute a term which is smaller than δ by the factor $(\|Z_0\|_2/\delta)^\theta$, and $\theta > 0$ exactly because the top mode grows strictly faster than the second one. For the second bound, we use $\int_0^{s_\delta} e^{-\rho_1 r} dr \le 1/\rho_1$, and the decaying part of the top mode contributes at most $\frac{1+\gamma}{2\rho_1}\|Z_0\|_2$. For the third bound, we use $\int_0^{s_\delta} e^{2\rho_1 r} dr \le e^{2\rho_1 s_\delta}/(2\rho_1)$ and $e^{2\rho_1 s_\delta} = \delta^2/\|Z_0\|_2^2$, so the nonlinear error contributes at most $\frac{C_4}{2\rho_1}\delta^2$. Adding the three contributions, we can conclude that

$$\|R\|_{\mathrm{F}} \le \frac{1+\gamma}{\rho_2}\, \delta \Big(\frac{\|Z_0\|_2}{\delta}\Big)^{\theta} + \frac{1+\gamma}{2\rho_1}\|Z_0\|_2 + \frac{C_4}{2\rho_1}\delta^2.$$

Now we put everything together. Substituting the splitting of $\int_0^{s_\delta} C_r\, dr$ into the expression for Ψ_{s_δ}, we get

$$\Psi_{s_\delta} - \bar{v}_1 z_{\Psi,\delta}^\top m = \Psi_0 + R\, m + \int_0^{s_\delta} f_r\, dr.$$

We bound the three terms on the right in $C^1(\mathcal{A})$. For Ψ_0, we use (120), which gives $\|\Psi_0\|_{C^1(\mathcal{A})} \le C_0\|Z_0\|_2$. For Rm, as R is a constant matrix, we have $\|Rm\|_{C^1(\mathcal{A})} \le \|R\|_{\mathrm{op}}\|m\|_{C^1(\mathcal{A})} \le \|R\|_{\mathrm{F}}\|m\|_{C^1(\mathcal{A})}$, and by the same computation as in the proof of Lemma 23 (value bounded by $S_0 R_1$ and first derivative bounded by $S_1 R_1^2$), we have $\|m\|_{C^1(\mathcal{A})} \le R_1(S_0 + S_1 R_1) \le R_1(S_0 + S_1 R_1 + S_2 R_1^2)$. For the last term, we already showed the bound $\frac{C_\Psi K^2}{\rho_1}\delta^2$. So, we can conclude that

$$\big\|\Psi_{s_\delta} - \bar{v}_1 z_{\Psi,\delta}^\top m\big\|_{C^1(\mathcal{A})} \le C_0\|Z_0\|_2 + \|R\|_{\mathrm{F}}\, R_1\big(S_0 + S_1 R_1 + S_2 R_1^2\big) + \frac{C_\Psi K^2}{\rho_1}\delta^2.$$

Now we insert the bound on $\|R\|_{\mathrm{F}}$ and collect the constants. We factor one power of δ out of every term, writing $\|Z_0\|_2 = \delta \cdot (\|Z_0\|_2/\delta)$ and $\delta^2 = \delta \cdot \delta$. This sorts the right-hand side into three groups. The first group carries $(\|Z_0\|_2/\delta)^\theta$ and comes from the lower modes; its coefficient is $\frac{1+\gamma}{\rho_2} R_1(S_0 + S_1 R_1 + S_2 R_1^2)$. The second group carries $\|Z_0\|_2/\delta$ and comes from the initial map Ψ_0 and the decaying part of the top mode; its coefficient is $C_0 + \frac{1+\gamma}{2\rho_1} R_1(S_0 + S_1 R_1 + S_2 R_1^2)$. The third group carries δ and comes from the nonlinear errors; its coefficient is $\frac{C_4}{2\rho_1} R_1(S_0 + S_1 R_1 + S_2 R_1^2) + \frac{C_\Psi K^2}{\rho_1}$.

Since $\rho_2 \le \rho_1$, we have $\frac{1}{2\rho_1} \le \frac{1}{\rho_2}$, and so each of these three coefficients, and also the coefficient $\frac{1+\gamma}{2\rho_1}$ in the bound on $\|z_{\Psi,\delta}\|_2$, is bounded by

$$C_\star := C_0 + \frac{(1+\gamma)\big(1 + R_1(S_0 + S_1R_1 + S_2R_1^2)\big)}{\rho_2} + \frac{C_4R_1(S_0 + S_1R_1 + S_2R_1^2) + 2C_\Psi K^2}{2\rho_1}.$$

Hence

$$\big\|\Psi_{s_\delta} - \bar v_1 z_{\Psi,\delta}^\top m\big\|_{C^1(\mathcal{A})} \le C_\star\,\delta\Big[\Big(\frac{\|Z_0\|_2}{\delta}\Big)^\theta + \frac{\|Z_0\|_2}{\delta} + \delta\Big], \qquad \|z_{\Psi,\delta}\|_2 \le C_\star\,\delta.$$

All three terms inside the bracket go to zero when $\|Z_0\|_2/\delta \to 0$ and $\delta \to 0$, so the decoder map at the exit time is, up to a small error, the rank-one map $\bar v_1 z_{\Psi,\delta}^\top m$ aligned with the top direction $\bar v_1$. This proves (125) and finishes the proof. □

M.2.3 Recovering ordinary PCA in a local regime

The direction selected in Theorem 10 is the top eigenvector of $\mathcal{C}$, not of Σ, because the first-layer features are nonlinear. We now show that when the features are in the linear range of σ the two directions agree up to a small error, which connects the nonlinear result to Theorem 9. We consider $\sigma \in C_b^2(\mathbb{R})$ from Assumption 1, with the properties $\sigma(0) = 0$ and $\sigma'(0) = 1$ (as for $\tanh$).

Suppose $w_1 = \eta z$ for a feature scale $\eta \in (0, 1]$ and a compactly supported z with $\|z\|_2 \le R_z$ and

$$\mathbb{E}[zz^\top] = \kappa I_d, \qquad \kappa > 0. \tag{145}$$

By Taylor's theorem with $\sigma(0) = 0$ and $\sigma'(0) = 1$, and since $|\sigma''| \le S_2$, we have $|\sigma(t) - t| \le \frac{1}{2}S_2t^2$ for all t. We apply this with $t = \eta\langle z, x\rangle$, for which $|t| \le \eta R_z M_x$, and use $\|x\|_2 \le M_x$ and $\mathbb{E}_x[\langle z, x\rangle x] = \Sigma z$. This gives

$$h(\eta z) = \mathbb{E}_x\big[\sigma(\eta\langle z, x\rangle)\,x\big] = \eta\,\Sigma z + e(z),\ \|e(z)\|_2 \le \mathbb{E}_x\big[\tfrac{1}{2}S_2\eta^2\langle z, x\rangle^2\|x\|_2\big] \le \tfrac{1}{2}S_2R_z^2M_x^3\,\eta^2. \tag{146}$$

So, in a small neighborhood around 0, the feature map h is the linear map $\eta\Sigma$ up to an error of second order in η. Now we insert this into $\mathcal{C} = \int h(w_1)h(w_1)^\top\rho_{W,0}(dw) = \mathbb{E}[h(\eta z)h(\eta z)^\top]$. Expanding the product and using (145) for the leading term, we get

$$\mathcal{C} = \eta^2\,\Sigma\,\mathbb{E}[zz^\top]\,\Sigma + E_\eta = \kappa\eta^2\Sigma^2 + E_\eta, \qquad E_\eta := \mathbb{E}\big[\eta\Sigma z\,e(z)^\top + e(z)(\eta\Sigma z)^\top + e(z)e(z)^\top\big].$$

To bound the error matrix we use $\|\eta\Sigma z\|_2 \le \eta\mu_1R_z$ and the bound on $\|e(z)\|_2$ from (146). The two cross terms contribute at most $2\cdot\eta\mu_1R_z\cdot\frac{1}{2}S_2R_z^2M_x^3\eta^2 = \mu_1S_2R_z^3M_x^3\,\eta^3$, and the last term at most $\big(\frac{1}{2}S_2R_z^2M_x^3\eta^2\big)^2 = \frac{1}{4}S_2^2R_z^4M_x^6\,\eta^4$. Since $\eta \le 1$, we can conclude that

$$\|E_\eta\|_{\mathrm{op}} \le C_E\,\eta^3, \qquad C_E := \mu_1S_2R_z^3M_x^3 + \tfrac{1}{4}S_2^2R_z^4M_x^6.$$

So the leading part of $\mathcal{C}$ is of order η^2, while the error is of order η^3.

The matrix Σ^2 has the same eigenvectors as Σ, with eigenvalues $\mu_1^2 > \mu_2^2 \ge \dots$, so the top eigenvector of $\kappa\eta^2\Sigma^2$ is v_1 and its spectral gap is $\kappa\eta^2(\mu_1^2 - \mu_2^2)$. If η is small enough that the perturbation is at most half of the gap, that is,

$$C_E\,\eta \le \frac{\kappa}{2}(\mu_1^2 - \mu_2^2),$$

then the Davis–Kahan $\sin\Theta$ theorem applied to the symmetric matrices $\kappa\eta^2\Sigma^2$ and $\mathcal{C}$ gives

$$\sin\angle(\bar v_1, v_1) \le \frac{2\|E_\eta\|_{\mathrm{op}}}{\kappa\eta^2(\mu_1^2 - \mu_2^2)} \le \frac{2C_E}{\kappa(\mu_1^2 - \mu_2^2)}\,\eta. \tag{147}$$

So the direction learned first by the nonlinear model is the first principal component up to an error of order η, and in the limit $\eta \to 0$ the two coincide. The eigenvalues move as well: by Weyl's inequality, $|\nu_i - \kappa\eta^2\mu_i^2| \le C_E\eta^3$ for every i, so $\nu_2/\nu_1 \to \mu_2^2/\mu_1^2$ as $\eta \to 0$, and the exponent $\theta = 1 - \sqrt{\nu_2/\nu_1}$ in (124) tends to the exponent $1 - \mu_2/\mu_1$ in (111). Therefore, the PCA-first result does not require an exactly linear activation as in Section M.1. So, in a small neighborhood around 0, a bounded activation such as $\tanh$ gives approximately the same leading direction as linear activation function (PCA).

Remark 7. *If in addition $\sigma \in C^3$ with $\sigma''(0) = 0$ (as for $\tanh$), the remainder in the Taylor expansion improves to $|\sigma(t) - t| \le \frac{1}{6}S_3|t|^3$, the error matrix satisfies $\|E_\eta\|_{\mathrm{op}} \le C'_E\eta^4$ with $C'_E := \frac{1}{3}S_3R_z^4M_x^4\mu_1 + \frac{1}{36}S_3^2R_z^6M_x^8$, and the angle bound* (147) *improves to order η^2.*