BEYOND ACCURACY: DIAGNOSING FLOW MATCHING POSTERIOR ESTIMATION THROUGH ENDPOINT VARIABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow matching posterior estimation (FMPE) offers a flexible and increasingly important approach to simulation-based Bayesian inference, enabling expressive posterior approximation through learned continuous-time transport. Yet its performance can be difficult to diagnose through conventional posterior-accuracy metrics, as they cannot reliably distinguish expected sample-size-driven variability from systematic errors induced by suboptimal configurations. In this work, we develop a theoretically grounded diagnostic based on the $N^{-1/2}$ scaling of endpoint variability across training-set sizes. Building on the predicted scaling law, we construct a statistical test for assessing whether the sample-size-driven behavior of FMPE is consistent with theoretical expectations, and develop a pseudo-dataset strategy that reduces the need for extensive repeated simulator runs. Experiments, including domain-specific evaluations with complex epidemiological simulators, support the validity of our proposed diagnostic and show that it can identify suboptimal FMPE configurations.

1 INTRODUCTION

Neural posterior estimation (NPE) has emerged as a powerful approach to simulation-based Bayesian inference, learning a direct mapping from observations to an approximate posterior using simulated training data (Papamakarios & Murray, 2016; Lueckmann et al., 2017). Such approaches have been increasingly applied to scientific inference problems across diverse domains, including epidemiology (Radev et al., 2021; Wieland et al., 2025), astronomy (Hansen et al., 2022; Duan et al., 2025), and biology (Gonçalves et al., 2020; Boelts et al., 2023). Flow matching posterior estimation (FMPE) (Dax et al., 2023) extends this idea by learning a conditional continuous-time vector field that transports samples from a simple reference distribution to the posterior, providing a more expressive alternative to standard parametric posterior approximations. Recent studies like Gebhard et al. (2024); Moss et al. (2025) empirically demonstrate the advantage of FMPE over standard NPE.

Conventional posterior-accuracy metrics face a fundamental diagnostic challenge for FMPE: when an FMPE model performs poorly on training data, *is the observed error simply the sample-size-driven variability expected from a well-behaved estimator, or does it reflect a suboptimal pipeline?* Standard accuracy metrics such as Wasserstein distance, KL divergence, or classifier two-sample tests do not directly answer this question. Their values can be noisy, their trends can be difficult to distinguish from sampling variation, and even a decreasing error curve does not by itself establish that the learning procedure is behaving as expected. Thus, posterior accuracy alone provides an incomplete diagnostic of FMPE. This diagnostic challenge is particularly consequential for FMPE because its flexibility can require large simulation budgets (Dax et al., 2023).

Recent theoretical work has begun to characterize the sampling variability of flow-based posterior estimators. In particular, prior work (Mena et al., 2025) has established endpoint central limit behavior for flow matching when the velocity field is estimated through nonparametric regression, while related analyses of diffusion models (Wang et al., 2026) derive refined approximations to endpoint variance under simplified linear-regression assumptions. These results suggest that endpoint variability contains predictable statistical structure, but they do not yet provide a direct diagnostic theory for neural FMPE trained through its actual learning procedure.

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.

We build on this insight by establishing a theoretical path from neural FMPE training to endpoint variability. Under a neural tangent kernel (NTK) limit for a regularized two-layer neural network (Jacot et al., 2020; McNamara et al., 2025), we show that the terminal flow output satisfies an endpoint central limit theorem. For a fixed observation and initial state, independently fitted FMPE models exhibit endpoint variability of order $N^{-1/2}$, where N is the number of simulated training samples. This provides a theoretically specified scaling law for the variability of the learned posterior transport.

Our contributions are threefold. First, we establish an endpoint central limit theorem for regularized neural flow matching by connecting the infinite-width training dynamics of a two-layer neural network to kernel ridge regression and propagating its sampling fluctuations through the learned flow. Second, we derive the resulting $N^{-1/2}$ scaling law for disagreement between independently trained flow-matching estimators. Third, we use this scaling law to construct a practical diagnostic and evaluate whether it can distinguish ordinary sample-size-driven error from suboptimal pipeline designs that may not be apparent from conventional posterior-accuracy metrics alone. Besides, a pseudo dataset strategy is designed to bypass extensive repeated simulator runs.

We demonstrate the practical value of our diagnostic through a series of experiments. We first show that conventional posterior-accuracy metrics, particularly W_1, can exhibit misleading trends under suboptimal FMPE configurations, whereas our diagnostic reveals departures from the theoretically predicted scaling. We then evaluate the proposed pseudo-dataset strategy and show that it preserves the diagnostic distinction while reducing the need for repeated simulator runs. Finally, through domain-specific evaluations with two complex epidemiological simulators, we show that our diagnostic can distinguish a well-tuned FMPE configuration from a problematic one even when the well-tuned FMPE initially has worse posterior accuracy than a simpler NPE and becomes competitive only after substantially more training data are available.

2 PRELIMINARIES

This section introduces the notation and mathematical setup used throughout the paper. We first review the FMPE formulation and define the population and fitted endpoints that form the basis of our diagnostic. We then introduce neural tangent kernel and the RKHS ridge-regression representation used to characterize the fitted velocity field and develop the theoretical analysis in Section 4, and close with a review of related work.

2.1 FLOW MATCHING POSTERIOR ESTIMATION

Flow matching posterior estimation (FMPE) (Dax et al., 2023) learns an observation-conditioned velocity field whose ordinary differential equation (ODE) transports base noise to posterior samples. The field is trained by conditional flow matching, a regression onto the velocity of a prescribed path from a base sample to a prior draw, which places no invertibility or Jacobian constraints on the network.

In FMPE, parameter–observation pairs are generated according to $\Theta \sim \pi, X \sim P(\cdot \mid \Theta)$, where $\Theta \in \mathbb{R}^d$ denotes the parameter and $X \in \mathbb{R}^m$ the corresponding observation. Let $Z \sim q_0$ be base noise, independent of (Θ, X). Set $W = (Z, \Theta, X), \Delta = \Theta - Z, Y(s) = (1-s)Z + s\Theta, U(s) = (s, Y(s), X), s \in [0,1]$. The path $Y(s)$ is the straight line from the noise Z to the parameter Θ. Its velocity Δ is constant in s and serves as the regression target of conditional flow matching. The triple $U(s)$ is what the velocity network sees at path time s: the time, the current position, and the observation it is conditioned on. We call W a training triple. During sampling, the velocity field f generates the trajectory $\phi_t^f = z + \int_0^t f(s, \phi_s^f, x)\, ds$ and endpoint $R_{z,x}(f) = \phi_1^f$ for initial state z and observation x. $R_{z,x}$ is a map from velocity fields to points in $\mathbb{R}^d$. In the following, we mainly consider the endpoints of the population and fitted velocity fields, i.e., $R_\lambda = R_{z,x}(v_\lambda)$ and $\hat{R}_{N,\lambda} = R_{z,x}(\hat{v}_{N,\lambda})$. Given N training triples $W_1, \ldots, W_N$, we parameterize f by a neural network $f_M(u;\omega)$ of width M with parameters ω and train it on the regularized conditional flow-matching loss

$$\mathcal{F}_{N,M}(\omega) = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 \left\|\Delta_i - f_M(U_i(s);\omega)\right\|^2 ds + \lambda \sum_{r=1}^{M} \|c_r\|^2, \tag{1}$$

where Δ_i and $U_i(s)$ are defined from W_i, $c_1, \ldots, c_M \in \mathbb{R}^d$ are the output-layer weights contained in ω, and $\lambda > 0$ is a ridge parameter. The ridge penalty on the output layer is the network-level counterpart of the penalty $\lambda\|f\|^2_{\mathcal{H}}$ in the RKHS objective of Section 2.3, and it is what allows the trained network to be identified with a ridge-regression solution in Section 4.

2.2 Neural tangent kernel

The neural tangent kernel (NTK) (Jacot et al., 2020) describes how a neural network evolves under gradient-based training when viewed as a function rather than as a set of weights. Let $f_M(u;\omega) \in \mathbb{R}^d$ be a network of width M with parameters ω, trained by the gradient flow $\dot{\omega}(\tau) = -\nabla_\omega \mathcal{L}(\omega(\tau))$ on a loss $\mathcal{L}(\omega) = \sum_{i=1}^N \ell_i\big(f_M(u_i;\omega)\big)$ over training inputs $u_1, \ldots, u_N$, where τ denotes training time. The NTK is the $d \times d$ matrix-valued kernel

$$K_M^\tau(u,u') = \nabla_\omega f_M(u;\omega(\tau))\, \nabla_\omega f_M(u';\omega(\tau))^\top, \tag{2}$$

the Gram matrix of the parameter gradients of the outputs at u and u', which measures how much a gradient step that corrects the output at u' also changes the output at u. By the chain rule, $\partial_\tau f_M(u;\omega(\tau)) = -\sum_{i=1}^N K_M^\tau(u,u_i)\, \nabla \ell_i\big(f_M(u_i;\omega(\tau))\big)$, so the weights enter the function-space dynamics only through the kernel. The central result of NTK theory is that, under suitably scaled random initialization, K_M^0 converges to a deterministic kernel K as $M \to \infty$, and K_M^τ stays close to K throughout training because a wide network moves only slightly in parameter space (Jacot et al., 2020; Lee et al., 2019; Chizat et al., 2019). In this limit the network trains as a linear model in its parameters, or equivalently as a kernel method in the reproducing kernel Hilbert space of K.

2.3 Ridge regression on RKHS

To facilitate the theoretical analysis of the FMPE velocity field, we formulate the corresponding regularized regression problem in a vector-valued reproducing kernel Hilbert space (RKHS). The space $\mathcal{H}$ is equipped with kernel $k(u,u')$ (scalar) and $K(u,u') = k(u,u')I_d$, where point $u = (t,\vartheta,x)$, which contains time t, location ϑ in parameter space, and observation x. On this space, the evaluation operator and its adjoint are $E_u : \mathcal{H} \to \mathbb{R}^d$ and $E_u^* : \mathbb{R}^d \to \mathcal{H}$, having property $E_u f = f(u)$, $E_u^* y = k(\cdot,u)y$, $\langle E_u f, y\rangle = \langle f, E_u^* y\rangle_{\mathcal{H}}$ and $E_u E_u^* = K(u,u')$. Using these operators, we can measure how a certain training triple W contributes to the ridge regression through $C_W = \int_0^1 E_{U(s)}^* E_{U(s)} ds$ and $g_W = \int_0^1 E_{U(s)}^* \Delta ds$. They have population and average expressions $C = \mathbb{E}C_W$, $g = \mathbb{E}g_W$, $C_N = \sum_{i=1}^N C_{W_i}/N$ and $g_N = \sum_{i=1}^N g_{W_i}/N$. Under the regularized objective $J(f) = \mathbb{E}\int_0^1 \|\Delta - f(U(s))\|^2 ds + \lambda\|f\|^2_{\mathcal{H}}, f \in \mathcal{H}$, the unique minimizer for population and current training dataset are $v_\lambda = (C+\lambda I)^{-1} g$ and $\hat{v}_{N,\lambda} = (C_N + \lambda I)^{-1} g_N$, respectively. We define $A_\lambda = (C+\lambda I)^{-1}$ and $A_{N,\lambda} = (C_N + \lambda I)^{-1}$ for convenience of subsequent parts.

2.4 Related work

Endpoint variability. Diffusion and flow models trained on disjoint data splits can map the same noise to strikingly similar samples (Kadkhodaie et al., 2024; Zhang et al., 2024), and a substantial part of the learned sampling map can already be predicted by a Gaussian linear score (Wang & Vastola, 2024). Building on this linear structure, Wang et al. (2026) use random matrix theory to characterize the mean and variance of the sampling map across data splits, isolating the effects of spectral anisotropy, input inhomogeneity, and dataset size. From a statistical perspective, Mena et al. (2025) establish regularity, convergence rates, and central limit theorems for rectified-flow map estimators based on nonparametric regression or density estimation by linearizing the ODE solution map around the population velocity field. Our endpoint central limit theorem follows a related linearization strategy, but starts from a neural network trained under the regularized FMPE objective, which we connect to ridge regression through the NTK.

Bounds on posterior-accuracy metrics. A complementary line of work derives upper bounds on the distance between generated and target distributions. For diffusion models, Chen et al. (2023) control the total-variation error through the score-estimation error, while Oko et al. (2023) obtain minimax-optimal rates in total variation and W_1. For flow matching, Benton et al. (2024) derive W_2 bounds in terms of velocity-field error and its Lipschitz constant, Gao et al. (2024) provide

non-asymptotic W_2 bounds for continuous normalizing flows, and Fukumizu et al. (2025) show that flow matching attains nearly minimax-optimal Wasserstein rates. Closest to our setting, Zhai et al. (2025) extend such guarantees to flow-matching posterior samplers and establish W_2 consistency of the learned posterior. These results characterize expected posterior accuracy as a function of N, whereas our diagnostic asks whether the observed endpoint variability follows the scaling rate predicted by theory.

3 ENDPOINT VARIABILITY DIAGNOSTIC

In this section, we further elaborate on why conventional posterior-accuracy metrics are insufficient for diagnosing FMPE, and then introduce our endpoint-variability diagnostic. We use the theoretically predicted $N^{-1/2}$ scaling of endpoint variability to construct a regression-based test, and finally develop a pseudo-dataset strategy that reduces the need for repeated simulator runs.

As discussed in Section 1, conventional posterior-accuracy metrics do not directly distinguish poor performance caused by limited training data from that caused by a suboptimal FMPE configuration. This distinction is particularly important for FMPE, whose flexibility can require larger training sets than standard NPE methods to achieve comparable posterior accuracy (Dax et al., 2023). Moreover, posterior-accuracy metrics inherit variability from both the training data and the fitting procedure, so their observed values and trends across training-set sizes can be noisy or misleading (Fournier & Guillin, 2013; Bouthillier et al., 2021; Bischoff et al., 2024). A decreasing error curve, for example, does not necessarily imply that the underlying FMPE pipeline is behaving as intended, and existing posterior diagnostics may be insensitive to particular forms of suboptimal inference behavior (Hermans et al., 2022; Lemos et al., 2023). We illustrate such a misleading pattern in our experiments (Section 5.1), where conventional posterior-accuracy metrics continue to decrease with increasing training-set size even under suboptimal FMPE configurations.

We therefore consider a complementary diagnostic based on the variability of the learned FMPE endpoint across independently trained models. The key idea is not to ask only whether posterior error is small, but whether the variability of the learned estimator changes with training-set size in the manner predicted by theory. In Section 4, we establish sufficient conditions under which the endpoint satisfies the central limit theorem

$$\sqrt{N}(\hat{R}_{N,\lambda} - R_\lambda) \Rightarrow \mathcal{N}(0, V_\lambda(z; x)) \tag{3}$$

Accordingly, for large N, the fitted endpoint $\hat{R}_{N,\lambda}$ is approximately normally distributed around its population counterpart R_λ, with endpoint fluctuations of order $N^{-1/2}$. This scaling law provides a theoretically specified reference behavior for FMPE.

Interpretation of our endpoint-variability diagnostic. Systematic departures from the predicted scaling should warrant further scrutiny of the FMPE configuration; conversely, agreement with the predicted scaling provides no evidence, from this diagnostic alone, that the observed endpoint variability is inconsistent with the theoretical regime, but does not establish that the fitted posterior is unbiased or globally optimal. Our experiments examine whether departures from this scaling can be used in practice to identify suboptimal FMPE configurations.

Regression-based scaling test. Motivated by the endpoint CLT, we construct a regression-based test of the predicted $N^{-1/2}$ scaling. For a one-dimensional endpoint and a fixed training-set size N, we independently fit the FMPE model n times and compute the sample variance of the resulting endpoint estimates, $S_N^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(\hat{R}_{N,\lambda}^{(j)} - \bar{R}_{N,\lambda}\right)^2$, where $\hat{R}_{N,\lambda}^{(j)}$ is the endpoint estimate from the jth independent fit and $\bar{R}_{N,\lambda}$ is their sample mean. Under Equation 3, the standard deviation of the endpoint scales as $N^{-1/2}$, and hence its variance scales as N^{-1}. To test this scaling empirically, we parameterize the endpoint standard deviation more generally as proportional to $N^{-\alpha}$, so that its variance is proportional to $N^{-2\alpha}$. This motivates the regression

$$\log S_N^2 = -2\alpha \log N + \log V + b_n + \epsilon_N, \tag{4}$$

where the theoretically predicted value is $\alpha = 1/2$. Here, b_n accounts for the mean of the log-transformed chi-squared sampling term, and ϵ_N denotes the corresponding centered sampling fluc-

tuation. For moderately large n (we use $n = 20$ in our experiments), this fluctuation can be approximated by a normal distribution, allowing us to test $H_0 : \alpha = \frac{1}{2}$.

For multidimensional endpoints, the same regression-based diagnostic can be applied separately to multiple endpoint components, allowing the scaling behavior of different inferred quantities to be assessed individually.

Pseudo-dataset strategy. The regression-based test described above requires fitting FMPE models to multiple independently simulated datasets at each training-set size, which can be computationally prohibitive when simulator evaluations are expensive or the simulator is not readily accessible (Krouglova et al., 2026). To reduce this requirement, we develop a pseudo-dataset strategy that evaluates the scaling behavior using datasets generated from an FMPE model fitted to the available simulations.

Algorithm 1 Pseudo-dataset regression test

Require: $D_N^{\text{true}} = \{(\theta_i, x_i)\}_{i=1}^N$; FMPE configuration $\mathcal{C}$; pseudo sizes $N_1 < \cdots < N_K$; replicates n; probe (z, x)

Train FMPE $\hat{v}$ on D_N^{true} under $\mathcal{C}$ ▷ learned $\hat{p}(\theta \mid x)$
for $k = 1, \ldots, K$; $j = 1, \ldots, n$ **do**
  $\tilde{D}_{N_k}^{(j)} \leftarrow \emptyset$
  for $i = 1, \ldots, N_k$ **do** ▷ no simulator call
    Draw $\tilde{x}_i$ from $\{x_1, \ldots, x_N\}$, $\tilde{z}_i \sim q_0$
    $\tilde{\theta}_i \leftarrow R_{\tilde{z}_i, \tilde{x}_i}(\hat{v})$
    $\tilde{D}_{N_k}^{(j)} \leftarrow \tilde{D}_{N_k}^{(j)} \cup \{(\tilde{\theta}_i, \tilde{x}_i)\}$
  Re-train FMPE $\hat{v}_{N_k}^{(j)}$ on $\tilde{D}_{N_k}^{(j)}$ under $\mathcal{C}$
  $\hat{R}_{N_k,\lambda}^{(j)} \leftarrow R_{z,x}(\hat{v}_{N_k}^{(j)})$
$S_{N_k}^2 \leftarrow \frac{1}{n-1} \sum_j \left(\hat{R}_{N_k,\lambda}^{(j)} - \bar{R}_{N_k,\lambda}\right)^2$ for all k
Fit $\log S_{N_k}^2 \sim -2\alpha \log N_k + c$; Test $H_0 : \alpha = 1/2$
return $\hat{\alpha}$, p-value; pass if H_0 is not rejected

The approach is motivated by the hypothesis that suboptimal FMPE configurations may exhibit inconsistent endpoint-scaling behavior even under pseudo-data generated from the fitted model. Specifically, we first train an FMPE model on the available dataset D_N^{true} to obtain an approximate posterior distribution. We then use the fitted model to generate independent pseudo-datasets $\tilde{D}_{N_k}^{(j)}$ across a range of training-set sizes N_k. FMPE models are re-trained on these pseudo-datasets, and the resulting endpoint estimates are used in the same regression-based scaling test described above (Algorithm 1). This procedure avoids additional simulator calls while allowing the endpoint-variability behavior of the FMPE configuration to be assessed across training-set sizes. In our experiments, the pseudo-dataset procedure preserves the empirical distinction between well-performing and suboptimal FMPE configurations.

4 Theory

In this section, we briefly introduce the theory underlying the endpoint CLT and outline the simplified proof path. The formal assumptions (B.1–B.3) used in the results below are stated in Appendix B; here we summarize the main proof ideas and refer to the appendix for the complete technical conditions. In Section 4.1, we connect FMPE training to ridge regression over the space $\mathcal{H}$. Through an NTK argument and convergence analysis, we show that a two-layer SiLU neural network converges, in the appropriate limiting regime, to the previously defined ridge-regression solution $\hat{v}_{N,\lambda} = (C_N + \lambda I)^{-1} g_N$. We then analyze this ridge-regression estimator in Section 4.2 and establish the endpoint CLT. Together, Sections 4.1 and 4.2 provide a theoretical path from FMPE training to the endpoint CLT.

4.1 Neural network training and ridge limit

Under the regularized objective, the trained FMPE network can be connected to the ridge-regression solution in $\mathcal{H}$ in the infinite-width and long-training-time limits. To establish this connection, we show that the training dynamics of an infinite-width two-layer neural network under the regularized FMPE loss can approximate the gradient flow of the corresponding ridge-regression objective in $\mathcal{H}$. As the training time increases, this kernel gradient flow converges to the ridge-regression solution $\hat{v}_{N,\lambda}$. We summarize the main ingredients of this argument below; the complete proof is provided in Appendix C.

We begin with the neural tangent kernel at initialization. We consider a two-layer SiLU neural network with zero output initialization, as specified in Appendix C. Its neural tangent kernel is

defined as $K_M(u, u') = \nabla_\omega f_M(u) \nabla_\omega f_M(u')^\top$, where M denotes the network width and ω the network parameters. As the width tends to infinity, this kernel converges to a fixed limiting kernel K at initialization.

Proposition 1. *Under Assumption B.2, the tangent kernel K_M at initialization converges almost surely, as $M \to \infty$, to a deterministic, continuous, positive semidefinite kernel K.*

The limiting kernel K is precisely the kernel defining the RKHS $\mathcal{H}$ used in the ridge-regression analysis of Section 4.2. We next show that the neural tangent kernel remains close to K during training, i.e., $K^\tau \approx K$, where τ denotes training time. The evolution of the neural network can be written as $f_{N,M,\tau}(u) = -2 \int_0^\tau e^{-2\lambda(\tau - t)} \int K^\tau \big(f_{N,M,t}(U) - \Delta\big)\, d\mu_N\, dt$. The corresponding functional $\tilde{v}_{N,\tau}$ in $\mathcal{H}$ evolves according to the same gradient-flow equation with K^τ replaced by its limiting kernel K. Thus, in the infinite-width limit, the neural-network dynamics follow the corresponding kernel gradient flow. The functional $\tilde{v}_{N,\tau}$ converges exponentially to the empirical ridge-regression solution $\hat{v}_{N,\lambda}$ as $\tau \to \infty$. Stability of the ODE solution then transfers this convergence from the velocity field to the resulting endpoint. Combining these results yields the following theorem.

Theorem 1. *Let Assumptions B.1–B.3 hold. For a two-layer SiLU neural network $f_{N,M,\tau}$ with zero output initialization trained under $\mathcal{F}_{N,M}(\omega)$, $\lim_{\tau\to\infty} \lim_{M\to\infty} R_{z,x}(f_{N,M,\tau}) = R_{z,x}(\hat{v}_{N,\lambda})$. Thus, in the infinite-width and long-training-time limits, the endpoint produced by the trained neural network converges to the endpoint produced by the empirical ridge-regression field.*

4.2 Ridge regression to endpoint CLT

For both the population model and the model fitted to a finite training dataset, the endpoint is obtained through the same sequence of operations. The population distribution and empirical training dataset first determine (C, g) and (C_N, g_N), respectively, as defined in Section 2.3. A ridge solve maps these quantities to the velocity fields v_λ and $\hat{v}_{N,\lambda}$, and an ODE solve maps the velocity fields to the endpoints R_λ and $\hat{R}_{N,\lambda}$. Our goal is to characterize how the sampling difference between (C_N, g_N) and (C, g) propagates through these steps. The argument has three parts. First, the sampling fluctuation is represented through an average score $\bar{\eta}_N$. Second, this score determines the velocity-field error $h_N = \hat{v}_{N,\lambda} - v_\lambda$ to first order. Third, linearization of the endpoint map propagates this error to $\hat{R}_{N,\lambda} - R_\lambda$. We summarize these steps below; rigorous statements and proofs are provided in Appendix D and the subsequent sections.

We define the score of a training triple at the population field as $\eta_\lambda(W) = g_W - C_W v_\lambda - \lambda v_\lambda$. It measures the contribution of an individual training triple to the sampling fluctuation around the population normal equation $(C + \lambda I)v_\lambda = g$.

Lemma 1. *Let Assumptions B.1 and B.3 hold. The score η_λ is a centred, square-integrable random element of $\mathcal{H}$: $\mathbb{E}\eta_\lambda(W) = 0, \mathbb{E}\|\bar{\eta}_N\|_{\mathcal{H}}^2 = \frac{s_\lambda^2}{N}$, where $s_\lambda^2 = \mathbb{E}\|\eta_\lambda(W)\|_{\mathcal{H}}^2 < \infty$. Thus, $\bar{\eta}_N = O_p(N^{-1/2})$.*

The average score is directly related to the velocity-field error. In particular, $h_N = A_{N,\lambda}\bar{\eta}_N$, and replacing the sample-dependent inverse $A_{N,\lambda}$ by its population counterpart A_λ introduces only a higher-order error.

Lemma 2. *Let Assumptions B.1 and B.3 hold. The field error satisfies $h_N = A_\lambda \bar{\eta}_N + o_p(N^{-1/2})$, and therefore $\|h_N\|_{\mathcal{H}} = O_p(N^{-1/2})$.*

We next propagate this field error through the ODE solve. The endpoint map is differentiable at the population field v_λ, with first-order derivative given by the bounded linear map L.

Lemma 3. *Let Assumptions B.1 and B.3 hold and fix (z, x). For $h \in \mathcal{H}$ sufficiently small, $R_{z,x}(v_\lambda + h) - R_{z,x}(v_\lambda) = Lh + r(h)$, where $\|r(h)\| = o(\|h\|_{\mathcal{H}})$ as $\|h\|_{\mathcal{H}} \to 0$. Consequently, when $h = h_N = O_p(N^{-1/2})$, the remainder is $o_p(N^{-1/2})$.*

Combining Lemmas 2 and 3 gives $\hat{R}_{N,\lambda} - R_\lambda = \frac{1}{N}\sum_{i=1}^N \xi_\lambda(W_i; z, x) + o_p(N^{-1/2})$, where $\xi_\lambda(W_i; z, x) = L A_\lambda \eta_\lambda(W_i)$ are i.i.d., centred, and square-integrable. The ordinary multivariate central limit theorem then gives our main result.

Theorem 2. *Let Assumptions B.1 and B.3 hold. For each fixed (z, x), $\sqrt{N}\left(\hat{R}_{N,\lambda} - R_\lambda\right) \Rightarrow \mathcal{N}\left(0, V_\lambda(z; x)\right)$. The covariance $V_\lambda(z; x) = \mathbb{E}\left[\xi_\lambda \xi_\lambda^\top\right]$ is finite and may be singular.*

5 EXPERIMENTS

In this section, we empirically evaluate the proposed endpoint-variability diagnostic. We first compare it with the Wasserstein-1 distance (W_1) (Villani, 2009), a commonly used sample-based measure of posterior discrepancy, and show that W_1 can exhibit misleading trends under suboptimal FMPE configurations whereas our diagnostic reveals departures from the theoretically predicted scaling (Section 5.1). We then examine more systematically how the diagnostic responds to different types and severities of suboptimal FMPE configurations (Section 5.2). Next, we evaluate the proposed pseudo-dataset strategy and assess whether it preserves the diagnostic behavior observed with independently simulated datasets (Section 5.3). Finally, we apply the diagnostic to two complex, domain-specific epidemiological simulation studies to examine whether it can distinguish limited training data from suboptimal FMPE configurations (Section 5.4). The detailed description (e.g., simulators, neural network architecture and regression test) of these experiments can be found in Appendix F.

5.1 COMPARISON OF W_1 AND OUR DIAGNOSTIC

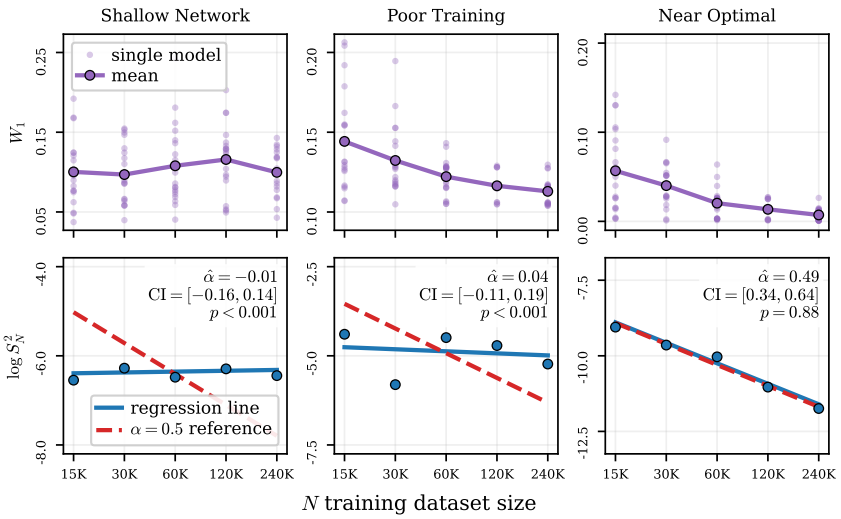

Figure 1: Comparison of W_1 and our endpoint-variability diagnostic across three FMPE configurations. Columns correspond to a shallow network, poor training, and a near-optimal configuration. The shallow-network setting uses a reduced network width and depth, while the poor-training setting uses fewer optimization steps.

As shown in Figure 1, the conventional metric W_1 can exhibit misleading trends under two suboptimal FMPE configurations, whereas our diagnostic reveals clear departures from the theoretically predicted scaling. The left column shows W_1 and $\log S_N^2$ as functions of training-set size for a shallow-network configuration, in which the network width is reduced to 32 and the depth to 2, compared with 6 layers of width 512 in the near-optimal configuration. In this setting, W_1 fluctuates around an approximately constant level with substantial variability. Although this plateau is apparent when averaging W_1 across repeated fits, a single W_1 value at each N, as is commonly reported, exhibits a decreasing trend with probability greater than 50% in our experiment. Thus, the variability of W_1 can obscure the underlying lack of improvement.

A similar issue arises under the poor-training configuration, where the model is trained for fewer optimization steps ($20{,}000$ versus $60{,}000$ in the near-optimal configuration). In this case, even the expected W_1 continues to decrease with training-set size despite the model reaching an optimization floor at larger N. Relative to the near-optimal configuration, incomplete optimization results in an approximately 0.1 increase in W_1. Without access to the near-optimal reference curve, the decreasing trend in W_1 alone would therefore provide little indication of the underlying training issue.

In contrast, our endpoint-variability diagnostic distinguishes these configurations through their scaling behavior. As shown in the bottom-left and bottom-center panels, both the shallow-network and

poor-training configurations strongly reject the null hypothesis $\alpha = 1/2$, with $p < 0.001$ in both cases. Their $\log S_N^2$ values remain approximately constant across training-set sizes, yielding fitted values of α close to zero. By comparison, the near-optimal FMPE exhibits a fitted rate of $\hat{\alpha} \approx 0.5$, consistent with the theoretically predicted $N^{-1/2}$ scaling.

5.2 Sensitivity to suboptimal FMPE configurations

Building on the examples in Section 5.1, we next examine more systematically how our diagnostic responds to different types and severities of suboptimal FMPE configurations. As shown in Table 1, we retain the high-severity poor-training and shallow-network settings considered above, and additionally evaluate milder perturbations as well as two further failure modes: an inaccurate ODE solver and an inappropriate flow design.

Severity	Shallow Network	Poor Training	Inaccurate ODE solver	Wrong Flow Design
Low	$0.37_{(0.09)}$	$0.41_{(0.24)}$	$0.44_{(0.43)}$	$0.48_{(0.79)}$
Middle	$0.11_{(<10^{-3})}$	$0.22_{(<10^{-3})}$	$0.33_{(0.03)}$	$0.35_{(0.05)}$
High	$0.01_{(<10^{-3})}$	$0.04_{(<10^{-3})}$	$-0.03_{(<10^{-3})}$	$0.10_{(<10^{-3})}$

Table 1: Sensitivity of the endpoint-variability diagnostic to suboptimal FMPE configurations. Each entry reports the fitted scaling parameter $\hat{\alpha}$, with the corresponding p-value for testing $H_0 : \alpha = 1/2$ shown in parentheses. Columns correspond to different types of suboptimal configuration, and rows indicate increasing severity.

For the ODE solver, we reduce numerical accuracy by decreasing the number of sampling steps to 100, 10, and 3 for the low-, middle-, and high-severity settings, respectively. For the flow-design perturbation, we consider target distributions with increasingly large variance (2, 10, and 100 for low, middle, and high severity) without standardizing the targets beforehand. For poor training, we use 50,000, 30,000, and 20,000 optimization steps for the low-, middle-, and high-severity settings. For network capacity, the low- and middle-severity settings use width-depth combinations of $(256, 5)$ and $(64, 3)$, respectively, while the high-severity setting corresponds to the shallow network considered in Section 5.1.

Across these perturbations, our diagnostic rejects the theoretical scaling for most middle- and high-severity configurations, with reported p-values below 0.05. In contrast, the fitted $\hat{\alpha}$ values for the low-severity settings are generally closer to the theoretical value of 0.5, consistent with these configurations being closer to the near-optimal setting. Importantly, this suggests that the diagnostic is not overly sensitive to small deviations from the idealized configuration.

More broadly, the fitted $\hat{\alpha}$ values tend to move toward 0.5 as the severity of the perturbation decreases, indicating a gradual transition from clear scaling violations toward behavior consistent with the theoretical regime. Although the shallow-network setting produces some of the strongest departures in Table 1, direct comparison of effect size across configuration classes should be interpreted cautiously because the severity levels are not calibrated to be equivalent across the different perturbations.

5.3 Pseudo-dataset strategy

We next evaluate whether the pseudo-dataset strategy preserves the diagnostic behavior observed with independently simulated training datasets. We assume that only 15,000 simulator-generated samples are initially available and apply Algorithm 1 to generate pseudo-datasets for training-set sizes $N \in \{15\text{K}, 30\text{K}, 60\text{K}, 120\text{K}, 240\text{K}\}$. For each N, we generate 20 independent pseudo-datasets, retrain the FMPE model on each dataset, and apply the same regression-based scaling test.

As shown in Figure 2, the pseudo-dataset procedure preserves the distinction between the near-optimal and high-severity poor-training configurations. The near-optimal configuration remains consistent with the theoretically predicted scaling, whereas the poor-training configuration shows a clear departure from it. Through this strategy, we achieve substantial computational savings. Repeated simulator calls account for approximately 4.9% of the total runtime in this experiments, and 38.8% in the following large-scale simulation. Replacing these calls with sam-

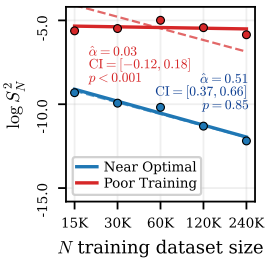

Figure 2: Pseudo-dataset tests. The FMPE models are retrained on pseudo-datasets using the same pipeline configuration.

pling from the fitted FMPE model reduces the overall runtime by approximately a factor of 1.04 for this experiment, and 1.5 for large-scale experiments.

5.4 Domain-specific evaluation with complex epidemiological simulators

We next evaluate our diagnostic using two more complex, domain-specific epidemiological simulators, CRKP (Chatha et al., 2026) and BDEISS-CT (Xie et al., 2026). CRKP is a stochastic, discrete-time Susceptible–Infected (SI) model for healthcare-associated transmission of carbapenem-resistant *Klebsiella pneumoniae*, with heterogeneous transmission determined by patient location within the hospital, including floor and room. BDEISS-CT is a multi-type birth–death phylodynamic model that combines Birth–Death Exposed Infectious (BDEI), Birth–Death with Super-Spreading (BDSS), and Birth–Death with Contact-Tracing (BD-CT) components. Together, these simulators provide a more challenging testbed for FMPE in settings with higher-dimensional state and parameter spaces and more complex posterior structure.

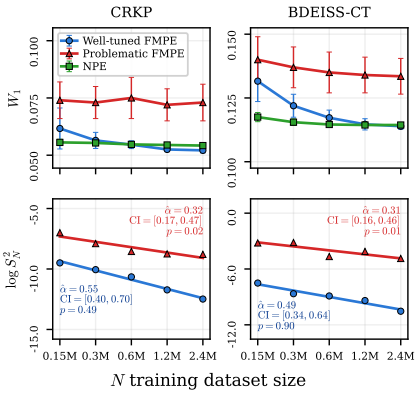

Figure 3: W_1 and our diagnostic across two complex epidemiological simulation studies. We compare a standard NPE, a well-tuned FMPE, and a problematic FMPE on datasets generated by CRKP (Chatha et al., 2026) and BDEISS-CT (Xie et al., 2026).

As shown in Figure 3, a well-tuned FMPE has higher W_1 than a standard NPE with a normalizing-flow posterior estimator at smaller training-set sizes. For CRKP, FMPE becomes competitive with and then slightly outperforms NPE only once the training-set size exceeds approximately 0.6M. For BDEISS-CT, a small improvement over NPE is observed only at the largest training-set size of 2.4M. These results illustrate that a well-configured FMPE can require substantially more training data before its posterior accuracy becomes competitive.

Our endpoint-variability diagnostic provides complementary information in this setting by helping distinguish limited-data effects from suboptimal FMPE configurations. Even at smaller training-set sizes, where posterior accuracy remains relatively poor, the well-tuned FMPE exhibits fitted scaling rates close to the theoretical value of $\alpha = 0.5$. In contrast, the problematic FMPE configurations show clear departures from this rate. This suggests that the diagnostic can indicate whether poor posterior accuracy is more consistent with insufficient training data than with a fundamentally suboptimal FMPE configuration.

6 Conclusion

We develop a theoretically motivated diagnostic for FMPE based on the scaling behavior of endpoint variability across training-set sizes. Rather than assessing posterior accuracy alone, our diagnostic asks whether endpoint variability follows the $N^{-1/2}$ rate predicted by theory, providing complementary information about whether poor performance is more consistent with limited training data or with a suboptimal FMPE configuration. Across controlled experiments and domain-specific epidemiological simulation studies, the diagnostic identifies departures induced by several forms of suboptimal configuration. The pseudo-dataset strategy further provides a practical approximation when repeated simulator calls are expensive.

Some limitations remain. The endpoint CLT is established for a regularized two-layer SiLU network in the infinite-width and long-training-time limits, and extending the theory to practical finite and deeper FMPE architectures remains an important direction. The pseudo-dataset strategy is motivated by a self-consistency hypothesis and is supported empirically, while a formal theoretical justification remains for future work.

More broadly, endpoint variability and posterior accuracy answer different questions. Our diagnostic is therefore intended to complement, rather than replace, established posterior checks and accuracy metrics. Combining endpoint-scaling diagnostics with conventional posterior evaluation may provide a more informative basis for deciding whether additional simulations are likely to improve inference or whether the FMPE configuration itself should be reconsidered.

AI DISCLOSURE

In this work, we use generative AI tools to polish the text of the paper and to reword and rearrange the presentation of the proofs. We also use generative AI tools to generate the Python code for parts of the experiment implementation, but all experiment designs are made by the authors. Generative AI tools are not used for any other task requiring disclosure. We have reviewed all AI-assisted text, reworded proofs and generated code, and we take responsibility for the final content of this work.

REFERENCES

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2305.16860.

Sebastian Bischoff, Alana Darcher, Michael Deistler, Richard Gao, Franziska Gerken, Manuel Gloeckler, Lisa Haxel, Jaivardhan Kapoor, Janne K Lappalainen, Jakob H Macke, Guy Moss, Matthijs Pals, Felix Pei, Rachel Rapp, A Erdem Sağtekin, Cornelius Schröder, Auguste Schulz, Zinovia Stefanidi, Shoji Toyota, Linda Ulmer, and Julius Vetter. A practical guide to sample-based statistical distances for evaluating generative models in science, 2024. URL https://arxiv.org/abs/2403.12636.

Jan Boelts, Philipp Harth, Richard Gao, Daniel Udvary, Felipe Yáñez, Daniel Baum, Hans-Christian Hege, Marcel Oberlaender, and Jakob H. Macke. Simulation-based inference for efficient identification of generative models in computational connectomics. *bioRxiv*, 2023. doi: 10.1101/2023.01.31.526269. URL https://www.biorxiv.org/content/early/2023/06/27/2023.01.31.526269.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks, 2021. URL https://arxiv.org/abs/2103.03098.

Prayag Chatha, Fan Bu, Jeffrey Regier, Evan Snitkin, and Jon Zelner. Neural posterior estimation for stochastic epidemic modeling, 2026. URL https://arxiv.org/abs/2412.12967.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.11215.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming, 2019. URL https://arxiv.org/abs/1812.07956.

Maximilian Dax, Jonas Wildberger, Simon Buchholz, Stephen R. Green, Jakob H. Macke, and Bernhard Schölkopf. Flow matching for scalable simulation-based inference, 2023. URL https://arxiv.org/abs/2305.17161.

Yicun Duan, Xinyue Li, Camille Avestruz, Jeffrey Regier, and LSST Dark Energy Science Collaboration. Neural posterior estimation for cataloging astronomical images from the legacy survey of space and time, 2025. URL https://arxiv.org/abs/2510.15315.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure, 2013. URL https://arxiv.org/abs/1312.2128.

Kenji Fukumizu, Taiji Suzuki, Noboru Isobe, Kazusato Oko, and Masanori Koyama. Flow matching achieves almost minimax optimal convergence. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2405.20879.

Yuan Gao, Jian Huang, Yuling Jiao, and Shurong Zheng. Convergence of continuous normalizing flows for learning probability distributions, 2024. URL https://arxiv.org/abs/2404.00551.

Timothy D. Gebhard, Jonas Wildberger, Maximilian Dax, Annalena Kofler, Daniel Angerhausen, Sascha P. Quanz, and Bernhard Schölkopf. Flow matching for atmospheric retrieval of exoplanets: Where reliability meets adaptive noise levels. *Astronomy & Astrophysics*, 693: A42, December 2024. ISSN 1432-0746. doi: 10.1051/0004-6361/202451861. URL http://dx.doi.org/10.1051/0004-6361/202451861.

Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P Vogels, David S Greenberg, and Jakob H Macke. Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLife*, 9:e56261, sep 2020. ISSN 2050-084X. doi: 10.7554/eLife.56261. URL https://doi.org/10.7554/eLife.56261.

Derek Hansen, Ismael Mendoza, Runjing Liu, Ziteng Pang, Zhe Zhao, Camille Avestruz, and Jeffrey Regier. Scalable bayesian inference for detection and deblending in astronomical images, 2022. URL https://arxiv.org/abs/2207.05642.

Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, Volodimir Begy, and Gilles Louppe. A trust crisis in simulation-based inference? your posterior approximations can be unfaithful, 2022. URL https://arxiv.org/abs/2110.06581.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020. URL https://arxiv.org/abs/1806.07572.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.02557.

Anastasia N. Krouglova, Hayden R. Johnson, Basile Confavreux, Michael Deistler, and Pedro J. Gonçalves. Multifidelity simulation-based inference for computationally expensive simulators, 2026. URL https://arxiv.org/abs/2502.08416.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin, 1991.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent, 2019. URL https://arxiv.org/abs/1902.06720.

Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-based accuracy testing of posterior estimators for general inference, 2023. URL https://arxiv.org/abs/2302.03026.

Jan-Matthis Lueckmann, Pedro J. Gonçalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H. Macke. Flexible statistical inference for mechanistic models of neural dynamics. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 1289–1299, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Declan McNamara, Jackson Loper, and Jeffrey Regier. Globally convergent variational inference, 2025. URL https://arxiv.org/abs/2501.08201.

Gonzalo Mena, Arun Kumar Kuchibhotla, and Larry Wasserman. Statistical properties of rectified flow, 2025. URL https://arxiv.org/abs/2511.03193.

Guy Moss, Leah Sophie Muhle, Reinhard Drews, Jakob H. Macke, and Cornelius Schröder. Fnope: Simulation-based inference on function spaces with fourier neural operators, 2025. URL https://arxiv.org/abs/2505.22573.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2303.01861.

George Papamakarios and Iain Murray. Fast
epsilon -free inference of simulation models with bayesian conditional density estimation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/6aca97005c68f1206823815f66102863-Paper.pdf.

Stefan T. Radev, Frederik Graw, Simiao Chen, Nico T. Mutters, Vanessa M. Eichel, Till Bärnighausen, and Ullrich Kothe. Outbreakflow: Model-based bayesian inference of disease outbreak dynamics with invertible neural networks and its application to the covid-19 pandemics in germany. *PLOS Computational Biology*, 17(10):e1009472, October 2021. doi: 10.1371/journal.pcbi.1009472. URL https://doi.org/10.1371/journal.pcbi.1009472.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, New York, 2008.

Cédric Villani. *The Wasserstein distances*, pp. 93–111. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-71050-9. doi: 10.1007/978-3-540-71050-9_6. URL https://doi.org/10.1007/978-3-540-71050-9_6.

Binxu Wang and John J. Vastola. The unreasonable effectiveness of gaussian score approximation for diffusion models and its applications. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2412.09726.

Binxu Wang, Jacob Zavatone-Veth, and Cengiz Pehlevan. A random matrix theory perspective on the consistency of diffusion models, 2026. URL https://arxiv.org/abs/2602.02908.

Vincent Wieland, Nils Wassmuth, Lorenzo Contento, Martin Kühn, and Jan Hasenauer. Assessment of simulation-based inference methods for stochastic compartmental models in epidemiological research, 2025. URL https://arxiv.org/abs/2512.02528.

Ruopeng Xie, Olivier Gascuel, and Anna Zhukova. Simpler is not always better: Phylodynamic misspecification and deep-learning corrections. *medRxiv*, 2026. doi: 10.64898/2026.05.07.26352661. URL https://www.medrxiv.org/content/early/2026/05/08/2026.05.07.26352661.

Percy S. Zhai, So Won Jeong, and Veronika Ročková. Conditional flow matching for bayesian posterior inference, 2025. URL https://arxiv.org/abs/2510.09534.

Huijie Zhang, Jinfan Zhou, Yifu Lu, Minzhe Guo, Peng Wang, Liyue Shen, and Qing Qu. The emergence of reproducibility and consistency in diffusion models. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2310.05264.

A SETUP AND NOTATION

We introduce the objects of the paper in the order in which the argument uses them: the training data and their flow-matching target; the kernel and the function space in which the velocity field is fitted; the per-sample operators whose averages summarise a training set; the ridge step, which turns those averages into a velocity field; the endpoint step, which turns the field into a generated sample; and finally the influence function that the main theorem is built from.

A.1 TRAINING DATA AND THE FLOW-MATCHING TARGET

The dimensions $d, m \geq 1$ are fixed. A random pair $(\Theta, X) \in \mathbb{R}^d \times \mathbb{R}^m$ is generated by the simulator: $\Theta \sim \pi$ is a parameter drawn from the prior and $X \sim P(\cdot \mid \Theta)$ is the observation the simulator produces for it. Let $Z \sim q_0$ be base noise, independent of (Θ, X). Set

$$W = (Z, \Theta, X), \quad \Delta = \Theta - Z, \quad Y(s) = (1-s)Z + s\Theta, \quad U(s) = (s, Y(s), X), \quad s \in [0,1]. \tag{5}$$

The path $Y(s)$ is the straight line from the noise Z to the parameter Θ; its velocity Δ is constant in s and is the regression target of conditional flow matching. The triple $U(s)$ is what a velocity network sees at path time s: the time, the current position, and the observation it is conditioned on. We call W a *training triple*.

The training set consists of N independent copies $W_1, \ldots, W_N$ of W, with Δ_i and $U_i(s)$ defined from W_i as in (5). Two probability laws on triples appear throughout: the *population law* $\nu_\star$, the law of W, and the *empirical law* $\nu_N = N^{-1}\sum_{i=1}^N \delta_{W_i}$. For a function q of a triple we write $\mathbb{E}_{\nu_\star} q = \mathbb{E}\, q(W)$ and $\mathbb{E}_{\nu_N} q = N^{-1}\sum_{i=1}^N q(W_i)$. Every object below that is built from the data has a population version and an empirical version, and the generic symbol $\nu \in \{\nu_\star, \nu_N\}$ lets one statement cover both. The scalar

$$S(W) = 1 + \|Z\|^2 + \|\Theta\|^2 + \|X\|^2 \tag{6}$$

measures the size of a triple; all moment conditions are stated through it.

A.2 INPUTS, THE KERNEL AND THE HYPOTHESIS SPACE

The input domain of a velocity field is $\mathcal{U} = [0,1] \times \mathbb{R}^d \times \mathbb{R}^m$, with generic element $u = (t, \vartheta, x)$: a time t, a location ϑ in parameter space at which the field is evaluated, and an observation x. The location ϑ is a deterministic argument and is not the random parameter Θ. The field does not see u directly but the *augmented input*

$$a(u) = (t, \vartheta, x, 1) \in \mathbb{R}^p, \qquad p = d + m + 2, \tag{7}$$

whose trailing coordinate 1 turns the last coordinate of a weight vector into a bias.

The activation is SiLU, $\sigma(q) = q/(1+e^{-q})$, and we write

$$L_\sigma = 1 + e^{-1} < 2, \tag{8}$$

which will be shown to bound $|\sigma'|$. Let $\gamma_p = \mathcal{N}(0, I_p)$ and define the *Gaussian feature map* $\Phi(a) \in L^2(\gamma_p)$ by $\Phi(a)(w) = \sigma(w^\top a)$, together with the scalar kernel

$$k(a,b) = \mathbb{E}_{w\sim\gamma_p}\big[\sigma(w^\top a)\,\sigma(w^\top b)\big] = \langle \Phi(a),\, \Phi(b)\rangle_{L^2(\gamma_p)}, \qquad K(u,u') = k\big(a(u), a(u')\big)\, I_d. \tag{9}$$

We abbreviate $k(u,u') = k(a(u), a(u'))$ when the arguments are inputs. The number $k(u,u')$ is the second moment $\mathbb{E}[\sigma(w^\top a(u))\,\sigma(w^\top a(u'))]$ of the outputs of a random SiLU neuron at u and at u', an inner product of feature vectors. Section C shows that K is the infinite-width neural tangent kernel of a two-layer SiLU network with zero output initialisation; for the main argument only the definition (9) is needed.

Let $\mathcal{H}_k$ be the reproducing kernel Hilbert space of k on $\mathcal{U}$ and let $\mathcal{H} = \mathcal{H}_k^d$ with $\langle f, g\rangle_{\mathcal{H}} = \sum_{j=1}^d \langle f_j, g_j\rangle_{\mathcal{H}_k}$ be the space of $\mathbb{R}^d$-valued fields whose coordinates lie in $\mathcal{H}_k$. This is the vector-valued reproducing kernel Hilbert space of K. Concretely, $\mathcal{H}$ consists of the fields $f(u) =$

$\mathbb{E}_{w\sim\gamma_p}[c(w)\,\sigma(w^\top a(u))]$ with $c \in L^2(\gamma_p;\mathbb{R}^d)$, and $\|f\|_{\mathcal{H}}$ is the smallest L^2 norm of a c that produces f: an infinitely wide two-layer network with frozen Gaussian hidden weights, whose norm is the size of its output layer. For $u \in \mathcal{U}$ the *evaluation operator* and its adjoint are

$$E_u : \mathcal{H} \to \mathbb{R}^d, \quad E_u f = f(u), \qquad\qquad E_u^* : \mathbb{R}^d \to \mathcal{H}, \quad E_u^* y = k(\cdot, u)\, y, \tag{10}$$

so that $\langle E_u f,\, y\rangle = \langle f,\, E_u^* y\rangle_{\mathcal{H}}$ and $E_u E_{u'}^* = K(u, u')$. Lemma D.1 collects the properties of k, Φ and E_u that the proofs use.

A.3 Per-sample operators and their averages

In ridge regression with a finite feature vector ψ, the data enter only through the averages of $\psi\psi^\top$ and of ψy. Here the features of a training triple are the evaluation operators along its path and the response is Δ. For a triple W define

$$C_W = \int_0^1 E_{U(s)}^* E_{U(s)}\,\mathrm{d}s, \qquad g_W = \int_0^1 E_{U(s)}^* \Delta\,\mathrm{d}s, \tag{11}$$

as Bochner integrals in the space of Hilbert–Schmidt operators on $\mathcal{H}$ and in $\mathcal{H}$ respectively; Lemma D.1 shows that they exist. Their meaning is the pair of identities

$$\langle f,\, C_W f\rangle_{\mathcal{H}} = \int_0^1 \|f(U(s))\|^2\,\mathrm{d}s, \qquad \langle f,\, g_W\rangle_{\mathcal{H}} = \int_0^1 \langle f(U(s)),\, \Delta\rangle\,\mathrm{d}s, \qquad f \in \mathcal{H}, \tag{12}$$

which follow from (10) because inner products pass through Bochner integrals. Consequently the flow-matching loss of one triple expands as

$$\int_0^1 \|\Delta - f(U(s))\|^2\,\mathrm{d}s = \|\Delta\|^2 - 2\langle f,\, g_W\rangle_{\mathcal{H}} + \langle f,\, C_W f\rangle_{\mathcal{H}}. \tag{13}$$

The averages of the per-sample objects are

$$\begin{aligned} &C_\nu = \mathbb{E}_\nu C_W, \qquad g_\nu = \mathbb{E}_\nu g_W; \qquad \text{thus} \\ &C_N = \frac{1}{N}\sum_{i=1}^N C_{W_i}, \qquad g_N = \frac{1}{N}\sum_{i=1}^N g_{W_i}, \qquad C = \mathbb{E}\, C_W, \qquad g = \mathbb{E}\, g_W. \end{aligned} \tag{14}$$

Thus C_N, g_N are the sample averages that summarise a training set and C, g are their expectations. Finally, the accumulated squared input norm along a training path,

$$Q(W) = \int_0^1 \|a(U(s))\|^2\,\mathrm{d}s, \tag{15}$$

is the single scalar through which the size of C_W and g_W is controlled.

A.4 The ridge step

For a law $\nu \in \{\nu_\star, \nu_N\}$ the regularised flow-matching objective is

$$J_\nu(f) = \mathbb{E}_\nu \int_0^1 \|\Delta - f(U(s))\|^2\,\mathrm{d}s + \lambda\|f\|_{\mathcal{H}}^2, \qquad f \in \mathcal{H}, \tag{16}$$

with a fixed ridge parameter $\lambda > 0$. Its minimiser is the *ridge field* v_ν; we write

$$v_\lambda = v_{\nu_\star} = \operatorname*{arg\,min}_{f\in\mathcal{H}} J_{\nu_\star}(f), \qquad \widehat{v}_{N,\lambda} = v_{\nu_N} = \operatorname*{arg\,min}_{f\in\mathcal{H}} J_{\nu_N}(f) \tag{17}$$

for the population and the empirical ridge fields. By (13), $J_\nu(f) = \langle f,\, (C_\nu + \lambda I) f\rangle_{\mathcal{H}} - 2\langle f,\, g_\nu\rangle_{\mathcal{H}} + \mathbb{E}_\nu\|\Delta\|^2$ is a quadratic in f, and Lemma D.2 shows that its unique minimiser solves the *normal equation* $(C_\nu + \lambda I)v_\nu = g_\nu$. We name the inverses

$$A_\nu = (C_\nu + \lambda I)^{-1}, \qquad A_\lambda = A_{\nu_\star} = (C + \lambda I)^{-1}, \qquad A_{N,\lambda} = A_{\nu_N} = (C_N + \lambda I)^{-1}, \tag{18}$$

so that $v_\nu = A_\nu g_\nu$. This is the operator form of $\widehat{\beta} = (X^\top X/N + \lambda I)^{-1} X^\top y/N$; the ridge term guarantees that the inverse exists and has operator norm at most $1/\lambda$, whatever the data. The

population ridge field v_λ is the regularised approximation of the marginal velocity field of flow matching; the gap between the two is a bias that does not vanish with N and is not the subject of this paper.

The *score* of a triple at the population field,

$$\eta_\lambda(W) = g_W - C_W v_\lambda - \lambda v_\lambda \in \mathcal{H}, \qquad \bar{\eta}_N = \frac{1}{N}\sum_{i=1}^{N} \eta_\lambda(W_i), \tag{19}$$

measures how far one triple, on its own, pushes the population normal equation out of balance; it equals minus one half of the gradient at v_λ of the regularised loss of that triple. Its mean is zero because v_λ solves the population normal equation. The *field error* and its size are

$$h_N = \widehat{v}_{N,\lambda} - v_\lambda \in \mathcal{H}, \qquad \delta_N = \|h_N\|_{\mathcal{H}}. \tag{20}$$

A.5 The endpoint step

Fix an initial state $z \in \mathbb{R}^d$ and an observation $x \in \mathbb{R}^m$. For a velocity field f the sampling trajectory and its endpoint are

$$\phi_t^f = z + \int_0^t f\big(s, \phi_s^f, x\big)\,\mathrm{d}s, \qquad R_{z,x}(f) = \phi_1^f, \tag{21}$$

so that $R_{z,x}$ is a map from fields to points of $\mathbb{R}^d$. The two endpoints compared in the main theorem are

$$R_\lambda = R_{z,x}(v_\lambda), \qquad \widehat{R}_{N,\lambda} = R_{z,x}(\widehat{v}_{N,\lambda}). \tag{22}$$

Lemma D.3 shows that both are well defined. Along the population trajectory $\phi = \phi^{v_\lambda}$ define

$$\begin{aligned} B(t) &= D_\vartheta v_\lambda(t, \phi_t, x), \qquad \partial_t \Gamma(t,s) = B(t)\Gamma(t,s), \quad \Gamma(s,s) = I_d, \\ Lh &= \int_0^1 \Gamma(1,t)\, h(t, \phi_t, x)\,\mathrm{d}t. \end{aligned} \tag{23}$$

Here $B(t)$ is the $d \times d$ Jacobian of the population field with respect to the state, evaluated along the population trajectory; $\Gamma(t,s)$ is the propagator of the linearised dynamics, which carries a small displacement of the trajectory at time s to time t; and Lh is the first-order response of the endpoint to a perturbation h of the field, obtained by adding up the nudges $h(t, \phi_t, x)\,\mathrm{d}t$ received along the trajectory, each transported to time 1. Lemma D.6 shows that L is the derivative of the endpoint map $R_{z,x}$ at v_λ.

A.6 The influence function and the limit covariance

The three ingredients combine into the *influence function* of a triple on the endpoint and its covariance,

$$\xi_\lambda(W; z, x) = L\, A_\lambda\, \eta_\lambda(W) \in \mathbb{R}^d, \qquad V_\lambda(z; x) = \mathbb{E}\big[\xi_\lambda \xi_\lambda^\top\big] \in \mathbb{R}^{d \times d}. \tag{24}$$

Read from right to left: a triple leaves residuals along its path (η_λ); the population ridge solve converts them into a change of the fitted field everywhere (A_λ); the linearised flow converts that change into a displacement of the endpoint (L). The main theorem states that $\widehat{R}_{N,\lambda} - R_\lambda$ is, to first order, the average of the N influence vectors.

A.7 The neural network

The following objects are used only in Section C. The two-layer SiLU network of width M is

$$f_M(u; \omega) = \frac{1}{\sqrt{M}} \sum_{r=1}^{M} c_r\, \sigma\big(w_r^\top a(u)\big), \qquad w_r \in \mathbb{R}^p,\ c_r \in \mathbb{R}^d, \qquad \omega = (w_r, c_r)_{r=1}^M, \tag{25}$$

with hidden weights w_r (whose last coordinate is the hidden bias) and output weights c_r. For a fixed training set it is trained by continuous-time Euclidean gradient flow on

$$\mathcal{F}_{N,M}(\omega) = \frac{1}{N}\sum_{i=1}^{N} \int_0^1 \|\Delta_i - f_M(U_i(s); \omega)\|^2\,\mathrm{d}s + \lambda \sum_{r=1}^{M} \|c_r\|^2, \qquad \dot{\omega}(\tau) = -\nabla_\omega \mathcal{F}_{N,M}(\omega(\tau)), \tag{26}$$

and we write $f_{N,M,\tau} = f_M(\cdot\,;\omega(\tau))$ for the network after optimisation time τ. The optimisation time $\tau \geq 0$ is distinct from the path time s and the sampling time t. The *tangent kernel* of the network is

$$[K_M(u,u')]_{ij} = \big\langle \nabla_\omega f_{M,i}(u),\, \nabla_\omega f_{M,j}(u') \big\rangle, \qquad 1 \leq i,j \leq d, \tag{27}$$

where $f_{M,i}$ is the i-th output coordinate.

A.8 Conventions

$\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^d$ or $\mathbb{R}^p$; $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_{\mathrm{HS}}$ are the operator and Hilbert–Schmidt norms of linear maps between Hilbert spaces, and $\|T\|_{\mathrm{op}} \leq \|T\|_{\mathrm{HS}}$. All integrals of Hilbert-space-valued or operator-valued functions are Bochner integrals; they commute with bounded linear maps and satisfy $\|\int F\| \leq \int \|F\|$. For random vectors, $X_N = O_p(r_N)$ means that X_N/r_N is bounded in probability, $X_N = o_p(r_N)$ means $X_N/r_N \to 0$ in probability, and $\Rightarrow$ denotes convergence in distribution. Unless stated otherwise, limits are taken as $N \to \infty$ with everything else fixed.

B Assumptions

Assumption B.1 (Sampling and moments)**.** The dimensions $d, m \geq 1$ are fixed. The pair (Θ, X) has a fixed joint simulation law with $\Theta \sim \pi$ and $X \mid \Theta \sim P(\cdot \mid \Theta)$, and $Z \sim q_0$ is independent of (Θ, X). The training triples $W_1, \ldots, W_N$ are independent copies of W, their law does not depend on N, and

$$\mathbb{E}\, S(W)^2 = \mathbb{E}\big(1 + \|Z\|^2 + \|\Theta\|^2 + \|X\|^2\big)^2 < \infty. \tag{28}$$

Condition (28) is a fourth-moment condition on the data. It is the natural one: C_W is quadratic in the features, the features grow linearly in the data, and the central limit theorem for the score requires C_W and g_W to have finite second moments. No boundedness or exponential-moment condition is imposed.

Assumption B.2 (Network architecture and initialisation)**.** The network is (25). At initialisation, $(w_r(0))_{r\geq 1}$ is an i.i.d. sequence with $w_r(0) \sim \mathcal{N}(0, I_p)$ and $c_r(0) = 0$ for every r, as in conditions (C2)/(D2) of McNamara et al. (2025). All coordinates, including the hidden bias, are trainable with the Euclidean parameter metric; there is no output bias.

Assumption B.3 (Ridge targets, training rule and endpoints)**.** The ridge parameter $\lambda > 0$, the initial state $z \in \mathbb{R}^d$ and the observation $x \in \mathbb{R}^m$ are fixed, independently of N. The estimator is the empirical ridge field $\widehat{v}_{N,\lambda}$ of (17) in the space $\mathcal{H}$ of the kernel (9). The network is trained by the gradient flow (26), whose penalty acts on the output weights only and carries no factor M^{-1}; the limits are taken in the order $M \to \infty$, then $\tau \to \infty$, then $N \to \infty$. Endpoints are computed from the exact flow (21).

The results of Section D use Assumption B.1 and the ridge part of Assumption B.3 only. Assumption B.2 and the training part of Assumption B.3 are used only in Section C, which explains why the ridge estimator is the object of interest.

C Network training and the ridge limit

This section connects the estimator analysed in the rest of the paper to a trained neural network. It is isolated from the statistical argument: nothing in Section D, nor in the proofs of its results in Sections E.1–E.8, depends on it, although its own proof draws on the regularity Lemmas D.1–D.3. Its two statements are that the tangent kernel of the network at initialisation converges to the kernel K of (9), and that for a fixed training set the network trained by (26) converges to the empirical ridge field $\widehat{v}_{N,\lambda}$ when the width and then the optimisation time tend to infinity, endpoints included.

Proposition C.1 (Limiting SiLU tangent kernel)**.** *Under Assumption B.2, the tangent kernel K_M of (27) at initialisation converges almost surely, as $M \to \infty$, to the kernel K of (9): for every finite set of inputs the convergence holds simultaneously for all pairs in the set, and it holds uniformly on every compact subset of $\mathcal{U} \times \mathcal{U}$ on a single event of probability one. The limit K is deterministic, continuous and positive semidefinite, with $0 \leq k(a,a) \leq \|a\|^2$.*

Theorem C.1 (Network training and the ridge limit)**.** *Let Assumptions B.1–B.3 hold. Fix a finite training set and let $f_{N,M,\tau}$ be the network trained by (26) from the initialisation of Assumption B.2. Define the kernel gradient-flow path*

$$\tilde{v}_{N,\tau} = \big(I - e^{-2\tau(C_N+\lambda I)}\big)\widehat{v}_{N,\lambda}. \tag{29}$$

Then, almost surely with respect to the Gaussian initialisation, for every $T < \infty$ and every compact $\mathcal{B} \subset \mathcal{U}$,

$$\sup_{0\le\tau\le T}\ \sup_{u\in\mathcal{B}} \|f_{N,M,\tau}(u) - \tilde{v}_{N,\tau}(u)\| \longrightarrow 0 \qquad (M\to\infty), \tag{30}$$

$$\|\tilde{v}_{N,\tau} - \widehat{v}_{N,\lambda}\|_{\mathcal{H}} \le e^{-2\lambda\tau}\,\|\widehat{v}_{N,\lambda}\|_{\mathcal{H}}, \tag{31}$$

and every finite-network endpoint $R_{z,x}(f_{N,M,\tau})$ is well defined, with

$$\lim_{\tau\to\infty}\ \lim_{M\to\infty} R_{z,x}(f_{N,M,\tau}) = R_{z,x}(\widehat{v}_{N,\lambda}). \tag{32}$$

Statement (30) is the lazy-training phenomenon: with zero output initialisation and an output-only penalty, the wide network follows the gradient flow of the ridge objective in $\mathcal{H}$. Statement (31) says that this flow reaches the ridge field exponentially fast, at a rate set by λ alone. Statement (32) combines the two at the level of endpoints. After these two limits the fitted field *equals* the ridge field, so no optimisation error and no finite-width error enters the central limit theorem of Section D. The proofs are in Section E.9.

D Ridge regression to endpoint CLT

D.1 Overview

The estimator is built in three steps: the training set is summarised by the averages (C_N, g_N); a ridge solve turns the averages into the field $\widehat{v}_{N,\lambda}$; an ODE solve turns the field into the endpoint $\widehat{R}_{N,\lambda}$. The population endpoint R_λ is produced by the same three steps applied to the law of W. Figure 4 shows the two chains and, between them, the difference at each stage.

The proof of the central limit theorem follows the middle row. Averages of independent terms obey the central limit theorem. The other two steps are smooth maps, so a fluctuation at their input is transmitted to their output through their derivatives; this is the delta method, carried out once in the Hilbert space $\mathcal{H}$ and once for the endpoint map. The derivative of the ridge solve at the population averages is the fixed operator A_λ, and the derivative of the endpoint map at the population field is the fixed operator L. That is why the influence function is $\xi_\lambda = LA_\lambda\eta_\lambda$: the three factors are the three arrows of the middle row, read from right to left.

The results below are organised accordingly. Lemmas D.1, D.2 and D.3 establish that every box and arrow in the figure is well defined and record the regularity that the later lemmas draw on. Lemma D.4 concerns the left node of the middle row, Lemma D.5 the arrow labelled A_λ, and Lemma D.6 the arrow labelled L. Proposition D.1 assembles them into the asymptotically linear expansion of the endpoint, and Theorem D.1 applies the multivariate central limit theorem.

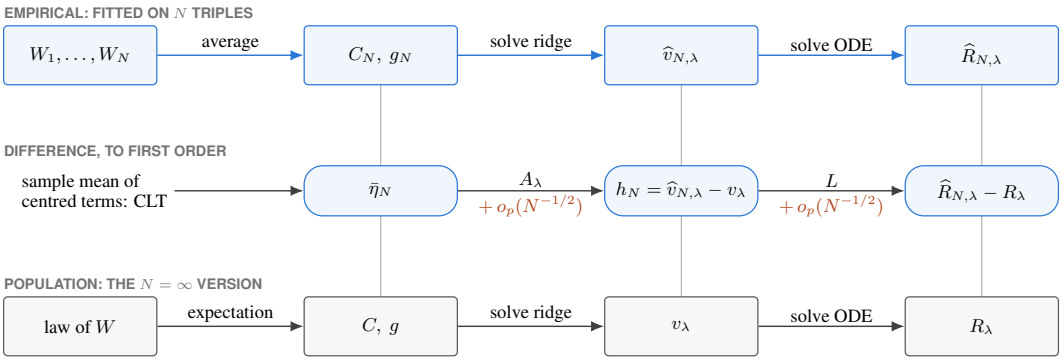

Figure 4: The proof path. The top and bottom rows run the same three steps on a finite training set and on the population law. The middle row tracks the difference between the two rows: it is carried from left to right by the two fixed linear maps A_λ (Lemma D.5) and L (Lemma D.6), each up to a remainder smaller than $N^{-1/2}$, and it starts from a centred sample mean (Lemma D.4). Lemmas D.1–D.3 make every box and arrow well defined. Blue objects are random because they depend on the training set; black objects are fixed; the orange terms are the remainders that the proofs show to be negligible.

D.2 Regularity along the path

The first lemma concerns the objects of Sections A.2 and A.3: the kernel, the evaluation operators and the per-sample operators. Its most used consequence is that closeness in $\mathcal{H}$-norm implies closeness at every point, with an explicit constant.

Lemma D.1 (Kernel, evaluation operators and per-sample objects)**.** *Parts* (a) *and* (b) *hold without any assumption; part* (c) *holds under Assumption B.1, and part* (d) *under Assumptions B.1 and B.3.*

(a) *The activation σ is smooth with $|\sigma(q)| \le |q|$ and $|\sigma'(q)| \le L_\sigma$ for all $q \in \mathbb{R}$. The map $\Phi : \mathbb{R}^p \to L^2(\gamma_p)$ is continuously differentiable and L_σ-Lipschitz, $\|\Phi(a) - \Phi(a')\|_{L^2(\gamma_p)} \le L_\sigma \|a - a'\|$. The kernel k is symmetric, positive semidefinite and continuously differentiable on $\mathbb{R}^p \times \mathbb{R}^p$, with $0 \le k(a,a) \le \|a\|^2$ and $\|\nabla_a k(a,b)\| \le L_\sigma \|b\|$. The space $\mathcal{H}$ is separable.*

(b) *For $u, u' \in \mathcal{U}$,*

$$\|E_u\|_{\mathrm{op}} = \sqrt{k(u,u)} \le \|a(u)\|, \qquad \|E_u - E_{u'}\|_{\mathrm{op}} \le L_\sigma \|a(u) - a(u')\|. \tag{33}$$

Consequently every $f \in \mathcal{H}$ is continuous on $\mathcal{U}$, with $\|f(u)\| \le \|f\|_{\mathcal{H}} \|a(u)\|$ and $\|f(u) - f(u')\| \le L_\sigma \|f\|_{\mathcal{H}} \|a(u) - a(u')\|$; in particular f is Lipschitz in the state ϑ with constant $L_\sigma \|f\|_{\mathcal{H}}$, uniformly in (t,x).

(c) *The scalar (15) equals*

$$Q(W) = \frac{4}{3} + \|X\|^2 + \left\| \frac{Z + \Theta}{2} \right\|^2 + \frac{\|\Delta\|^2}{12}, \qquad \text{and} \qquad Q \le \tfrac{4}{3} S, \quad \|\Delta\|^2 \le 12 Q. \tag{34}$$

The Bochner integrals (11) exist, $W \mapsto C_W$ and $W \mapsto g_W$ are continuous into the Hilbert–Schmidt operators and into $\mathcal{H}$, and

$$\|C_W\|_{\mathrm{op}} \le Q, \qquad \|C_W\|_{\mathrm{HS}} \le \sqrt{d}\, Q, \qquad \|g_W\|_{\mathcal{H}} \le \sqrt{Q}\, \|\Delta\|. \tag{35}$$

Hence $\mathbb{E}\|C_W\|_{\mathrm{HS}}^2 < \infty$ and $\mathbb{E}\|g_W\|_{\mathcal{H}}^2 < \infty$, and the expectations $C = \mathbb{E} C_W$ and $g = \mathbb{E} g_W$ in (14) exist.

(d) *For $\nu \in \{\nu_\star, \nu_N\}$ the operator C_ν is self-adjoint and positive semidefinite, $C_\nu + \lambda I$ is invertible, and $\|A_\nu\|_{\mathrm{op}} \le 1/\lambda$.*

The second lemma is the ridge step: it makes both ridge fields well defined and introduces the Lipschitz constant that every Gronwall argument below relies on.

Lemma D.2 (The ridge step is well posed). *Let Assumptions B.1 and B.3 hold and let $\nu \in \{\nu_\star, \nu_N\}$.*

(i) *J_ν has the unique minimiser $v_\nu = A_\nu g_\nu$, which is the unique solution of the normal equation $(C_\nu + \lambda I)v_\nu = g_\nu$, and $\|v_\nu\|_{\mathcal{H}} \le \|g_\nu\|_{\mathcal{H}}/\lambda$. For every $f \in \mathcal{H}$, $J_\nu(f) - J_\nu(v_\nu) = \langle f - v_\nu, (C_\nu + \lambda I)(f - v_\nu)\rangle_{\mathcal{H}} \ge \lambda \|f - v_\nu\|_{\mathcal{H}}^2$.*

(ii) *v_ν has a jointly continuous state derivative $D_\vartheta v_\nu$ on $\mathcal{U}$, and*

$$\sup_{u \in \mathcal{U}} \|D_\vartheta v_\nu(u)\|_{\mathrm{op}} \le \ell_\nu := \frac{L_\sigma}{\lambda} \sqrt{\mathbb{E}_\nu Q \cdot \mathbb{E}_\nu \|\Delta\|^2} < \infty. \tag{36}$$

Thus v_ν is globally Lipschitz in the state with constant ℓ_ν, uniformly in (t, x), and it grows at most linearly, $\|v_\nu(u)\| \le \|v_\nu\|_{\mathcal{H}} \|a(u)\|$.

(iii) *For fixed N, the empirical ridge field $\widehat{v}_{N,\lambda}$ is a continuous function of the training set $(W_1, \dots, W_N)$ into $\mathcal{H}$.*

For the population law we write $\ell = \ell_{\nu_\star}$; it is a deterministic constant. For the empirical law, ℓ_{ν_N} is a finite random variable.

The third lemma is the endpoint step. It makes the two endpoints well defined, states the stability estimate that transfers closeness of fields to closeness of endpoints, and defines the operator L rigorously.

Lemma D.3 (The endpoint step is well posed). *Let Assumptions B.1 and B.3 hold and fix (z, x).*

(i) *For $\nu \in \{\nu_\star, \nu_N\}$ the flow (21) with $f = v_\nu$ has a unique solution on $[0, 1]$. In particular R_λ and $\widehat{R}_{N,\lambda}$ are well defined.*

(ii) *(Endpoint stability.) Let $v, v' : \mathcal{U} \to \mathbb{R}^d$ be continuous, let v be globally Lipschitz in the state with constant ℓ', uniformly in t (the observation x is fixed), and let ϕ' be a solution of (21) for v' on $[0, 1]$. Then the flow ϕ of v exists on $[0, 1]$ and*

$$\|R_{z,x}(v) - R_{z,x}(v')\| \le e^{\ell'} \int_0^1 \|v(t, \phi'_t, x) - v'(t, \phi'_t, x)\| \, \mathrm{d}t. \tag{37}$$

(iii) *For fixed N, the empirical endpoint $\widehat{R}_{N,\lambda}$ is a continuous, hence measurable, function of the training set.*

(iv) *The Jacobian B of (23) is continuous on $[0, 1]$ with $\|B(t)\|_{\mathrm{op}} \le \ell$; the propagator Γ exists, is unique, and satisfies $\|\Gamma(t, s)\|_{\mathrm{op}} \le e^{\ell(t-s)}$ for $0 \le s \le t \le 1$; and $L : \mathcal{H} \to \mathbb{R}^d$ is a bounded linear map with*

$$\|L\|_{\mathrm{op}} \le B_* e^{\ell}, \qquad B_* = \sup_{0 \le t \le 1} \|a(t, \phi_t, x)\| < \infty. \tag{38}$$

D.3 The middle row

We now follow the difference between the empirical and the population chain, one arrow at a time.

Lemma D.4 (The score). *Let Assumptions B.1 and B.3 hold. The score η_λ of (19) is a centred, square-integrable random element of $\mathcal{H}$: there is a deterministic constant c_η with*

$$\mathbb{E}\, \eta_\lambda(W) = 0, \qquad \|\eta_\lambda(W)\|_{\mathcal{H}} \le c_\eta \, S(W), \qquad s_\lambda^2 := \mathbb{E}\|\eta_\lambda(W)\|_{\mathcal{H}}^2 < \infty. \tag{39}$$

Consequently $\mathbb{E}\|\bar{\eta}_N\|_{\mathcal{H}}^2 = s_\lambda^2/N$ and $\|\bar{\eta}_N\|_{\mathcal{H}} = O_p(N^{-1/2})$.

Remark. Since $\mathcal{H}$ is separable, the central limit theorem for i.i.d. centred square-integrable random elements of a Hilbert space (Ledoux & Talagrand, 1991) shows that $\sqrt{N}\, \bar{\eta}_N$ converges in distribution in $\mathcal{H}$ to a centred Gaussian element. The proof below does not use this: it applies the central limit theorem in $\mathbb{R}^d$ after the two linear maps of the middle row, which requires only the rate in Lemma D.4.

Lemma D.5 (The ridge step: from $\bar{\eta}_N$ to h_N)**.** *Let Assumptions B.1 and B.3 hold. The field error satisfies the exact identity*

$$(C_N + \lambda I)\, h_N = \bar{\eta}_N, \qquad \textit{that is} \qquad h_N = A_{N,\lambda}\bar{\eta}_N, \tag{40}$$

so that $\delta_N \le \|\bar{\eta}_N\|_{\mathcal{H}}/\lambda$, $\mathbb{E}\,\delta_N^2 \le s_\lambda^2/(\lambda^2 N)$ *and* $\delta_N = O_p(N^{-1/2})$. *Moreover* $\|C_N - C\|_{\mathrm{HS}} \to 0$ *almost surely, and*

$$\begin{aligned} h_N &= A_\lambda \bar{\eta}_N - \zeta_N, \qquad \zeta_N = A_{N,\lambda}(C_N - C)A_\lambda \bar{\eta}_N, \\ \|\zeta_N\|_{\mathcal{H}} &\le \frac{\|C_N - C\|_{\mathrm{op}} \|\bar{\eta}_N\|_{\mathcal{H}}}{\lambda^2} = o_p(N^{-1/2}). \end{aligned} \tag{41}$$

The next lemma is deterministic: it says that the endpoint map is differentiable at the population field, with derivative L and an explicit remainder. Randomness enters only when it is applied to $h = h_N$.

Lemma D.6 (The endpoint step: the derivative of $R_{z,x}$ at v_λ)**.** *Let Assumptions B.1 and B.3 hold and fix* (z, x). *Let* $c_0 = B_* e^{\ell+1}$, *let*

$$\mathcal{T} = \big\{(t, \vartheta) : t \in [0, 1],\ \|\vartheta - \phi_t\| \le c_0\big\}$$

be the closed tube of radius c_0 *around the population trajectory, and let*

$$\varrho(\epsilon) = \sup\big\{\|D_\vartheta v_\lambda(t, \vartheta, x) - D_\vartheta v_\lambda(t, \vartheta', x)\|_{\mathrm{op}} :\ (t, \vartheta), (t, \vartheta') \in \mathcal{T},\ \|\vartheta - \vartheta'\| \le \epsilon\big\}$$

be the modulus of continuity of the state derivative on the tube, so that $\varrho(\epsilon) \to 0$ *as* $\epsilon \downarrow 0$. *Then for every* $h \in \mathcal{H}$ *with* $\|h\|_{\mathcal{H}} \le 1$, *the flow of* $v_\lambda + h$ *from* z *exists on* $[0, 1]$, *stays in* $\mathcal{T}$, *and*

$$R_{z,x}(v_\lambda + h) - R_{z,x}(v_\lambda) = Lh + \mathsf{r}(h), \qquad \|\mathsf{r}(h)\| \le e^\ell c_0\big(\varrho(c_0\|h\|_{\mathcal{H}}) + L_\sigma \|h\|_{\mathcal{H}}\big)\|h\|_{\mathcal{H}}. \tag{42}$$

In particular $R_{z,x}$ *is Fréchet differentiable at* v_λ *with derivative* L.

D.4 Main results

Proposition D.1 (Asymptotic linearity of the endpoint)**.** *Let Assumptions B.1 and B.3 hold and fix* (z, x). *The influence vectors* $\xi_\lambda(W_i; z, x)$ *of (24) are i.i.d., centred and square-integrable, with* $\mathbb{E}\|\xi_\lambda\|^2 \le \|L\|_{\mathrm{op}}^2 s_\lambda^2/\lambda^2$, *and*

$$\widehat{R}_{N,\lambda} - R_\lambda = \frac{1}{N}\sum_{i=1}^{N} \xi_\lambda(W_i; z, x) + o_p(N^{-1/2}). \tag{43}$$

Theorem D.1 (Endpoint central limit theorem)**.** *Let Assumptions B.1 and B.3 hold. For each fixed* (z, x),

$$\sqrt{N}\,\big(\widehat{R}_{N,\lambda} - R_\lambda\big) \ \Rightarrow\ \mathcal{N}\big(0,\, V_\lambda(z; x)\big). \tag{44}$$

The covariance $V_\lambda(z; x) = \mathbb{E}[\xi_\lambda \xi_\lambda^\top]$ *is finite and may be singular.*

D.5 Reading the proof path

Each result corresponds to one element of Figure 4.

The left node. The score $\eta_\lambda(W_i)$ is what remains of the difference between the top and bottom rows at the first stage after it has been pushed through the population normal equation: $g_N - (C_N + \lambda I)v_\lambda = \bar{\eta}_N$. It is a sample mean of i.i.d. centred terms (Lemma D.4), which is the only place where the central limit theorem is invoked. The fourth-moment condition (28) is exactly what makes these terms square-integrable.

The arrow A_λ. The empirical normal equation gives the exact identity $h_N = A_{N,\lambda}\bar{\eta}_N$. The inverse $A_{N,\lambda}$ depends on the whole sample, so this is not yet a sum of independent terms; the resolvent identity replaces $A_{N,\lambda}$ by the fixed operator A_λ at the cost of a remainder that is a product of $\|C_N - C\|_{\mathrm{op}} = o_p(1)$ and $\|\bar{\eta}_N\| = O_p(N^{-1/2})$ (Lemma D.5). Only consistency of C_N is needed, not a rate. The uniform bound $\|A_{N,\lambda}\|_{\mathrm{op}} \le 1/\lambda$ is why $\lambda > 0$ is held fixed.

The arrow L. The endpoint map is a nonlinear function of the field, and Lemma D.6 is its first-order Taylor expansion at v_λ. The remainder has two sources: the Taylor error of the population field along the perturbed trajectory, controlled by the uniform continuity of $D_\vartheta v_\lambda$ on a compact tube, and the evaluation of the perturbation h on the perturbed trajectory instead of the population one, controlled by the Lipschitz property of the feature map. Both are $o(\|h\|_{\mathcal{H}})$, and $\delta_N = O_p(N^{-1/2})$ turns them into $o_p(N^{-1/2})$. Only the population Lipschitz constant ℓ enters; the random Lipschitz constant of $\widehat{v}_{N,\lambda}$ is never needed.

The right node. Proposition D.1 chains the three pieces, and Theorem D.1 is the ordinary multivariate central limit theorem for the i.i.d. vectors $\xi_\lambda(W_i)$, with Slutsky's lemma absorbing the remainders.

Three structural facts make the argument work. The estimator has a closed form, so the first linearisation is an algebraic identity rather than an approximation. The ridge bounds every inverse by $1/\lambda$ and gives the population field a finite Lipschitz constant proportional to $1/\lambda$. The reproducing property turns an abstract rate in $\mathcal{H}$ into a concrete statement about velocities along one trajectory.

E PROOFS

Sections E.1–E.8 prove the results of Section D in the order of the proof path; Section E.9 proves the isolated results of Section C.

E.1 A GRONWALL INEQUALITY

We use the following form of Gronwall's inequality.

Lemma E.1 (Gronwall). *Let $t_0 < t_1$, let $\psi : [t_0, t_1] \to [0, \infty)$ be continuous, $\alpha \geq 0$, $\gamma_0 \geq 0$, and $\beta : [t_0, t_1] \to [0, \infty)$ integrable, with $\psi(t) \leq \gamma_0 + \int_{t_0}^t \big(\alpha\psi(s) + \beta(s)\big)\,\mathrm{d}s$ for all $t \in [t_0, t_1]$. Then $\psi(t) \leq \big(\gamma_0 + \int_{t_0}^t \beta(s)\,\mathrm{d}s\big)e^{\alpha(t-t_0)}$. In particular, if $\beta \equiv \beta_0$ is constant and $\gamma_0 = 0$, then $\psi(t) \leq \beta_0\,(t - t_0)\,e^{\alpha(t-t_0)}$.*

Proof. Let $\Psi(t) = \gamma_0 + \int_{t_0}^t (\alpha\psi + \beta)$. Then Ψ is absolutely continuous, $\psi \leq \Psi$, and $\Psi' = \alpha\psi + \beta \leq \alpha\Psi + \beta$ almost everywhere, so $\big(e^{-\alpha(t-t_0)}\Psi(t)\big)' \leq e^{-\alpha(t-t_0)}\beta(t)$. Integrating from t_0, $e^{-\alpha(t-t_0)}\Psi(t) \leq \gamma_0 + \int_{t_0}^t e^{-\alpha(s-t_0)}\beta(s)\,\mathrm{d}s \leq \gamma_0 + \int_{t_0}^t \beta$. □

E.2 PROOF OF LEMMA D.1

Proof of Lemma D.1. (a) Write $\sigma(q) = q\,\varsigma(q)$ with the logistic function $\varsigma(q) = (1 + e^{-q})^{-1} \in (0, 1)$. Then $|\sigma(q)| \leq |q|$, and $\sigma'(q) = \varsigma(q) + q\,\varsigma(q)(1 - \varsigma(q))$ with $\varsigma(1 - \varsigma) \leq e^{-|q|}$, so $|\sigma'(q)| \leq 1 + |q|e^{-|q|} \leq 1 + e^{-1} = L_\sigma$. The function σ is smooth and σ' is continuous and bounded.

Since $\mathbb{E}(w^\top a)^2 = \|a\|^2$ for $w \sim \gamma_p$, we get $\|\Phi(a)\|_{L^2}^2 = \mathbb{E}\sigma(w^\top a)^2 \leq \|a\|^2$, so $\Phi(a) \in L^2(\gamma_p)$, and

$$\|\Phi(a) - \Phi(a')\|_{L^2}^2 = \mathbb{E}\big(\sigma(w^\top a) - \sigma(w^\top a')\big)^2 \leq L_\sigma^2\,\mathbb{E}\big(w^\top(a - a')\big)^2 = L_\sigma^2\|a - a'\|^2.$$

For $h \in \mathbb{R}^p$ put $D\Phi(a)[h](w) = \sigma'(w^\top a)\,w^\top h$. Then $\|D\Phi(a)[h]\|_{L^2} \leq L_\sigma\|h\|$, and by the fundamental theorem of calculus

$$\Phi(a + h) - \Phi(a) - D\Phi(a)[h] = \int_0^1 \big(\sigma'(w^\top(a + \theta h)) - \sigma'(w^\top a)\big)\,w^\top h\,\mathrm{d}\theta,$$

whose L^2 norm is at most $\|h\|\big(\mathbb{E}\sup_{\theta\in[0,1]}(\sigma'(w^\top(a + \theta h)) - \sigma'(w^\top a))^2\|w\|^2\big)^{1/2} = o(\|h\|)$ by dominated convergence, the integrand being dominated by $4L_\sigma^2\|w\|^2$. The same dominated convergence argument shows that $a \mapsto D\Phi(a)$ is continuous in operator norm. Hence Φ is continuously differentiable.

Consequently $k(a, b) = \langle\Phi(a), \Phi(b)\rangle_{L^2}$ is symmetric, continuously differentiable, and positive semidefinite, since $\sum_{i,j}\alpha_i\alpha_j k(a_i, a_j) = \|\sum_i \alpha_i\Phi(a_i)\|_{L^2}^2 \geq 0$. Also $0 \leq k(a, a) = \|\Phi(a)\|^2 \leq$

$\|a\|^2$, and $D_a k(a,b)[h] = \langle D\Phi(a)[h],\, \Phi(b)\rangle \le L_\sigma \|h\| \sqrt{k(b,b)} \le L_\sigma \|h\| \|b\|$. The scalar space $\mathcal{H}_k$ is isometric to the closed linear span of $\{\Phi(a(u)) : u \in \mathcal{U}\}$ in $L^2(\gamma_p)$ (Steinwart & Christmann, 2008, Theorem 4.21), which is separable because $L^2(\gamma_p)$ is; hence $\mathcal{H} = \mathcal{H}_k^d$ is separable.

(b) For $f \in \mathcal{H}$ and $u \in \mathcal{U}$ the reproducing property gives $f_j(u) = \langle f_j,\, k(\cdot,u)\rangle_{\mathcal{H}_k}$ and $\|k(\cdot,u)\|^2_{\mathcal{H}_k} = k(u,u)$, so $\|f(u)\|^2 = \sum_j f_j(u)^2 \le k(u,u)\sum_j \|f_j\|^2 = k(u,u)\|f\|^2_{\mathcal{H}}$, with equality for $f = k(\cdot,u)y$, $\|y\| = 1$. This proves $\|E_u\|_{\mathrm{op}} = \sqrt{k(u,u)} \le \|a(u)\|$. Likewise $(E_u - E_{u'})f = f(u) - f(u')$ and $f_j(u) - f_j(u') = \langle f_j,\, k(\cdot,u) - k(\cdot,u')\rangle$, where

$$\begin{aligned}\|k(\cdot,u) - k(\cdot,u')\|^2_{\mathcal{H}_k} &= k(u,u) + k(u',u') - 2k(u,u') \\ &= \|\Phi(a(u)) - \Phi(a(u'))\|^2_{L^2} \le L_\sigma^2 \|a(u) - a(u')\|^2,\end{aligned}$$

which gives the second bound in (33). The consequences for f follow by applying the two bounds to f; since $u \mapsto a(u)$ is continuous, f is continuous, and $a(u) - a(u') = (0, \vartheta - \vartheta', 0, 0)$ when u, u' differ only in the state.

(c) Write $Y(s) = \frac{Z+\Theta}{2} + (s - \frac12)\Delta$. Since $\int_0^1 (s-\frac12)\,\mathrm{d}s = 0$ and $\int_0^1 (s-\frac12)^2\,\mathrm{d}s = \frac{1}{12}$,

$$Q(W) = \int_0^1 \big(s^2 + \|Y(s)\|^2 + \|X\|^2 + 1\big)\,\mathrm{d}s = \frac{4}{3} + \|X\|^2 + \Big\|\frac{Z+\Theta}{2}\Big\|^2 + \frac{\|\Delta\|^2}{12}.$$

The bounds $\|Z+\Theta\|^2 \le 2\|Z\|^2 + 2\|\Theta\|^2$ and $\|\Delta\|^2 \le 2\|Z\|^2 + 2\|\Theta\|^2$ give $Q \le \frac43 + \|X\|^2 + \frac23(\|Z\|^2 + \|\Theta\|^2) \le \frac43 S$, and $\|\Delta\|^2 \le 12Q$ is immediate.

The map $s \mapsto E_{U(s)}$ is continuous in operator norm by (33). The operator $E_u^* E_u$ has rank at most d and $E_u^*E_u - E_{u'}^*E_{u'} = E_u^*(E_u - E_{u'}) + (E_u - E_{u'})^* E_{u'}$, each term having rank at most d, so $s \mapsto E_{U(s)}^* E_{U(s)}$ is continuous in Hilbert–Schmidt norm and $s \mapsto E_{U(s)}^*\Delta$ is continuous in $\mathcal{H}$. Both integrals in (11) therefore exist as Bochner integrals of continuous functions on $[0,1]$. Since $E_u E_u^* = k(u,u) I_d$, the nonzero eigenvalues of $E_u^* E_u$ are $k(u,u)$ with multiplicity at most d, so $\|E_u^* E_u\|_{\mathrm{op}} = k(u,u) \le \|a(u)\|^2$ and $\|E_u^* E_u\|_{\mathrm{HS}} \le \sqrt{d}\, k(u,u)$. Integrating gives the first two bounds in (35), and $\|g_W\|_{\mathcal{H}} \le \|\Delta\| \int_0^1 \|a(U(s))\|\,\mathrm{d}s \le \|\Delta\|\sqrt{Q}$ by Cauchy–Schwarz. Continuity of $W \mapsto (C_W, g_W)$ follows from the same rank-d decomposition and (33), since $U(s)$ depends continuously on W uniformly in s. Finally $\mathbb{E}\|C_W\|^2_{\mathrm{HS}} \le d\,\mathbb{E}Q^2 \le \frac{16}{9} d\,\mathbb{E}S^2 < \infty$ and $\mathbb{E}\|g_W\|^2_{\mathcal{H}} \le \mathbb{E}[Q\|\Delta\|^2] \le 12\,\mathbb{E}Q^2 < \infty$ by (28); the expectations exist as Bochner integrals in the separable spaces concerned.

(d) Because inner products pass through Bochner integrals and expectations, (12) gives $\langle f,\, C_\nu f\rangle_{\mathcal{H}} = \mathbb{E}_\nu \int_0^1 \|f(U(s))\|^2\,\mathrm{d}s \ge 0$ and $\langle f,\, C_\nu g\rangle_{\mathcal{H}} = \mathbb{E}_\nu \int_0^1 \langle f(U(s)),\, g(U(s))\rangle\,\mathrm{d}s = \langle C_\nu f,\, g\rangle_{\mathcal{H}}$, so C_ν is self-adjoint and positive semidefinite. Its spectrum lies in $[0,\infty)$, the spectrum of $C_\nu + \lambda I$ lies in $[\lambda,\infty)$, and $\|A_\nu\|_{\mathrm{op}} \le 1/\lambda$. □

E.3 Proof of Lemma D.2

Proof of Lemma D.2. (i) By (13) and (14), $J_\nu(f) = \langle f,\, (C_\nu + \lambda I)f\rangle - 2\langle f,\, g_\nu\rangle + \mathbb{E}_\nu\|\Delta\|^2$. Put $v_\nu = A_\nu g_\nu$, so that $(C_\nu + \lambda I)v_\nu = g_\nu$. For $f = v_\nu + h$, expanding the quadratic and using that $C_\nu + \lambda I$ is self-adjoint,

$J_\nu(v_\nu + h) - J_\nu(v_\nu) = 2\langle h,\, (C_\nu + \lambda I)v_\nu - g_\nu\rangle + \langle h,\, (C_\nu + \lambda I)h\rangle = \langle h,\, (C_\nu + \lambda I)h\rangle \ge \lambda\|h\|^2$,

which is strictly positive unless $h = 0$. Hence v_ν is the unique minimiser, and $\|v_\nu\| \le \|A_\nu\|_{\mathrm{op}}\|g_\nu\| \le \|g_\nu\|/\lambda$.

(ii) Write $r_\nu(s,W) = \Delta - v_\nu(U(s))$ for the residual. Comparing $J_\nu(v_\nu) \le J_\nu(0) = \mathbb{E}_\nu\|\Delta\|^2$ with $J_\nu(v_\nu) \ge \mathbb{E}_\nu \int_0^1 \|r_\nu\|^2\,\mathrm{d}s$ gives

$$\mathbb{E}_\nu \int_0^1 \|r_\nu(s,W)\|^2\,\mathrm{d}s \le \mathbb{E}_\nu\|\Delta\|^2.$$

Rearranging the normal equation as $\lambda v_\nu = g_\nu - C_\nu v_\nu = \mathbb{E}_\nu \int_0^1 E_{U(s)}^* r_\nu(s,W)\,\mathrm{d}s$ and evaluating at u, using $E_u E_{U(s)}^* = k(u,U(s))I_d$ and the fact that E_u passes through the integrals,

$$\lambda\, v_\nu(t,\vartheta,x) = \mathbb{E}_\nu \int_0^1 k\big((t,\vartheta,x,1),\, a_s\big)\, r_\nu(s,W)\,\mathrm{d}s, \qquad \text{where } a_s = a(U(s)). \tag{45}$$

By Lemma D.1(a), the integrand is continuously differentiable in ϑ, with derivative the $d \times d$ matrix $r_\nu(s, W)\, \nabla_\vartheta k((t, \vartheta, x, 1), a_s)^\top$ whose operator norm is at most $L_\sigma \|a_s\|\, \|r_\nu(s, W)\|$. This dominating function does not depend on u and is integrable:

$$\mathbb{E}_\nu \int_0^1 \|a_s\| \|r_\nu(s, W)\|\, \mathrm{d}s \le \sqrt{\mathbb{E}_\nu Q \cdot \mathbb{E}_\nu \int_0^1 \|r_\nu\|^2\, \mathrm{d}s} \le \sqrt{\mathbb{E}_\nu Q \cdot \mathbb{E}_\nu \|\Delta\|^2} < \infty$$

by Cauchy–Schwarz, (34) and (28). Dominated convergence applied to the difference quotients of (45) shows that $D_\vartheta v_\nu$ exists and equals $\lambda^{-1} \mathbb{E}_\nu \int_0^1 r_\nu(s, W) \nabla_\vartheta k(\cdot, a_s)^\top\, \mathrm{d}s$; applied along convergent sequences of u, together with the continuity of $\nabla_a k$, it shows that $D_\vartheta v_\nu$ is jointly continuous. Taking norms gives (36). The linear growth bound is Lemma D.1(b).

(iii) Let training sets $D_n \to D$ in $(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^m)^N$, with averages $(C_n, g_n) \to (C_N, g_N)$ in Hilbert–Schmidt norm and in $\mathcal{H}$ by Lemma D.1(c). With $A_n = (C_n + \lambda I)^{-1}$, the resolvent identity $A_n - A_{N,\lambda} = -A_n(C_n - C_N)A_{N,\lambda}$ gives

$$\begin{aligned} \|A_n g_n - A_{N,\lambda} g_N\| &\le \|A_n\|_{\mathrm{op}} \|g_n - g_N\| + \|A_n - A_{N,\lambda}\|_{\mathrm{op}} \|g_N\| \\ &\le \frac{\|g_n - g_N\|}{\lambda} + \frac{\|C_n - C_N\|_{\mathrm{op}} \|g_N\|}{\lambda^2} \longrightarrow 0. \end{aligned}$$

□

E.4 Proof of Lemma D.3

Proof of Lemma D.3. (i) By Lemma D.2(ii), v_ν is continuous on $\mathcal{U}$ and globally Lipschitz in the state with constant ℓ_ν, uniformly in t. The Picard–Lindelöf theorem gives a unique local solution of (21), and the linear growth $\|v_\nu(t, \vartheta, x)\| \le \|v_\nu(t, 0, x)\| + \ell_\nu \|\vartheta\|$ with $\sup_t \|v_\nu(t, 0, x)\| < \infty$ prevents blow-up, so the solution exists on $[0, 1]$.

(ii) The flow ϕ of v exists on $[0, 1]$ by the same argument. Subtracting the two integral equations,

$$\begin{aligned} \|\phi_t - \phi'_t\| &\le \int_0^t \Big(\|v(s, \phi_s, x) - v(s, \phi'_s, x)\| + \|v(s, \phi'_s, x) - v'(s, \phi'_s, x)\| \Big)\, \mathrm{d}s \\ &\le \int_0^t \Big(\ell' \|\phi_s - \phi'_s\| + \beta(s) \Big)\, \mathrm{d}s \end{aligned}$$

with $\beta(s) = \|v(s, \phi'_s, x) - v'(s, \phi'_s, x)\|$, and Lemma E.1 on $[0, 1]$ with $\gamma_0 = 0$ gives (37) at $t = 1$.

(iii) Let $D_n \to D$ be training sets with empirical ridge fields $\hat{v}_n \to \widehat{v}_{N,\lambda}$ in $\mathcal{H}$ (Lemma D.2(iii)). Apply (ii) with $v = \hat{v}_n$, whose Lipschitz constant $\ell_n = \ell_{\nu_N}(D_n)$ is a continuous function of the training set by (36) and therefore bounded along the sequence, and with $v' = \widehat{v}_{N,\lambda}$, whose trajectory $\hat{\phi}$ is fixed. By Lemma D.1(b),

$$\begin{aligned} \|R_{z,x}(\hat{v}_n) - R_{z,x}(\widehat{v}_{N,\lambda})\| &\le e^{\ell_n} \int_0^1 \|(\hat{v}_n - \widehat{v}_{N,\lambda})(t, \hat{\phi}_t, x)\|\, \mathrm{d}t \\ &\le e^{\ell_n}\, \|\hat{v}_n - \widehat{v}_{N,\lambda}\|_{\mathcal{H}} \sup_t \|a(t, \hat{\phi}_t, x)\| \longrightarrow 0. \end{aligned}$$

(iv) The population trajectory ϕ of part (i) is continuous, and $D_\vartheta v_\lambda$ is jointly continuous by Lemma D.2(ii), so $B(t) = D_\vartheta v_\lambda(t, \phi_t, x)$ is continuous on $[0, 1]$ with $\|B(t)\|_{\mathrm{op}} \le \ell$. The linear equation $\partial_t \Gamma(t, s) = B(t)\Gamma(t, s)$, $\Gamma(s, s) = I_d$, has a unique solution on $[0, 1]$, and for $y \in \mathbb{R}^d$ we have $\|\Gamma(t, s)y\| \le \|y\| + \int_s^t \ell\, \|\Gamma(r, s)y\|\, \mathrm{d}r$ for $t \ge s$, so Lemma E.1 on $[s, 1]$ with $\gamma_0 = \|y\|$ gives $\|\Gamma(t, s)\|_{\mathrm{op}} \le e^{\ell(t-s)}$. Since ϕ is continuous, $B_* < \infty$, and for $h \in \mathcal{H}$ the function $t \mapsto \Gamma(1, t) h(t, \phi_t, x)$ is continuous, with

$$\|Lh\| \le \int_0^1 \|\Gamma(1, t)\|_{\mathrm{op}} \|h(t, \phi_t, x)\|\, \mathrm{d}t \le e^{\ell} \|h\|_{\mathcal{H}} \int_0^1 \|a(t, \phi_t, x)\|\, \mathrm{d}t \le B_* e^{\ell} \|h\|_{\mathcal{H}}$$

by Lemma D.1(b). Linearity of L is clear. □

E.5 Proof of Lemma D.4

Proof of Lemma D.4. Expectations of Bochner-integrable random elements commute with bounded linear maps, so $\mathbb{E}[C_W v_\lambda] = (\mathbb{E}C_W)v_\lambda = Cv_\lambda$ and $\mathbb{E}g_W = g$; hence $\mathbb{E}\eta_\lambda = g - Cv_\lambda - \lambda v_\lambda = 0$ by the population normal equation. By (35) and (34),

$$\begin{aligned}\|\eta_\lambda(W)\|_{\mathcal{H}} &\le \|g_W\| + \|C_W\|_{\mathrm{op}}\|v_\lambda\| + \lambda\|v_\lambda\| \le \sqrt{Q}\|\Delta\| + (Q+\lambda)\|v_\lambda\| \\ &\le \sqrt{12}\,Q + (Q+\lambda)\|v_\lambda\| \le c_\eta S(W)\end{aligned}$$

with $c_\eta = \frac{4}{3}(\sqrt{12} + \|v_\lambda\|_{\mathcal{H}}) + \lambda\|v_\lambda\|_{\mathcal{H}}$, using $S \ge 1$. Thus $s_\lambda^2 \le c_\eta^2\,\mathbb{E}S^2 < \infty$ by (28). The terms $\eta_\lambda(W_i)$ are i.i.d. and centred, so $\mathbb{E}\langle\eta_\lambda(W_i), \eta_\lambda(W_j)\rangle = 0$ for $i \ne j$ and $\mathbb{E}\|\bar\eta_N\|^2 = N^{-2}\sum_i \mathbb{E}\|\eta_\lambda(W_i)\|^2 = s_\lambda^2/N$. Chebyshev's inequality gives $\|\bar\eta_N\| = O_p(N^{-1/2})$. □

E.6 Proof of Lemma D.5

Proof of Lemma D.5. Subtract $(C_N + \lambda I)v_\lambda$ from both sides of the empirical normal equation $(C_N + \lambda I)\widehat{v}_{N,\lambda} = g_N$:

$$(C_N + \lambda I)h_N = g_N - C_N v_\lambda - \lambda v_\lambda = \frac{1}{N}\sum_{i=1}^{N}\big(g_{W_i} - C_{W_i}v_\lambda - \lambda v_\lambda\big) = \bar\eta_N,$$

which is (40). With $\|A_{N,\lambda}\|_{\mathrm{op}} \le 1/\lambda$ (Lemma D.1(d)) and Lemma D.4, $\delta_N \le \|\bar\eta_N\|/\lambda$, $\mathbb{E}\delta_N^2 \le s_\lambda^2/(\lambda^2 N)$ and $\delta_N = O_p(N^{-1/2})$.

The operators C_{W_i} are i.i.d. random elements of the separable Hilbert space of Hilbert–Schmidt operators with $\mathbb{E}\|C_W\|_{\mathrm{HS}} < \infty$ (Lemma D.1(c)), so the strong law of large numbers in separable Banach spaces (Ledoux & Talagrand, 1991) gives $\|C_N - C\|_{\mathrm{HS}} \to 0$ almost surely, hence $\|C_N - C\|_{\mathrm{op}} \to 0$. The resolvent identity $A_{N,\lambda} - A_\lambda = A_{N,\lambda}\big[(C+\lambda I) - (C_N+\lambda I)\big]A_\lambda = -A_{N,\lambda}(C_N - C)A_\lambda$ gives $h_N = A_{N,\lambda}\bar\eta_N = A_\lambda\bar\eta_N - \zeta_N$ with ζ_N as in (41), and $\|\zeta_N\| \le \lambda^{-2}\|C_N - C\|_{\mathrm{op}}\|\bar\eta_N\| = o_p(1)\,O_p(N^{-1/2}) = o_p(N^{-1/2})$. □

E.7 Proof of Lemma D.6

Proof of Lemma D.6. Fix $h \in \mathcal{H}$ with $\|h\|_{\mathcal{H}} \le 1$ and put $v = v_\lambda + h$. By Lemma D.2(ii) and Lemma D.1(b), v is continuous and globally Lipschitz in the state with constant $\ell + L_\sigma\|h\|_{\mathcal{H}}$, so its flow $\hat\phi$ from z exists on $[0,1]$ by the argument of Lemma D.3(i): continuity, a global state-Lipschitz constant and linear growth. The tube $\mathcal{T}$ is compact and $(t,\vartheta) \mapsto D_\vartheta v_\lambda(t,\vartheta,x)$ is continuous (Lemma D.2(ii)), hence uniformly continuous on $\mathcal{T}$, which is the statement $\varrho(\epsilon) \to 0$; note that ϱ is nondecreasing. Since v_λ, ϕ, ℓ and B_* are deterministic, so are c_0, $\mathcal{T}$ and ϱ.

Step 1: the perturbed trajectory stays in the tube. Let $u_h(t) = \hat\phi_t - \phi_t$ be the gap between the perturbed and the population trajectory. Subtracting the two flow equations and writing $v = v_\lambda + h$,

$$\dot u_h(t) = \big[v_\lambda(t,\hat\phi_t,x) - v_\lambda(t,\phi_t,x)\big] + h(t,\hat\phi_t,x).$$

The first bracket has norm at most $\ell\|u_h(t)\|$. For the second term, Lemma D.1(b) gives $\|h(t,\hat\phi_t,x)\| \le \|h\|_{\mathcal{H}}\|a(t,\hat\phi_t,x)\| \le \|h\|_{\mathcal{H}}(B_* + \|u_h(t)\|)$, because $a(t,\hat\phi_t,x)$ and $a(t,\phi_t,x)$ differ only in the state coordinates. Hence

$$\|u_h(t)\| \le \int_0^t \Big((\ell + \|h\|_{\mathcal{H}})\|u_h(s)\| + B_*\|h\|_{\mathcal{H}}\Big)\,\mathrm{d}s,$$

and Lemma E.1 gives $\sup_{t\le1}\|u_h(t)\| \le B_*\|h\|_{\mathcal{H}}e^{\ell+\|h\|_{\mathcal{H}}} \le c_0\|h\|_{\mathcal{H}} \le c_0$, where the restriction $\|h\|_{\mathcal{H}} \le 1$ was used to bound $e^{\|h\|_{\mathcal{H}}}$ by e. Thus $(t,\hat\phi_t) \in \mathcal{T}$ for all t, and so is every point of the segment between ϕ_t and $\hat\phi_t$.

Step 2: the linearised gap is Lh. Let y solve $\dot y(t) = B(t)y(t) + h(t,\phi_t,x)$, $y(0) = 0$; the coefficient B and the forcing term are continuous, so y exists uniquely, and variation of constants gives $y(1) = \int_0^1 \Gamma(1,t)h(t,\phi_t,x)\,\mathrm{d}t = Lh$.

Step 3: the discrepancy solves a linear equation driven by two remainders. Let $e = u_h - y$. Writing $v_\lambda(t, \hat{\phi}_t, x) - v_\lambda(t, \phi_t, x) = B(t)u_h(t) + a_h(t)$ and $h(t, \hat{\phi}_t, x) = h(t, \phi_t, x) + b_h(t)$ with

$$a_h(t) = v_\lambda(t, \phi_t + u_h(t), x) - v_\lambda(t, \phi_t, x) - B(t)u_h(t), \qquad b_h(t) = h(t, \phi_t + u_h(t), x) - h(t, \phi_t, x),$$

the equation for u_h becomes $\dot{u}_h = Bu_h + h(t, \phi_t, x) + a_h + b_h$, and subtracting the equation for y gives $\dot{e} = Be + a_h + b_h$ with $e(0) = 0$. Variation of constants and Lemma D.3(iv) yield

$$e(1) = \int_0^1 \Gamma(1, t)\big(a_h(t) + b_h(t)\big)\,\mathrm{d}t, \qquad \|e(1)\| \le e^\ell \int_0^1 \big(\|a_h(t)\| + \|b_h(t)\|\big)\,\mathrm{d}t.$$

Step 4: both remainders are small. By the mean value theorem along the segment from ϕ_t to $\phi_t + u_h(t)$, which lies in $\mathcal{T}$ by Step 1,

$$a_h(t) = \int_0^1 \big(D_\vartheta v_\lambda(t, \phi_t + \theta u_h(t), x) - B(t)\big)u_h(t)\,\mathrm{d}\theta,$$
$$\|a_h(t)\| \le \varrho(\|u_h(t)\|)\,\|u_h(t)\| \le \varrho(c_0\|h\|_{\mathcal{H}})\,c_0\|h\|_{\mathcal{H}},$$

using that ϱ is nondecreasing. By Lemma D.1(b) applied to h,

$$\|b_h(t)\| \le L_\sigma\|h\|_{\mathcal{H}}\|u_h(t)\| \le L_\sigma c_0\|h\|_{\mathcal{H}}^2.$$

Since $u_h(1) = R_{z,x}(v_\lambda + h) - R_{z,x}(v_\lambda)$ and $u_h(1) = y(1) + e(1) = Lh + e(1)$, the claim (42) holds with $\mathsf{r}(h) = e(1)$ and

$$\|\mathsf{r}(h)\| \le e^\ell\big(\varrho(c_0\|h\|_{\mathcal{H}})c_0\|h\|_{\mathcal{H}} + L_\sigma c_0\|h\|_{\mathcal{H}}^2\big) = e^\ell c_0\big(\varrho(c_0\|h\|_{\mathcal{H}}) + L_\sigma\|h\|_{\mathcal{H}}\big)\|h\|_{\mathcal{H}}.$$

As $\|h\|_{\mathcal{H}} \to 0$ the bracket tends to zero, so $\|\mathsf{r}(h)\| = o(\|h\|_{\mathcal{H}})$: $R_{z,x}$ is Fréchet differentiable at v_λ with derivative L, which is bounded by Lemma D.3(iv). □

E.8 Proofs of Proposition D.1 and Theorem D.1

Proof of Proposition D.1. The maps $A_\lambda : \mathcal{H} \to \mathcal{H}$ and $L : \mathcal{H} \to \mathbb{R}^d$ are bounded and linear (Lemmas D.1(d) and D.3(iv)), so $\xi_\lambda(W_i) = LA_\lambda\eta_\lambda(W_i)$ are i.i.d., $\mathbb{E}\xi_\lambda = LA_\lambda\mathbb{E}\eta_\lambda = 0$, and $\mathbb{E}\|\xi_\lambda\|^2 \le \|L\|_{\mathrm{op}}^2\lambda^{-2}s_\lambda^2 < \infty$ by Lemma D.4. By linearity, $LA_\lambda\bar{\eta}_N = N^{-1}\sum_i \xi_\lambda(W_i)$.

Let $\Omega_N = \{\delta_N \le 1\}$; by Lemma D.5, $\mathbb{P}(\Omega_N^c) \le \mathbb{E}\delta_N^2 \le s_\lambda^2/(\lambda^2 N) \to 0$. On Ω_N, Lemma D.6 applied with $h = h_N$ gives, since $v_\lambda + h_N = \widehat{v}_{N,\lambda}$,

$$\widehat{R}_{N,\lambda} - R_\lambda = Lh_N + \mathsf{r}(h_N), \qquad \|\mathsf{r}(h_N)\| \le e^\ell c_0\big(\varrho(c_0\delta_N) + L_\sigma\delta_N\big)\delta_N.$$

Here $\delta_N = O_p(N^{-1/2})$, and $\varrho(c_0\delta_N) \to 0$ in probability because $\delta_N \to 0$ in probability and ϱ is a deterministic nondecreasing function with $\varrho(\epsilon) \to 0$ as $\epsilon \downarrow 0$; hence $\|\mathsf{r}(h_N)\| = o_p(1)\,O_p(N^{-1/2}) + O_p(N^{-1}) = o_p(N^{-1/2})$. Inserting (41),

$$Lh_N = LA_\lambda\bar{\eta}_N - L\zeta_N = \frac{1}{N}\sum_{i=1}^N \xi_\lambda(W_i) - L\zeta_N, \qquad \|L\zeta_N\| \le \|L\|_{\mathrm{op}}\|\zeta_N\|_{\mathcal{H}} = o_p(N^{-1/2}).$$

The vector $\widehat{R}_{N,\lambda} - R_\lambda - N^{-1}\sum_i \xi_\lambda(W_i)$ is a well-defined random vector by Lemma D.3(iii), equals $\mathsf{r}(h_N) - L\zeta_N$ on Ω_N, and is therefore $o_p(N^{-1/2})$. □

Proof of Theorem D.1. By Proposition D.1, $\sqrt{N}(\widehat{R}_{N,\lambda} - R_\lambda) = N^{-1/2}\sum_{i=1}^N \xi_\lambda(W_i) + o_p(1)$. The multivariate central limit theorem in $\mathbb{R}^d$ applies to the i.i.d. centred square-integrable vectors $\xi_\lambda(W_i)$ and gives $N^{-1/2}\sum_i \xi_\lambda(W_i) \Rightarrow \mathcal{N}(0, V_\lambda(z; x))$; Slutsky's lemma removes the $o_p(1)$ term. The covariance $V_\lambda = \mathbb{E}[\xi_\lambda\xi_\lambda^\top]$ is finite because $\mathbb{E}\|\xi_\lambda\|^2 < \infty$; nothing in the argument forces it to be nonsingular. □

E.9 Proofs of the results of Section C

Proof of Proposition C.1. Let $a = a(u)$, $b = a(u')$, and let e_i be the i-th coordinate vector of $\mathbb{R}^d$. Differentiating (25),

$$\nabla_{c_r} f_{M,i}(u) = M^{-1/2}\sigma(w_r^\top a)\, e_i, \qquad \nabla_{w_r} f_{M,i}(u) = M^{-1/2} c_{r,i}\, \sigma'(w_r^\top a)\, a,$$

where the last coordinate of the second gradient is the derivative with respect to the hidden bias. Taking inner products and summing over r,

$$K_M(u,u') = \frac{1}{M}\sum_{r=1}^{M}\Big\{\sigma(w_r^\top a)\sigma(w_r^\top b)\, I_d + (a^\top b)\,\sigma'(w_r^\top a)\sigma'(w_r^\top b)\, c_r c_r^\top\Big\}. \tag{46}$$

At initialisation $c_r = 0$, so only the first term remains, and it is an average of i.i.d. integrable terms because $|\sigma(w^\top a)\sigma(w^\top b)| \le |w^\top a||w^\top b|$. The strong law of large numbers gives $K_M(u,u') \to \mathbb{E}[\sigma(w^\top a)\sigma(w^\top b)]I_d = K(u,u')$ almost surely, simultaneously over any fixed finite collection of input pairs. The limit is deterministic, continuous and positive semidefinite with $0 \le k(a,a) \le \|a\|^2$ by Lemma D.1(a). On an augmented-input ball of radius R_0 the summand is continuous and dominated in matrix norm by $R_0^2\|w\|^2$, and Proposition 2 of McNamara et al. (2025) gives uniform almost-sure convergence on each compact set. A countable exhaustion of $\mathcal{U}\times\mathcal{U}$ by compact balls gives a single probability-one event on which all these convergences hold. □

Proof of Theorem C.1. Throughout, the training set is fixed. Let $\mu_N = \nu_N \otimes \mathrm{Leb}[0,1]$ be the probability measure on path points, so that for a function q of a path point

$$\int q\,\mathrm{d}\mu_N = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 q(s,W_i)\,\mathrm{d}s, \qquad \text{and put} \qquad d_N = \int \|\Delta\|^2\,\mathrm{d}\mu_N.$$

The objective (26) reads $\mathcal{F}_{N,M}(\omega) = \int \|f_M(U;\omega) - \Delta\|^2\,\mathrm{d}\mu_N + \lambda\sum_r \|c_r\|^2$.

Step 1: energy identity and a priori bounds. The training paths are compact and σ is smooth, so differentiation under the integrals shows that $\mathcal{F}_{N,M}$ is smooth in ω with a locally Lipschitz gradient; hence the gradient flow has a unique local solution. Along it, the chain rule gives $\frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{F}_{N,M}(\omega(\tau)) = -\|\dot\omega(\tau)\|^2$, and since $\omega(0)$ has $c_r(0) = 0$ the initial value is $\mathcal{F}_{N,M}(\omega(0)) = d_N$. Integrating,

$$\mathcal{F}_{N,M}(\omega(\tau)) + \int_0^\tau \|\dot\omega(r)\|^2\,\mathrm{d}r = d_N.$$

By Cauchy–Schwarz the displacement on $[0,T]$ is at most $\sqrt{Td_N}$, so the flow extends to every finite time, and both parts of $\mathcal{F}_{N,M}$ being nonnegative, for $0 \le \tau \le T$,

$$\int \|f_{N,M,\tau}(U) - \Delta\|^2\,\mathrm{d}\mu_N \le d_N, \qquad \sum_r \|c_r(\tau)\|^2 \le \frac{d_N}{\lambda}, \qquad \sum_r \|w_r(\tau) - w_r(0)\|^2 \le T d_N. \tag{47}$$

Set $b_{M,\tau} = M^{-1}\sum_r \|w_r(\tau)\|^2$. The strong law of large numbers and the last bound in (47) give

$$b_{M,0} \longrightarrow p \ \text{ almost surely}, \qquad \sup_{\tau\le T}\sqrt{b_{M,\tau}} \le \sqrt{b_{M,0}} + \sqrt{Td_N/M}. \tag{48}$$

In words: the total squared movement of the hidden weights is bounded independently of the width, while their total squared size grows like Mp, so the hidden layer barely moves relative to its initialisation.

Step 2: the two-time feature kernel. For $\tau, r \in [0,T]$ define

$$H_M^{\tau,r}(u,v) = \frac{1}{M}\sum_{j=1}^{M}\sigma\big(w_j(\tau)^\top a(u)\big)\,\sigma\big(w_j(r)^\top a(v)\big)\, I_d.$$

Fix $T < \infty$ and a compact $\mathcal{B} \subset \mathcal{U}$. Since (30) for a larger compact set implies it for a smaller one, we may and do assume that $\mathcal{B}$ contains the finitely many training paths and the compact set $\mathcal{K}$ of

Step 7 below, which does not depend on $\mathcal{B}$. Let $R_{\mathcal{B}} = \sup_{u\in\mathcal{B}}\|a(u)\|$, write $a = a(u)$, $b = a(v)$ and split

$$\begin{aligned}\sigma(w_j(\tau)^\top a)\sigma(w_j(r)^\top b) - \sigma(w_j(0)^\top a)\sigma(w_j(0)^\top b)\\ = \big(\sigma(w_j(\tau)^\top a) - \sigma(w_j(0)^\top a)\big)\sigma(w_j(r)^\top b)\\ + \sigma(w_j(0)^\top a)\big(\sigma(w_j(r)^\top b) - \sigma(w_j(0)^\top b)\big).\end{aligned}$$

Using $|\sigma(q)| \le |q|$, $|\sigma(q) - \sigma(q')| \le L_\sigma|q - q'|$ and Cauchy–Schwarz over j, and then (47)–(48),

$$\sup_{\tau,r\le T}\sup_{u,v\in\mathcal{B}}\|H_M^{\tau,r}(u,v) - H_M^{0,0}(u,v)\|_{\mathrm{op}} \le L_\sigma R_{\mathcal{B}}^2\Big(2\sqrt{b_{M,0}} + \sqrt{Td_N/M}\Big)\sqrt{Td_N/M} \longrightarrow 0.$$

Since $H_M^{0,0} = K_M$ at initialisation, Proposition C.1 gives uniform almost-sure convergence of $H_M^{0,0}$ to K on $\mathcal{B}$, on a probability-one event common to all compact balls; the estimates above hold pathwise for the fixed training set and its unique gradient-flow solution. Together,

$$\varepsilon_M := \sup_{\tau,r\le T}\sup_{u,v\in\mathcal{B}}\|H_M^{\tau,r}(u,v) - K(u,v)\|_{\mathrm{op}} \longrightarrow 0 \quad \text{almost surely.} \tag{49}$$

Step 3: the network solves an integral equation. The gradient-flow equation for the j-th output weight is

$$\dot c_j(\tau) = -\frac{2}{\sqrt{M}}\int \sigma\big(w_j(\tau)^\top a(U)\big)\big\{f_{N,M,\tau}(U) - \Delta\big\}\,\mathrm{d}\mu_N - 2\lambda c_j(\tau),$$

where the damping term is exactly $-2\lambda c_j$ because the penalty acts on the output weights only and carries no factor M^{-1}. Since $c_j(0) = 0$, variation of constants solves this linear equation, and substituting the solution into the network representation (25) gives

$$f_{N,M,\tau}(u) = -2\int_0^\tau e^{-2\lambda(\tau-r)}\int H_M^{\tau,r}(u,U)\big\{f_{N,M,r}(U) - \Delta\big\}\,\mathrm{d}\mu_N\,\mathrm{d}r. \tag{50}$$

The hidden-weight dynamics are never solved: they enter only through $H_M^{\tau,r}$, which Step 2 controls.

Step 4: kernel gradient flow solves the same equation with K. The path (29) satisfies $\tilde v_{N,0} = 0$ and $\frac{\mathrm{d}}{\mathrm{d}\tau}\tilde v_{N,\tau} = -2(C_N + \lambda I)\tilde v_{N,\tau} + 2g_N$, which is gradient flow on J_{ν_N} in $\mathcal{H}$ started from zero. Evaluated at u, using $(C_N f)(u) = \int K(u,U)f(U)\,\mathrm{d}\mu_N$ and $g_N(u) = \int K(u,U)\Delta\,\mathrm{d}\mu_N$, variation of constants gives (50) with K in place of $H_M^{\tau,r}$ and $\tilde v_{N,r}$ in place of $f_{N,M,r}$.

Step 5: proof of (30). Put $\kappa_{\mathcal{B}} = \sup_{u,v\in\mathcal{B}}\|K(u,v)\|_{\mathrm{op}}$ and $G_M(\tau) = \sup_{u\in\mathcal{B}}\|f_{N,M,\tau}(u) - \tilde v_{N,\tau}(u)\|$; recall that $\mathcal{B}$ contains the training paths. Subtracting the two integral equations and splitting the integrand as $(H_M^{\tau,r} - K)(f_{N,M,r} - \Delta) + K(f_{N,M,r} - \tilde v_{N,r})$, the first part is bounded by $\varepsilon_M\int\|f_{N,M,r} - \Delta\|\,\mathrm{d}\mu_N \le \varepsilon_M\sqrt{d_N}$ by Cauchy–Schwarz and (47), and the second by $\kappa_{\mathcal{B}}G_M(r)$. With $e^{-2\lambda(\tau-r)} \le 1$,

$$G_M(\tau) \le 2\int_0^\tau\big(\varepsilon_M\sqrt{d_N} + \kappa_{\mathcal{B}}G_M(r)\big)\,\mathrm{d}r, \qquad \text{so} \qquad \sup_{\tau\le T}G_M(\tau) \le 2T\sqrt{d_N}\,\varepsilon_M\,e^{2\kappa_{\mathcal{B}}T}$$

by Lemma E.1 on $[0,T]$, which tends to zero by (49). This proves (30).

Step 6: proof of (31). From (29), $\tilde v_{N,\tau} - \widehat v_{N,\lambda} = -e^{-2\tau(C_N+\lambda I)}\widehat v_{N,\lambda}$. The operator C_N is bounded, self-adjoint and positive semidefinite (Lemma D.1(c),(d)), so by the spectral theorem $\|e^{-2\tau(C_N+\lambda I)}\|_{\mathrm{op}} = e^{-2\lambda\tau}\|e^{-2\tau C_N}\|_{\mathrm{op}} \le e^{-2\lambda\tau}$, which gives (31). The same argument gives $\|\tilde v_{N,\tau}\|_{\mathcal{H}} \le \|\widehat v_{N,\lambda}\|_{\mathcal{H}}$ for all $\tau \ge 0$.

Step 7: proof of (32). For arbitrary $u, v \in \mathcal{U}$ the network representation, $|\sigma(q)-\sigma(q')| \le L_\sigma|q-q'|$, $|\sigma(q)| \le |q|$ and Cauchy–Schwarz over neurons with (47) give

$$\|f_{N,M,\tau}(u) - f_{N,M,\tau}(v)\| \le L_\sigma\sqrt{\frac{d_N}{\lambda}}\sqrt{b_{M,\tau}}\,\|a(u) - a(v)\|,$$

$$\|f_{N,M,\tau}(u)\| \le \sqrt{\frac{d_N}{\lambda}}\sqrt{b_{M,\tau}}\,\|a(u)\|.$$

Hence each $f_{N,M,\tau}$ is continuous and globally Lipschitz in the state, so its flow from z exists uniquely on $[0,1]$ and every finite-network endpoint is well defined; by (48) its Lipschitz constant is bounded uniformly in M (almost surely, for all large M) and in $\tau \le T$. Next, $\|\tilde{v}_{N,\tau}\|_{\mathcal{H}} \le \|\widehat{v}_{N,\lambda}\|_{\mathcal{H}}$ and Lemma D.1(b) give the linear growth bound $\|\tilde{v}_{N,\tau}(t,\vartheta,x)\| \le \|\widehat{v}_{N,\lambda}\|_{\mathcal{H}}(\sqrt{2+\|x\|^2}+\|\vartheta\|)$ for all $\tau \ge 0$, so Lemma E.1 places the trajectories of $\tilde{v}_{N,\tau}$, for all $\tau \ge 0$, in one compact set $\mathcal{K} \subset \mathcal{U}$, which is the set that $\mathcal{B}$ was chosen to contain in Step 2. Fix τ and apply the stability estimate (37) with $v = f_{N,M,\tau}$ and $v' = \tilde{v}_{N,\tau}$: the integrand is at most $\sup_{u\in\mathcal{B}}\|f_{N,M,\tau}(u) - \tilde{v}_{N,\tau}(u)\|$, which tends to zero by (30), while the Lipschitz constant of v stays bounded. Hence $R_{z,x}(f_{N,M,\tau}) \to R_{z,x}(\tilde{v}_{N,\tau})$ as $M \to \infty$. Next apply (37) with $v = \widehat{v}_{N,\lambda}$, whose Lipschitz constant is ℓ_{ν_N} by Lemma D.2(ii), and $v' = \tilde{v}_{N,\tau}$: by Lemma D.1(b) and (31),

$$\begin{aligned}\|R_{z,x}(\widehat{v}_{N,\lambda}) - R_{z,x}(\tilde{v}_{N,\tau})\| &\le e^{\ell_{\nu_N}} \int_0^1 \|(\widehat{v}_{N,\lambda} - \tilde{v}_{N,\tau})(t, \phi_t^{\tilde{v}_{N,\tau}}, x)\|\,\mathrm{d}t \\ &\le e^{\ell_{\nu_N}}\, e^{-2\lambda\tau}\|\widehat{v}_{N,\lambda}\|_{\mathcal{H}}\, R_{\mathcal{B}} \to 0\end{aligned}$$

as $\tau \to \infty$. Combining the two limits proves (32). □

F Experimental setting

This appendix documents the experimental setting of Section 5 in the detail needed to reproduce it. Section F.1 describes the three simulators: the scalar bimodal simulator, whose posterior is available exactly and which underlies the controlled experiments of Sections 5.1–5.3, and the two epidemiological simulators, CRKP and BDEISS-CT, behind the domain-specific evaluation of Section 5.4. Section F.2 specifies the velocity network and the training procedure of the near-optimal FMPE configuration from which the suboptimal configurations of Section 5.2 are derived. Section F.3 gives the statistical details of the regression-based scaling test: the log-variance regression behind the reported $\hat{\alpha}$ and p-values, its design-constant standard error, and the regime checks applied before the test is read.

F.1 Simulators

Each simulator specifies a prior π on the parameter Θ and a sampling law $P(\cdot \mid \Theta)$ for the observation X. FMPE is trained on N independent joint draws (Θ_i, X_i), the base state is $Z \sim q_0 = \mathcal{N}(0, I)$, and the observation enters the velocity network as conditioning only. Table 2 summarizes the three simulators.

	Bimodal	CRKP	BDEISS-CT
Used in	Sections 5.1–5.3	Section 5.4	Section 5.4
Parameter	$\theta \in \mathbb{R}$	$\beta \in \mathbb{R}^7_{>0}$	$(R, d, f_E, f_S, X_S, \upsilon, X_C) \in \mathbb{R}^7$
Prior	$\mathcal{N}(0,1)$	independent lognormal	independent uniform
Observation	scalar	seven weekly incidence series	218 tree summary statistics
Forward model	closed form	stochastic discrete-time SI	multi-type birth–death tree
Exact posterior	by quadrature	not in closed form	not in closed form

Table 2: Simulators used in the experiments of Section 5.

Bimodal simulator. Sections 5.1–5.3 use a scalar simulator with a Gaussian prior and a quadratic observation model,

$$\Theta \sim \mathcal{N}(0,1), \qquad X = \Theta^2 + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,\sigma^2), \qquad \sigma = 0.25. \tag{51}$$

The observation determines θ only up to sign, so the posterior

$$p(\theta \mid x) = \frac{1}{C(x)} \exp\Big[-\frac{\theta^2}{2} - \frac{(x-\theta^2)^2}{2\sigma^2}\Big] \tag{52}$$

is symmetric about zero. Its log density is concave in $u = \theta^2$ with maximum at $u = x - \sigma^2/2$, so the posterior has two modes at $\pm\sqrt{x-\sigma^2/2}$ separated by a valley at zero whenever $x > \sigma^2/2 =$

0.03125, and a single mode at zero otherwise. Along the observation axis the posterior thus passes from unimodal to bimodal, with modes that separate further as x grows. Despite its low dimension the simulator is a demanding target for a transport map: the population velocity field is nonlinear in the state, and a fitted flow that carries base mass across the symmetry axis has placed it in the wrong mode.

All reference quantities are computed by the simulator itself. The marginal law of X is a χ^2_1 distribution blurred by the noise, $\mathbb{P}(X \le x) = \mathbb{E}\,\Phi\big((x-\Theta^2)/\sigma\big)$, evaluated by adaptive quadrature over the prior and inverted by root finding; the observation probes at which endpoints are recorded are placed at its quantiles (the primary probe is $z = 0.5$ at the marginal median $x \approx 0.511$), so that they sit where observations actually fall. The normalizer $C(x)$, the posterior distribution function F_x and its moments are obtained by mode-centred adaptive quadrature, and posterior quantiles by root finding on the integrated tail, with tolerances tightened until quantiles and moments are stable to 10^{-6}. In one dimension the flow of a smooth velocity field cannot cross itself, so for each x its time-one map is increasing, and the unique increasing map sending $\mathcal{N}(0,1)$ to $p(\cdot \mid x)$ is the quantile transport. The population endpoint at a probe (z, x) is therefore the posterior quantile $R^\star(z; x) = F_x^{-1}(\Phi(z))$, and reference posterior samples for the Wasserstein-1 distance of Figure 1 are drawn by inverting F_x at uniform variates.

CRKP simulator. The first epidemiological simulator of Section 5.4 is the stochastic, discrete-time susceptible–infected (SI) model of Chatha et al. (2026) for healthcare-associated transmission of carbapenem-resistant *Klebsiella pneumoniae* (CRKP) in a long-term acute care hospital. Colonization is persistent and confers no immunity, so a patient never recovers within a stay. A facility with P floors and R rooms hosts N patients over T weekly steps, and $X_t^{(i)} \in \{0,1\}$ records whether patient i is colonized at the start of week t. Transmission is spatially heterogeneous: the parameter is a vector of positive rates $\beta = (\beta_0, \beta_1, \ldots, \beta_P, \beta_{P+1})$, where β_0 is the facility-wide rate of transmission-producing contacts per colonized patient and week, β_p the additional rate towards floormates on floor p, and β_{P+1} the additional rate towards roommates. A colonized donor j contributes to the force of infection on a susceptible recipient i through

$$\lambda_{i\leftarrow j} = \frac{\beta_0}{N} + \frac{\beta_{F(i)}}{N_{F(i)}}\,\mathbf{1}[i,j \text{ floormates}] + \frac{\beta_{P+1}}{N_R}\,\mathbf{1}[i,j \text{ roommates}], \qquad \lambda_i(t) = \sum_{j:\, X_{t-1}^{(j)}=1} \lambda_{i\leftarrow j}, \tag{53}$$

where $F(i)$ is the floor of patient i, $N_{F(i)}$ its population and N_R the room capacity. Over a week a susceptible patient acquires CRKP with probability $1 - e^{-\lambda_i(t)}$, independently across patients, and a newly admitted patient cannot acquire it before the following week. Patient turnover is taken from the patient trace data of the reference study: admission and discharge times, the floor and room occupied in each week, and the screening result on admission, which fixes imported cases, are supplied as fixed inputs, so randomness enters only through within-facility acquisition. The reference data set is a 2012–2013 intervention study in a Chicago-area facility with 890 patients, five floors, 95 rooms and $T = 53$ weeks, in which every patient was screened on admission and every other week thereafter, so that cases are essentially fully observed. Following Chatha et al. (2026), β receives an independent lognormal prior, $\log\beta \sim \mathcal{N}(\mu, I_7)$, with a larger prior mean for the facility rate ($\mu_0 = -3$) than for the floor and room rates (-4 and -5), which shrinks towards a mainly homogeneous model in the absence of evidence to the contrary. The observation is not the $N \times T$ status matrix but seven weekly summary series, each of length T: the facility-wide proportion of colonized patients, the five floor-specific proportions, and the proportion of rooms in which both roommates are colonized. The posterior over β is unimodal but far from isotropic: the facility-level and room-level rates are strongly negatively correlated and the sublocation rates are only weakly identified from a single outbreak.

BDEISS-CT simulator. The second epidemiological simulator is the Birth–Death Exposed–Infectious model with Super-Spreading and Contact Tracing (BDEISS-CT) of Xie et al. (2026), a multi-type birth–death phylodynamic model whose observation is a pathogen phylogeny rather than an incidence series. A newly infected individual enters the exposed state E and becomes infectious at rate μ, so the incubation period has mean $d_{\text{inc}} = 1/\mu$; on becoming infectious, a fraction f_S become super-spreaders (S), who transmit at rate $X_S\lambda$ with $X_S > 1$, and the rest become regular spreaders (I), who transmit at rate λ. Infectious individuals are removed at rate ψ, so the infec-

tious period has mean $d_{\text{inf}} = 1/\psi$; at removal the pathogen is sampled with probability ρ, and only sampled removals become tips of the tree. At removal an individual also notifies each contact (their infector and the people they infected) independently with probability υ; a notified contact moves to the contact-traced version of its state (E_C, I_C or S_C) and, once infectious, is removed at the accelerated rate $X_C\psi$ with $X_C > 1$. Contact tracing therefore shortens infectious periods and lowers the realized reproduction number, and for the contact-traced models neither R nor d is a closed-form function of the rates. The model is parametrized by $\Theta = (R, d, \rho, f_E, f_S, X_S, \upsilon, X_C)$: the average reproduction number R, equal to $(1 + (X_S - 1)f_S)\lambda d_{\text{inf}}$ without contact tracing; the average infection duration $d = d_{\text{inc}} + d_{\text{inf}}$; the sampling probability ρ; the incubation fraction $f_E = d_{\text{inc}}/d$; the super-spreader fraction f_S and transmission increase X_S; and the contact-tracing probability υ and removal speed-up X_C. Setting $\upsilon = 0$ or $X_C = 1$ removes contact tracing, $f_E = 0$ removes incubation, and $f_S = 0$ or $X_S = 1$ removes super-spreading, so BDEISS-CT nests seven simpler models down to the three-parameter BD model (R, d, ρ), and near these boundaries the corresponding parameters are only weakly identified.

Simulation follows Xie et al. (2026). Parameters are drawn independently and uniformly, $R \in [1, 10)$, $d \in [1, 31)$, $\rho \in [0.05, 0.75)$, $f_E \in [0, 1)$, $f_S \in [0, 0.5)$, $X_S \in [1, 25)$, $\upsilon \in [0, 0.75)$ and $X_C \in [1, 100)$, ranges wide enough to cover epidemics from HIV-1 to SARS-CoV-2 and Ebola. Given the parameters, a transmission tree is simulated forward with the `treesimulator` package until a target number of sampled tips, itself drawn uniformly from a prescribed range (the reference study uses ranges from 200–500 up to 2,000–5,000 tips), and the time-scaled tree on the sampled tips is the observation. Each tree is rescaled to unit mean branch length, with the target d rescaled accordingly, and encoded as a standardized vector of 218 summary statistics describing branch lengths, tree topology and the lineages-through-time curve, augmented with statistics sensitive to contact tracing and including the sampling probability ρ, which is needed for identifiability and treated as known; the seven remaining parameters are the target of inference. The likelihood of a tree under a multi-type birth–death model has no closed form, and the wide prior, the degenerate sub-models at the boundary and the 218-dimensional observation make BDEISS-CT the most demanding of the three simulators. Consistently, Section 5.4 shows that FMPE needs the largest training sets on this simulator before its posterior accuracy becomes competitive with a normalizing-flow NPE.

F.2 Neural network architecture and optimizer

This section describes the velocity network and the training procedure used for the bimodal simulator in Sections 5.1–5.3. The near-optimal configuration is the one described below; the suboptimal configurations of Section 5.2 change one factor of it at a time, as described there.

Velocity network. The velocity field $v(t, y, x)$ acts on the scalar path state $y \in \mathbb{R}$ and is conditioned on the time $t \in [0, 1]$ and the observation x. Time enters through a fixed Fourier embedding $e(t) = (t, \sin 2\pi kt, \cos 2\pi kt)_{k\in\{1,2,4\}} \in \mathbb{R}^7$, which retains the raw t so that the endpoints of the path are distinguished. The observation is standardized with fixed prior-predictive constants, $\tilde{x} = (x - m)/s$ with $m = \mathbb{E}X = 1$ and $s^2 = \operatorname{Var} X = 2 + \sigma^2 = 2.0625$ for the bimodal simulator; these are population constants owned by the simulator, not statistics of the training data. A linear input layer maps $(y, \tilde{x}, e(t))$ to a hidden state of width 512, and a two-layer SiLU network maps $(\tilde{x}, e(t))$ to a context vector $c \in \mathbb{R}^{512}$. The hidden state then passes through six conditioned residual blocks,

$$h \;\leftarrow\; h + \tfrac{1}{\sqrt{6}}\, W_2\, \mathrm{SiLU}\big(W_1\, \mathrm{SiLU}\big((1 + \gamma(c)) \odot \mathrm{LN}(h) + \beta(c)\big)\big), \tag{54}$$

where LN is layer normalization ($\epsilon = 10^{-5}$) and γ, β are linear maps of the context, so the conditioning acts by feature-wise affine modulation. The output head applies layer normalization, a SiLU and a linear map to $\mathbb{R}$, and a linear skip connection on $(y, \tilde{x}, t)$ is added. The network output is the residual field $g(t, y, x; \omega)$, and the velocity is $v(t, y, x) = \hat{\mu} + g(t, y, x; \omega)$, where the buffer $\hat{\mu} = N^{-1} \sum_{i=1}^{N} \Theta_i$ is the sample mean of the training parameters, fixed at fit time. Under the standard normal prior the constant mode of the target velocity, $\mathbb{E}[\Theta - Z]$, is zero, so $\hat{\mu}$ carries exactly the $1/N$ fluctuation of the prior mean and the parameters learn the zero-mean remainder, including every observation-dependent part of the level. Weights are initialized with Xavier-uniform draws and zero biases; the second linear map of every block is scaled by 10^{-2} and the modulation maps γ, β start at zero, so each block is close to the identity at initialization. No dropout is used, so the

fitted network is a deterministic function of its training data and seeds. The network has $6{,}583{,}813$ parameters.

Training. The network is trained on the FMPE loss (1) along the linear path $Y_t = (1-t)Z + t\Theta$ with target velocity $\Theta - Z$, conditioned on the simulation's X, by minibatch stochastic optimization with an Adam-type optimizer. The time integral is estimated by Monte Carlo over uniformly sampled times, and the base state Z is drawn afresh whenever a simulation is visited, so the fitted flow depends on the N simulations alone. Regularization is a decoupled weight decay, the finite-network counterpart of the ridge penalty $\lambda\|f\|_{\mathcal{H}}^2$ in Assumption B.3. Each fit runs for the fixed number of optimization steps stated in Section 5.1, without early stopping or validation-based selection, and endpoints of the trained field are computed with a fixed step ODE solver.

Epidemiological simulators. For the CRKP and BDEISS-CT experiments of Section 5.4, we adapt the neural networks provided by the papers introducing the two simulators (Chatha et al., 2026; Xie et al., 2026) as the velocity network of FMPE, with the summary statistics of Section F.1 as the conditioning input.

F.3 Regression-based scaling test

This section gives the statistical details of the regression-based scaling test of Section 3. Fix a probe (z, x) and training-set sizes $N_1 < \cdots < N_K$. At each size, n FMPE models are fitted to n independent training sets, their endpoints $\hat{R}_{N_k}^{(j)}$, $j = 1, \dots, n$, are computed with the same solver, and $S_{N_k}^2$ is their sample variance with $n-1$ degrees of freedom. In the experiments $K = 5$, the sizes run from $15{,}000$ to $240{,}000$ a factor of two apart, and $n = 20$.

The regression. Under Theorem D.1 with a nondegenerate one-dimensional limit variance $V = V_\lambda(z; x)$, the endpoints at size N are approximately $\mathcal{N}(R_\lambda, V/N)$, so $(n-1)S_N^2/(V/N)$ is approximately χ_{n-1}^2 and

$$\log S_N^2 = \log V - 2\alpha \log N + b_n + \varepsilon_N, \qquad b_n = \psi\big(\tfrac{n-1}{2}\big) - \log \tfrac{n-1}{2}, \qquad \mathrm{Var}(\varepsilon_N) = \psi'\big(\tfrac{n-1}{2}\big), \tag{55}$$

with $\alpha = 1/2$, where ψ and ψ' are the digamma and trigamma functions, b_n is the mean of $\log\big(\chi_{n-1}^2/(n-1)\big)$ and ε_N its centred fluctuation. For $n = 20$, $b_n = -0.054$ and the standard deviation of ε_N is 0.333; both are known before any data exist and neither depends on V. The exponent is estimated by ordinary least squares of $\log S_{N_k}^2 - b_n$ on $\log N_k$ with an intercept, $\hat{\alpha} = -\hat{\beta}/2$ for the fitted slope $\hat{\beta}$. Since the error variance is known and the same at every size, least squares is efficient and the standard error is a design constant fixed by the allocation alone,

$$\mathrm{SE}(\hat{\alpha}) = \frac{\sqrt{\psi'((n-1)/2)}}{2\sqrt{\sum_k (\log N_k - \overline{\log N})^2}} = 0.076 \tag{56}$$

for the sizes above, so the 95% interval $\hat{\alpha} \pm 1.96\,\mathrm{SE}$ has half-width 0.149 and the exponents $1/2$ and 1 lie 6.6 standard errors apart. The test of $H_0\colon \alpha = 1/2$ uses $z = (\hat{\alpha} - 1/2)/\mathrm{SE}(\hat{\alpha})$ with a two-sided normal p-value; these are the $\hat{\alpha}$ and p-values reported in Section 5. The intercept is an unbiased estimate of $\log V$ and is kept on the log scale. Because the residual variance is known, the residuals $\hat{\varepsilon}_k$ also support a lack-of-fit statistic $X^2 = \sum_k \hat{\varepsilon}_k^2/\psi'((n-1)/2)$, which is χ_{K-2}^2 under the power-law model; it responds to curvature, a break or a variance floor, not to a different exponent, which is still a power law.

Regime checks. The regression presumes the Gaussian regime of Theorem D.1 at every size. For a multimodal posterior this can fail at small N when some fits transport z to the wrong mode: the variance then measures mode assignment rather than Gaussian fluctuation, and the endpoint map is effectively discontinuous in the data. A pre-registered gate is therefore applied before the confirmatory fit. A size enters only if every endpoint is finite, no fitted endpoint lies on the far side of the posterior's symmetry axis from the reference endpoint (for the bimodal simulator the axis is $\theta = 0$ whenever the posterior at x is bimodal), and the endpoints pass a Shapiro–Wilk normality test at level 0.01. Excluded sizes are reported with their diagnostics, and $\hat{\alpha}$ is refitted on the admissible

sizes with the lever arm and standard error recomputed; no other choice of fit window is allowed. Alongside the test we report the Shapiro–Wilk p-values of the endpoints at each size combined across sizes by Fisher's method, and a Brown–Forsythe test of equal variance of the $\sqrt{N}$-scaled endpoints across sizes, a distribution-free cross-check of $\alpha = 1/2$ that does not rely on the Gaussian approximation.

Probes and multiplicity. Confirmatory inference uses the pre-registered primary probe only, $z = 0.5$ at the marginal median of the observation. The remaining probes, sixteen base states on $[-0.5, 0.5]$ at five observation quantiles, are exploratory: the per-probe exponents are mapped under Benjamini–Hochberg control at level 0.05, and the exponent of the probe-averaged $\log S_N^2$, with a delete-one-model jackknife standard error, is reported as a pooled summary. All probes share the same n models and the limit variance $V(z; x)$ varies across them, so the pooled summary is a different estimand and is labelled as such. The pseudo-dataset test of Algorithm 1 applies exactly the same statistic to models retrained on pseudo-datasets.