Attention as In-Context Empirical Bayes: A Two-Stage View via Particle Dynamics

Matthew Smart[1,4,*] **Soumya Ganguly**[2,*] **Nilava Metya**[2] **Alexandre V. Morozov**[3] **Anirvan M. Sengupta**[3,5]

Abstract

We study minimal attention-only transformers under all-token corruption and show they admit a two-stage empirical Bayes interpretation. A single attention step computes a kernel-weighted posterior mean with respect to the empirical distribution defined by the context. Depth refines this distribution through particle dynamics (Stage 1), while a long-range skip-connection carries the noisy input as a query for posterior inference (Stage 2), revealing distinct statistical roles for depth and attention residuals. The framework isolates a minimal setting in which the context itself induces a depth-dependent energy landscape governing in-context inference. We show that effective denoising can emerge without an explicit noise schedule: a fixed kernel bandwidth and finite integration horizon suffice, yielding a principled depth-noise relationship. We further establish a posterior-mean recovery guarantee for a class of well-behaved priors, where the empirical estimator converges to the Bayes-optimal predictor under asymptotic conditions. Connecting these dynamics to reverse-diffusion limits, our results provide a statistical interpretation of attention as in-context inference via sample-based posterior estimation, without explicit density modeling.

1. Introduction

Transformers (Vaswani et al., 2017) have achieved remarkable empirical success across language, vision, and generative modeling, yet the mechanisms underlying this flexibility are only partially understood. Several distinct perspectives on their defining ingredient — *attention* — have emerged. Geshkovski et al. (2023); Rigollet (2025); Bruno et al. (2025) treat multilayer attention as interacting particle systems; Rosu et al. (2025) connect attention to score-based denoising; and Ramsauer et al. (2021) interpret attention as memory retrieval in dense associative memory networks (Krotov & Hopfield, 2016). This latter connection raises the question of what task would naturally motivate identity-scaled weights and single-step energy minimization. Smart et al. (2025) proposed *in-context denoising* as such a task, showing that a single attention layer provably solves Bayes-optimal denoising when context tokens are clean. However, these perspectives have largely been studied in isolation, and their relationship — as well as the architectural consequences they imply — remains unclear.

Motivated by this lineage, we study transformer dynamics through a collective in-context denoising task: N tokens drawn i.i.d. from an unknown prior ρ_0 are corrupted by isotropic Gaussian noise of known variance σ^2 and processed jointly by multilayer self-attention, with the goal of recovering the clean tokens. This generalizes Smart et al. (2025) to a regime in which the prior itself must be inferred from corrupted samples, so that depth becomes necessary. Our central claim is that attention-only transformers in this setting implement a finite-sample empirical Bayes (EB) denoiser, decomposable into two stages (Fig. 1). *Stage 1* (self-attention across depth) iteratively refines a particle approximation (Del Moral, 2013; Sander et al., 2022) to the prior ρ_0 from the corrupted samples. *Stage 2* (a final cross-attention step from the noisy input, enabled by a long-range residual connection) computes a posterior mean against the refined prior. This decomposition gives the architectural elements of depth and attention residuals statistically derivable roles.

[*]Equal contribution [1] Lewis-Sigler Institute for Integrative Genomics, Princeton University, NJ [2] Department of Mathematics, Rutgers University, NJ [3] Department of Physics and Astronomy, Rutgers University, NJ [4] Center for Computational Biology, Flatiron Institute [5] Center for Computational Quantum Physics and Center for Computational Mathematics, Flatiron Institute, NY
Correspondence: mattsmart@princeton.edu, soumya.ganguly@rutgers.edu, nilava.metya@rutgers.edu, morozov@physics.rutgers.edu, anirvans.physics@gmail.com
Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

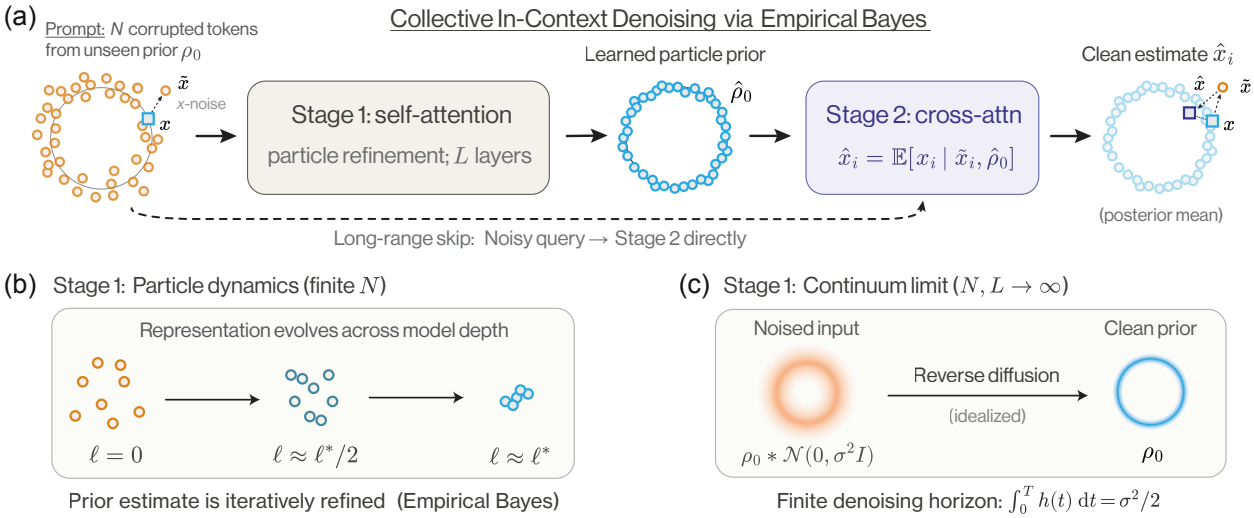

Figure 1. (a) **Collective in-context denoising**: Multilayer attention implements a two-stage EB procedure. Depth iteratively refines a particle prior (Stage 1), while a long-range skip— acting as an Attention Residual (AttnRes)—performs posterior averaging against the initial noisy input (Stage 2). (b) **Particle dynamics**: Discrete self-attention updates move corrupted tokens from their initial noised distribution toward high-density regions of the underlying clean prior. (c) **Theoretical limit**: In the large-context, continuous-depth regime, these dynamics recover an anti-diffusive denoising operator consistent with a reverse diffusion process.

This framework has roots in classical empirical Bayes (Robbins, 1956; Miyasawa, 1961; Efron, 2011; Johnstone & Silverman, 2005; Raphan & Simoncelli, 2011), where multiple observations from a common unknown prior are used to estimate said prior before performing posterior inference. Under Gaussian corruption, the posterior mean connects naturally to score-based denoising through Tweedie's formula, linking our perspective to modern diffusion modeling (Ho et al., 2020; Song et al., 2021). Our approach complements existing perspectives on attention: unlike recent score-based denoising treatments (Rosu et al., 2025), we provide an explicit empirical Bayes interpretation, and we do not impose a noise schedule. By connecting mean-field particle dynamics (Geshkovski et al., 2023; Rigollet, 2025; Bruno et al., 2025) to a denoising task with a finite time horizon, our work contextualizes the long-time regimes those works emphasize. Finally, our two-stage decomposition reveals a *dynamic* associative memory landscape (Ramsauer et al., 2021; Smart et al., 2025) that evolves in-context with model depth and governs the network's inference step.

Our main contributions are as follows:

1. We generalize *in-context denoising* (Smart et al., 2025) to the all-token corruption setting, showing that attention-only transformers admit a two-stage empirical Bayes interpretation: **depth** iteratively refines a particle approximation of the prior from noisy samples alone (Stage 1), while **attention residuals** use the original noisy input to query the resulting energy landscape for posterior inference (Stage 2). This decomposition clarifies the provenance of stored patterns in associative-memory views of attention (Ramsauer et al., 2021).

2. In the continuous-depth and large-context regime, we show that self-attention dynamics approximate an anti-diffusive denoising operator acting on a particle system, yielding a continuous-time flow that refines a particle-based prior in a nonparametric, in-context manner. This connects multilayer attention to score-based denoising and reverse diffusion.

3. We show that effective denoising does not require a noise or parameter schedule: it can be achieved via a fixed attention kernel bandwidth β and a finite integration horizon $T^* \approx \sigma^2/2$, providing a principled depth–noise relationship that informs architectural design.

4. We prove a sequential posterior-mean recovery theorem for an admissible class of priors $\mathcal{A}_\tau$ characterized by stability of the mean-field flow under hard truncation. The theorem shows that the truncated particle-flow estimator converges, through particular sequential limits, to the Bayes posterior mean uniformly on compact query sets. We verify that Gaussian priors belong to $\mathcal{A}_\tau$. In contrast to Bruno et al. (2025), whose particle system is compact via spherical layer norm, our analysis is set on $\mathbb{R}^d$ and must control the dynamics through hard truncation, requiring stability estimates for the truncated mean-field flow.

2. Problem formulation and background

Collective denoising task. We assume we are given a collection of noisy samples $\{\tilde{x}_i\}_{i=1}^N \subset \mathbb{R}^d$ obtained by adding noise to uncorrupted datapoints $\{x_i\}_{i=1}^N$ independently sampled from an unknown prior distribution P_0 with density ρ_0 (Fig. 1). We have $\tilde{x}_i = x_i + \epsilon_i, \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I_d)$ where I_d is the d-dimensional identity matrix. The task is to approximately recover the uncorrupted data under the x-prediction mean squared error (MSE) objective $\mathbb{E}\big[\|\hat{x}_i - x_i\|^2\big]$. We generate denoised estimates $\hat{x}_i$ using an empirical Bayes approach which is divided into two stages:

Stage 1. We perform iterative refinement on $\{\tilde{x}_i\}$ to construct a particle approximation to the underlying prior distribution ρ_0 from which $\{x_i\}$ were sampled.

Stage 2. Given the particle approximation for the prior from Stage 1, we denoise each $\tilde{x}_i$ by computing the posterior mean under the reconstructed prior.

At first glance, this problem appears challenging, as the prior ρ_0 is not known *a priori* and must be inferred from the corrupted samples alone. Nevertheless, we will show that simple nonparametric schemes can achieve near Bayes-optimal performance in certain settings, and that these schemes align naturally with the structure of attention-based architectures. The problem generalizes the single-token denoising task studied in (Smart et al., 2025), which was shown to be solvable by a *single* attention layer (corresponding to Stage 2 above). In contrast, our all-token corruption setting necessitates depth, as the prior itself must now be estimated (Stage 1).

Below, we begin by drawing an explicit connection between Stage 1 and reverse diffusion. This leads to an attention-based construction of the score (log-density gradient) that can be applied directly to noisy data in a non-parametric manner. This connection motivates a discrete particle update (Alg. 1), which admits a continuous-time limit (Alg. 2). For analytical tractability, we introduce a truncated variant (Alg. 3), whose behavior can be studied rigorously. We provide empirical demonstrations in a finite-sample setting before formally analyzing the resulting particle dynamics and establishing posterior recovery guarantees. We emphasize that the reverse-diffusion connection is heuristic, and several steps in this chain are not controlled in full generality.

3. Reverse diffusion as an attention-like particle dynamics for Stage 1

A connection between attention and score-based methods has been noted (Rosu et al., 2025; Ilin & Sushko, 2026), with classical roots in kernel-based estimation of density gradients (Fukunaga & Hostetler, 1975; Comaniciu & Meer, 1999). We revisit this correspondence to set up the subsequent analysis. Here we provide an informal proposition showing that reverse diffusion can be approximated by a particle system whose update rule takes the form of a leaky residual averaging step. The resulting dynamics closely resembles attention mechanisms (Vaswani et al., 2017) and kernel regression (Nadaraya, 1964; Watson, 1964).

Proposition 3.1. (Informal) *Assume that the typical density of the corrupted data distribution is $\bar{\rho}$. Under the setup described in section 2, we start with noisy outputs $\{\tilde{x}_i\}_{i=1}^N \subset \mathbb{R}^d$ and run the following dynamics*

$$\begin{aligned} z_i^{(\ell+1)} &= (1-\eta) z_i^{(\ell)} + \eta \frac{\sum_j e^{-\frac{\beta}{2}\|z_i^{(\ell)} - z_j^{(\ell)}\|^2} z_j^{(\ell)}}{\sum_j e^{-\frac{\beta}{2}\|z_i^{(\ell)} - z_j^{(\ell)}\|^2}} \\ &\equiv (1-\eta) z_i^{(\ell)} + \eta \sum_j a_{ij}(\{z_i^{(\ell)}\}) z_j^{(\ell)}, \end{aligned} \tag{1}$$

for $l = 0, 1, \ldots, L-1$, starting with $z_i(0) = \tilde{x}_i$. Here, $\eta \in (0,1)$ is the step size and $L \approx \beta\sigma^2/2\eta$ is the number of steps or transformation layers. If we let $N \to \infty$ and $\beta \to \infty$ in such a way that $\frac{N\bar{\rho}}{\beta^{d/2}} \to \infty$ and also let $\eta \to 0$, then $\{z_i^{(L)}\}_{i=1}^N \subset \mathbb{R}^d$ forms a good particle approximation of P_0.

Proof. The details of the arguments are provided in Appendix A. □

The update Eq. (1) admits three equivalent interpretations: (i) reverse diffusion via local denoising, (ii) Nadaraya–Watson kernel regression, (iii) attention-like averaging with Gaussian similarity weights.

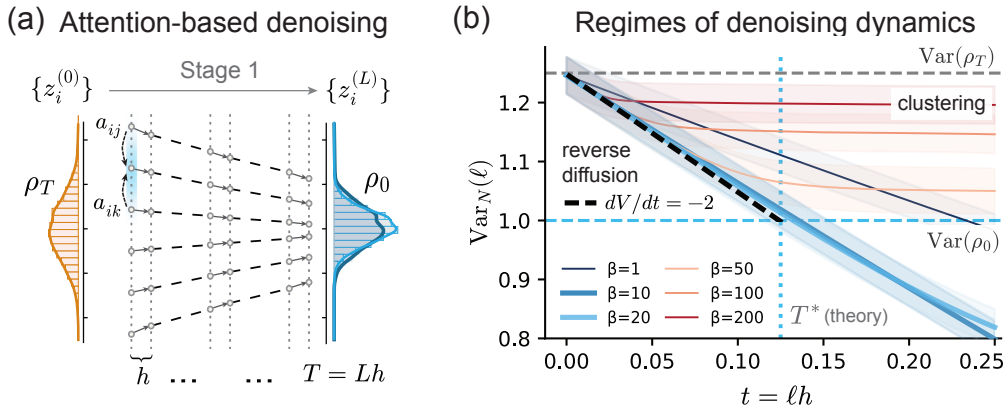

Figure 2. **Attention as discretized reverse diffusion.** (a) Particles initialized from a noised distribution ρ_T are iteratively updated across layers using Gaussian attention steps ($t = \ell h$) to approximately recover the clean prior $\rho_0 = \mathcal{N}(0, 1)$. (b) Variance dynamics for Gaussian denoising under different kernel bandwidths β. Parameters: 20 seeds, $N = 5000$, $L_0 = 200$, $\sigma^2 = 0.25$.

In the infinite-particle, continuous-depth limit of these dynamics, the Stage 1 integration time remains $\sigma^2/2$ at leading order in β, with a first-order correction of size $O(\beta^{-1})$.

Proposition 3.2. (Informal) *For sufficiently regular priors P_0 in the limit $N \to \infty$ and finite β, the denoising time is $\frac{\sigma^2}{2} + O(\beta^{-1})$.*

Proof. See Lemmas F.3 and F.4 for the formal statements and proofs. □

Leveraging insights from Propositions 3.1 and 3.2, we next demonstrate that multilayer attention schemes can implement a finite-sample empirical Bayes procedure that dynamically refines a particle representation to improve posterior inference with depth, before presenting our asymptotic analysis.

Algorithm 1 Two-stage denoising via multilayer attention

Require: Noisy tokens $\{\tilde{x}_i\}_{i=1}^N \subset \mathbb{R}^d$, noise variance σ^2, bandwidth β, step size η, depth L.

Initialize $z_i^{(0)} \leftarrow \tilde{x}_i$.
for $\ell = 0, \dots, L-1$ **do** {Stage 1 (particle refinement)}
  $a_{ij}^{(\ell)} = \dfrac{\exp\big(-\frac{\beta}{2}\|z_i^{(\ell)} - z_j^{(\ell)}\|^2\big)}{\sum_k \exp\big(-\frac{\beta}{2}\|z_i^{(\ell)} - z_k^{(\ell)}\|^2\big)}$
  $z_i^{(\ell+1)} \leftarrow (1-\eta) z_i^{(\ell)} + \eta \sum_j a_{ij}^{(\ell)} z_j^{(\ell)}$
end for
$\beta_c \leftarrow 1/\sigma^2$
for $i = 1, \dots, N$ **do** {Stage 2 (posterior mean)}
  $b_{ij} = \dfrac{\exp\big(-\frac{\beta_c}{2}\|\tilde{x}_i - z_j^{(L)}\|^2\big)}{\sum_k \exp\big(-\frac{\beta_c}{2}\|\tilde{x}_i - z_k^{(L)}\|^2\big)}$
  $\hat{x}_i \leftarrow \sum_j b_{ij} z_j^{(L)}$
end for

4. Attention architecture for collective denoising

Algorithm 1 outlines a multilayer attention architecture implementing the collective denoising approach to Fig. 1 (details in Appendix C). We use this for finite-sample, finite-depth empirical demonstrations before presenting an asymptotic variant that connects to our theoretical analysis.

Fig. 2 studies Stage 1 variance dynamics for the canonical problem of denoising a Gaussian distribution. The predicted reverse-diffusion behavior is linear variance decay $dv/dt = -2$ over the horizon $[0, T^*]$ with $T^* = \sigma^2/2$. Empirically, only intermediate values of kernel bandwidth β recover the diffusion-like linear variance decay and stopping time

predicted by theory. Large β induces premature clustering at finite N, while small β produces overly global averaging that misses the local denoising structure and leads to slower denoising. The slowed $\beta = 1$ behavior is captured by the solution to an ODE $v'(t) = -2v/(v + \beta^{-1})$ arising from a perturbative analysis (Appendix F, Theorem F.1).

Unlike typical diffusion modeling and related work (Rosu et al., 2025), which invoke time-dependent noise schedules, our iterative attention mechanism uses a single time-independent bandwidth β. Correct denoising is recovered provided the depth L and step size η are scaled according to noise σ^2.

4.1. In-context learning of particle priors and energy landscape of posterior inference

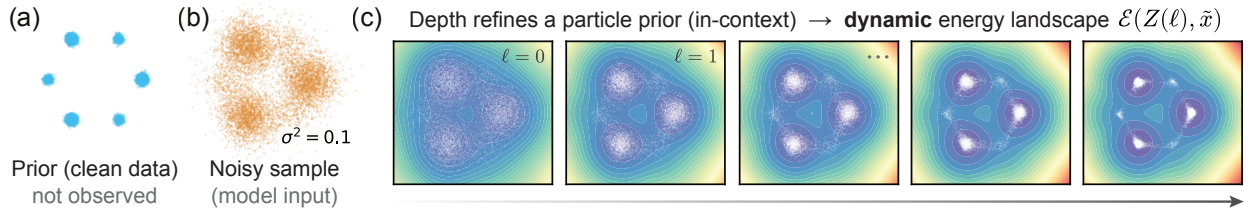

Figure 3. **Dynamic energy landscape.** (a) Ground truth GMM prior. (b) Corrupted input prompt ($N = 5,000$). (c) Dynamic refinement: As model depth ℓ increases, Stage 1 self-attention iteratively sharpens the particle prior (white cloud). This process dynamically sculpts an associative memory landscape $\mathcal{E}(Z(\ell), \tilde{x})$ which determines the posterior averaging step in Stage 2; Eq. (2).

Fig. 2(b) provides numerical indication that multilayer attention can denoise distributions (Stage 1) when model depth and parameters are in line with the theoretical predictions. We next study the downstream role of Stage 2 for sample denoising (x-prediction). For such posterior inference tasks, note that the prediction step of Alg. 1 (details in App. C) can be interpreted as in (Smart et al., 2025): each noisy query $\tilde{x}_i$ is updated by taking a gradient step on an energy landscape defined by the refined particle approximation

$$\hat{x}_i = \tilde{x}_i - \nabla_q \mathcal{E}(Z(\ell); q)\big|_{q=\tilde{x}_i}, \tag{2}$$

where

$$\mathcal{E}(Z(\ell); q) = -\frac{1}{\beta_c} \log \sum_j e^{-\frac{\beta_c}{2}\|q - z_j^{(\ell)}\|^2}. \tag{3}$$

This defines a dense associative memory (Krotov & Hopfield, 2016; Ramsauer et al., 2021) induced by the context. In contrast to Smart et al. (2025), where the memories are fixed during inference, here Stage 1 dynamically *sculpts* the landscape across depth: as the particle prior is refined, the energy surface on which the noisy query descends sharpens correspondingly (Fig. 3). We next check that approaching Bayes optimality on the all-token corruption task requires depth-dependent refinement.

4.2. Roles of context length, depth, and cross-attention for denoising all-token corruption

While Stage 1 can recover the underlying structure of the data, it does not by itself yield the Bayes-optimal estimator. Using a symmetric two-component Gaussian mixture prior, we observe that improvements from depth are pronounced in regimes where the posterior distribution is ambiguous across multiple modes (Fig. 4). In this minimal setting, the Bayes optimal predictor (exact posterior mean given the clean prior) is analytically tractable (see Appendix D). This example establishes a key validation: the combination of Stage 1 prior refinement and Stage 2 posterior averaging approaches the Bayes optimal bound as a function of model depth. Additional numerics in Fig. 5 of App. C.2.

4.3. Attention dynamics corresponding to asymptotic analysis (hard truncation)

Algorithm 1 and the associated numeric demonstrations are equivalent to using Euler steps on a neural ODE (Chen et al., 2018) acting on a finite context length. This discrete dynamics admits a continuous-time mean-field limit, which we formalize via an ODE governing particle evolution. Algorithm 2 is the corresponding practical full-sample version. For technical reasons, our analysis is carried out on a truncated version of this flow to ensure compact support. Algorithm 3 is the theorem-certified version. For Gaussian noisy laws, the tail estimate shows that taking $R = R_N$ slightly larger

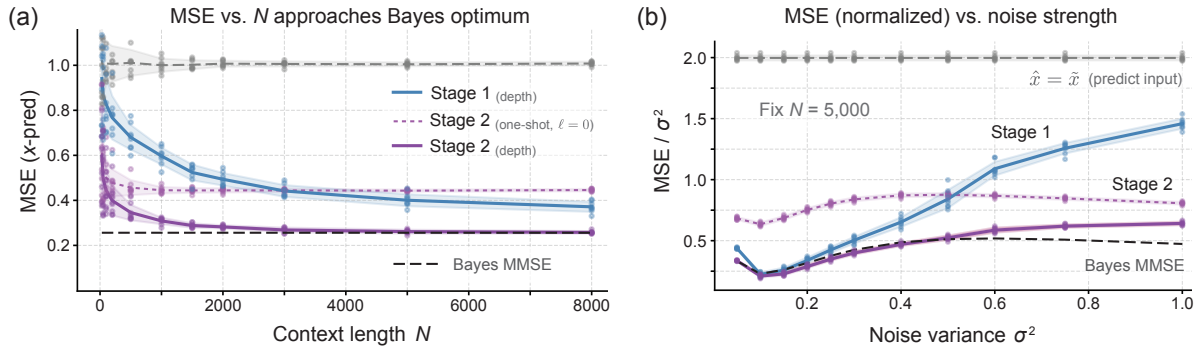

Figure 4. Multilayer attention-based denoiser approaches Bayes-optimal x-pred MSE with increasing context and depth. (a) MSE vs. context length (particle count) N. (b) MSE (normalized) vs. noise variance σ^2. Solid lines show the mean (bands ± 1 standard deviation) over 8 seeds (with subsampling preserved across N). Prior: Symmetric Gaussian mixture with $\mu = (\pm 1, 0)$ and $a = 0.1$. Parameters: (a) $\beta = 20$, $\sigma^2 = 0.5$, $L_0 = 200$; (b) $N = 5000$ is fixed for varying σ^2. Details in Appendix C.

than $\sqrt{\log N}$ makes the full-sample and truncated procedures agree with high probability. We refer to Appendix E.1 for the precise algorithms.

5. Sequential posterior-mean recovery

We now state the recovery guarantee underlying the Stage 2 posterior-mean estimator. Recall that the clean law is denoted by P_0, and the observation model is

$$Y = X + \sqrt{\tau}\, Z, \quad X \sim P_0, \quad Z \sim \mathcal{N}(0, I_d), \quad \tau = \sigma^2.$$

Thus, the noisy one-particle law is

$$f_0 = \gamma_\tau * P_0, \qquad \gamma_\tau(z) = (2\pi\tau)^{-d/2} e^{-|z|^2/(2\tau)}.$$

For any candidate prior ν, define the Gaussian posterior-mean map

$$m_\nu(y) := \frac{\int_{\mathbb{R}^d} x\, \gamma_\tau(y - x)\, \nu(dx)}{\int_{\mathbb{R}^d} \gamma_\tau(y - x)\, \nu(dx)}. \tag{4}$$

For $\nu = P_0$, this is the Bayes-optimal square-loss denoiser,

$$m_{P_0}(y) = \mathbb{E}[X \mid Y = y].$$

The goal of the sequential construction is therefore to recover the function m_{P_0} asymptotically from noisy samples. Let

$$G_\beta(z) = \left(\frac{\beta}{2\pi}\right)^{d/2} e^{-\frac{\beta}{2}|z|^2} = \mathcal{N}(0, \beta^{-1} I_d),$$

and define the exact Gaussian attention drift

$$X_\beta[\mu](x) := \frac{1}{\beta} \nabla \log(G_\beta * \mu)(x). \tag{5}$$

Equivalently,

$$\begin{aligned} X_\beta[\mu](x) &= F_{\beta,\mu}(x) - x, \\ F_{\beta,\mu}(x) &:= \frac{\int z\, G_\beta(x - z)\, \mu(dz)}{\int G_\beta(x - z)\, \mu(dz)}. \end{aligned} \tag{6}$$

Thus, $X_\beta[\mu]$ is precisely the continuous-depth Gaussian-attention barycentric update.

Appendix E.4.3 defines a class $\mathcal{A}_\tau \subset \mathcal{P}_1(\mathbb{R}^d)$ of *recoverable admissible priors*. In other words, a prior $P_0 \in \mathcal{A}_\tau$ is one whose noisy law $f_0 = \gamma_\tau * P_0$ has a well-defined noncompact mean-field flow

$$\partial_t f_t^\beta + \nabla \cdot \big(f_t^\beta X_\beta[f_t^\beta]\big) = 0, \qquad f_{t=0}^\beta = f_0,$$

which is obtained as the $R \to \infty$ limit of the corresponding hard-truncated compactly supported flows, and which recovers the clean prior at a β-scaled denoising time, $T_\beta := \beta\tau/2$:

$$W_1(f_{T_\beta}^\beta, P_0) \longrightarrow 0 \qquad \text{as } \beta \to \infty.$$

This class packages the two requirements needed for the theorem below: propagation of chaos after hard truncation and deterministic recovery of the clean law in the large-β denoising limit.

The following theorem is the main-text version of the hard-truncated recovery result proved as Theorem E.9 in Appendix E.

Theorem 5.1 (Sequential posterior-mean recovery)**.** *Let $P_0 \in \mathcal{A}_\tau$, and define*

$$f_0^{[R]} := \frac{\mathbf{1}_{B_R} f_0}{f_0(B_R)}.$$

Initialize N particles independently from $f_0^{[R]}$, and evolve them by

$$\dot{X}_i(t) = X_\beta[\mu_t^{N,\beta,[R]}](X_i(t)), \qquad \mu_t^{N,\beta,[R]} := \frac{1}{N}\sum_{j=1}^N \delta_{X_j(t)}.$$

Then, for every bounded observation region $\{|y| \le M\}$,

$$\lim_{\beta\to\infty} \limsup_{R\to\infty} \limsup_{N\to\infty} \mathbb{E} \sup_{|y|\le M} \left| m_{\mu_{T_\beta}^{N,\beta,[R]}}(y) - m_{P_0}(y) \right| = 0. \tag{7}$$

Consequently, the same convergence holds in probability.

The order of limits is part of the statement: first $N \to \infty$ at fixed (β, R), then $R \to \infty$ at fixed β, and finally $\beta \to \infty$. Thus this is a sequential consistency theorem rather than a joint scaling law in (N, R, β). It says that after evolving the noisy samples by the exact Gaussian-attention dynamics up to depth $T_\beta = \beta\tau/2$, the posterior mean computed from the evolved empirical measure converges uniformly on compact observation regions to the Bayes-optimal denoiser m_{P_0}.

The class $\mathcal{A}_\tau$ is nonempty. In particular, Appendix E.6 contains a proof that every Gaussian prior

$$P_0 = \mathcal{N}(a, \Sigma_0),$$

with $a \in \mathbb{R}^d$ and Σ_0 symmetric nonnegative semidefinite, belongs to $\mathcal{A}_\tau$. Indeed, if

$$f_0 = \gamma_\tau * P_0 = \mathcal{N}(a, \Gamma_0), \qquad \Gamma_0 = \Sigma_0 + \tau I_d,$$

then the deterministic mean-field flow remains Gaussian:

$$f_t^\beta = \mathcal{N}(a, \Gamma_t), \qquad \dot{\Gamma}_t = -2\Gamma_t(I + \beta\Gamma_t)^{-1}.$$

At $T_\beta = \beta\tau/2$, $\Gamma_{T_\beta} \to \Sigma_0$, hence $f_{T_\beta}^\beta \to P_0 \in W_1$. Therefore, Gaussian priors satisfy the full sequential posterior-mean recovery theorem.

For Gaussian noisy data, the hard-truncation loss is explicit. If $f_0 = \mathcal{N}(a, \Gamma_0)$ and $\lambda_+ = \lambda_{\max}(\Gamma_0)$, then for $R \ge 1 + 2|a|$,

$$f_0(B_R^c) \le C(1+R)^d \exp\left(-\frac{R^2}{8\lambda_+}\right).$$

Therefore, for untruncated noisy samples $Y_1, \dots, Y_N \sim f_0$,

$$\mathbb{P}(Y_1, \dots, Y_N \in B_R) \ge 1 - CN(1+R)^d \exp\left(-\frac{R^2}{8\lambda_+}\right).$$

Choosing $R = R_N$ slightly larger than $\sqrt{\log N}$ makes this loss probability vanish. This relates the practical full-sample construction in Alg. 2 to the theorem-certified truncated construction in Alg. 3.

6. Related work

Regimes of in-context denoising. Smart et al. (2025) introduced in-context denoising for attention networks, showing that a single layer solves it when context tokens are clean. The present work extends this to the *all-token corruption* regime, where every token is noisy and representations evolve jointly across depth, requiring the network to infer the prior from corrupted samples. Concurrent work by Rosu et al. (2025) studies architectural equivalences between attention and denoising algorithms, including manifold denoising (Hein & Maier, 2006; Belkin & Niyogi, 2003). Our analysis is complementary: a finite-sample empirical Bayes interpretation, a two-stage decomposition in which a long-range residual enables posterior averaging, and an explicit depth–noise relation $T^* = \sigma^2/2$.

Dynamics of attention layers. A growing body of work analyzes multilayer attention as an interacting particle system, including mean-field limits (Geshkovski et al., 2023; 2025; Rigollet, 2025; Bruno et al., 2025; Burger et al., 2025), particle-transport formulations (Sander et al., 2022), and unrolled-denoising analyses (Wang et al., 2025). These works emphasize long-time clustering and metastability. We complement this perspective by tying particle dynamics to a *collective* in-context denoising task with a finite horizon. Also, many of these works consider dynamics on a compact manifold, while we have to contend with an initial distribution supported on the whole of $\mathbb{R}^d$.

Empirical Bayes and score-based denoising. Classical empirical Bayes methods study posterior estimation under unknown priors from repeated observations (Robbins, 1956; Miyasawa, 1961; Stein, 1981; Efron, 2011; Johnstone & Silverman, 2005; Raphan & Simoncelli, 2011). Under Gaussian corruption, these estimators are closely connected to score-based denoising and diffusion formulations through Tweedie's formula (Robbins, 1956; Miyasawa, 1961; Efron, 2011; Vincent, 2011; Ho et al., 2020; Song et al., 2021). Kernel approximation of this connection (Fukunaga & Hostetler, 1975) can be represented by Gaussian attention, which is leveraged here as well as in recent work (Rosu et al., 2025; Ilin & Sushko, 2026). Related transport-based perspectives include constrained EB denoising via optimal transport (Jaffe et al., 2025) and stochastic interpolants for generative modeling from corrupted observations (Modi et al., 2025). Our contribution within this lineage is to identify a specifically in-context empirical Bayes role for multilayer attention, where depth performs finite-sample refinement of a particle prior and attention residuals implement posterior averaging against the refined representation.

Transformers and empirical Bayes. Recent work shows that transformers can be trained to solve empirical Bayes problems in an in-context manner (Teh et al., 2025). Our approach is complementary: we provide a mechanistic interpretation of multilayer attention as implementing internal prior refinement followed by posterior averaging via a long-range attention residual connection. More broadly, our collective in-context denoising formulation instantiates the perspective that transformer depth executes an implicit algorithm in the forward pass (Von Oswald et al., 2023).

7. Discussion

Our collective in-context denoising framework decomposes attention-based denoising into two stages: iterative refinement of a particle prior (Stage 1) and posterior averaging via a long-range attention residual (Stage 2). The cross-attention from the noisy input to a dynamically refined particle prior distinguishes empirical Bayes denoising from iterative score-based denoising (Rosu et al., 2025; Bruno et al., 2025) and from the clean-context setting of Smart et al. (2025), where Stage 1 is unnecessary and Stage 2 recovers the exact Bayes posterior mean. In the population limit, one-step Tweedie estimators are themselves Bayes-optimal; depth in our framework therefore plays the specific role of correcting finite-sample error. Stage 1 alone is valuable for representation learning and distribution-level denoising (Hein & Maier, 2006).

This setting connects to but differs from diffusion models such as DDPM (Ho et al., 2020) in two ways. First, our denoiser is nonparametric and context-dependent — the score is estimated directly in-context via attention (no time-dependent score is learned). Second, where diffusion models traverse a full trajectory of intermediate distributions ρ_t via a noise schedule, we target recovery from a single corruption level σ^2, for which a fixed-bandwidth kernel and finite horizon $T^* = \sigma^2/2$ suffice. The convergence of these two paradigms — transformer-based sequence models and diffusion-style denoising — is increasingly visible in practical architectures (see DiT (Peebles & Xie, 2023), SiT (Ma et al., 2024), and JiT (Li & He, 2025)). Our framework offers one statistical account of why this convergence is principled, showing that attention natively implements a finite-sample empirical Bayes operator for iterative denoising.

A growing body of work interprets attention through the lens of associative memory and energy-based models (Ramsauer et al., 2021; Smart et al., 2025), drawing on ideas from statistical physics (Krotov & Hopfield, 2016). From the perspective of a single attention head, however, it is not clear where the "memories" encoded in keys and values

should come from. We show that a deep, single-head attention network can construct such memories *in-context*: the particle prior refined by Stage 1 acts as a working memory that is itself shaped by collective dynamics, then read out by Stage 2 for posterior inference. This picture — where the energy landscape is not fixed but emerges from coupled interactions among many agents — has parallels in multicellular self-organization (Smart & Zilman, 2023), and connects to recent looped- and energy-transformer formulations (Yang et al., 2024; Saunshi et al., 2025; Hoover et al., 2023; Dehmamy et al., 2026; Gladstone et al., 2026). An open question is whether the finite denoising horizon analyzed here can inform depth selection for recurrent attention networks more broadly.

Outlook. Several practical directions are natural follow-ups. Efficient approximations to the kernel sum — such as Nyström-style schemes (Xiong et al., 2021; Rosu et al., 2025) or fast Gaussian-kernel approximations (Greengard & Strain, 1991) — could broaden applicability, while lower-dimensional embeddings analogous to latent diffusion (Rombach et al., 2022) could improve convergence rates. Unlike standard settings, however, the kernel geometry here is not fixed: the particle prior and associated energy landscape evolve in-context with depth. Extending acceleration and compression schemes to this dynamically evolving collective representation remains an interesting problem.

More conceptually, the two-stage decomposition suggests an in-context amortized inference perspective: once the particle prior is refined through Stage 1, new queries can be processed cheaply via Stage 2 without recomputing the full dynamics. The value of cross-depth attention in our framework also resonates with recent empirical work showing that depth-wise attention residuals improve training stability and downstream performance (Team et al., 2026); our framework provides one statistical reason for this benefit, since the original corrupted input carries token-identity information that posterior inference preserves. More broadly, the empirical Bayes lens on attention — in which depth refines a particle approximation to an unknown prior — may offer insight into the dynamic latent representations that emerge in deep generative models more generally.

Limitations

Theoretical considerations. Our analysis relies on structural assumptions on the prior, formalized through the admissible class $\mathcal{A}_\tau$ that may be restrictive in practice. The theory does not establish propagation of chaos in full generality, and instead proceeds via a truncated mean-field construction with sequential limits. In particular, we do not obtain finite-sample convergence rates, joint scaling laws in (N, R, β), or guarantees for the practical untruncated algorithm. The posterior-recovery theorem is proved only for priors in $\mathcal{A}_\tau$ Gaussian priors are verified as a concrete case, while heavy-tailed, singular, or more general priors remain open. The proof relies on a hard-truncated compact-support flow with sequential limits $N \to \infty$ at fixed (β, R), then $R \to \infty$, and finally $\beta \to \infty$; although Gaussian tail bounds make truncation negligible with high probability for $R \gtrsim \sqrt{\log N}$. Extending these results to broader classes of priors, establishing finite-sample rates, and developing joint scaling laws are natural directions for future work.

Architectural scope. The analysis applies to a minimal attention architecture: single-head, scaled-identity weights, tied parameters across depth, and no MLP blocks or positional embedding. We further neglect layer norm which motivates our use of the Gaussian attention kernel (Chen et al., 2021). Beyond these and other design choices, practical transformers learn their enormous weight vectors by training on fixed datasets. We conjecture the empirical Bayes lens extends to such settings, with learned weights potentially implementing approximations to the kernel structure analyzed here. A rigorous treatment would first require extending the analysis to anisotropic weights (cf. (Bruno et al., 2025; Rosu et al., 2025)) which should then also vary with depth. It will be particularly interesting to characterize how MLP layers augment the finite-N particle dynamics studied here and elsewhere.

Extensions left to future work. Several architectural extensions are not addressed here. Allowing parameters to vary across depth opens the door to noise-schedule designs from diffusion models. Multi-head architectures with distinct scales $\{\beta_h\}$ would enable multi-scale particle refinement. Our framework takes the noise level σ^2 as input, setting both the integration time and the Stage 2 scale $\beta_c = 1/\sigma^2$; inferring σ^2 from corrupted samples in-context is a separate challenge. Finally, the theoretical horizon $T^* = \sigma^2/2$ is a leading-order prediction. Required depth in finite-N settings depends on N, β, and prior structure, motivating adaptive depth-selection schemes; such schemes could be based on input characteristics or leverage limited clean samples from the prior if available.

Acknowledgements

We thank Alberto Bietti, Joan Bruna, Bob Carpenter, Suryanarayana Maddu, Ilya Nemenman, and Gautam Reddy for valuable discussions and feedback. M.S. acknowledges support from the Simons Foundation. A.V.M. and A.M.S. acknowledge financial and logistical support from the Center for Quantitative Biology, Rutgers University.

References

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008. ISBN 978-3-7643-8721-1.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Bihari, I. A generalization of a lemma of Bellman and its application to uniqueness problems of differential equations. *Acta Math. Acad. Sci. Hungar.*, 7:81–94, 1956. ISSN 0001-5954,1588-2632. doi: 10.1007/BF02022967. URL https://doi.org/10.1007/BF02022967.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Bruno, G., Pasqualotto, F., and Agazzi, A. A multiscale analysis of mean-field transformers in the moderate interaction regime. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=WCRPgBpbcA.

Burger, M., Kabri, S., Korolev, Y., Roith, T., and Weigand, L. Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 383(2298):20240233, 06 2025. ISSN 1364-503X. doi: 10.1098/rsta.2024.0233. URL https://doi.org/10.1098/rsta.2024.0233.

Chaintron, L.-P. and Diez, A. Propagation of chaos: a review of models, methods and applications. I. Models and methods. *Kinet. Relat. Models*, 15(6):895–1015, 2022. ISSN 1937-5093,1937-5077. doi: 10.3934/krm.2022017. URL https://doi.org/10.3934/krm.2022017.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Chen, Y., Zeng, Q., Ji, H., and Yang, Y. Skyformer: Remodel self-attention with gaussian kernel and nyström method. *Advances in Neural Information Processing Systems*, 34:2122–2135, 2021.

Comaniciu, D. and Meer, P. Mean shift analysis and applications. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pp. 1197–1203 vol.2, 1999. doi: 10.1109/ICCV.1999.790416.

Dehmamy, N., Hoover, B., Saha, B., Kozachkov, L., Slotine, J.-J., and Krotov, D. NRGPT: An energy-based alternative for GPT. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=B3Muyi2zgo.

Del Moral, P. Mean field simulation for monte carlo integration. *Monographs on Statistics and Applied Probability*, 126(26):6, 2013.

Dobrushin, R. L. Vlasov equations. *Functional Analysis and Its Applications*, 13(2):115–123, 1979.

Efron, B. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162(3-4):707–738, 2015. ISSN 0178-8051,1432-2064. doi: 10.1007/s00440-014-0583-7. URL https://doi.org/10.1007/s00440-014-0583-7.

Fukunaga, K. and Hostetler, L. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975. doi: 10.1109/TIT.1975.1055330.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36:57026–57037, 2023.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025.

Gladstone, A., Nanduru, G., Islam, M. M., Han, P., Ha, H., Chadha, A., Du, Y., Ji, H., Li, J., and Iqbal, T. Energy-based transformers are scalable learners and thinkers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZBj3Qp1bYg.

Greengard, L. and Strain, J. The fast gauss transform. *SIAM Journal on Scientific and Statistical Computing*, 12(1): 79–94, 1991.

Hein, M. and Maier, M. Manifold denoising. *Advances in neural information processing systems*, 19, 2006.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hoover, B., Liang, Y., Pham, B., Panda, R., Strobelt, H., Chau, D. H., Zaki, M. J., and Krotov, D. Energy transformer. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=MbwVNEx9KS.

Ilin, V. and Sushko, P. Discoformer: Plug-in density and score estimation with transformers, 2026. URL https://arxiv.org/abs/2511.05924.

Jaffe, A. Q., Ignatiadis, N., and Sen, B. Constrained denoising, empirical bayes, and optimal transport. *arXiv preprint arXiv:2506.09986*, 2025.

Johnstone, I. M. and Silverman, B. W. Empirical bayes selection of wavelet thresholds. *Annals of Statistics*, 33, 2005. ISSN 00905364. doi: 10.1214/009053605000000345.

Krotov, D. and Hopfield, J. J. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

Li, T. and He, K. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Ma, N., Goldstein, M., Albergo, M. S., Boffi, N. M., Vanden-Eijnden, E., and Xie, S. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXVII*, pp. 23–40, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72979-9. doi: 10.1007/978-3-031-72980-5_2. URL https://doi.org/10.1007/978-3-031-72980-5_2.

Miyasawa, K. An empirical bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist*, 38 (181-188):1–2, 1961.

Modi, C., Han, J., Vanden-Eijnden, E., and Bruna, J. Generative modeling from black-box corruptions via self-consistent stochastic interpolants. *arXiv preprint arXiv:2512.10857*, 2025.

Nadaraya, E. On estimating regression. theor. *Probab. Appl*, 9(1), 1964.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Adler, T., Kreil, D. P., Kopp, M. K., Klambauer, G., Brandstetter, J., and Hochreiter, S. Hopfield networks is all you need. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Raphan, M. and Simoncelli, E. P. Least squares estimation without priors or supervision. *Neural computation*, 23(2): 374–420, 2011.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Rigollet, P. The mean-field dynamics of transformers. *arXiv preprint arXiv:2512.01868*, 2025.

Robbins, H. E. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pp. 388–394. Springer, 1956.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Rosu, P., Carin, L., and Cheng, X. From softmax to score: Transformers can effectively implement in-context denoising steps. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4QRoLzD11x.

Sander, M. E., Ablin, P., Blondel, M., and Peyré, G. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 3515–3530. PMLR, 2022.

Saunshi, N., Dikkala, N., Li, Z., Kumar, S., and Reddi, S. J. Reasoning with latent thoughts: On the power of looped transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=din0lGfZFd.

Smart, M. and Zilman, A. Emergent properties of collective gene-expression patterns in multicellular systems. *Cell Reports Physical Science*, 4(2), 2023.

Smart, M., Bietti, A., and Sengupta, A. M. In-context denoising with one-layer transformers: Connections between attention and associative memory retrieval. In *International Conference on Machine Learning*, pp. 55950–55971. PMLR, 2025.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Stein, C. M. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pp. 1135–1151, 1981.

Sznitman, A.-S. Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX—1989*, volume 1464 of *Lecture Notes in Math.*, pp. 165–251. Springer, Berlin, 1991. ISBN 3-540-53841-0. doi: 10.1007/BFb0085169. URL https://doi.org/10.1007/BFb0085169.

Team, K., Chen, G., Zhang, Y., Su, J., Xu, W., Pan, S., Wang, Y., Wang, Y., Chen, G., Yin, B., Chen, Y., Yan, J., Wei, M., Zhang, Y., Meng, F., Hong, C., Xie, X., Liu, S., Lu, E., Tai, Y., Chen, Y., Men, X., Guo, H., Charles, Y., Lu, H., Sui, L., Zhu, J., Zhou, Z., He, W., Huang, W., Xu, X., Wang, Y., Lai, G., Du, Y., Wu, Y., Yang, Z., and Zhou, X. Attention residuals, 2026. URL https://arxiv.org/abs/2603.15031.

Teh, A., Jabbour, M., and Polyanskiy, Y. Solving empirical bayes via transformers. *arXiv preprint arXiv:2502.09844*, 2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Villani, C. *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9. URL https://doi.org/10.1007/978-3-540-71050-9. Old and new.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Wang, P., Lu, Y., Yu, Y., Pai, D., Qu, Q., and Ma, Y. Attention-only transformers via unrolled subspace denoising. In *International Conference on Machine Learning*, pp. 63840–63859. PMLR, 2025.

Watson, G. S. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372, 1964.

Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., and Singh, V. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14138–14148, 2021.

Yang, L., Lee, K., Nowak, R. D., and Papailiopoulos, D. Looped transformers are better at learning learning algorithms. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=HHbRxoDTxE.

A. The informal argument for Empirical Bayes, Stage 1, via reverse diffusion

Before we give the informal arguments for the Proposition 3.1, let us indicate its relationship to the formal Theorem 5.1:

- **Proposition 3.1 (Informal particle dynamics).** Provides a heuristic description of attention updates as a finite-step particle system, where each update corresponds to kernel-weighted averaging. This perspective connects attention to reverse diffusion, kernel regression, and associative memory dynamics.
- **Algorithm 1 (Discrete attention dynamics).** Instantiates the informal dynamics of Proposition 3.1 as a finite-depth, finite-sample procedure. It can be viewed as an Euler discretization of an underlying continuous-time particle flow, with step size η and depth L determining the effective integration horizon.
- **Algorithm 2 (Continuous-time flow).** Provides a continuum formulation of the dynamics underlying Algorithm 1, in which the empirical particle distribution evolves according to a Gaussian-attention drift $X_\beta[\mu] = \beta^{-1}\nabla \log(G_\beta * \mu)$, where μ is the empirical distribution of the particles.
- **Algorithm 3 (Truncated, theorem-certified dynamics).** Introduces the hard-truncated version of the empirical flow, obtained by restricting the initial noisy sample to a radius-R ball and then evolving by the same Gaussian-attention drift. The rigorous theorem is stated for N i.i.d. samples from the normalized truncated law $f_0^{[R]}$; the algorithmic filtering version is the practical finite-sample analogue. Compact support is preserved by the flow, which makes the propagation-of-chaos analysis available.
- **Theorem 5.1 (Posterior-mean recovery).** Establishes that, for recoverable admissible priors and under appropriate sequential limits the posterior mean computed from the hard-truncated evolved empirical measure converges uniformly on compact observation regions to the Bayes-optimal predictor $m_{P_0}(y) = \mathbb{E}[X \mid Y = y]$.

Now we discuss the steps towards Proposition 3.1.

A.1. Noise addition as forward diffusion

Let $X_0 \sim P_0$ be a distribution on $\mathbb{R}^d$ with density ρ_0. Consider the forward diffusion over the time interval $s \in [0, T]$:

$$\partial_s \rho_s = \Delta \rho_s, \tag{8}$$

where ρ_s is the density at time s, with the initial condition $\rho_{s=0} = \rho_0$. Eq. (8) is solved by

$$\rho_s = \rho_0 * \mathcal{N}(0, 2sI). \tag{9}$$

A.2. Backward-time formulation

We introduce backward time:

$$t = T - s, \qquad f_t := \rho_{T-t}. \tag{10}$$

Then f_t satisfies

$$\partial_t f_t = -\Delta f_t = -\nabla \cdot \big(f_t \nabla \log f_t\big). \tag{11}$$

Comparing Eq. (11) with the continuity equation, $\partial_t f_t + \nabla(f_t v_t) = 0$, gives the velocity field as $v_t(x) = \nabla \log f_t(x)$. Thus we may view the diffusive flow as the transport of a cloud of particles with instantaneous velocities determined by $v_t = \nabla \log f_t$. If a particle has position Z_t at time t, with the probability density f_t, its trajectory follows the deterministic probability flow (Song et al., 2021):

$$\dot{Z}_t = \nabla \log f_t(Z_t), \tag{12}$$

which transports $f_0 = \rho_T$ backwards in time toward $f_T = \rho_0$. Under this formulation, the drift is governed by the score function $\nabla \log f_t$. The set of trajectories $\{Z_t\}$ constitutes a particle approximation to the backward density evolution (Del Moral, 2013), effectively acting as a normalizing flow (Rezende & Mohamed, 2015) which maps the diffused measure back to the data distribution.

We introduce the Gaussian-kernel dynamics:

$$\partial_t f_t = -\nabla \cdot \left(f_t \frac{\nabla (G_\beta * f_t)}{G_\beta * f_t} \right) = -\nabla \cdot \left(f_t \nabla \log (G_\beta * f_t) \right), \qquad t \geq 0$$

where G_β is the Gaussian distribution with mean 0 and covariance $\beta^{-1} I$. This is an intermediate model because the original backward-time dynamics is expressed in terms of the exact score field $\nabla \log f_t$, which depends on the unknown population density itself and is therefore not directly accessible to a finite attention mechanism. In this case, the velocity field is $v_t = \nabla \log (G_\beta * f_t)$; replacing $\nabla \log f_t$ with the smoothed score $\nabla \log (G_\beta * f_t)$ rewrites the denoising flow in terms of Gaussian local averages.

By Proposition E.2, the corresponding flow equation (Eq. (12)) becomes

$$\dot{Z}_t = \beta \left(F_{\beta, f_t}(Z_t) - Z_t \right), \tag{13}$$

where

$$F_{\beta,\rho}(y) = \frac{\int_{\mathbb{R}^d} x \exp\left(-\frac{\beta \|y-x\|^2}{2} \right) \rho(x) dx}{\int_{\mathbb{R}^d} \exp\left(-\frac{\beta \|y-x\|^2}{2} \right) \rho(x) dx}. \tag{14}$$

When the observation-noise variance is $\tau = \beta^{-1}$, $F_{\tau^{-1},\rho} = m_\rho$, the Gaussian posterior mean for prior ρ.

Next, we treat the f_t's as the empirical measure on the time-evolved particles in discrete time. Consider N particles at time t with positions $\{Z_t^{(i)}\}_{i=1}^N$ and the corresponding empirical measure $\nu_N^{(t)} = \dfrac{1}{N} \displaystyle\sum_{i=1}^N \delta_{Z_i^{(t)}}$. Then the particle positions at time $t + \eta/\beta$ are given by:

$$Z_{t+\eta/\beta}^{(i)} - Z_t^{(i)} = \eta \left(F_{\beta, \nu_N^{(t)}}(Z_t^{(i)}) - Z_t^{(i)} \right). \tag{15}$$

If the total time budget is T and passing through one layer corresponds to moving $h = \eta/\beta$ in time, $L\eta/\beta = T$, so that the total number of layers is $L = T\beta/\eta$ and the layer-wise dynamics becomes

$$Z_{\ell+1}^{(i)} = (1-\eta) Z_\ell^{(i)} + \eta F_{\beta, \nu_N^{(\ell)}}(Z_\ell^{(i)}).$$

In this way, the Gaussian-kernel formulation preserves the denoising interpretation while making the dynamics compatible with an empirical particle approximation. It also regularizes the ideal backward-heat flow, which is formally anti-diffusive and unstable, by replacing the raw score with a modified one. Finally, in the large-β regime, the Gaussian attention barycenter satisfies $F_{\beta,\rho}(x) - x \approx \beta^{-1} \nabla \log \rho(x)$, and the corresponding nonlocal PDE reduces, after rescaling time, to the original backward-heat equation. Thus the Gaussian-kernel dynamics is meant to provide the correct bridge from the exact continuum denoising law to the finite-sample attention dynamics implemented via particles.

A.3. Tweedie's formula

Let $\tilde{X} = X + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ and $X \sim \rho$. Then

$$\mathbb{E}[X \mid \tilde{X} = \tilde{x}] = \tilde{x} - \sigma^2 \nabla \log \tilde{\rho}(\tilde{x}),$$

where $\mathbb{E}[X \mid \tilde{X} = \tilde{x}] = F_{\sigma^{-2},\tilde{\rho}}(\tilde{x})$, and

$$\nabla \log \tilde{\rho}(\tilde{x}) = \frac{1}{\gamma^2} \big(\tilde{x} - \mathbb{E}[X \mid \tilde{X} = \tilde{x}] \big).$$

Thus the score is a denoising residual.

A.4. Local denoising along the diffusion

Fix backward time t and let $s = T - t$. Over a short forward time increment $\delta > 0$,

$$X_{s+\delta} = X_s + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 2\delta I).$$

Equivalently,

$$f_t = f_{t-\delta} * \mathcal{N}(0, 2\delta I).$$

Applying Tweedie's formula yields the local identity

$$\nabla \log f_t(x) = \frac{1}{2\delta}\Big(x - \mathbb{E}[Y \mid X = x]\Big),$$

where $Y \sim f_{t-\delta}$ and $X = Y + \varepsilon \sim f_t$.

A.5. Particle approximation

We discretize backward time:

$$t_\ell = \ell h, \qquad h = \frac{T}{L}.$$

Let $\{z_i^{(\ell)}\}_{i=1}^N$ approximate f_{t_ℓ}. The reverse flow gives

$$z_i^{(\ell+1)} = z_i^{(\ell)} + h\nabla \log f_{t_\ell}(z_i^{(\ell)}).$$

Using the denoising identity, we instead write the update as

$$z_i^{(\ell+1)} = (1-\eta) z_i^{(\ell)} + \eta\, \mathbb{E}[Y \mid X = z_i^{(\ell)}], \qquad \eta = \frac{h}{2\delta}.$$

A.6. Kernel approximation

Since $X = Y + \varepsilon$ with Gaussian noise, we have:

$$\mathbb{E}[Y \mid X = z] = \frac{\int y\, e^{-\frac{\|z-y\|^2}{4\delta}} f_{t-\delta}(y)dy}{\int e^{-\frac{\|z-y\|^2}{4\delta}} f_{t-\delta}(y)dy}.$$

Approximating $f_{t-\delta}$ by f_t, with further approximation of f_t with particles, yields

$$\mathbb{E}[Y \mid X = z] \approx \frac{\sum_j e^{-\frac{\|z-z_j^{(\ell)}\|^2}{4\delta}} z_j^{(\ell)}}{\sum_j e^{-\frac{\|z-z_j^{(\ell)}\|^2}{4\delta}}}.$$

Let $\beta = 1/2\delta$. The update becomes

$$z_i^{(\ell+1)} = (1-\eta) z_i^{(\ell)} + \eta \frac{\sum_j e^{-\frac{\beta}{2}\|z_i^{(\ell)} - z_j^{(\ell)}\|^2} z_j^{(\ell)}}{\sum_j e^{-\frac{\beta}{2}\|z_i^{(\ell)} - z_j^{(\ell)}\|^2}}. \tag{16}$$

In this replacement of the integral by the sum, the condition $\frac{N\bar{\rho}}{\beta^{d/2}} \gg 1$ becomes important (here, $\bar{\rho}$ denotes a typical density, e.g. $\bar{\rho} \simeq \int \rho(x)^2 dx$). Since the Gaussian centered around a particle position is effectively supported by a spherical volume $\sim \beta^{-d/2}$, this condition corresponds to the number of other particles in this volume being large, making the replacement by the sum justifiable.

B. Stability and well-definedness of the Gaussian kernelized score field

B.1. Stability under smooth score field

The advantage of replacing the raw score $\nabla \log f_t$ by $\nabla \log(K_\beta * f_t)$ is stability and well-definedness. It holds for any smooth positive kernel K_β.

Let $K_\beta : \mathbb{R}^d \to (0, \infty)$ be a $\mathcal{C}^1$ probability density, and define

$$u_f(x) := (K_\beta * f)(x), \qquad b_f(x) := \nabla \log(K_\beta * f)(x) = \frac{\nabla(K_\beta * f)(x)}{(K_\beta * f)(x)}.$$

The raw score $\nabla \log f$ is unstable because it may be undefined when f vanishes, and differentiation amplifies microscopic oscillations. It vanishes at most points even if f is the empirical measure. If f is empirical, $\nabla \log f$ is not a classical vector field.

If K_β is $\mathcal{C}^1$ and positive, then for probability measure with density f,

$$K_\beta * f \in \mathcal{C}^1(\mathbb{R}^d), \qquad \nabla(K_\beta * f) = (\nabla K_\beta) * f,$$

and $K_\beta * f > 0$. Hence b_f is globally well-defined.

We also have the following heuristic properties:

1. On any compact set $\mathcal{B}_R$,
$$b_f - b_g = \frac{\nabla u_f - \nabla u_g}{u_f} + \nabla u_g \left(\frac{1}{u_f} - \frac{1}{u_g} \right),$$
so if u_f and u_g are bounded below on $\mathcal{B}_R$, then
$$\|b_f - b_g\|_{L^\infty(B_R)} \lesssim \|\nabla K_\beta * f - \nabla K_\beta * g\|_{L^\infty(B_R)} + \|K_\beta * f - K_\beta * g\|_{L^\infty(B_R)}.$$
Thus stability of the score field reduces to stability of two linear convolutions.

2. If K_β and ∇K_β are Lipschitz, then for probability measures f, g,
$$\|K_\beta * f - K_\beta * g\|_{L^\infty(B_R)} + \|\nabla K_\beta * f - \nabla K_\beta * g\|_{L^\infty(B_R)} \lesssim W_1(f, g),$$
hence $\|b_f - b_g\|_{L^\infty(B_R)} \lesssim W_1(f, g)$ whenever the denominators stay uniformly positive on $\mathcal{B}_R$.

3. For empirical measures
$$\nu^N = \frac{1}{N} \sum_{j=1}^N \delta_{z_j},$$
one has
$$(K_\beta * \nu^N)(x) = \frac{1}{N} \sum_{j=1}^N K_\beta(x - z_j), \qquad \nabla(K_\beta * \nu^N)(x) = \frac{1}{N} \sum_{j=1}^N \nabla K_\beta(x - z_j),$$
so
$$\nabla \log(K_\beta * \nu^N)(x) = \frac{\sum_{j=1}^N \nabla K_\beta(x - z_j)}{\sum_{j=1}^N K_\beta(x - z_j)}.$$
Thus the field is well-defined for empirical laws and varies continuously under perturbations of the particle cloud.

4. Finally, if $K_\beta \in \mathcal{C}^2$, then on any compact region where $K_\beta * f$ is bounded below, the field b_f is locally Lipschitz in space, since
$$\nabla b_f = \frac{\nabla^2(K_\beta * f)}{K_\beta * f} - \frac{\nabla(K_\beta * f) \otimes \nabla(K_\beta * f)}{(K_\beta * f)^2}.$$
Hence the ODE
$$\dot{Z}_t = \nabla \log(K_\beta * f_t)(Z_t)$$
is locally well-posed.

In summary, for any smooth positive kernel K_β, the map

$$f \mapsto \nabla \log(K_\beta * f)$$

is a stable regularization of the raw score: it is well-defined for empirical measures, continuous under perturbations of the law, and suitable for particle and mean-field limits.

The Gaussian is special only because it adds further exact structures, not because basic stability is unique to it. This is shown in the next subsection.

B.2. Importance of the Gaussian kernel for smoothing

Continuing from the previous discussion, let us say that we use a smooth density K_β.

We focus on the following two properties:

(A.I) There is a positive definite matrix C_β (for every positive β) such that $\mathbb{E}[X \mid X + e_\beta = y] = y + C_\beta \nabla \log(K_\beta * f)(y)$ for every prior f on X, where $e_\beta \sim K_\beta$.

(A.II) There is a positive definite matrix C_β such that for every empirical measure

$$\nu^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{z_j},$$

the field looks like

$$\nabla \log(K_\beta * \nu_N)(x) = C_\beta^{-1}\left[\sum_{j=1}^{N} w_j(x) z_j - x\right], \qquad w_j(x) = \frac{K_\beta(x - z_j)}{\sum_{i=1}^{N} K_\beta(x - z_i)}.$$

Both of these expressions appear in the paper and the clean formulae help us derive the attention dynamics. Now, assuming either one of (A.I) or (A.II) above holds, then K_β is forced to be a Gaussian with covariance matrix C_β.

Now that we have K_β is Gaussian, we still need $K_\beta * f \to f$ (either in L^1 convergence or almost everywhere convergence) if $\beta \to \infty$. This forces all eigenvalues of C_β (which are all taken as β^{-1} in the paper) to approach 0 as $\beta \to \infty$. This still leaves us with a Gaussian with a covariance matrix C_β. For simplicity, we use an isotropic Gaussian with $C_\beta = r_\beta I$, where $r_\beta = \beta^{-1}$. Note that by the last paragraph, $\lim_{\beta\to\infty} r_\beta = 0$.

C. Attention architecture for two-stage collective denoising

Here we detail the multilayer attention architecture (Alg. 1) that implements the two stages in Fig. 1:

Stage 1. Given a collection of noisy input tokens $\{\tilde{x}_i\}_{i=1}^N$, the role of Stage 1 is to construct a particle approximation to the prior using Eq. (1). This approximation serves as the model's internal representation of the unobserved data distribution that supports downstream inference. Eq. (1) can be interpreted as applying a Gaussian self-attention kernel (Chen et al., 2021) with scaled-identity weights $W_Q = W_K = \sqrt{\beta} I_d$ and $W_V = I_d$. Repeated application of the attention operation produces a dynamic internal representation $Z(\ell) = \{z_i^{(\ell)}\}_{i=1}^N$, which we use interchangeably with its $N \times d$ matrix form when convenient. $Z(\ell)$ serves as the model's particle estimate of the prior at depth ℓ; it provides an estimate of $\hat{\rho}_0$ at depth L in Fig. 1.

The first term of Eq. (1) differs slightly from the canonical attention update $z_i \leftarrow z_i + \mathrm{Attn}(W_V Z, W_K Z, W_Q z_i)$. Since $\eta \in (0, 1)$, this *leaky* residual update ensures forward invariance of the convex hull of $\{z_i^{(\ell)}\}_{i=1}^N$. Assuming that η is fixed across layers (tied), the continuous-depth limit, $\eta \to 0$, $L \to \infty$ with $\eta L = T$ fixed, can be viewed as a neural ODE (Chen et al., 2018):

$$\dot{z}_i = -z_i + \sum_{j=1}^{N} a_{ij}(Z) z_j, \tag{17}$$

which is integrated from $t = 0$ to $t = T$ with the initial condition $z_i(0) = \tilde{x}_i \; \forall i$. In practice, we use Eq. (1) with a fixed step size, corresponding to the forward Euler discretization of Eq. (17).

Stage 2. Once the approximate prior is obtained, the denoised estimates $\hat{x}_i$ are computed for each noised token $\tilde{x}_i$ using cross-attention with scale $\beta_c = 1/\sigma^2$ between the corrupted input $z_i(0) = \tilde{x}_i$ and the refined representation after ℓ steps, $Z(\ell)$, where ℓ is chosen to approximate the theoretical stopping depth L:

$$\hat{x}_i(\ell) = m_{\beta_c, Z(\ell)}(\tilde{x}_i) = \frac{\sum_j \exp\big(-\frac{\beta_c \|z_j^{(\ell)} - \tilde{x}_i\|^2}{2}\big) z_j(\ell)}{\sum_k \exp\big(-\frac{\beta_c \|z_k^{(\ell)} - \tilde{x}_i\|^2}{2}\big)}. \tag{18}$$

As an architectural element, this is enabled by an additional skip connection from the input to layer ℓ (Fig. 1(a), dashed path). This connects to recent work identifying the empirical benefits of residual pathways in attention models, with particular emphasis on AttnRes connections to the embedding layer (Team et al., 2026). Here, the preservation of the original noisy token plays a specific statistical role: it serves as the query in the final posterior estimation step. The long-range skip connection implements the separation between data (query) and learned prior (memory) required by the empirical Bayes formulation.

C.1. Remarks on the two-stage architecture and numerics

Remark (Choice of prior). Gaussian mixture models admit closed-form Bayes-optimal estimators (Appendix D), which provides an unambiguous reference point against which to validate Stage 2 inference. Low-dimensional instantiations permit direct visualization of the internal energy landscape (Fig. 3), allowing us to illustrate the two-stage mechanism. We emphasize however that our framework is not specific to low-dimensional Gaussian mixtures.

Remark (Finite discretization and parameter scaling). In the finite-depth setting, the parameters (σ, L, β) cannot be chosen independently if the dynamics are to approximate the reverse diffusion flow. Fixing a noise level σ^2 determines the total (finite) time horizon $T^* = \sigma^2/2$. Discretizing with L steps gives a step size $h = T^*/L = \sigma^2/(2L)$. Considering *infinitesimal* denoising between adjacent time slices in the reverse diffusion introduces a parameter $\delta = (\beta\sigma^2)^{-1} \ll 1$, requiring β to be sufficiently large, and induces an effective step size $\eta = h/\sigma^2\delta$. Substituting δ gives

$$\eta = \beta h = \frac{\beta\sigma^2}{2L} < 1, \tag{19}$$

which should not be treated as an independent parameter. In particular, consistent approximation of the continuous reverse flow corresponds to regimes with large L, large β, and $\beta/L \to 0$, so that $\eta \to 0$ while $\delta \to 0$. In practice, the numerics fix large L_0 and integrate to $3L_0$ using $h = T^*/L_0$.

Remark (Annealing schedule). Allowing depth-dependent weights $\eta(t), \beta(t)$ induces a time-inhomogeneous flow, analogous to the noise (or annealing) schedules commonly used in diffusion models. In this interpretation, $\beta(t)$ controls the locality of the particle-based score approximation via $\delta(t) = (\beta(t)\sigma^2)^{-1}$, while the discrete flow step $h(t) = \eta(t)/\beta(t)$ satisfies $\int_0^T h(t)\,dt \approx \sigma^2/2$. In the finite-depth setting, this implies $\sum_{t=1}^{L} \frac{\eta(t)}{\beta(t)} \approx T$. We emphasize that the denoising behavior is governed not by $\beta(t)$ or $\eta(t)$ individually, but by their ratio through this finite-time horizon constraint.

Remark (Convex hull invariance). The convex hull of the particle set $\{z_i(t)\}_{i=1}^N$ is positively invariant under Eq. (1) for any $\eta \in (0, 1)$, $\beta > 0$. To see this, observe that each particle updates its position as a convex combination of its current position and the attention-weighted barycenter $\sum_j a_{ij}(t) z_j(t)$, which itself lies in the convex hull.

Remark (Layer-norm). We study simplified dynamics without layer-norm/spherical projection. Off sphere, distances are relevant and the Gaussian kernel is appropriate. If one projects to the sphere between attention layers, then the Gaussian kernel reduces to softmax dot-product attention.

Remark (One-shot Tweedie-like predictor). Note that Eq. (18) can in principle be used without iterating Stage 1, yielding an empirical Tweedie-like estimator,

$$\hat{x}_i(0) = \tilde{x}_i + \frac{\sum_j b_{ij}(\tilde{x}_j - \tilde{x}_i)}{\sum_k b_{ik}} \approx \tilde{x}_i + \sigma^2 \nabla \log p_{\tilde{X}}(\tilde{x}_i), \tag{20}$$

via the Gaussian kernel representation of the score. In the large N limit, this estimator coincides with the Bayes-optimal posterior mean; in practically relevant finite-sample settings, it provides a baseline for the role of depth in the recovery of the uncorrupted data distribution.

Remark (Tweedie & Finite-N). The gap between one-shot and iterative estimators arises from finite-sample effects; in the population limit, both coincide with the Bayes-optimal estimator.

Remark (Depth/stepsize). While we intentionally use small steps (many layers) to match the theory, fewer layers could be used in practice, provided the step size per layer is increased according to T^*.

Remark (Posterior sufficiency and clustering). While Stage 1 is motivated by recovery of the clean prior ρ_0, we empirically observe that accurate posterior inference can be achieved even when the particle distribution concentrates into discrete clusters. This suggests that, for Stage 2 cross-attention, it may not be necessary to reconstruct the full prior density. Instead, it appears sufficient to concentrate particle mass near high-density regions of the data, forming a discrete surrogate representation that preserves the relevant local energy landscape. This provides a possible functional interpretation of the clustering behavior emphasized in prior analyses of attention dynamics, suggesting that it could enable sparse in-context representations rather than a geometric artifact or pathology.

C.2. Additional numerics

Figure 5 indicates that careful tuning of β and sufficient depth can allow self-attention kernel denoising (Stage 1) alone to approach Bayes optimality. However, consistent with the empirical Bayes approach, it further emphasizes that the combination of depth and cross-attention to the input (Stage 2) provides a more robust mechanism for posterior averaging across regimes, requiring significantly less depth or context length to achieve the same error. This decomposition naturally points towards an amortized inference design: once the model refines the particle prior (Stage 1), it can be "cached" so that novel queries can be processed cheaply (Stage 2) via a gradient descent step on Eq. (2).

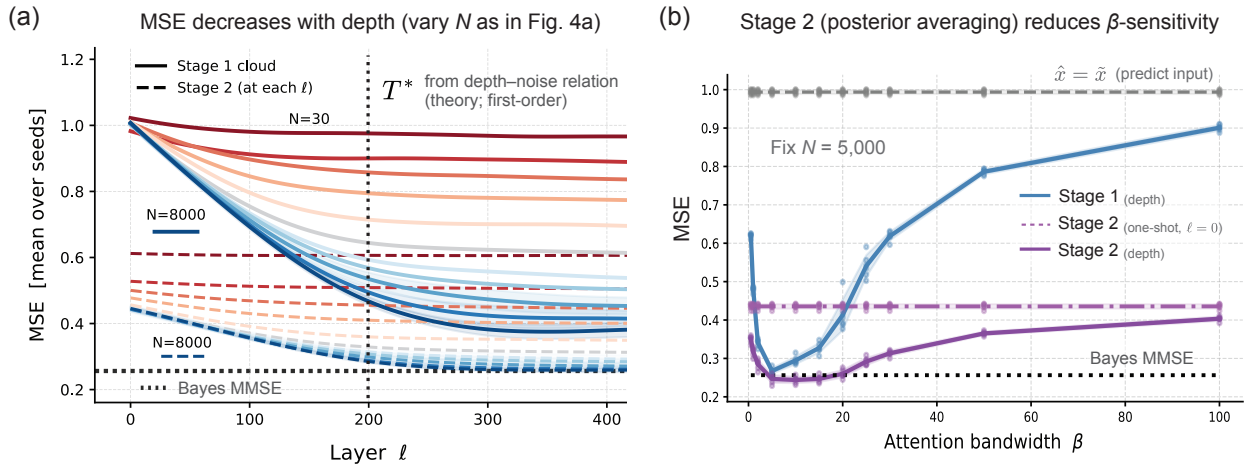

Figure 5. (a) Mean squared error (MSE) as a function of depth ℓ for varying context size N. Increasing depth improves performance up to a finite horizon T^* predicted by the depth–noise relation, beyond which gains saturate. Larger context sizes yield better particle priors and lower error, approaching the Bayes MMSE. (b) MSE as a function of attention bandwidth β for fixed N. While Stage 1 (particle refinement) is sensitive to kernel bandwidth, Stage 2 (posterior averaging) substantially reduces this sensitivity, yielding robust performance across a wide range of β. These results highlight the complementary roles of depth (prior refinement) and cross-attention (posterior inference). Same parameters used as Fig. 4.

Figure 5(a) shows the evolution of MSE with depth (i.e., in-context during a single forward pass) for varying context sizes N. Consistent with the empirical Bayes interpretation, refinement of the particle approximation to the prior facilitates improved posterior estimates. The improvement saturates near a finite depth T^*, in agreement with the predicted depth–noise relationship. Larger context sizes reduce finite-sample error and allow the estimator to approach the Bayes optimal bound more closely.

Figure 5(b) examines sensitivity to the attention bandwidth β. We emphasize that, for moderate noise strength, there should be no expectation that the particle prior alone be used for x-pred denoising. Posterior averaging (Stage 2) is the correct step for this task in the framework, and the empirics indicate its general robustness. Indeed, while

Stage 1 performance depends relatively strongly on β (consistent with Fig. 2(b) in the main text), posterior averaging substantially mitigates this sensitivity.

D. Posterior mean for Gaussian mixture prior

We recall the posterior mean for Gaussian mixture priors in order to make explicit its connection to the attention-based estimator used in the main text (see e.g. Bishop (2006) for background). The point-mass case arises as a limiting regime.

Recall that under squared error loss (MSE), the Bayes-optimal estimator is the posterior mean

$$\hat{x}(y) = \mathbb{E}[X \mid \tilde{X} = y] = \int x\, p_{X|\tilde{X}}(x \mid y)\, dx.$$

Under Gaussian mixture priors, $\rho_0 = \sum_{i=1}^{K} \pi_i \mathcal{N}(\mu_i, \Sigma_i)$, this expression has a closed form. For observations corrupted by additive Gaussian noise, $\tilde{X} = X + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$, the posterior distribution $p(X \mid \tilde{X} = y)$ is also a mixture, but with observation-dependent weights:

$$w_i(y) := P(\text{component} = i \mid y) = \frac{\pi_i\, \mathcal{N}(y \mid \mu_i, \Sigma_i + \sigma^2 I_d)}{\sum_{j=1}^{K} \pi_j\, \mathcal{N}(y \mid \mu_j,\, \Sigma_j + \sigma^2 I_d)}.$$

Evaluating the posterior mean $\int x\, p_{X|\tilde{X}}(x \mid y)\, dx$ yields:

$$\mathbb{E}[X \mid \tilde{X} = y] = \sum_{i=1}^{K} w_i(y) \left[\mu_i + \Sigma_i(\Sigma_i + \sigma^2 I_d)^{-1}(y - \mu_i)\right], \tag{21}$$

which is a weighted combination of component-wise shrinkage estimators. The weights are determined by the likelihood of each mixture component under the noisy observation y. Intuitively, the estimate moves the observation toward nearby component means, combining them according to their likelihood and shrinking the displacement according to the signal-to-noise ratio. E.g., for a single isotropic component with zero mean and covariance $\Sigma = a^2 I_d$, this reduces to the familiar shrinkage estimator $\mathbb{E}[X \mid \tilde{X} = y] = \frac{a^2}{a^2+\sigma^2}\, y$.

Point-mass mixture. In the limiting case where the prior is a point mass mixture, $\rho_0 = \sum_{i=1}^{K} \pi_i \delta_{\mu_i}$, the posterior weights reduce to $w_i(y) \propto \pi_i \exp\left(-\frac{\|y-\mu_i\|^2}{2\sigma^2}\right)$ and the posterior mean becomes

$$\mathbb{E}[X \mid \tilde{X} = y] = \sum_{i=1}^{K} w_i(y)\, \mu_i = \frac{\sum_{i=1}^{K} \pi_i \exp\left(-\frac{\|y-\mu_i\|^2}{2\sigma^2}\right) \mu_i}{\sum_{j=1}^{K} \pi_j \exp\left(-\frac{\|y-\mu_j\|^2}{2\sigma^2}\right)}.$$

Thus, in the point-mass limit, the posterior mean reduces exactly to a Gaussian kernel-weighted average over support points, coinciding with the attention operation in Algorithm 1 and yielding a probabilistic interpretation of finite-sample attention as posterior averaging under a discrete prior.

Bayes MMSE. Eq. (21) defines the Bayes optimal estimator for the x-prediction MSE objective when the prior is a known Gaussian mixture. As this is the best estimate one could make for $\hat{x}$ when the prior is known, this serves as the key baseline in Fig. 4. In contrast, our framework operates in the regime where the prior itself is unknown and must be jointly estimated from corrupted observations.

E. Exact Gaussian Drift, Hard Truncation, and Sequential Posterior-Mean Recovery

E.1. Pseudocode: ODE-based continuous-time sequential posterior-mean estimator

We record the continuous-depth analogues of the two-stage estimator in Algorithm 1. The notation is chosen to match the main text: the observed noisy tokens are $\tilde{x}_i$, the evolving particles are z_i, and y denotes a query point. To denoise

the original prompt one takes $y = \tilde{x}_i$ for each input token. The observation-noise variance is denoted by τ in the analysis below; it is the same quantity as σ^2 in the main text, so $T_\beta = \beta\tau/2 = \beta\sigma^2/2$. Algorithm 2 is the full empirical flow corresponding to the finite-sample attention dynamics, while Algorithm 3 is the compact-support version used in the proof for fixed R.

Algorithm 2 Full-sample continuous-depth posterior-mean estimator

Require: Noisy tokens $\tilde{x}_1, \dots, \tilde{x}_N \sim f_0 = \gamma_\tau * P_0$, inverse temperature β, terminal time $T_\beta = \beta\tau/2$, query y

1: Initialize $z_i(0) \leftarrow \tilde{x}_i$ and $\mu_t^N := \frac{1}{N}\sum_{j=1}^N \delta_{z_j(t)}$

2: Evolve, for $0 \le t \le T_\beta$ and $i = 1, \dots, N$,

$$\dot{z}_i(t) = X_\beta[\mu_t^N](z_i(t)) = \frac{\sum_{j=1}^N z_j(t) G_\beta(z_i(t) - z_j(t))}{\sum_{j=1}^N G_\beta(z_i(t) - z_j(t))} - z_i(t).$$

3: Output the posterior mean against the refined particle prior:

$$\widehat{m}_N(y) := m_{\mu_{T_\beta}^N}(y) = \frac{\sum_{i=1}^N z_i(T_\beta)\gamma_\tau(y - z_i(T_\beta))}{\sum_{i=1}^N \gamma_\tau(y - z_i(T_\beta))}.$$

Algorithm 3 Empirical hard-truncation continuous-depth posterior-mean estimator

Require: Noisy tokens $\tilde{x}_1, \dots, \tilde{x}_N \sim f_0 = \gamma_\tau * P_0$, radius R, inverse temperature β, terminal time $T_\beta = \beta\tau/2$, query y

1: Retain the tokens in the radius-R ball and relabel them as

$$\{z_1(0), \dots, z_{N_R}(0)\} := \{\tilde{x}_i : \ |\tilde{x}_i| \le R\}.$$

2: **if** $N_R = 0$ **then**

3: Increase R or resample.

4: **end if**

5: Set $\mu_t^{N_R,\beta,[R]} := \frac{1}{N_R}\sum_{j=1}^{N_R} \delta_{z_j(t)}$

6: Evolve, for $0 \le t \le T_\beta$ and $i = 1, \dots, N_R$,

$$\dot{z}_i(t) = X_\beta[\mu_t^{N_R,\beta,[R]}](z_i(t)) = \frac{\sum_{j=1}^{N_R} z_j(t) G_\beta(z_i(t) - z_j(t))}{\sum_{j=1}^{N_R} G_\beta(z_i(t) - z_j(t))} - z_i(t).$$

7: Output the posterior mean against the truncated refined particle prior:

$$\widehat{m}_{N_R,R}(y) := m_{\mu_{T_\beta}^{N_R,\beta,[R]}}(y) = \frac{\sum_{i=1}^{N_R} z_i(T_\beta)\gamma_\tau(y - z_i(T_\beta))}{\sum_{i=1}^{N_R} \gamma_\tau(y - z_i(T_\beta))}.$$

E.2. Mathematical preliminaries

This section proves a sequential particle approximation theorem for Gaussian-prior posterior means under the exact Gaussian logarithmic drift. The argument has three steps. First, at fixed β, we establish the compact-support mean-field theory, including the barycentric form of the drift, invariant-ball estimates, deterministic well-posedness, stability, and pointwise-in-time propagation of chaos. Second, we formulate a noncompact admissibility criterion based on hard truncation and show that it transfers compact-support particle approximation to noncompact data. Third, we introduce an admissible-recoverable class of priors and verify that Gaussian priors belong to it.

We will use the following background results: the Kantorovich–Rubinstein duality formula for W_1 (Villani, 2009, Particular Case 5.16), the characterization of Wasserstein convergence by weak convergence plus moment convergence

(Villani, 2009, Definition 6.8 and Theorem 6.9), the continuity-equation formulation of absolutely continuous Wasserstein curves (Ambrosio et al., 2008, Sections 8.1–8.3), the Dobrushin fixed-point and stability method for Vlasov-type mean-field equations (Dobrushin, 1979, main Theorem, Propositions 2–4, and Section 6), the standard coupling method for propagation of chaos and the nonlinear-process viewpoint (Sznitman, 1991, Chapter I, Sections 1–2) (Chaintron & Diez, 2022, Section 4.1), the empirical W_1 estimate obtained from (Fournier & Guillin, 2015, Theorem 1, with $p = 1$), the Bihari–Osgood form of the generalized Bellman lemma (Bihari, 1956, Sections 3–4).

We shall also use the following elementary form of Gronwall's inequality. If $u : [0, T] \to [0, \infty)$ is continuous and

$$u(t) \le a(t) + L \int_0^t u(s)\, ds, \qquad 0 \le t \le T,$$

where $L \ge 0$ and $a : [0, T] \to [0, \infty)$ is nondecreasing, then

$$u(t) \le a(t) e^{Lt}, \qquad 0 \le t \le T.$$

In particular, if $u(t) \le L \int_0^t u(s)\, ds$, then $u \equiv 0$. The only nonlinear variant used below is the Osgood–Bihari argument in Theorem E.20.

All noncompact particle statements below are sequential: first $N \to \infty$ at fixed (β, R), then $R \to \infty$ at fixed β, and finally $\beta \to \infty$.

All existence assertions are made on a prescribed finite time interval. For compactly supported data this causes no loss, because the invariant-ball and Lipschitz estimates below give global-in-time characteristic solutions on every finite horizon. For general noncompact data we do not assert a general global theory; instead, admissibility is a finite-horizon hypothesis. In the Gaussian case this hypothesis is verified directly by an explicit positive-definite covariance flow.

E.3. Fixed-β mean-field theory for the exact Gaussian drift

Throughout we write $\mathcal{P}(\mathbb{R}^d)$ for the set of Borel probability measures on $\mathbb{R}^d$, and

$$\mathcal{P}_1(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} |x|\, \mu(dx) < \infty \right\}.$$

For $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$, we denote by $W_1(\mu, \nu)$ the 1-Wasserstein distance. If $E \subset \mathbb{R}^d$ and $F : E \to \mathbb{R}^k$, we write

$$\mathrm{Lip}_E(F) := \sup_{\substack{x, x' \in E \\ x \ne x'}} \frac{|F(x) - F(x')|}{|x - x'|}$$

for the Lipschitz constant of F on E. When $E = \mathbb{R}^d$, we write simply

$$\mathrm{Lip}(F) := \mathrm{Lip}_{\mathbb{R}^d}(F).$$

If $F = F(x, \theta)$ depends on a spatial variable $x \in \mathbb{R}^d$ and on an auxiliary parameter θ, then

$$\mathrm{Lip}_x\, F(\cdot, \theta)$$

denotes the Lipschitz constant of the map $x \mapsto F(x, \theta)$, with θ held fixed.

We shall use the Kantorovich–Rubinstein formula (Villani, 2009, Particular Case 5.16)

$$W_1(\mu, \nu) = \sup_{\substack{\varphi : \mathbb{R}^d \to \mathbb{R} \\ \mathrm{Lip}(\varphi) \le 1}} \int_{\mathbb{R}^d} \varphi(x)\, (\mu - \nu)(dx). \tag{22}$$

For $R > 0$, we write

$$B_R := \{ x \in \mathbb{R}^d : |x| \le R \}$$

for the closed Euclidean ball of radius R centered at the origin, and

$$\mathcal{P}(B_R) := \{ \mu \in \mathcal{P}(\mathbb{R}^d) : \operatorname{supp} \mu \subset B_R \}.$$

On $\mathcal{P}(B_R)$, weak convergence and W_1-convergence agree by (Villani, 2009, Theorem 6.9); since B_R is compact, $\mathcal{P}(B_R)$ is weakly compact, hence $(\mathcal{P}(B_R), W_1)$ is compact and therefore complete.

E.3.1. Notation and basic objects

The compact-support results in this section are stated for initial laws already supported in a fixed ball. No truncation is used in their formulation. Truncations enter only in Section E.4, where noncompact initial laws are approximated by compactly supported laws.

Throughout this section, $|\cdot|$ refers to Euclidean vector norm. For $\beta > 0$, set

$$G_\beta(z) := \Big(\frac{\beta}{2\pi}\Big)^{d/2} e^{-\frac{\beta}{2}|z|^2}, \qquad z \in \mathbb{R}^d.$$

For $\mu \in \mathcal{P}_1(\mathbb{R}^d)$, define

$$D_\beta[\mu](x) := (G_\beta * \mu)(x) = \int_{\mathbb{R}^d} G_\beta(x-y)\,\mu(dy).$$

Since $G_\beta > 0$, the logarithmic drift

$$X_\beta[\mu](x) := \frac{1}{\beta}\frac{\nabla(G_\beta * \mu)(x)}{(G_\beta * \mu)(x)} \tag{23}$$

is well defined for every $x \in \mathbb{R}^d$. For a fixed observation-noise variance $\tau > 0$, define the Gaussian posterior-mean map

$$m_\mu(y) := \frac{\int_{\mathbb{R}^d} x\,\gamma_\tau(y-x)\,\mu(dx)}{\int_{\mathbb{R}^d} \gamma_\tau(y-x)\,\mu(dx)}, \qquad \gamma_\tau(z) := (2\pi\tau)^{-d/2} e^{-|z|^2/(2\tau)}. \tag{24}$$

For a matrix $A \in \mathbb{R}^{d\times d}$, we write

$$\|A\|_{\mathrm{HS}} := \big(\mathrm{Tr}(A^\top A)\big)^{1/2} = \left(\sum_{i,j=1}^{d} A_{ij}^2\right)^{1/2}$$

for its Hilbert–Schmidt, equivalently Frobenius, norm.

E.3.2. Posterior means under W_1-convergence

Proposition E.1 (Local W_1-continuity of the Gaussian posterior mean)**.** *Fix $\tau > 0$. Let $\nu_n, \nu \in \mathcal{P}_1(\mathbb{R}^d)$, and assume*

$$W_1(\nu_n, \nu) \to 0.$$

Then, for every $M < \infty$,

$$\sup_{|y|\le M} |m_{\nu_n}(y) - m_\nu(y)| \to 0.$$

Proof. Write

$$Z_\lambda(y) := \int \gamma_\tau(y-x)\,\lambda(dx), \qquad N_\lambda(y) := \int x\gamma_\tau(y-x)\,\lambda(dx).$$

For $|y| \le M$, the functions $x \mapsto \gamma_\tau(y-x)$ are globally Lipschitz with a uniform Lipschitz constant. Therefore

$$\sup_{|y|\le M} |Z_{\nu_n}(y) - Z_\nu(y)| \le C_{\tau,M} W_1(\nu_n, \nu) \to 0.$$

Similarly, each coordinate of $x \mapsto x\gamma_\tau(y-x)$ is bounded and globally Lipschitz uniformly for $|y| \le M$, because polynomial factors are dominated by the Gaussian. Hence

$$\sup_{|y|\le M} |N_{\nu_n}(y) - N_\nu(y)| \to 0.$$

Since Z_ν is continuous and strictly positive on the compact set B_M, it has a positive lower bound there. Uniform convergence of numerator and denominator then gives uniform convergence of the quotient $N_{\nu_n}/Z_{\nu_n} \to N_\nu/Z_\nu$ on B_M. □

E.3.3. Exact barycentric form of the drift

Proposition E.2 (Exact drift as a Gaussian barycenter minus the identity)**.** *For $\mu \in \mathcal{P}_1(\mathbb{R}^d)$, define*

$$A_\beta[\mu](x) := \int_{\mathbb{R}^d} y\, G_\beta(x-y)\, \mu(dy), \qquad F_{\beta,\mu}(x) := \frac{A_\beta[\mu](x)}{D_\beta[\mu](x)}.$$

Then, for every $x \in \mathbb{R}^d$,

$$X_\beta[\mu](x) = F_{\beta,\mu}(x) - x. \tag{25}$$

In particular, if $\mu \in \mathcal{P}(B_R)$, then

$$F_{\beta,\mu}(x) \in \overline{\mathrm{co}}(\operatorname{supp}\mu) \subset B_R \qquad \textit{for every } x \in \mathbb{R}^d.$$

Moreover, if $\beta = \tau^{-1}$, then

$$X_{\tau^{-1}}[\mu](y) = m_\mu(y) - y \qquad (y \in \mathbb{R}^d). \tag{26}$$

Proof. Since $\nabla G_\beta(z) = -\beta z G_\beta(z)$,

$$\frac{1}{\beta}\nabla(G_\beta * \mu)(x) = -\int_{\mathbb{R}^d} (x-y) G_\beta(x-y)\, \mu(dy).$$

Dividing by $D_\beta[\mu](x) > 0$ gives

$$X_\beta[\mu](x) = \frac{\int y G_\beta(x-y)\, \mu(dy)}{\int G_\beta(x-y)\, \mu(dy)} - x = F_{\beta,\mu}(x) - x.$$

The normalized measure

$$\frac{G_\beta(x-y)}{D_\beta[\mu](x)}\, \mu(dy)$$

is a probability measure supported on $\operatorname{supp}\mu$, so its barycenter lies in $\overline{\mathrm{co}}(\operatorname{supp}\mu)$. If $\beta = \tau^{-1}$, then $G_\beta = \gamma_\tau$, and the identity becomes $X_{\tau^{-1}}[\mu] = m_\mu - \mathrm{Id}$. □

E.3.4. Compact-support geometry and invariant balls

Proposition E.3 (Invariant ball for the exact drift)**.** *Fix $\beta > 0$ and $R > 0$.*

(i) If $\mu \in \mathcal{P}(B_R)$, then for every $x \in \mathbb{R}^d$,

$$\langle x, X_\beta[\mu](x)\rangle \le R|x| - |x|^2. \tag{27}$$

In particular, $\langle x, X_\beta[\mu](x)\rangle \le 0$ whenever $|x| \ge R$.

(ii) Let $(\mu_t)_{t\in[0,T]} \subset \mathcal{P}(B_R)$ be measurable, and let x_t solve

$$\dot{x}_t = X_\beta[\mu_t](x_t), \qquad x_0 \in B_R.$$

Then $x_t \in B_R$ for every $t \in [0,T]$.

(iii) If the initial data of the exact N-particle system

$$\dot{X}_t^{i,N} = X_\beta[\mu_t^N](X_t^{i,N}), \qquad \mu_t^N := \frac{1}{N}\sum_{j=1}^N \delta_{X_t^{j,N}}, \qquad i = 1,\dots,N,$$

all belong to B_R, then every particle stays in B_R for all later times.

Proof. By Proposition E.2,

$$X_\beta[\mu](x) = F_{\beta,\mu}(x) - x, \qquad F_{\beta,\mu}(x) \in B_R$$

whenever $\mu \in \mathcal{P}(B_R)$. Hence

$$\langle x, X_\beta[\mu](x)\rangle = \langle x, F_{\beta,\mu}(x)\rangle - |x|^2 \le R|x| - |x|^2,$$

which proves (i). Parts (ii) and (iii) follow by the standard first-exit argument applied to $t \mapsto |x_t|^2/2$: on the exterior region $\{|x| > R\}$, the radial derivative is strictly negative. Therefore a trajectory starting in B_R cannot exit B_R. The particle assertion is the same argument applied to each particle, since the empirical measure is supported in B_R as long as the particles are. □

E.3.5. Lipschitz control on a bounded support class

Proposition E.4 (Exact drift is Lipschitz on an invariant ball)**.** *Fix $\beta > 0$ and $R > 0$. Then there exists $L_{\beta,R} > 0$ such that for all $x, x' \in B_R$ and all $\mu, \nu \in \mathcal{P}(B_R)$,*

$$|X_\beta[\mu](x) - X_\beta[\nu](x')| \le L_{\beta,R}\big(|x - x'| + W_1(\mu,\nu)\big).$$

Proof. Write $D_\mu = D_\beta[\mu]$, $A_\mu = A_\beta[\mu]$, and $F_{\beta,\mu} = A_\mu/D_\mu$. By Proposition E.2, $X_\beta[\mu] = F_{\beta,\mu} - \mathrm{Id}$, so it suffices to estimate $F_{\beta,\mu}$.

For $x, y \in B_R$,

$$D_\mu(x) \ge c_{\beta,R} := \inf_{u,v\in B_R} G_\beta(u - v) > 0.$$

Moreover,

$$|A_\mu(x)| \le R\|G_\beta\|_\infty.$$

The maps $y \mapsto G_\beta(x - y)$ and $y \mapsto yG_\beta(x - y)$, restricted to B_R, have Lipschitz constants bounded uniformly for $x \in B_R$. Hence, by the Kantorovich–Rubinstein formula (Villani, 2009, Particular Case 5.16),

$$|D_\mu(x) - D_\nu(x')| + |A_\mu(x) - A_\nu(x')| \le C_{\beta,R}\big(|x - x'| + W_1(\mu,\nu)\big).$$

Using the quotient identity

$$\frac{A_\mu(x)}{D_\mu(x)} - \frac{A_\nu(x')}{D_\nu(x')} = \frac{A_\mu(x) - A_\nu(x')}{D_\mu(x)} + A_\nu(x')\frac{D_\nu(x') - D_\mu(x)}{D_\mu(x)D_\nu(x')}$$

and the lower bound $D_\mu, D_\nu \ge c_{\beta,R}$, we obtain

$$|F_{\beta,\mu}(x) - F_{\beta,\nu}(x')| \le C_{\beta,R}\big(|x - x'| + W_1(\mu,\nu)\big).$$

Since $X_\beta[\mu](x) = F_{\beta,\mu}(x) - x$, the claim follows after increasing the constant. □

Proposition E.5 (Characteristic well-posedness and stability on the invariant ball)**.** *Fix $\beta > 0$, $R > 0$, and $T > 0$, and define*

$$b(x,\mu) := X_\beta[\mu](x), \qquad x \in \mathbb{R}^d,\ \mu \in \mathcal{P}(B_R).$$

Let $L_{\beta,R} > 0$ be the constant from Proposition E.4.

Then the following hold.

(i) For every continuous path

$$q \in C([0,T];\mathcal{P}(B_R))$$

and every $x \in B_R$, there exists a unique solution

$$z^{q,x} \in C^1([0,T];\mathbb{R}^d)$$

of

$$\dot z_t = b(z_t, q_t), \qquad z_0 = x.$$

Moreover,

$$z_t^{q,x} \in B_R \qquad \textit{for every } t \in [0,T].$$

We define

$$\Phi_t^q(x) := z_t^{q,x}, \qquad x \in B_R,\ t \in [0,T].$$

(ii) For every $f_0 \in \mathcal{P}(B_R)$, the map

$$\Gamma : C([0,T];\mathcal{P}(B_R)) \to C([0,T];\mathcal{P}(B_R)), \qquad (\Gamma q)_t := (\Phi_t^q)_\# f_0,$$

is well defined and has a unique fixed point

$$f \in C([0,T];\mathcal{P}(B_R)).$$

Equivalently, there exist a unique curve

$$f \in C([0,T];\mathcal{P}(B_R))$$

and a unique family of maps

$$\Phi_t : B_R \to B_R, \qquad t \in [0,T],$$

such that, for every $x \in B_R$, the trajectory

$$t \longmapsto \Phi_t(x)$$

is the unique C^1 solution of

$$\dot{z}_t = b(z_t, f_t), \qquad z_0 = x,$$

and

$$f_t = (\Phi_t)_\# f_0 \qquad \textit{for every } t \in [0,T].$$

(iii) If $f, g \in C([0,T];\mathcal{P}(B_R))$ are two solutions corresponding to initial data $f_0, g_0 \in \mathcal{P}(B_R)$, then for every $t \in [0,T]$,

$$W_1(f_t, g_t) \le e^{2L_{\beta,R}t} W_1(f_0, g_0). \tag{28}$$

Proof. For a prescribed path $q \in C([0,T];\mathcal{P}(B_R))$, the vector field $x \mapsto b(x, q_t)$ is continuous in t, Lipschitz in x on B_R uniformly in t, and points inward at ∂B_R by Proposition E.3. Hence the characteristic equation is globally well posed on $[0,T]$, and the flow $\Phi_t^q : B_R \to B_R$ is well defined.

The nonlinear law is obtained as the fixed point of

$$\Gamma q := \big((\Phi_t^q)_\# f_0\big)_{0 \le t \le T}.$$

By Proposition E.4, for two paths q, r,

$$|\Phi_t^q(x) - \Phi_t^r(x)| \le L_{\beta,R} \int_0^t e^{L_{\beta,R}(t-s)} W_1(q_s, r_s)\, ds.$$

Therefore, for the exponentially weighted metric

$$d_\alpha(q,r) := \sup_{0 \le t \le T} e^{-\alpha t} W_1(q_t, r_t),$$

one has

$$d_\alpha(\Gamma q, \Gamma r) \le \frac{L_{\beta,R}}{\alpha - L_{\beta,R}}\, d_\alpha(q,r).$$

Choosing $\alpha > 2L_{\beta,R}$, Γ is a contraction on the complete space $C([0,T];\mathcal{P}(B_R))$. This gives existence and uniqueness of f and the characteristic flow. This is the usual Dobrushin fixed-point argument in the Kantorovich–Rubinstein metric; compare (Dobrushin, 1979, Propositions 2–3 and Section 6).

For stability, let f, g be two solutions and couple their initial data by an arbitrary coupling π_0. If X_t, Y_t are the corresponding nonlinear characteristics, then

$$|X_t - Y_t| \le |X_0 - Y_0| + L_{\beta,R} \int_0^t \big(|X_s - Y_s| + W_1(f_s, g_s)\big)\, ds.$$

Taking expectations, using $W_1(f_s, g_s) \le \mathbb{E}|X_s - Y_s|$, and applying Gronwall gives

$$W_1(f_t, g_t) \le e^{2L_{\beta,R}t} W_1(f_0, g_0).$$

This is the standard Dobrushin stability estimate; compare (Dobrushin, 1979, Proposition 4). □

E.3.6. Weak formulations for the later noncompact passage

The compact-support solutions constructed above are characteristic solutions. In the noncompact truncation limit, however, compactness only gives convergence of measure-valued curves. We therefore record the weak continuity-equation formulation used later. This is the usual weak formulation of the continuity equation in $\mathbb{R}^d$; compare (Ambrosio et al., 2008, Sections 8.1–8.3). Compactly supported characteristic solutions are special cases of this definition, by the chain rule along characteristics.

Definition E.1 (Finite-action weak solutions)**.** Fix $\beta > 0$ and $T > 0$. A curve

$$f \in C([0,T]; \mathcal{P}_1(\mathbb{R}^d))$$

is called a finite-action weak solution of

$$\partial_t f_t + \nabla \cdot (f_t X_\beta[f_t]) = 0 \tag{29}$$

on $[0,T]$, with initial condition f_0, if $f_{t=0} = f_0$,

$$\int_0^T \int_{\mathbb{R}^d} |X_\beta[f_t](x)|\, f_t(dx)\, dt < \infty,$$

and, for every $\phi \in C_c^1(\mathbb{R}^d)$,

$$\int \phi\, df_t - \int \phi\, df_0 = \int_0^t \int \nabla\phi(x) \cdot X_\beta[f_s](x)\, f_s(dx)\, ds \tag{30}$$

for every $t \in [0,T]$.

This definition is not an additional existence theorem. It is only the formulation in which the noncompact limits below will be identified. We now return to the compactly supported particle system and prove the pointwise propagation-of-chaos estimate needed for the truncation argument.

E.3.7. Fixed-β compact-support pointwise propagation of chaos

The next theorem uses propagation of chaos in the pointwise-in-time sense: for each fixed t and each fixed k, the k-particle marginal converges to $f_t^{\otimes k}$.

Theorem E.6 (Compactly supported fixed-β propagation of chaos)**.** *Fix $\beta > 0$, $T > 0$, and $R > 0$, and assume*

$$f_0 \in \mathcal{P}(B_R).$$

Let $(X_0^i)_{i\ge1}$ be an i.i.d. sequence with common law f_0. For each $N \ge 1$, use $(X_0^1, \dots, X_0^N)$ as particle initial data and let

$$\dot{X}_t^{i,N} = X_\beta[\mu_t^N](X_t^{i,N}), \qquad \mu_t^N := \frac{1}{N}\sum_{j=1}^N \delta_{X_t^{j,N}}, \qquad i = 1, \dots, N,$$

be the exact particle system. Let

$$f \in C([0,T]; \mathcal{P}(B_R))$$

and

$$\Phi_t : B_R \to B_R$$

be the unique nonlinear law path and nonlinear characteristic flow from Proposition E.5, and define

$$\bar{X}_t^i := \Phi_t(X_0^i), \qquad i \geq 1, \qquad \bar{\mu}_t^N := \frac{1}{N}\sum_{j=1}^N \delta_{\bar{X}_t^j}.$$

Then:

(i) the particle system is well posed on $[0,T]$, and the nonlinear law path f_t is well posed on $[0,T]$, and

$$\operatorname{supp} \mu_t^N \subset B_R, \qquad \operatorname{supp} f_t \subset B_R \qquad \text{for every } t \in [0,T];$$

(ii) the random variables $(\bar{X}_t^i)_{i\geq 1}$ are i.i.d. with common law f_t, and for every $t \in [0,T]$,

$$\mathbb{E}|X_t^{1,N} - \bar{X}_t^1| \leq L_{\beta,R} e^{2L_{\beta,R}t} \int_0^t \mathbb{E}\big[W_1(\bar{\mu}_s^N, f_s)\big]\, ds;$$

(iii) for every fixed $t \in [0,T]$,

$$\mathbb{E}\big[W_1(\mu_t^N, f_t)\big] \longrightarrow 0 \qquad \text{as } N \to \infty;$$

(iv) if $F_t^{N,k}$ denotes the law of $(X_t^{1,N}, \dots, X_t^{k,N})$, then for each fixed $k \geq 1$,

$$W_1\big(F_t^{N,k}, f_t^{\otimes k}\big) \longrightarrow 0 \qquad \text{as } N \to \infty,$$

where W_1 on $(\mathbb{R}^d)^k$ is taken with respect to the cost

$$(x_1, \dots, x_k), (y_1, \dots, y_k) \longmapsto \sum_{i=1}^k |x_i - y_i|.$$

Proof. The particle vector field is smooth on $(\mathbb{R}^d)^N$, since

$$F_{N,i}(x_1, \dots, x_N) = \frac{\sum_{j=1}^N x_j G_\beta(x_i - x_j)}{\sum_{j=1}^N G_\beta(x_i - x_j)} - x_i$$

has a strictly positive denominator. Proposition E.3 gives global existence on $[0,T]$ and invariance of B_R^N. The nonlinear law path and flow are given by Proposition E.5.

Let

$$\bar{X}_t^i := \Phi_t(X_0^i), \qquad \bar{\mu}_t^N := \frac{1}{N}\sum_{j=1}^N \delta_{\bar{X}_t^j}.$$

Then $(\bar{X}_t^i)_{i\geq 1}$ are i.i.d. with common law f_t. Set

$$\nu_N(t) := \mathbb{E}|X_t^{1,N} - \bar{X}_t^1|.$$

We spell out the coupling step. For each time s, the measure

$$\Pi_s^N := \frac{1}{N}\sum_{j=1}^N \delta_{(X_s^{j,N}, \bar{X}_s^j)}$$

is a coupling of μ_s^N and $\bar{\mu}_s^N$. Hence

$$W_1(\mu_s^N, \bar{\mu}_s^N) \leq \frac{1}{N}\sum_{j=1}^N |X_s^{j,N} - \bar{X}_s^j|.$$

Taking expectations and using exchangeability of the particle system and of the coupled nonlinear characteristics gives

$$\mathbb{E}W_1(\mu_s^N, \bar{\mu}_s^N) \leq \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}|X_s^{j,N} - \bar{X}_s^j| = \nu_N(s). \tag{31}$$

On the other hand, the Lipschitz estimate of Proposition E.4 gives, for the first particle,

$$|X_t^{1,N} - \bar{X}_t^1| \leq L_{\beta,R} \int_0^t \left(|X_s^{1,N} - \bar{X}_s^1| + W_1(\mu_s^N, f_s)\right) ds.$$

Using

$$W_1(\mu_s^N, f_s) \leq W_1(\mu_s^N, \bar{\mu}_s^N) + W_1(\bar{\mu}_s^N, f_s)$$

and then (31), we obtain

$$\nu_N(t) \leq 2L_{\beta,R} \int_0^t \nu_N(s)\, ds + L_{\beta,R} \int_0^t \mathbb{E}W_1(\bar{\mu}_s^N, f_s)\, ds.$$

Applying the Gronwall estimate stated at the beginning of the appendix yields

$$\nu_N(t) \leq L_{\beta,R} e^{2L_{\beta,R}t} \int_0^t \mathbb{E}W_1(\bar{\mu}_s^N, f_s)\, ds.$$

This is the usual coupling estimate behind propagation of chaos; compare the classical nonlinear-process construction in (Sznitman, 1991, Chapter I, Section 1) and the modern discussion of coupling methods in (Chaintron & Diez, 2022, Section 4.1).

For fixed s, the random variables $(\bar{X}_s^j)_{j\geq 1}$ are i.i.d. with law $f_s \in \mathcal{P}(B_R)$. The empirical-measure estimate (Fournier & Guillin, 2015, Theorem 1, with $p = 1$) gives

$$\mathbb{E}W_1(\bar{\mu}_s^N, f_s) \to 0 \qquad \text{for every fixed } s,$$

because the required moment condition $M_q(f_s) < \infty$ for some $q > 1$ is automatic from the support bound $f_s \in \mathcal{P}(B_R)$. Since $0 \leq W_1(\bar{\mu}_s^N, f_s) \leq 2R$, dominated convergence gives

$$\int_0^t \mathbb{E}W_1(\bar{\mu}_s^N, f_s)\, ds \to 0.$$

Therefore $\nu_N(t) \to 0$. Hence

$$\mathbb{E}W_1(\mu_t^N, f_t) \leq \mathbb{E}W_1(\mu_t^N, \bar{\mu}_t^N) + \mathbb{E}W_1(\bar{\mu}_t^N, f_t) \leq \nu_N(t) + \mathbb{E}W_1(\bar{\mu}_t^N, f_t) \to 0.$$

Finally, coupling $(X_t^{1,N}, \ldots, X_t^{k,N})$ with $(\bar{X}_t^1, \ldots, \bar{X}_t^k)$ gives

$$W_1(F_t^{N,k}, f_t^{\otimes k}) \leq \sum_{i=1}^{k} \mathbb{E}|X_t^{i,N} - \bar{X}_t^i| = k\nu_N(t) \to 0.$$

□

E.4. Noncompact admissibility and hard-truncated particle approximation

We next formulate the deterministic hypothesis that allows compactly supported particle systems to approximate noncompact initial data. The condition is a stability statement for the noncompact mean-field equation under hard truncation of the initial law. With this in hand, the compact-support propagation-of-chaos theorem applies at fixed radius, and the radius can then be removed at the deterministic level.

E.4.1. Hard truncation as a noncompact approximation

For $R > 0$ and $\mu \in \mathcal{P}_1(\mathbb{R}^d)$, set

$$p_R(\mu) := \mu(B_R), \qquad q_R(\mu) := 1 - p_R(\mu).$$

Whenever $p_R(\mu) > 0$, define the hard truncation of μ by

$$\mu^{[R]} := \frac{\mathbf{1}_{B_R}\mu}{\mu(B_R)}. \tag{32}$$

Thus $\mu^{[R]} \in \mathcal{P}(B_R)$.

Lemma E.7 (Hard truncation in W_1)**.** *Let $\mu \in \mathcal{P}_1(\mathbb{R}^d)$, and assume that $p_R(\mu) > 0$. Let $\mu^{[R]}$ be defined by (32). Then*

$$W_1(\mu^{[R]}, \mu) \le \int_{\{|x|>R\}} |x|\, \mu(dx) + \frac{q_R(\mu)}{p_R(\mu)} \int_{B_R} |x|\, \mu(dx). \tag{33}$$

In particular,

$$W_1(\mu^{[R]}, \mu) \longrightarrow 0 \qquad \text{as } R \to \infty.$$

Proof. By Kantorovich–Rubinstein duality, it suffices to test against 1-Lipschitz φ with $\varphi(0) = 0$, so that $|\varphi(x)| \le |x|$. Since

$$\mu^{[R]} - \mu = \frac{q_R(\mu)}{p_R(\mu)} \mathbf{1}_{B_R}\mu - \mathbf{1}_{B_R^c}\mu,$$

we get

$$\left| \int \varphi \, d(\mu^{[R]} - \mu) \right| \le \frac{q_R(\mu)}{p_R(\mu)} \int_{B_R} |x|\, \mu(dx) + \int_{B_R^c} |x|\, \mu(dx).$$

Taking the supremum over $\mathrm{Lip}(\varphi) \le 1$ proves the bound. The right-hand side tends to zero because $p_R(\mu) \to 1$, $q_R(\mu) \to 0$, and $\mu \in \mathcal{P}_1(\mathbb{R}^d)$. □

Remark (Untruncated samples versus empirical hard truncation). Let $\mu \in \mathcal{P}_1(\mathbb{R}^d)$, let $Y_1, \dots, Y_N$ be i.i.d. with law μ, and set

$$\nu_N := \frac{1}{N} \sum_{i=1}^N \delta_{Y_i}, \qquad N_R := \sum_{i=1}^N \mathbf{1}_{\{|Y_i| \le R\}}, \qquad q_R(\mu) := \mu(B_R^c).$$

By Hoeffding's inequality, for every $\eta > 0$,

$$\mathbb{P}\left(\left| \frac{N_R}{N} - (1 - q_R(\mu)) \right| > \eta \right) \le 2e^{-2N\eta^2}.$$

Thus the random retained fraction N_R/N is close to $1 - q_R(\mu)$ with high probability. On the event $\{N_R > 0\}$, define the renormalized empirical truncation

$$\nu_N^{\mathrm{emp},[R]} := \frac{1}{N_R} \sum_{i:\, |Y_i| \le R} \delta_{Y_i}.$$

Then

$$\mathbb{P}(Y_1, \dots, Y_N \in B_R) = (1 - q_R(\mu))^N \ge 1 - N q_R(\mu).$$

Thus this event has high probability whenever $N q_R(\mu) \ll 1$.

Moreover, on $\{N_R > 0\}$,

$$W_1(\nu_N, \nu_N^{\mathrm{emp},[R]}) \le \frac{1}{N} \sum_{i=1}^N |Y_i| \mathbf{1}_{\{|Y_i|>R\}} + R \frac{N - N_R}{N}.$$

Consequently, if

$$A_R(\mu) := \int_{\{|x|>R\}} |x|\, \mu(dx),$$

then for every $\varepsilon > 0$,

$$\mathbb{P}\left(N_R > 0,\, W_1(\nu_N, \nu_N^{\mathrm{emp},[R]}) > \varepsilon\right) \le \frac{A_R(\mu) + Rq_R(\mu)}{\varepsilon}.$$

Since $Rq_R(\mu) \le A_R(\mu)$, the right-hand side tends to 0 as $R \to \infty$. Therefore, for any fixed $\varepsilon > 0$,

$$\lim_{R\to\infty} \limsup_{N\to\infty} \mathbb{P}\left(N_R > 0,\, W_1(\nu_N, \nu_N^{\mathrm{emp},[R]}) > \varepsilon\right) = 0.$$

This is a sampling-level comparison only; the main particle theorem is still formulated with i.i.d. initialization from the deterministic truncated law $\mu^{[R]}$.

Definition E.2 (Noncompact admissibility)**.** Fix $\beta > 0$ and $T > 0$. A law $f_0 \in \mathcal{P}_1(\mathbb{R}^d)$ is called (β, T)-admissible for the exact Gaussian drift if the following two conditions hold.

(i) There exists a finite-action weak solution

$$f^\beta \in C([0,T]; \mathcal{P}_1(\mathbb{R}^d))$$

of (29) with initial datum f_0. This solution is part of the admissibility datum and is called the associated admissible solution.

(ii) For all sufficiently large R, let

$$f_0^{[R]} := \frac{\mathbf{1}_{B_R} f_0}{f_0(B_R)},$$

and let $f^{\beta,[R]} \in C([0,T]; \mathcal{P}(B_R))$ be the compact-support solution constructed in Proposition E.5, initialized from $f_0^{[R]}$. Then

$$\sup_{0\le t\le T} W_1(f_t^{\beta,[R]}, f_t^\beta) \longrightarrow 0 \qquad \text{as } R \to \infty. \tag{34}$$

Remark (Uniqueness of the associated admissible flow). For fixed (β, T) and f_0, the associated admissible flow is unique whenever it exists. Indeed, suppose f^β and g^β both satisfy Definition E.2 for the same initial law f_0. Let $f^{\beta,[R]}$ denote the compact-support solution initialized from $f_0^{[R]}$. Then, for every $t \in [0,T]$,

$$W_1(f_t^\beta, g_t^\beta) \le W_1(f_t^\beta, f_t^{\beta,[R]}) + W_1(f_t^{\beta,[R]}, g_t^\beta).$$

Taking the supremum over $t \in [0,T]$ and then letting $R \to \infty$, the two terms on the right-hand side vanish by the admissibility condition. Hence $f_t^\beta = g_t^\beta$ for all $t \in [0,T]$.

E.4.2. From noncompact admissibility to hard-truncated particles

Theorem E.8 (Particle approximation for noncompact admissible data)**.** *Fix $\beta > 0$, $T > 0$, and $t \in [0,T]$. Let $f_0 \in \mathcal{P}_1(\mathbb{R}^d)$ be (β, T)-admissible, with admissible noncompact flow $(f_s^\beta)_{0\le s\le T}$. For each sufficiently large R, let*

$$X_0^{1,N,\beta,[R]}, \dots, X_0^{N,N,\beta,[R]}$$

be i.i.d. with law

$$f_0^{[R]} = \frac{\mathbf{1}_{B_R} f_0}{f_0(B_R)}.$$

For $j = 1, \dots, N$, let $X_s^{j,N,\beta,[R]}$ denote the solution of the exact N-particle system with inverse temperature β and truncation radius R:

$$\dot{X}_s^{j,N,\beta,[R]} = X_\beta[\mu_s^{N,\beta,[R]}]\big(X_s^{j,N,\beta,[R]}\big), \qquad \mu_s^{N,\beta,[R]} := \frac{1}{N}\sum_{\ell=1}^N \delta_{X_s^{\ell,N,\beta,[R]}}.$$

Thus the superscript $[R]$ records the hard-truncated initial law, while β records the drift parameter. The only randomness in this theorem is the i.i.d. initial sample; conditional on that sample, the particle ODE is deterministic. Then, for every $M < \infty$,

$$\lim_{R\to\infty} \limsup_{N\to\infty} \mathbb{E} \sup_{|y|\le M} \left| m_{\mu_t^{N,\beta,[R]}}(y) - m_{f_t^\beta}(y) \right| = 0. \tag{35}$$

Consequently, the same convergence holds in probability.

Proof. Fix R. Since $f_0^{[R]} \in \mathcal{P}(B_R)$, the compact-support propagation of chaos theorem gives

$$\mathbb{E}W_1(\mu_t^{N,\beta,[R]}, f_t^{\beta,[R]}) \to 0.$$

Hence $W_1(\mu_t^{N,\beta,[R]}, f_t^{\beta,[R]}) \to 0$ in probability. By Proposition E.1,

$$\sup_{|y|\le M} \left| m_{\mu_t^{N,\beta,[R]}}(y) - m_{f_t^{\beta,[R]}}(y) \right| \to 0$$

in probability. Moreover both measures are supported in B_R, so both posterior means take values in B_R, and the last display is bounded by $2R$. Since the convergence is in probability and the sequence is uniformly bounded, the expectations also converge to zero:

$$\lim_{N\to\infty} \mathbb{E} \sup_{|y|\le M} \left| m_{\mu_t^{N,\beta,[R]}}(y) - m_{f_t^{\beta,[R]}}(y) \right| = 0.$$

Therefore

$$\begin{aligned} &\limsup_{N\to\infty} \mathbb{E} \sup_{|y|\le M} \left| m_{\mu_t^{N,\beta,[R]}}(y) - m_{f_t^{\beta}}(y) \right| \\ &\quad \le \sup_{|y|\le M} \left| m_{f_t^{\beta,[R]}}(y) - m_{f_t^{\beta}}(y) \right|. \end{aligned}$$

By admissibility,

$$W_1(f_t^{\beta,[R]}, f_t^{\beta}) \to 0.$$

Proposition E.1 then implies that the right-hand side tends to zero as $R \to \infty$.

To make the last assertion explicit, set

$$Z_{N,R} := \sup_{|y|\le M} \left| m_{\mu_t^{N,\beta,[R]}}(y) - m_{f_t^{\beta}}(y) \right|.$$

The preceding argument proves

$$\lim_{R\to\infty} \limsup_{N\to\infty} \mathbb{E} Z_{N,R} = 0.$$

Therefore, for every $\varepsilon > 0$, Markov's inequality gives

$$\lim_{R\to\infty} \limsup_{N\to\infty} \mathbb{P}(Z_{N,R} > \varepsilon) \le \lim_{R\to\infty} \limsup_{N\to\infty} \frac{\mathbb{E} Z_{N,R}}{\varepsilon} = 0.$$

This is the claimed sequential convergence in probability. □

E.4.3. Recoverable admissible priors

The previous theorem separates particle approximation from the large-β recovery step. We now record the corresponding class of priors for which the full sequential posterior-mean recovery theorem follows.

Definition E.3 (Recoverable admissible priors)**.** Fix $\tau > 0$. We measure time in the raw attention-depth variable associated with the normalized drift $X_\beta = \beta^{-1}\nabla \log(G_\beta * \cdot)$. Equivalently, if $s = t/\beta$ denotes the unnormalized score-flow time, then the denoising horizon $s = \tau/2$ corresponds to

$$T_\beta := \frac{\beta\tau}{2}.$$

A prior $P_0 \in \mathcal{P}_1(\mathbb{R}^d)$ belongs to the class $\mathcal{A}_\tau$ if, with

$$f_0 := \gamma_\tau * P_0,$$

the following two conditions hold.

(i) For every $\beta > 0$, there exists an admissible flow

$$(f_t^\beta)_{0\le t\le T_\beta}$$

such that the noisy law f_0 is (β, T_β)-admissible in the sense of Definition E.2, with associated admissible solution $(f_t^\beta)_{0\le t\le T_\beta}$.

(ii) This choice of admissible flow recovers the clean prior at the denoising time:

$$W_1(f_{T_\beta}^\beta, P_0) \longrightarrow 0 \qquad \text{as } \beta \to \infty. \tag{36}$$

Remark (Role of the class $\mathcal{A}_\tau$). The class $\mathcal{A}_\tau$ is an admissible-recoverable class, not an a priori standard regularity class of priors. Its first condition is the noncompact mean-field admissibility needed to pass from hard-truncated compactly supported particles to the noncompact noisy law f_0. Its second condition is the large-β denoising condition: after the raw attention-depth time $T_\beta = \beta\tau/2$, the associated deterministic noncompact flow recovers the clean prior P_0 in W_1.

Thus the theorem below is a transfer result for priors satisfying these two conditions. The concrete case needed here is the Gaussian one, verified later in this appendix.

Theorem E.9 (Hard-truncated posterior-mean recovery for recoverable priors)**.** *Fix $\tau > 0$, and let $P_0 \in \mathcal{A}_\tau$. Set*

$$f_0 := \gamma_\tau * P_0, \qquad T_\beta := \frac{\beta\tau}{2}.$$

For $R > 0$, define

$$f_0^{[R]} := \frac{\mathbf{1}_{B_R} f_0}{f_0(B_R)}.$$

Let $\mu_t^{N,\beta,[R]}$ be the empirical measure of the exact particle system initialized i.i.d. from $f_0^{[R]}$:

$$\dot{X}_i^{N,\beta,[R]}(t) = X_\beta[\mu_t^{N,\beta,[R]}]\big(X_i^{N,\beta,[R]}(t)\big), \qquad \mu_t^{N,\beta,[R]} := \frac{1}{N}\sum_{j=1}^N \delta_{X_j^{N,\beta,[R]}(t)}.$$

Then, for every $M < \infty$,

$$\lim_{\beta\to\infty} \limsup_{R\to\infty} \limsup_{N\to\infty} \mathbb{E} \sup_{|y|\le M} \left| m_{\mu_{T_\beta}^{N,\beta,[R]}}(y) - m_{P_0}(y) \right| = 0. \tag{37}$$

The order of limits in (37) is part of the statement: first $N \to \infty$ at fixed (β, R), then $R \to \infty$ at fixed β, and finally $\beta \to \infty$. In particular, this is not a joint scaling statement in (N, R, β). Consequently, the same convergence holds in probability.

Proof. Fix $\beta > 0$. Since $P_0 \in \mathcal{A}_\tau$, the noisy law $f_0 = \gamma_\tau * P_0$ is (β, T_β)-admissible. Applying Theorem E.8 with $T = T_\beta$ and $t = T_\beta$, we obtain

$$\limsup_{R\to\infty} \limsup_{N\to\infty} \mathbb{E} \sup_{|y|\le M} \left| m_{\mu_{T_\beta}^{N,\beta,[R]}}(y) - m_{f_{T_\beta}^\beta}(y) \right| = 0.$$

Therefore

$$\begin{aligned} &\limsup_{R\to\infty} \limsup_{N\to\infty} \mathbb{E} \sup_{|y|\le M} \left| m_{\mu_{T_\beta}^{N,\beta,[R]}}(y) - m_{P_0}(y) \right| \\ &\qquad \le \sup_{|y|\le M} \left| m_{f_{T_\beta}^\beta}(y) - m_{P_0}(y) \right|. \end{aligned}$$

By the defining recovery condition (36),

$$W_1(f_{T_\beta}^\beta, P_0) \to 0 \qquad \text{as } \beta \to \infty.$$

Proposition E.1 gives

$$\sup_{|y|\le M}\left|m_{f^{\beta}_{T_\beta}}(y)-m_{P_0}(y)\right|\to 0.$$

This proves (37). If

$$Z_{N,R,\beta}:=\sup_{|y|\le M}\left|m_{\mu^{N,\beta,[R]}_{T_\beta}}(y)-m_{P_0}(y)\right|,$$

then for every $\varepsilon>0$, Markov's inequality gives

$$\mathbb{P}(Z_{N,R,\beta}>\varepsilon)\le \varepsilon^{-1}\mathbb{E}Z_{N,R,\beta},$$

and the same ordered limits therefore give convergence in probability. □

E.5. Gaussian-prior noisy laws are noncompact admissible

We now verify admissibility for noisy laws obtained by convolving a Gaussian prior with a Gaussian observation kernel. The proof uses the explicit Gaussian solution, compactness of the hard-truncated compact-support flows, and weak–strong uniqueness around the Gaussian solution.

Fix $\tau>0$. Let $m\in\mathbb{R}^d$, let Σ_0 be a symmetric nonnegative semidefinite matrix, and set

$$P_0=\mathcal{N}(m,\Sigma_0),\qquad f_0:=\gamma_\tau * P_0=\mathcal{N}(m,\Gamma_0),\qquad \Gamma_0:=\Sigma_0+\tau I.$$

Here $\mathcal{N}(m,\Sigma_0)$ is understood in the usual possibly degenerate sense: if Σ_0 is singular, it is the law of $m+\Sigma_0^{1/2}Z$ with $Z\sim\mathcal{N}(0,I_d)$, supported on $m+\operatorname{Ran}\Sigma_0^{1/2}$. Thus Γ_0 is strictly positive definite.

Remark (Sampling interpretation of hard truncation). The sequential theorem below initializes the particle system directly from the normalized truncation $f_0^{[R]}$. This is different from drawing $Y_1,\dots,Y_N$ from the untruncated law f_0 and then retaining only the particles in B_R. For the latter interpretation,

$$\mathbb{P}(Y_1,\dots,Y_N\in B_R)=f_0(B_R)^N=(1-q_R(f_0))^N\ge 1-Nq_R(f_0),\qquad q_R(f_0):=f_0(B_R^c).$$

If $f_0=\mathcal{N}(m,\Gamma_0)$, with $\lambda_+=\lambda_{\max}(\Gamma_0)$, then standard Gaussian tail bounds give, for $R\ge 1+2|m|$,

$$q_R(f_0)\le C(1+R)^d\exp\left(-\frac{R^2}{8\lambda_+}\right).$$

Thus $\mathbb{P}(Y_1,\dots,Y_N\in B_R)\to 1$ whenever

$$N(1+R)^d\exp\left(-\frac{R^2}{8\lambda_+}\right)\to 0.$$

This sampling interpretation is not used in the sequential particle theorem.

E.5.1. The explicit noncompact Gaussian flow

Lemma E.10 (Gaussian invariance and covariance equation)**.** *Fix $\beta>0$. There is a unique global positive definite solution Γ_t of the covariance equation below. The Gaussian curve*

$$f_t^\beta=\mathcal{N}(m,\Gamma_t)$$

solves

$$\partial_t f_t+\nabla\cdot(f_tX_\beta[f_t])=0,\qquad f_{t=0}=f_0, \tag{38}$$

where Γ_t is the positive definite solution of

$$\dot\Gamma_t=-2\Gamma_t(I+\beta\Gamma_t)^{-1},\qquad \Gamma_0=\Sigma_0+\tau I. \tag{39}$$

Equivalently, after diagonalizing Γ_0, each eigenvalue $\lambda_i(t)$ of Γ_t solves

$$\dot\lambda_i(t)=-\frac{2\lambda_i(t)}{1+\beta\lambda_i(t)}. \tag{40}$$

Proof. We first record why the covariance ODE is globally well posed and remains positive definite. Diagonalize Γ_0, and for each initial eigenvalue $\lambda_i(0) > 0$ consider

$$\dot{\lambda}_i(t) = -\frac{2\lambda_i(t)}{1+\beta\lambda_i(t)}.$$

The function

$$F_\beta(r) := r + \beta^{-1}\log r, \qquad r > 0,$$

is strictly increasing and maps $(0,\infty)$ onto $\mathbb{R}$. Along a solution,

$$\frac{d}{dt}F_\beta(\lambda_i(t)) = -\frac{2}{\beta},$$

so

$$F_\beta(\lambda_i(t)) = F_\beta(\lambda_i(0)) - \frac{2t}{\beta}.$$

For every finite t, this equation has a unique solution $\lambda_i(t) > 0$. Hence Γ_t is uniquely defined, remains positive definite for all finite t, and has the same eigenspaces as Γ_0.

If f is the Gaussian law $\mathcal{N}(m,\Gamma)$, with $\Gamma > 0$, then $G_\beta * f$ is the density of the Gaussian law

$$\mathcal{N}(m, \Gamma + \beta^{-1}I).$$

Hence

$$\nabla \log(G_\beta * f)(x) = -(\Gamma + \beta^{-1}I)^{-1}(x-m),$$

and therefore

$$X_\beta[f](x) = -\frac{1}{\beta}(\Gamma + \beta^{-1}I)^{-1}(x-m) = -(I+\beta\Gamma)^{-1}(x-m).$$

This velocity field is affine. Its flow fixes the mean m. The covariance satisfies

$$\dot{\Gamma}_t = -(I+\beta\Gamma_t)^{-1}\Gamma_t - \Gamma_t(I+\beta\Gamma_t)^{-1}.$$

Since $(I+\beta\Gamma_t)^{-1}$ is a matrix function of Γ_t, the two matrices commute. This gives (39). Conversely, if Γ_t solves (39), then the affine characteristic flow $\dot{x}_t = -(I+\beta\Gamma_t)^{-1}(x_t - m)$ pushes $\mathcal{N}(m,\Gamma_0)$ forward to $\mathcal{N}(m,\Gamma_t)$. Therefore the Gaussian curve satisfies (38) in weak form. The eigenvalue equation follows by diagonalizing the spectral ODE. □

Lemma E.11 (Gaussian verification of the recovery condition)**.** *Let*

$$T_\beta := \frac{\beta\tau}{2}.$$

Then

$$W_1(f^\beta_{T_\beta}, P_0) \longrightarrow 0 \qquad \textit{as } \beta \to \infty.$$

Proof. Let $s_i \geq 0$ be the eigenvalues of Σ_0. The eigenvalues $\lambda_i(t)$ of Γ_t solve

$$\dot{\lambda}_i(t) = -\frac{2\lambda_i(t)}{1+\beta\lambda_i(t)}, \qquad \lambda_i(0) = s_i + \tau.$$

Hence

$$\frac{d}{dt}\left(\lambda_i(t) + \frac{1}{\beta}\log\lambda_i(t)\right) = -\frac{2}{\beta}.$$

At $T_\beta = \beta\tau/2$,

$$\lambda_i(T_\beta) + \frac{1}{\beta}\log\lambda_i(T_\beta) = s_i + \frac{1}{\beta}\log(s_i+\tau).$$

If $s_i > 0$, the functions

$$F_\beta(r) := r + \beta^{-1}\log r$$

are strictly increasing and converge locally uniformly to r near s_i. Thus $\lambda_i(T_\beta) \to s_i$.

If $s_i = 0$, then $0 < \lambda_i(T_\beta) \le \tau$. If a subsequence converged to $\ell > 0$, then passing to the limit in the preceding identity would give $\ell = 0$, a contradiction. Hence $\lambda_i(T_\beta) \to 0$.

The covariance ODE is spectral, so the eigenspaces are fixed in time. Therefore

$$\Gamma_{T_\beta} \to \Sigma_0$$

as symmetric nonnegative matrices. Couple

$$Y_\beta = m + \Gamma_{T_\beta}^{1/2} Z, \qquad Y_0 = m + \Sigma_0^{1/2} Z, \qquad Z \sim \mathcal{N}(0, I_d).$$

By continuity of the matrix square-root on the positive semidefinite cone,

$$W_1(f_{T_\beta}^\beta, P_0) \le \mathbb{E}|Y_\beta - Y_0| \le (\mathbb{E}|Z|^2)^{1/2} \|\Gamma_{T_\beta}^{1/2} - \Sigma_0^{1/2}\|_{\mathrm{HS}} \to 0.$$

□

Remark (Time normalization). The drift is normalized as

$$X_\beta[\rho] = F_{\beta,\rho} - \mathrm{Id} = \beta^{-1} \nabla \log(G_\beta * \rho).$$

Thus the raw attention-depth time t is related to the unnormalized score time s by $s = t/\beta$. The terminal time

$$T_\beta = \frac{\beta\tau}{2}$$

therefore corresponds to the usual backward-heat denoising time $s = \tau/2$. If the observation-noise variance is denoted by σ^2, then $\tau = \sigma^2$, and this becomes $T_\beta = \beta\sigma^2/2$ and $s = \sigma^2/2$.

E.5.2. Uniform estimates for the hard-truncated Gaussian flows

For $R > 0$, define

$$p_R := f_0(B_R), \qquad f_0^{[R]} := \frac{\mathbf{1}_{B_R} f_0}{p_R},$$

and let $f_t^{\beta,[R]}$ be the compact-support solution starting from $f_0^{[R]}$:

$$\partial_t f_t^{\beta,[R]} + \nabla \cdot \big(f_t^{\beta,[R]} X_\beta[f_t^{\beta,[R]}]\big) = 0, \qquad f_{t=0}^{\beta,[R]} = f_0^{[R]}. \tag{41}$$

For fixed β and R, this is precisely the compact-support flow constructed earlier.

Lemma E.12 (Barycentric action estimate)**.** *For every $\mu \in \mathcal{P}(\mathbb{R}^d)$ with finite second moment,*

$$\int_{\mathbb{R}^d} |X_\beta[\mu](x)|^2 \, \mu(dx) \le 2 \int_{\mathbb{R}^d} |x|^2 \, \mu(dx).$$

Proof. Fix $x \in \mathbb{R}^d$, and define the probability measure

$$\pi_x^\mu(dy) := \frac{G_\beta(x-y)\,\mu(dy)}{\int_{\mathbb{R}^d} G_\beta(x-z)\,\mu(dz)}.$$

By the barycentric formula,

$$X_\beta[\mu](x) = \int_{\mathbb{R}^d} (y - x)\, \pi_x^\mu(dy).$$

Jensen's inequality gives

$$|X_\beta[\mu](x)|^2 \le \int_{\mathbb{R}^d} |y - x|^2 \, \pi_x^\mu(dy).$$

Put $r(y) = |y - x|^2$ and $w(y) = e^{-\beta r(y)/2}$. Since w is a decreasing function of r, for independent $Y, Y' \sim \mu$,

$$\mathbb{E}\big[(r(Y) - r(Y'))(w(Y) - w(Y'))\big] \le 0.$$

Expanding this inequality gives

$$\frac{\int r(y)w(y)\,\mu(dy)}{\int w(y)\,\mu(dy)} \le \int r(y)\,\mu(dy).$$

Therefore

$$|X_\beta[\mu](x)|^2 \le \int_{\mathbb{R}^d} |y - x|^2\,\mu(dy).$$

Integrating in x against $\mu(dx)$, we obtain

$$\int |X_\beta[\mu](x)|^2\,\mu(dx) \le \iint |y - x|^2\,\mu(dy)\mu(dx).$$

If $\bar{x}_\mu := \int x\,\mu(dx)$, then the double integral is computed exactly as

$$\iint |y - x|^2\,\mu(dy)\mu(dx) = 2\int |x|^2\,\mu(dx) - 2|\bar{x}_\mu|^2 \le 2\int |x|^2\,\mu(dx).$$

This proves the estimate with the stated constant. □

Proposition E.13 (Uniform second moment and action)**.** *Fix $\beta > 0$ and $T > 0$. There exists a constant $C_{\beta,T,f_0} < \infty$, independent of R, such that, for all sufficiently large R,*

$$\sup_{0\le t\le T} \int_{\mathbb{R}^d} |x|^2\,f_t^{\beta,[R]}(dx) \le C_{\beta,T,f_0},$$

and

$$\int_0^T \int_{\mathbb{R}^d} |X_\beta[f_t^{\beta,[R]}](x)|^2\,f_t^{\beta,[R]}(dx)\,dt \le C_{\beta,T,f_0}.$$

Proof. For each fixed R, the compact-support construction in Proposition E.5 gives a characteristic flow $\Phi_t^R : B_R \to B_R$ such that

$$f_t^{\beta,[R]} = (\Phi_t^R)_\# f_0^{[R]}, \qquad \frac{d}{dt}\Phi_t^R(x) = X_\beta[f_t^{\beta,[R]}](\Phi_t^R(x)).$$

The map $(t, x) \mapsto \Phi_t^R(x)$ is continuously differentiable in t and bounded on $[0, T] \times B_R$. Therefore

$$M_R(t) := \int |x|^2\,f_t^{\beta,[R]}(dx) = \int_{B_R} |\Phi_t^R(x)|^2\,f_0^{[R]}(dx)$$

is absolutely continuous and, for a.e. t,

$$\frac{d}{dt}M_R(t) = 2\int \Phi_t^R(x)\cdot X_\beta[f_t^{\beta,[R]}](\Phi_t^R(x))\,f_0^{[R]}(dx).$$

Equivalently,

$$\frac{d}{dt}M_R(t) = 2\int y\cdot X_\beta[f_t^{\beta,[R]}](y)\,f_t^{\beta,[R]}(dy).$$

Cauchy–Schwarz and Lemma E.12 give

$$\frac{d}{dt}M_R(t) \le 2M_R(t)^{1/2}\left(\int |X_\beta[f_t^{\beta,[R]}]|^2\,df_t^{\beta,[R]}\right)^{1/2} \le 2\sqrt{2}\,M_R(t).$$

Hence

$$M_R(t) \le e^{2\sqrt{2}t}M_R(0), \qquad 0 \le t \le T.$$

Since f_0 is Gaussian,

$$M_R(0) = \frac{\int_{B_R} |x|^2 \, f_0(dx)}{f_0(B_R)} \le 2 \int_{\mathbb{R}^d} |x|^2 \, f_0(dx)$$

for all sufficiently large R. This proves the uniform second-moment bound. Integrating the barycentric action estimate in time gives

$$\int_0^T \int |X_\beta[f_t^{\beta,[R]}]|^2 \, df_t^{\beta,[R]} \, dt \le 2 \int_0^T M_R(t) \, dt \le 2T \sup_{0 \le t \le T} M_R(t),$$

which is bounded independently of R. □

E.5.3. Compactness and identification of noncompact limits

Theorem E.14 (Compactness and PDE identification)**.** *Fix $\beta > 0$ and $T > 0$. Let $R_k \to \infty$. Then there exists a subsequence, not relabeled, and a curve*

$$\rho \in C([0,T]; \mathcal{P}_1(\mathbb{R}^d))$$

such that

$$\sup_{0 \le t \le T} W_1(f_t^{\beta,[R_k]}, \rho_t) \to 0.$$

Moreover, $\rho_{t=0} = f_0$, and ρ is a finite-action weak solution of

$$\partial_t \rho_t + \nabla \cdot (\rho_t X_\beta[\rho_t]) = 0$$

on $[0,T]$. In addition,

$$\sup_{0 \le t \le T} \int |x|^2 \, \rho_t(dx) < \infty, \qquad \int_0^T \int |X_\beta[\rho_t](x)|^2 \, \rho_t(dx) \, dt < \infty.$$

Proof. By Proposition E.13,

$$\sup_k \sup_{0 \le t \le T} \int |x|^2 \, f_t^{\beta,[R_k]}(dx) < \infty.$$

Hence

$$\sup_k \sup_{0 \le t \le T} \int_{|x|>A} |x| \, f_t^{\beta,[R_k]}(dx) \le \frac{C}{A} \to 0.$$

The same proposition gives a uniform L^2-action bound. Therefore, for $0 \le s < t \le T$, Kantorovich–Rubinstein duality and Cauchy–Schwarz give

$$W_1(f_t^{\beta,[R_k]}, f_s^{\beta,[R_k]}) \le \int_s^t \int |X_\beta[f_r^{\beta,[R_k]}]| \, df_r^{\beta,[R_k]} \, dr \le C|t-s|^{1/2},$$

uniformly in k. The preceding tail estimate gives tightness and uniform integrability of first moments, hence relative compactness of the time slices in W_1 by Prokhorov's theorem and (Villani, 2009, Definition 6.8 and Theorem 6.9). Together with the uniform W_1-equicontinuity, the metric Arzelà–Ascoli theorem yields a subsequence, not relabeled, and a curve $\rho \in C([0,T]; \mathcal{P}_1(\mathbb{R}^d))$ such that:

$$\sup_{0 \le t \le T} W_1(f_t^{\beta,[R_k]}, \rho_t) \to 0.$$

The initial condition follows from Lemma E.7. Lower semicontinuity of the second moment gives

$$\sup_{0 \le t \le T} \int |x|^2 \, \rho_t(dx) < \infty,$$

and Lemma E.12 gives the finite L^2-action bound.

It remains to pass to the limit in the weak formulation. Let $\phi \in C_c^1(\mathbb{R}^d)$ and $K = \operatorname{supp} \nabla\phi$. If $\mu_n \to \mu$ in W_1, then $G_\beta * \mu_n \to G_\beta * \mu$ and $\nabla G_\beta * \mu_n \to \nabla G_\beta * \mu$ uniformly on K, because the translated kernels are bounded Lipschitz uniformly on K. Since $G_\beta * \mu$ is strictly positive on K, it follows that

$$X_\beta[\mu_n] \to X_\beta[\mu] \quad \text{uniformly on } K.$$

Thus the tested fluxes converge pointwise in time. The uniform L^2-action bound gives uniform integrability in time, so Vitali's theorem permits passage to the limit in the time integral. For each R_k, the compact-support characteristic solution satisfies the weak formulation by the chain rule along characteristics. Passing to the limit in that weak formulation gives the finite-action weak formulation for ρ, in the sense of Definition E.1. □

E.5.4. Weak–strong uniqueness around the Gaussian flow

Fix $\beta > 0$ and $T > 0$, and write

$$g_t := f_t^\beta = \mathcal{N}(m, \Gamma_t).$$

On $[0, T]$, the matrices Γ_t are uniformly positive definite and uniformly bounded. For $t \in [0, T]$ and $x \in \mathbb{R}^d$, define

$$r_t(x) := |\Gamma_t^{-1/2}(x - m)|.$$

The function r_t is uniformly Lipschitz in x, and $1 + r_t(x)$ is uniformly equivalent to $1 + |x|$ on $[0, T] \times \mathbb{R}^d$.

Lemma E.15 (Reference Gaussian convolution bounds)**.** *Fix $\beta > 0$ and $T > 0$. There exist constants*

$$c > 0, \qquad C < \infty, \qquad \alpha \in (0, 1),$$

depending only on (β, T, Γ_0, d), such that, for every $t \in [0, T]$, with

$$D_{g_t} := G_\beta * g_t, \qquad H_{g_t} := \nabla(G_\beta * g_t),$$

one has

$$D_{g_t}(x) \geq c e^{-\alpha r_t(x)^2/2}, \qquad \frac{|H_{g_t}(x)|}{D_{g_t}(x)} \leq C(1 + r_t(x))$$

for every $x \in \mathbb{R}^d$. Moreover, r_t is uniformly Lipschitz in x for $t \in [0, T]$, and if $X \sim g_t$, then $r_t(X)$ has the same law as $|Z|$, where $Z \sim \mathcal{N}(0, I_d)$.

Proof. Since $g_t = \mathcal{N}(m, \Gamma_t)$, the convolution $D_{g_t} = G_\beta * g_t$ is the Gaussian density with mean m and covariance $\Gamma_t + \beta^{-1} I$. Thus

$$D_{g_t}(x) = C_t \exp\left(-\frac{1}{2}\left\langle (\Gamma_t + \beta^{-1} I)^{-1}(x - m), x - m \right\rangle\right),$$

where C_t is bounded above and below by positive constants on $[0, T]$, because Γ_t stays uniformly positive definite and uniformly bounded.

Writing $u = \Gamma_t^{-1/2}(x - m)$, the exponent becomes

$$\left\langle \Gamma_t^{1/2}(\Gamma_t + \beta^{-1} I)^{-1} \Gamma_t^{1/2} u, u \right\rangle = \left\langle \Gamma_t(\Gamma_t + \beta^{-1} I)^{-1} u, u \right\rangle.$$

The eigenvalues of

$$\Gamma_t(\Gamma_t + \beta^{-1} I)^{-1}$$

are

$$\frac{\lambda_i(t)}{\lambda_i(t) + \beta^{-1}}.$$

Since $0 < \sup_{0 \leq t \leq T, i} \lambda_i(t) \leq \Lambda_T < \infty$, these eigenvalues are bounded above by

$$\frac{\Lambda_T}{\Lambda_T + \beta^{-1}} < 1.$$

Thus the strict gap from 1 is uniform for $t \in [0, T]$. Hence there exists $\alpha \in (0, 1)$ such that

$$\left\langle (\Gamma_t + \beta^{-1} I)^{-1}(x - m), x - m \right\rangle \leq \alpha r_t(x)^2.$$

This gives the lower bound for D_{g_t}.

For the numerator,

$$H_{g_t}(x) = \nabla D_{g_t}(x) = -D_{g_t}(x)(\Gamma_t + \beta^{-1} I)^{-1}(x - m).$$

Therefore

$$\frac{|H_{g_t}(x)|}{D_{g_t}(x)} \leq C|\Gamma_t^{-1/2}(x - m)| = C r_t(x).$$

Increasing C gives $C(1 + r_t(x))$.

The uniform Lipschitz bound for r_t follows from the uniform upper bound on $\|\Gamma_t^{-1/2}\|$. Finally, if $X \sim \mathcal{N}(m, \Gamma_t)$, then

$$\Gamma_t^{-1/2}(X - m) \sim \mathcal{N}(0, I_d),$$

so $r_t(X)$ has the same law as $|Z|$. □

Lemma E.16 (Elementary logarithmic cutoff estimates)**.** *Let* $0 < \alpha < 1$*, set*

$$\kappa := \frac{1 + \alpha}{\alpha} > 2,$$

and let $C_0 \geq e$*. For* $0 < \delta < 1$*, define*

$$A_\delta := \sqrt{\kappa \log \frac{C_0}{\delta}}.$$

Then, for every $q \geq 0$*, there exist constants* $C < \infty$ *and* $\delta_0 \in (0, 1)$*, depending only on* (α, q, C_0)*, such that for* $0 < \delta \leq \delta_0$*,*

$$(1 + A_\delta)^q \delta e^{\alpha(A_\delta + 1)^2/2} \leq C$$

and

$$(1 + A_\delta)^q e^{-(A_\delta - 1)^2/2} \leq C\delta.$$

Proof. The first estimate follows from

$$\delta e^{\alpha(A_\delta + 1)^2/2} = \delta e^{\alpha A_\delta^2/2} e^{\alpha A_\delta + \alpha/2} = C\, \delta^{(1-\alpha)/2} e^{\alpha A_\delta}.$$

Since

$$A_\delta = \sqrt{\kappa \log(C_0/\delta)},$$

the factor $e^{\alpha A_\delta}$ grows sub-polynomially in δ^{-1}. More precisely, for every $\eta > 0$, after decreasing δ_0,

$$e^{\alpha A_\delta} \leq C_\eta \delta^{-\eta}.$$

Also $(1 + A_\delta)^q \leq C_\eta \delta^{-\eta}$ for every $\eta > 0$, again after decreasing δ_0. Choosing $\eta > 0$ sufficiently small gives

$$(1 + A_\delta)^q \delta^{(1-\alpha)/2} e^{\alpha A_\delta} \leq C,$$

which proves the first estimate.

For the second estimate,

$$e^{-(A_\delta - 1)^2/2} = e^{-A_\delta^2/2} e^{A_\delta - 1/2} = C\, \delta^{\kappa/2} e^{A_\delta}.$$

Because $\kappa/2 > 1$, we can choose $\eta > 0$ so small that

$$\kappa/2 - \eta > 1.$$

Using again the sub-polynomial bounds

$$e^{A_\delta} \leq C_\eta \delta^{-\eta}, \qquad (1 + A_\delta)^q \leq C_\eta \delta^{-\eta},$$

and decreasing η, if necessary, gives

$$(1 + A_\delta)^q e^{-(A_\delta - 1)^2/2} \leq C\delta.$$

□

Lemma E.17 (Transferring Gaussian radial tails by W_1)**.** *Fix $\beta > 0$, $T > 0$, and $M_2 < \infty$. There exist constants $C < \infty$ and $q < \infty$, depending only on $(\beta, T, m, \Gamma_0, M_2, d)$, such that the following holds. If $t \in [0, T]$, $\mu \in \mathcal{P}(\mathbb{R}^d)$,*

$$\int |x|^2 \, \mu(dx) \le M_2, \qquad \delta := W_1(\mu, g_t),$$

then for every $a \ge 1$,

$$\int_{\{r_t > a\}} (1 + r_t) \, d\mu \le C(1 + a)\delta + C(1 + a)^q e^{-(a-1)^2/2}.$$

Proof. Let

$$L_r := \sup_{0 \le t \le T} \mathrm{Lip}(r_t) < \infty.$$

For $a \ge 1$, choose $\theta_a : \mathbb{R} \to [0, 1]$ such that

$$\theta_a(s) = 0 \quad (s \le a - 1), \qquad \theta_a(s) = 1 \quad (s \ge a), \qquad \mathrm{Lip}(\theta_a) \le 2.$$

Then

$$\mathbf{1}_{\{r_t > a\}} \le \theta_a(r_t) \le \mathbf{1}_{\{r_t > a-1\}}, \qquad \mathrm{Lip}(\theta_a \circ r_t) \le 2L_r.$$

By Kantorovich–Rubinstein duality,

$$\mu\{r_t > a\} \le g_t\{r_t > a - 1\} + C\delta.$$

Similarly, the function $(r_t - a)_+$ is L_r-Lipschitz. Since it is unbounded, we apply Kantorovich–Rubinstein duality first to bounded Lipschitz truncations $\min\{(r_t - a)_+, M\}$, and then let $M \to \infty$ using monotone convergence and the finite first moments. Thus

$$\int (r_t - a)_+ \, d\mu \le \int (r_t - a)_+ \, dg_t + C\delta.$$

Therefore

$$\begin{aligned}\int_{\{r_t > a\}} (1 + r_t) \, d\mu &\le (1 + a)\mu\{r_t > a\} + \int (r_t - a)_+ \, d\mu \\ &\le C(1 + a)\delta + (1 + a)g_t\{r_t > a - 1\} + \int (r_t - a)_+ \, dg_t.\end{aligned}$$

Under g_t, the variable r_t has the same law as $|Z|$ with $Z \sim \mathcal{N}(0, I_d)$. Hence the standard Gaussian radial tail estimate gives, for some $q = q(d) < \infty$,

$$(1 + a)g_t\{r_t > a - 1\} + \int (r_t - a)_+ \, dg_t \le C(1 + a)^q e^{-(a-1)^2/2}.$$

Indeed, the radial density of $|Z|$ is

$$c_d s^{d-1} e^{-s^2/2} \mathbf{1}_{s \ge 0},$$

which gives the stated polynomial-times-Gaussian tail bound. □

Lemma E.18 (Gaussian-reference logarithmic stability estimate)**.** *Fix $\beta > 0$, $T > 0$, and $M_2 < \infty$. There exists $C = C(\beta, T, m, \Gamma_0, M_2, d) < \infty$ such that the following holds for every $t \in [0, T]$ and every $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfying*

$$\int |x|^2 \, \mu(dx) \le M_2.$$

Let

$$\delta := W_1(\mu, g_t).$$

Then

$$\int_{\mathbb{R}^d} |X_\beta[\mu](x) - X_\beta[g_t](x)| \, \mu(dx) \le C \, \delta \left(1 + \sqrt{\log \frac{C}{\delta}}\right), \tag{42}$$

with the convention that the right-hand side is zero when $\delta = 0$. The constant C is chosen large enough that $C/\delta \ge e$ whenever the estimate is nontrivial.

Proof. If $\delta = 0$, then $\mu = g_t$, and the estimate is immediate.

We first dispose of the large-δ range. By Lemma E.12,

$$\int |X_\beta[\mu]|\, d\mu \le (2M_2)^{1/2}.$$

The reference field is affine:

$$X_\beta[g_t](x) = -(I + \beta\Gamma_t)^{-1}(x - m),$$

and the matrices Γ_t are uniformly positive definite and uniformly bounded on $[0, T]$. Hence

$$\int |X_\beta[g_t](x)|\, \mu(dx) \le C(1 + M_2^{1/2}).$$

Also

$$\delta = W_1(\mu, g_t) \le \int |x|\, d\mu + \int |x|\, dg_t \le C(1 + M_2^{1/2}).$$

Thus the left-hand side of (42) is bounded by a constant depending only on the fixed parameters. Therefore, once the estimate is proved for $0 < \delta \le \delta_0$, the remaining range $\delta \ge \delta_0$ follows by increasing C. It remains to prove the estimate for sufficiently small δ.

Set

$$D_\nu(x) := (G_\beta * \nu)(x), \qquad H_\nu(x) := \nabla(G_\beta * \nu)(x).$$

The functions $y \mapsto G_\beta(x - y)$ and $y \mapsto \partial_i G_\beta(x - y)$ are bounded and globally Lipschitz, with bounds independent of x. Therefore Kantorovich–Rubinstein duality gives

$$\|D_\mu - D_{g_t}\|_{L^\infty(\mathbb{R}^d)} + \|H_\mu - H_{g_t}\|_{L^\infty(\mathbb{R}^d)} \le C\delta. \tag{43}$$

Let c, C, α be as in Lemma E.15. Choose $C_0 \ge e$, and define

$$\kappa := \frac{1+\alpha}{\alpha}, \qquad A_\delta := \sqrt{\kappa \log \frac{C_0}{\delta}}, \qquad E_\delta(t) := \{x : \ r_t(x) \le A_\delta\}.$$

We split the integral into the core $E_\delta(t)$ and its complement.

On $E_\delta(t)$, Lemma E.15 gives

$$D_{g_t}(x) \ge c e^{-\alpha A_\delta^2/2}.$$

By (43),

$$\frac{|D_\mu(x) - D_{g_t}(x)|}{D_{g_t}(x)} \le C\delta e^{\alpha A_\delta^2/2} = C C_0^{(1+\alpha)/2} \delta^{(1-\alpha)/2}.$$

After fixing C_0, choose $\delta_0 > 0$ so small that this last quantity is at most $1/2$ for every $0 < \delta \le \delta_0$. Then

$$D_\mu(x) \ge \frac{1}{2} D_{g_t}(x) \qquad \text{for } x \in E_\delta(t). \tag{44}$$

Using $X_\beta[\nu] = \beta^{-1} H_\nu / D_\nu$, (43), (44), and Lemma E.15, we obtain on $E_\delta(t)$

$$\begin{aligned} |X_\beta[\mu](x) - X_\beta[g_t](x)| &\le \frac{1}{\beta} \frac{|H_\mu - H_{g_t}|}{D_\mu} + \frac{1}{\beta} |H_{g_t}| \frac{|D_\mu - D_{g_t}|}{D_\mu D_{g_t}} \\ &\le C\delta \frac{1 + |H_{g_t}|/D_{g_t}}{D_{g_t}} \\ &\le C\delta(1 + r_t(x)) e^{\alpha r_t(x)^2/2}. \end{aligned}$$

It remains to show that the last weight has bounded μ-integral on $E_\delta(t)$, uniformly in t and small δ.

Let $\chi \in C^1(\mathbb{R})$ satisfy

$$0 \le \chi \le 1, \qquad \chi(s) = 1 \text{ for } s \le 0, \qquad \chi(s) = 0 \text{ for } s \ge 1, \qquad |\chi'| \le 2.$$

Define

$$h_{\delta,t}(x) := (1 + r_t(x))e^{\alpha r_t(x)^2/2}\chi(r_t(x) - A_\delta).$$

Then $h_{\delta,t}$ dominates

$$(1 + r_t)e^{\alpha r_t^2/2}\mathbf{1}_{E_\delta(t)}$$

and is supported in $\{r_t \le A_\delta + 1\}$. Moreover,

$$\mathrm{Lip}(h_{\delta,t}) \le C(1 + A_\delta)^2 e^{\alpha(A_\delta+1)^2/2}. \tag{45}$$

Indeed, write

$$F_A(r) := (1 + r)e^{\alpha r^2/2}\chi(r - A).$$

On the support of F_A, one has $0 \le r \le A + 1$, and

$$\frac{d}{dr}\big((1 + r)e^{\alpha r^2/2}\big) = e^{\alpha r^2/2}\big(1 + \alpha r(1 + r)\big).$$

Since $|\chi| \le 1$ and $|\chi'| \le 2$,

$$\sup_{r\ge 0} |F_A'(r)| \le C(1 + A)^2 e^{\alpha(A+1)^2/2}.$$

Composing with the uniformly Lipschitz functions r_t gives (45).

By Kantorovich–Rubinstein duality,

$$\int h_{\delta,t}\, d\mu \le \int h_{\delta,t}\, dg_t + \mathrm{Lip}(h_{\delta,t})\, \delta.$$

Under g_t, r_t has the same law as $|Z|$, $Z \sim \mathcal{N}(0, I_d)$. Since $\alpha < 1$,

$$\sup_{t\in[0,T]} \int (1 + r_t)e^{\alpha r_t^2/2}\, dg_t < \infty.$$

Together with (45) and Lemma E.16, this gives

$$\int h_{\delta,t}\, d\mu \le C$$

uniformly in t and $0 < \delta \le \delta_0$. Hence

$$\int_{E_\delta(t)} |X_\beta[\mu] - X_\beta[g_t]|\, d\mu \le C\delta. \tag{46}$$

We now estimate the complement. The pointwise estimate used in Lemma E.12 gives

$$|X_\beta[\mu](x)|^2 \le \int |y - x|^2\, \mu(dy) \le 2M_2 + 2|x|^2.$$

The affine formula for $X_\beta[g_t]$ gives the same linear-growth bound for the reference field. Since $1 + |x|$ and $1 + r_t(x)$ are uniformly equivalent on $[0, T] \times \mathbb{R}^d$, we have

$$|X_\beta[\mu](x)| + |X_\beta[g_t](x)| \le C(1 + r_t(x)). \tag{47}$$

Using Lemma E.17 with $a = A_\delta$, and then Lemma E.16, we obtain

$$\begin{aligned}\int_{\mathbb{R}^d\setminus E_\delta(t)} |X_\beta[\mu] - X_\beta[g_t]|\, d\mu &\le C \int_{\{r_t > A_\delta\}} (1 + r_t)\, d\mu \\ &\le C(1 + A_\delta)\delta + C(1 + A_\delta)^q e^{-(A_\delta-1)^2/2} \\ &\le C\delta(1 + A_\delta).\end{aligned}$$

Together with (46), this yields

$$\int_{\mathbb{R}^d} |X_\beta[\mu] - X_\beta[g_t]| \, d\mu \le C\delta(1 + A_\delta).$$

Since

$$A_\delta = \sqrt{\kappa \log \frac{C_0}{\delta}} \le C\sqrt{\log \frac{C}{\delta}}$$

after increasing C, the desired logarithmic estimate follows for $0 < \delta \le \delta_0$. As explained at the beginning of the proof, increasing C gives the full range of δ. □

Lemma E.19 (Stability against an affine reference field)**.** *Let $\rho_t, \eta_t \in C([0,T]; \mathcal{P}_1(\mathbb{R}^d))$ solve*

$$\partial_t \rho_t + \nabla \cdot (\rho_t b_t) = 0, \qquad \partial_t \eta_t + \nabla \cdot (\eta_t c_t) = 0$$

in the weak sense, with finite transport integrals:

$$\int_0^T \int |b_t(x)| \, \rho_t(dx) \, dt < \infty, \qquad \int_0^T \int |c_t(x)| \, \eta_t(dx) \, dt < \infty.$$

Assume that the reference field is affine,

$$c_t(x) = A_t x + a_t,$$

where $A \in C([0,T]; \mathbb{R}^{d\times d})$ and $a \in C([0,T]; \mathbb{R}^d)$. Set

$$L(t) := \|A_t\|, \qquad \Lambda(s,t) := \exp\left(\int_s^t L(r) \, dr\right).$$

Assume also

$$\int_0^T \int |b_t(x) - c_t(x)| \, \rho_t(dx) \, dt < \infty.$$

Then, for every $t \in [0,T]$,

$$W_1(\rho_t, \eta_t) \le \Lambda(0,t) W_1(\rho_{t=0}, \eta_{t=0}) + \int_0^t \Lambda(s,t) \int |b_s(x) - c_s(x)| \, \rho_s(dx) \, ds. \tag{48}$$

Proof. Let $\Phi_{s,t}$ be the affine flow generated by c_t. Then

$$|\Phi_{s,t}(x) - \Phi_{s,t}(y)| \le \Lambda(s,t)|x - y|.$$

For a 1-Lipschitz test function ψ, set

$$\phi_s(x) := \psi(\Phi_{s,t}(x)).$$

Then $\partial_s \phi_s + c_s \cdot \nabla \phi_s = 0$ and $\mathrm{Lip}(\phi_s) \le \Lambda(s,t)$. Since ψ is only Lipschitz and ϕ_s need not be compactly supported, the testing argument is justified by the standard approximation of Lipschitz functions by smooth compactly supported functions: first truncate in space, then mollify, and finally pass to the limit using the finite first moments and the finite transport integrals. Testing the two continuity equations against these approximants, subtracting, and passing to the limit gives, by the Kantorovich–Rubinstein formula,

$$W_1(\rho_t, \eta_t) \le \Lambda(0,t) W_1(\rho_{t=0}, \eta_{t=0}) + \int_0^t \Lambda(s,t) \int |b_s - c_s| \, d\rho_s \, ds.$$

This is the Dobrushin duality argument adapted to the present affine-reference setting; compare (Dobrushin, 1979, Proposition 4). □

The weak–strong uniqueness argument has three ingredients. We compare an arbitrary finite-action solution with the explicit Gaussian solution by transporting test functions along the affine Gaussian flow. The logarithmic stability estimate controls the nonlinear velocity error by an Osgood modulus of $W_1(\rho_t, g_t)$. Bihari's lemma then forces this distance to remain zero, since the two solutions have the same initial law.

Theorem E.20 (Weak–strong uniqueness around the Gaussian solution)**.** *Fix $\beta > 0$ and $T > 0$. Let $g_t = f_t^\beta = \mathcal{N}(m, \Gamma_t)$ be the Gaussian solution from Lemma E.10. Let*

$$\rho \in C([0,T]; \mathcal{P}_1(\mathbb{R}^d))$$

be a finite-action weak solution of

$$\partial_t \rho_t + \nabla \cdot (\rho_t X_\beta[\rho_t]) = 0$$

with $\rho_{t=0} = g_{t=0}$. Assume additionally that

$$\sup_{0 \le t \le T} \int |x|^2 \, \rho_t(dx) < \infty.$$

Then

$$\rho_t = g_t \qquad \text{for every } t \in [0,T].$$

Proof. Set

$$w(t) := W_1(\rho_t, g_t).$$

Apply Lemma E.19 with

$$b_t = X_\beta[\rho_t], \qquad c_t = X_\beta[g_t].$$

The hypotheses of Lemma E.19 are satisfied: ρ is finite-action, $X_\beta[g_t]$ is affine in x with uniformly bounded linear part on $[0,T]$, and the second moments of both ρ_t and g_t are uniformly bounded on $[0,T]$. Since $\rho_0 = g_0$, the affine-reference stability estimate gives a constant C_{st}, depending only on $(\beta, T, m, \Gamma_0, d)$, such that

$$w(t) \le C_{\mathrm{st}} \int_0^t \int_{\mathbb{R}^d} |X_\beta[\rho_s](x) - X_\beta[g_s](x)| \, \rho_s(dx) \, ds.$$

Let

$$M_2 := \sup_{0 \le s \le T} \int_{\mathbb{R}^d} |x|^2 \, \rho_s(dx) < \infty.$$

By Lemma E.18, there exists a constant C_{G}, depending only on

$$(\beta, T, m, \Gamma_0, M_2, d),$$

such that, for every $s \in [0,T]$,

$$\int_{\mathbb{R}^d} |X_\beta[\rho_s](x) - X_\beta[g_s](x)| \, \rho_s(dx) \le C_{\mathrm{G}} \, w(s) \left(1 + \sqrt{\log \frac{C_{\mathrm{G}}}{w(s)}}\right),$$

with the right-hand side interpreted as 0 when $w(s) = 0$.

The quantity $\sup_{0 \le s \le T} w(s)$ is finite and is controlled by the same data. Indeed,

$$\sup_{0 \le s \le T} w(s) \le \sup_{0 \le s \le T} \int |x| \, \rho_s(dx) + \sup_{0 \le s \le T} \int |x| \, g_s(dx) \le M_2^{1/2} + C(\beta, T, m, \Gamma_0, d).$$

Thus we may enlarge C_{G}, if necessary, so that

$$C_{\mathrm{G}} \ge e \sup_{0 \le s \le T} w(s).$$

This does not invalidate the previous estimate, because increasing C_{G} only increases the right-hand side and preserves the same parameter dependence. Hence, whenever $w(s) > 0$,

$$0 < w(s) \le \frac{C_{\mathrm{G}}}{e}.$$

Therefore

$$w(t) \le C_{\mathrm{st}} C_{\mathrm{G}} \int_0^t w(s) \left(1 + \sqrt{\log \frac{C_{\mathrm{G}}}{w(s)}}\right) ds.$$

Equivalently,

$$w(t) \le A \int_0^t \omega_B(w(s))\, ds, \qquad A := C_{\mathrm{st}} C_{\mathrm{G}},$$

where

$$\omega_B(r) := \begin{cases} r\left(1 + \sqrt{\log(B/r)}\right), & 0 < r \le B/e, \\ 2r, & r > B/e, \\ 0, & r = 0, \end{cases} \qquad B := C_{\mathrm{G}}.$$

The modulus ω_B satisfies the Osgood condition

$$\int_0^{B/e} \frac{dr}{\omega_B(r)} = \infty.$$

Bihari's generalized Bellman lemma (Bihari, 1956, Sections 3–4) therefore gives $w \equiv 0$ on $[0, T]$. Hence

$$\rho_t = g_t \qquad \text{for every } t \in [0, T].$$

□

Corollary E.21 (Convergence of hard-truncated Gaussian flows)**.** *Fix $\beta > 0$ and $T > 0$. Then*

$$\sup_{0 \le t \le T} W_1(f_t^{\beta,[R]}, f_t^{\beta}) \longrightarrow 0 \qquad \textit{as } R \to \infty.$$

Proof. We prove convergence by contradiction. Suppose that the asserted convergence fails. Then there exist $\varepsilon_0 > 0$ and a sequence $R_k \to \infty$ such that

$$\sup_{0 \le t \le T} W_1(f_t^{\beta,[R_k]}, f_t^{\beta}) \ge \varepsilon_0 \qquad \text{for every } k.$$

By Theorem E.14, after passing to a subsequence, not relabeled, there exists

$$\rho \in C([0, T]; \mathcal{P}_1(\mathbb{R}^d))$$

such that

$$\sup_{0 \le t \le T} W_1(f_t^{\beta,[R_k]}, \rho_t) \to 0.$$

Moreover, $\rho_{t=0} = f_0$, ρ is a finite-action weak solution of

$$\partial_t \rho_t + \nabla \cdot (\rho_t X_\beta[\rho_t]) = 0$$

on $[0, T]$, and

$$\sup_{0 \le t \le T} \int |x|^2 \, \rho_t(dx) < \infty.$$

Since $f_0 = g_0$, where $g_t = f_t^{\beta}$ is the explicit Gaussian solution, Theorem E.20 gives

$$\rho_t = g_t = f_t^{\beta} \qquad \text{for every } t \in [0, T].$$

Consequently,

$$\sup_{0 \le t \le T} W_1(f_t^{\beta,[R_k]}, f_t^{\beta}) \to 0,$$

contradicting the choice of R_k. Therefore

$$\sup_{0 \le t \le T} W_1(f_t^{\beta,[R]}, f_t^{\beta}) \to 0 \qquad \text{as } R \to \infty.$$

□

Theorem E.22 (Gaussian-prior noisy laws are noncompact admissible)**.** *Fix $\tau > 0$. Let $m \in \mathbb{R}^d$, let Σ_0 be a symmetric nonnegative semidefinite matrix, and set*

$$P_0 = \mathcal{N}(m, \Sigma_0), \qquad f_0 := \gamma_\tau * P_0 = \mathcal{N}(m, \Sigma_0 + \tau I).$$

Then, for every fixed $\beta > 0$ and $T > 0$, the law f_0 is (β, T)-admissible in the sense of Definition E.2. The admissible noncompact flow is the explicit Gaussian flow

$$f_t^\beta = \mathcal{N}(m, \Gamma_t), \qquad \dot{\Gamma}_t = -2\Gamma_t(I + \beta\Gamma_t)^{-1}.$$

Proof. Since

$$f_0 = \mathcal{N}(m, \Sigma_0 + \tau I)$$

and $\tau > 0$, the initial law f_0 has a smooth strictly positive Gaussian density and finite moments of all orders. In particular, its hard truncations $f_0^{[R]}$ are well-defined for every $R > 0$, and

$$W_1(f_0^{[R]}, f_0) \to 0 \qquad \text{as } R \to \infty$$

by Lemma E.7.

The explicit Gaussian curve

$$f_t^\beta = \mathcal{N}(m, \Gamma_t), \qquad \dot{\Gamma}_t = -2\Gamma_t(I + \beta\Gamma_t)^{-1},$$

is a weak solution by Lemma E.10. It has finite action on every finite time interval. Indeed,

$$X_\beta[f_t^\beta](x) = -(I + \beta\Gamma_t)^{-1}(x - m),$$

and Γ_t remains uniformly positive definite and uniformly bounded on $[0, T]$. Hence

$$\int_0^T \int |X_\beta[f_t^\beta](x)|^2 \, f_t^\beta(dx) \, dt < \infty.$$

Thus the first condition in Definition E.2 holds with admissible flow $(f_t^\beta)_{0 \le t \le T}$.

The second admissibility condition is exactly Corollary E.21, namely

$$\sup_{0 \le t \le T} W_1(f_t^{\beta,[R]}, f_t^\beta) \to 0 \qquad \text{as } R \to \infty.$$

Therefore f_0 is (β, T)-admissible. □

Corollary E.23 (Gaussian priors are recoverable admissible)**.** *Fix $\tau > 0$. Let*

$$P_0 = \mathcal{N}(m, \Sigma_0),$$

where $m \in \mathbb{R}^d$ and Σ_0 is symmetric nonnegative semidefinite. Then

$$P_0 \in \mathcal{A}_\tau.$$

In particular, the class $\mathcal{A}_\tau$ is nonempty.

Proof. Set

$$f_0 := \gamma_\tau * P_0 = \mathcal{N}(m, \Sigma_0 + \tau I).$$

By Theorem E.22, for every $\beta > 0$ and $T > 0$, the law f_0 is (β, T)-admissible. In particular, it is (β, T_β)-admissible for

$$T_\beta = \frac{\beta\tau}{2}.$$

The recovery condition

$$W_1(f_{T_\beta}^\beta, P_0) \to 0 \qquad \text{as } \beta \to \infty$$

is exactly Lemma E.11. Hence $P_0 \in \mathcal{A}_\tau$. □

E.6. Gaussian-prior posterior-mean recovery

Theorem E.24 (Hard-truncated Gaussian posterior-mean recovery)**.** *Fix $\tau > 0$. Let $m \in \mathbb{R}^d$, let Σ_0 be symmetric nonnegative semidefinite, and set*

$$P_0 = \mathcal{N}(m, \Sigma_0), \qquad f_0 := \gamma_\tau * P_0 = \mathcal{N}(m, \Sigma_0 + \tau I).$$

For $R > 0$, define

$$f_0^{[R]} := \frac{\mathbf{1}_{B_R} f_0}{f_0(B_R)}.$$

Let $\mu_t^{N,\beta,[R]}$ be the empirical measure of the exact particle system initialized i.i.d. from $f_0^{[R]}$. Set

$$T_\beta := \frac{\beta\tau}{2}.$$

Then, for every $M < \infty$,

$$\lim_{\beta\to\infty} \limsup_{R\to\infty} \limsup_{N\to\infty} \mathbb{E} \sup_{|y|\le M} \left| m_{\mu_{T_\beta}^{N,\beta,[R]}}(y) - m_{P_0}(y) \right| = 0. \tag{49}$$

Consequently, the same convergence holds in probability.

Proof. By Corollary E.23, the Gaussian prior P_0 belongs to $\mathcal{A}_\tau$. The result is therefore an immediate application of Theorem E.9. □

F. Stopping time for noise removal

The goal of this appendix is to explain the slowed denoising observed for small β in Fig. 2(b). We do this in the simplest solvable setting: an isolated Gaussian cluster evolving under the Stage-1 Gaussian-attention dynamics. This yields an explicit scalar ODE for the cluster variance, from which one can read off both the slowdown at finite β and the first-order correction to the denoising time. This calculation is only a calibration model, but it explains the qualitative regimes observed numerically and motivates the finite-β correction to the denoising time.

Throughout this section we use the effective denoising time $t := \frac{\eta\ell}{\beta}$, where ℓ is layer depth and η is the residual step size in Algorithm 1. This is the normalization for which the ideal $\beta = \infty$ limit has stopping time $\sigma^2/2$. Equivalently, if $s := \eta\ell$ denotes raw layer time, then $t = s/\beta$. All ODEs below are written in the $t-$variable.

Theorem F.1 (Variance ODE and denoising time for isotropic Gaussian prior)**.** *Let G_β be the centered Gaussian kernel on $\mathbb{R}^d$ with covariance $\beta^{-1} I_d$, and consider the kernelized evolution*

$$\partial_t f_t = -\nabla \cdot \left(f_t \nabla \log(G_\beta * f_t) \right), \qquad t \ge 0$$

with initial condition $f_0(x) = \mathcal{N}(m, (\tau_a^2 + \sigma^2) I_d)(x)$, where $m \in \mathbb{R}^d$, $\tau_a > 0$, and $\sigma > 0$.

Then for all $t \ge 0$, there is a well-defined solution, with a solution being the Gaussian $f_t(x) = \mathcal{N}(m, v(t) I_d)(x)$, where the variance parameter $v(t)$ solves

$$v'(t) = -\frac{2\beta v(t)}{\beta v(t) + 1}, \qquad v(0) = \tau_a^2 + \sigma^2.$$

For any target variance $v_* \in (0, \tau_a^2 + \sigma^2]$, the hitting time $T_a(v_*) := \inf\{t \ge 0 \mid v(t) = v_*\}$ is given exactly by

$$T_a(v_*) = \frac{\tau_a^2 + \sigma^2 - v_*}{2} + \frac{1}{2\beta} \log\left(\frac{\tau_a^2 + \sigma^2}{v_*} \right).$$

In particular, the exact time to denoise back to the clean variance τ_a^2 is

$$T_a^* = \frac{\sigma^2}{2} + \frac{1}{2\beta} \log\left(1 + \frac{\sigma^2}{\tau_a^2} \right).$$

Proof. Verification that $f_t = \gamma_{v(t)}$ solves this ODE is done in Lemma E.10 (in a different time scale) and it yields the ODE

$$v'(t) = -\frac{2v(t)}{v(t) + \beta^{-1}}, \qquad v(0) = \tau_a^2 + \sigma^2$$

for the variance.

We now solve for the hitting time. Rearranging,

$$dt = -\left(\frac{1}{2} + \frac{1}{2\beta v}\right) dv.$$

Integrating from the initial variance $v_0 = \tau_a^2 + \sigma^2$ down to a target variance v_* gives

$$T_a(v_*) = \int_{v_*}^{v_0} \left(\frac{1}{2} + \frac{1}{2\beta v}\right) dv = \frac{\tau_a^2 + \sigma^2 - v_*}{2} + \frac{1}{2\beta} \log\left(\frac{\tau_a^2 + \sigma^2}{v_*}\right).$$

Setting $v_* = \tau_a^2$ gives

$$T_a^* = \frac{(\tau_a^2 + \sigma^2) - \tau_a^2}{2} + \frac{1}{2\beta} \log\left(\frac{\tau_a^2 + \sigma^2}{\tau_a^2}\right) = \frac{\sigma^2}{2} + \frac{1}{2\beta} \log\left(1 + \frac{\sigma^2}{\tau_a^2}\right).$$

This proves the claim. □

With the observation that the point mass measure δ_z at point z is just $\mathcal{N}(z, 0)$, we immediately get the following result if the clean distribution on X is δ_z.

Corollary F.2 (Denoising time for Dirac delta)**.** *Consider the kernelized evolution*

$$\partial_t f_t + \nabla \cdot (f_t \nabla \log(G_\beta * f_t)), \qquad f_0 = \mathcal{N}(z, \sigma^2 I_d), \qquad t \geq 0.$$

Assume the same regularity assumptions hold as Theorem F.1. Then for all $t \geq 0$ the system is uniquely solved by $f_t(x) = \mathcal{N}(z, v(t) I_d)$ where $v(t)$ is the solution of

$$v'(t) = -\frac{2\beta v(t)}{\beta v(t) + 1}, \qquad v(0) = \sigma^2.$$

The time to reach a target variance v_* is

$$T(v_*) = \frac{\sigma^2 - v_*}{2} + \frac{1}{2\beta} \log\left(\frac{\sigma^2}{v_*}\right).$$

In particular, the time to denoise completely is $T(0) = +\infty$ and denoising to variance β^{-1} is

$$T(\beta^{-1}) = \frac{\sigma^2}{2} - \frac{1}{2\beta} + \frac{\log \sigma}{\beta} + \frac{\log \beta}{2\beta}.$$

Proof. This is same as Theorem F.1 where the initial distribution is $\delta_z = \mathcal{N}(z, 0)$. So take $\tau_a \downarrow 0, m = z$ in Theorem F.1. □

Now we move on to the stopping times for more general distributions. We discuss here that under some assumptions, the stopping time is $\sigma^2/2$ with a correction term of order β^{-1}.

Lemma F.3. *Let*

$$\partial_t f_t^{(\beta)} = -\nabla \cdot \left(f_t^{(\beta)} \nabla \log(G_\beta * f_t^{(\beta)})\right), \qquad f_0^{(\beta)} = \mu * \gamma_{\sigma^2},$$

on $\mathbb{R}^d$, where $\sigma > 0$ and $\beta \geq 1$. Define

$$T_0 := \frac{\sigma^2}{2}, \qquad g_t := \mu * \gamma_{\sigma^2 - 2t}, \qquad 0 \leq t \leq T_0.$$

Then g_t solves the backward heat equation

$$\partial_t g_t = -\Delta g_t, \qquad g_0 = \mu * \gamma_{\sigma^2}, \qquad g_{T_0} = \mu.$$

Fix $s > d/2 + 4$. Assume:

1. *μ is strictly positive and belongs to $H^{s+6}(\mathbb{R}^d)$;*
2. *the map $N_\beta(f) = -\nabla \cdot (f \nabla \log(G_\beta * f))$ admits the expansion*

$$\mathcal{N}_\beta(f) = -\Delta f - \frac{1}{2\beta} Q[f] + \frac{1}{\beta^2} \mathcal{E}_\beta[f],$$

with

$$Q[f] := \nabla \cdot \left(f \, \nabla \left(\frac{\Delta f}{f} \right) \right),$$

and with a uniform bound

$$\sup_{0 \le t \le T_0} \|\mathcal{E}_\beta[g_t]\|_{H^s} \le C_0$$

for all sufficiently large β;

3. *the linear backward heat propagator on the interval $[0, T_0]$ is bounded on H^s along this orbit, in the sense that for every $r \in L^1([0, T_0]; H^s)$, the solution u of*

$$\partial_t u = -\Delta u + r, \qquad u_0 = 0,$$

satisfies

$$\sup_{0 \le t \le T_0} \|u_t\|_{H^s} \le C_1 \int_0^{T_0} \|r_s\|_{H^s} ds.$$

Then there exists a function $h_t \in C([0, T_0]; H^s)$ solving

$$\partial_t h_t = -\Delta h_t - \frac{1}{2} Q[g_t], \qquad h_0 = 0,$$

such that

$$f_t^{(\beta)} = g_t + \frac{1}{\beta} h_t + r_t^{(\beta)}, \qquad 0 \le t \le T_0,$$

with remainder satisfying

$$\sup_{0 \le t \le T_0} \|r_t^{(\beta)}\|_{H^s} \le \frac{C}{\beta^2}$$

for some constant C independent of β.

In particular, at the denoising time $T_0 = \sigma^2/2$,

$$f_{T_0}^{(\beta)} = \mu + \frac{1}{\beta} h_{T_0} + O_{H^s}(\beta^{-2}),$$

and hence

$$\left\| f_{T_0}^{(\beta)} - \mu \right\|_{H^s} \le \frac{\|h_{T_0}\|_{H^s}}{\beta} + \frac{C}{\beta^2}.$$

Proof. We have $\mathcal{N}_\beta(f) := -\nabla \cdot (f \nabla \log(G_\beta * f))$. By construction, $g_t = \mu * \gamma_{\sigma^2 - 2t}$ solves the PDE for $\beta = +\infty$ with the agreement that $G_\infty * f = f$. So $f_t^{(\infty)} = g_t$.

We seek an expansion of the form

$$f_t^{(\beta)} = g_t + \beta^{-1} h_t + r_t^{(\beta)}.$$

Substituting into the equation and using the expansion of $\mathcal{N}_\beta$ gives

$$\partial_t \left(g_t + \beta^{-1} h_t + r_t^{(\beta)}\right) = -\Delta \left(g_t + \beta^{-1} h_t + r_t^{(\beta)}\right) - \frac{1}{2\beta} Q[g_t] + \frac{1}{\beta^2} \mathcal{E}_\beta[g_t] + \mathcal{R}_t^{(\beta)},$$

where $\mathcal{R}_t^\beta$ contains the nonlinear Taylor remainder coming from replacing g_t by $g_t + \beta^{-1} h_t + r_t^\beta$. Since g_t solves $\partial_t g_t = -\Delta g_t$, the $O(1)$ terms cancel. Matching the $O(\beta^{-1})$ terms leads to

$$\partial_t h_t = -\Delta h_t - \frac{1}{2} Q[g_t], \qquad h_0 = 0.$$

This determines h_t uniquely.

Now define the remainder by

$$r_t^\beta := f_t^\beta - g_t - \beta^{-1} h_t.$$

Subtracting the equations for f_t^β, g_t, and h_t yields

$$\partial_t r_t^{(\beta)} = -\Delta r_t^\beta + \frac{1}{\beta^2} \mathcal{E}_\beta[g_t] + \mathcal{R}_t^\beta, \qquad r_0^\beta = 0.$$

By assumption, $\|\mathcal{E}_\beta[g_t]\|_{H^s} \le C_0$ uniformly on $[0, T_0]$. Moreover, because $s > d/2 + 4$, Sobolev multiplication and composition estimates imply that the nonlinear Taylor remainder satisfies

$$\|\mathcal{R}_t^\beta\|_{H^s} \le C_2 \left(\beta^{-2} + \|r_t^\beta\|_{H^s}^2 + \beta^{-1} \|r_t^\beta\|_{H^s}\right)$$

for all $t \in [0, T_0]$, provided β is large enough.

Applying the assumed backward heat estimate to the remainder equation gives

$$\sup_{0 \le t \le T_0} \|r_t^\beta\|_{H^s} \le C_1 \int_0^{T_0} \left(\frac{C_0}{\beta^2} + C_2 \left(\beta^{-2} + \|r_s^\beta\|_{H^s}^2 + \beta^{-1} \|r_s^\beta\|_{H^s}\right)\right) ds.$$

A standard bootstrap argument now shows that

$$\sup_{0 \le t \le T_0} \|r_t^\beta\|_{H^s} \le \frac{C}{\beta^2}$$

for all sufficiently large β.

Evaluating at $t = T_0$ and using $g_{T_0} = \mu$ gives

$$f_{T_0}^{(\beta)} = \mu + \frac{1}{\beta} h_{T_0} + r_{T_0}^{(\beta)},$$

hence

$$\|f_{T_0}^\beta - \mu\|_{H^s} \le \frac{\|h_{T_0}\|_{H^s}}{\beta} + \frac{C}{\beta^2}.$$

□

The above perturbative expansion of $f_{T_0}^{(\beta)}$ immediately gives a first order correction for the stopping time. The next lemma should be read as a conditional perturbative criterion near the backward-heat horizon $T_0 = \sigma^2/2$; in particular, it assumes the existence of a sufficiently regular backward-heat continuation in the chosen function space instead of an H^s-like criterion as in the above lemma.

Lemma F.4 (Conditional perturbative criterion)**.** *Let X be a Banach space, and let $T_0 := \frac{\sigma^2}{2}$. Assume that for some $\delta > 0$ and all sufficiently large β, the solution f_t^β admits an expansion*

$$f_t^{(\beta)} = g_t + \beta^{-1} h_t + r_t^{(\beta)}, \qquad T_0 - \delta \le t \le T_0 + \delta,$$

where $g_t = \mu * \gamma_{\sigma^2-2t}, g_{T_0} = \mu$, *and*

$$\sup_{|t-T_0|\le\delta} \|h_t\|_X \le C_h, \qquad \sup_{|t-T_0|\le\delta} \|r_t^{(\beta)}\|_X \le C_r\beta^{-2}.$$

For $t > T_0$, *the notation* g_t *is understood as a chosen* $X-$*valued backward-heat continuation of the curve* $t \mapsto \mu*\gamma_{\sigma^2-2t}$, *not as a literal Gaussian convolution. Assume also that* $t \mapsto g_t$ *is* $\mathcal{C}^1$ *as an* X*-valued map on* $[T_0-\delta, T_0+\delta]$, *and that*

$$\partial_t g_t\big|_{t=T_0} = -\Delta\mu \neq 0 \quad \textit{in } X.$$

Define

$$T^*_{(\beta)} = \inf\{\arg\min_{|t-T_0|\le\delta} \|f_t^{(\beta)} - \mu\|_X\}.$$

Then $\exists C > 0$, *independent of* β, *such that* $|T^*_{(\beta)} - T_0| \le \frac{C}{\beta}$ *for sufficiently large* β. *In particular,* $T^*_{(\beta)} = \frac{\sigma^2}{2} + O(\beta^{-1})$.

Proof. Since $g_{T_0} = \mu$, the expansion at $t = T_0$ gives

$$f^\beta_{T_0} - \mu = \beta^{-1}h_{T_0} + r^\beta_{T_0}.$$

Hence

$$\|f^\beta_{T_0} - \mu\|_X \le \beta^{-1}\|h_{T_0}\|_X + \|r^\beta_{T_0}\|_X \le \frac{C_h}{\beta} + \frac{C_r}{\beta^2}.$$

Therefore there exists $C_0 > 0$ such that

$$\|f^\beta_{T_0} - \mu\|_X \le \frac{C_0}{\beta}$$

for all sufficiently large β.

Since g_t is $\mathcal{C}^1$ in X and

$$g_{T_0} = \mu, \qquad \partial_t g_t\big|_{t=T_0} = -\Delta\mu \neq 0,$$

the Fréchet differentiability of $t \mapsto g_t$ at T_0 implies

$$g_t - \mu = (t - T_0)\,\partial_t g_t\big|_{t=T_0} + \omega(t),$$

where $\lim_{t\to T_0} \frac{\|\omega(t)\|_X}{|t-T_0|} = 0$. Hence there exists $0 < \delta_1 \le \delta$ such that for all $|t - T_0| \le \delta_1$,

$$\|\omega(t)\|_X \le \frac{1}{2}\|\Delta\mu\|_X\,|t-T_0|.$$

Therefore, using $\partial_t g_t|_{t=T_0} = -\Delta\mu$,

$$\|g_t - \mu\|_X \ge |t-T_0|\|\Delta\mu\|_X - \|\omega(t)\|_X \ge \frac{1}{2}\|\Delta\mu\|_X\,|t-T_0|.$$

Set

$$c := \frac{1}{2}\|\Delta\mu\|_X > 0.$$

Then

$$\|g_t - \mu\|_X \ge c|t-T_0| \qquad \text{for all } |t-T_0| \le \delta_1.$$

For $|t - T_0| \le \delta_1$, the expansion gives

$$f^\beta_t - g_t = \beta^{-1}h_t + r^\beta_t \implies \|f^\beta_t - g_t\|_X \le \frac{C_h}{\beta} + \frac{C_r}{\beta^2} \le \frac{C_1}{\beta}$$

for some constant $C_1 > 0$ and all sufficiently large β.

Hence, for $|t - T_0| \le \delta_1$,

$$\|f_t^\beta - \mu\|_X \ge \|g_t - \mu\|_X - \|f_t^\beta - g_t\|_X \ge c|t - T_0| - \frac{C_1}{\beta}.$$

Now let

$$M := \frac{C_0 + C_1 + 1}{c}.$$

If $|t - T_0| \ge M/\beta$ and also $|t - T_0| \le \delta_1$, then

$$\|f_t^\beta - \mu\|_X \ge c\frac{M}{\beta} - \frac{C_1}{\beta} = \frac{C_0 + 1}{\beta} > \frac{C_0}{\beta} \ge \|f_{T_0}^\beta - \mu\|_X.$$

Therefore no minimizer of $t \mapsto \|f_t^\beta - \mu\|_X$ over $|t - T_0| \le \delta_1$ can lie outside the interval

$$|t - T_0| < \frac{M}{\beta}.$$

We have $M/\beta < \delta_1$ for sufficiently large β, so $|T^*_{(\beta)} - T_0| \le \frac{M}{\beta}$. □