ROBUSTNESS OF DIFFUSION MODELS UNDER DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Score-based diffusion models are increasingly considered in settings where the underlying data distribution may differ from the training distribution, yet existing theoretical guarantees largely focus on the no-shift setting. In this work, we study robust score estimation under Wasserstein perturbations of a reference distribution. For the Ornstein–Uhlenbeck diffusion, we show that robust estimation decomposes into two fundamental components: the statistical cost of learning the reference distribution and the intrinsic cost of distribution shift. The latter scales quadratically with the Wasserstein radius, and this dependence is minimax optimal. We construct an explicit finite-sample estimator achieving the resulting robust minimax rate without knowing the shift radius. When the reference distribution lies on an unknown low-dimensional subspace, the statistical term adapts to the intrinsic dimension while the shift cost remains unchanged. Finally, we show that the same decomposition governs positive-time reverse sampling and obtain matching minimax guarantees in KL divergence. Together, these results characterize how finite data, intrinsic dimension, and distribution shift affect the robustness of score-based diffusion models.

1 INTRODUCTION

Diffusion models have become central to modern generative modeling (Dhariwal & Nichol, 2021; Ho et al., 2020; Song et al., 2021a; Xu et al., 2022). A common formulation is the score-based diffusion model (SDM), which constructs a generative process by reversing a stochastic differential equation (SDE) (Song et al., 2021b). At each diffusion time, the reverse dynamics rely on the score function of the corresponding noisy data distribution. Since the score is unknown, it must be estimated from data, typically using a neural network in practical implementations.

A substantial body of theory has studied the statistical properties of learned scores and their impact on the reverse generative process[1]. Existing results provide guarantees for score estimation and sampling under a fixed data distribution, where the estimator is evaluated on the same distribution used for training (see, e.g., Gao et al. (2025); Chen et al. (2023b); Li et al. (2024); Yu & Yu (2026a)). However, pretrained diffusion models are increasingly reused or adapted across changing data sources and new target domains (Rombach et al., 2022; Ouyang et al., 2024; Gao et al., 2023). Such distribution shifts arise from domain changes, temporal drift, sampling bias, and variations in data collection (Moreno-Torres et al., 2012; Koh et al., 2021; Ben-David et al., 2010; Torralba & Efros, 2011; Cheng et al., 2026). This motivates a fundamental question: how robust are diffusion models when the deployment distribution differs from the reference distribution?

We study robustness through the score function, which governs the reverse diffusion dynamics and thus provides a natural entry point for analyzing generative robustness. Unlike standard score estimation, distribution shift introduces a mismatch between the distribution used for learning and the distribution encountered after deployment. The resulting robust error contains two components: the statistical error from learning the reference distribution and the additional error caused by the unknown shift. Our goal is to characterize these contributions and their optimal scaling.

To quantify distribution shifts, we model the allowed perturbations through Wasserstein neighborhoods of the reference distribution. Wasserstein distance has been widely used in diffusion sampling

[1] Due to space constraints, we defer a detailed discussion of additional related work to Appendix A.

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

and stability analyses (see, e.g., Gao et al. (2025); Yu & Yu (2025); Beyler & Bach (2025)). At the same time, Wasserstein proximity does not directly imply proximity of score functions, which depend on local variations of the underlying density. Forward diffusion regularizes this discrepancy through Gaussian smoothing, but the regularization becomes weaker at early diffusion times. Quantifying how a Wasserstein perturbation of the initial distribution propagates to the score along the diffusion trajectory is therefore a key step toward a statistical theory of diffusion robustness.

Specifically, let $\mathcal{Q}_\varepsilon(p_0)$ denote the Wasserstein-2 neighborhood of radius ε around the reference distribution p_0, where each $q_0 \in \mathcal{Q}_\varepsilon(p_0)$ represents a possible shifted distribution. Applying the same forward diffusion to p_0 and q_0 gives the corresponding noisy marginals p_t and q_t. We study score estimation over diffusion times $t \in [t_0, T]$, where $t_0 > 0$ specifies the earliest diffusion time considered and avoids the singular zero-noise regime. A score estimator $\hat{s}(x,t)$ is learned from samples of p_0, but under distribution shift it is evaluated on q_t and compared with the shifted score $\nabla \log q_t$. We quantify robustness through the worst-case time-averaged score risk

$$\mathcal{R}_{\varepsilon,p_0}(\hat{s}) := \sup_{q_0 \in \mathcal{Q}_\varepsilon(p_0)} \mathbb{E}\left[\frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{X\sim q_t}\left[\|\hat{s}(X,t) - \nabla \log q_t(X)\|_2^2\right] \mathrm{d}t\right],$$

where the outer expectation is over the training data and any randomness in the estimator.

When $\hat{s}$ is learned from finite samples of p_0, this robust risk captures two sources of uncertainty: the statistical error from learning the reference distribution and the shift error arising from the unknown deployment distribution q_0, which changes both the evaluation distribution and the target score. Understanding their interaction, the intrinsic cost of distribution shift, and whether such effects extend to generation remains beyond existing score-estimation theory.

These considerations lead to the central question of this work:

What are the statistical and intrinsic costs of distribution shift in diffusion models, and how do they propagate from score estimation to generation?

To our knowledge, this is the first work to develop a minimax theory for robust SDMs under distribution shift. We focus on the Ornstein–Uhlenbeck (OU) process as a canonical setting in which these costs can be characterized sharply. We develop a minimax theory that reveals a fundamental decomposition of the robust error into a statistical estimation cost and an intrinsic distribution-shift cost. The latter is quadratic in the Wasserstein radius and minimax optimal, while the former adapts to low-dimensional structure. We further show that this decomposition extends to positive-time reverse sampling with matching KL guarantees. Our main contributions are summarized as follows.

- **Intrinsic cost of distribution shift.** We establish a comparison principle for score estimators induced by estimated initial laws under the OU semigroup. The resulting bound separates reference distribution estimation error from an intrinsic shift cost that scales quadratically with the Wasserstein radius. A matching minimax lower bound shows that this dependence is unavoidable.
- **Finite-sample robust score estimation.** We construct an explicit estimator from n i.i.d. samples from the reference distribution and prove, under compact support and for fixed dimension,
$$\mathcal{R}_{\varepsilon,p_0}(\hat{s}_n) \lesssim \frac{1}{n(T-t_0)t_0^{d/2}} + \frac{\varepsilon^2}{t_0(T-t_0)}.$$
The estimator does not require knowledge of the shift radius and achieves the optimal robust rate.
- **Adaptation to intrinsic dimension.** We extend the theory to reference distributions supported on an unknown k-dimensional linear subspace of an ambient d-dimensional space. In this setting, we show that the statistical term depends on k rather than the ambient dimension d, while the shift cost remains unchanged. Matching lower bounds establish the minimax optimality of this intrinsic-dimensional rate.
- **Robust generation.** Finally, we extend the score estimation guarantees to positive-time reverse sampling. The resulting KL error admits the same decomposition into a finite-sample estimation term and a quadratic distribution shift term, with matching minimax

lower bounds. Thus, the statistical and intrinsic costs at the score level persist in the generated distribution.

Together, these results provide a unified characterization of diffusion robustness under Wasserstein shifts: more data and structural assumptions reduce the statistical difficulty of learning the reference distribution, but cannot remove the intrinsic uncertainty caused by the distribution shift. A controlled numerical study in Appendix H illustrates the predicted finite-sample decay and quadratic shift behavior for both the time-averaged score risk and the generation error in KL divergence at time t_0. All proofs are deferred to the appendix.

Notation. Let $\mathbb{R}^d$ be the d-dimensional Euclidean space and I_d the identity matrix. For a measurable $f : A \to \mathbb{R}^k$, write $\|f\|^2_{L^2(A)} := \int_A \|f(x)\|_2^2 \, \mathrm{d}x$ and $\|f\|_2 := \|f\|_{L^2(\mathbb{R}^d)}$; for a measure μ, $\|f\|^2_{L^2(\mu)} := \int \|f(x)\|_2^2 \, \mu(\mathrm{d}x)$. Let $\mathcal{P}_q(\mathbb{R}^d)$ denote the set of probability measures with finite q-th moment: $\mathcal{P}_q(\mathbb{R}^d) := \{\mu : \int \|x\|_2^q \, \mu(\mathrm{d}x) < \infty\}$. For $\mu, \nu \in \mathcal{P}_q(\mathbb{R}^d)$, define the Wasserstein distance $\mathsf{W}_q(\mu, \nu) := \big(\inf_{\varrho \in \Gamma(\mu,\nu)} \int \|x - y\|_2^q \, \varrho(\mathrm{d}x, \mathrm{d}y)\big)^{1/q}$, where $\Gamma(\mu, \nu)$ denotes the set of couplings of μ and ν; define $\mathrm{KL}(\mu\|\nu) := \int \log(\mathrm{d}\mu/\mathrm{d}\nu) \, \mathrm{d}\mu$ if $\mu \ll \nu$, and ∞ otherwise. For $x \in \mathbb{R}^d$ and $r > 0$, let $\mathcal{B}(x, r)$ denote the closed Euclidean ball centered at x, and let δ_x denote the Dirac measure at x.

2 Setting and Problem Formulation

We introduce the framework for studying score estimation under distribution shift, where p_0 is the reference distribution used to learn the score, and q_0 is a possible shifted distribution at deployment.

2.1 Diffusion Dynamics and Score Approximation

We focus on the Ornstein-Uhlenbeck (OU) forward process,

$$\mathrm{d}X_t = -\frac{1}{2} X_t \, \mathrm{d}t + \mathrm{d}B_t, \qquad X_0 \sim p_0,$$

where B_t is a standard d-dimensional Brownian motion. Its transition kernel is Gaussian:

$$X_t \mid X_0 = x_0 \sim \mathcal{N}(a_t x_0, \sigma_t^2 I_d), \qquad a_t = e^{-t/2}, \qquad \sigma_t^2 = 1 - e^{-t}.$$

Thus, if P_t denotes the OU semigroup, then $p_t := P_t \, p_0$ is the law of

$$a_t X_0 + \sigma_t Z, \qquad X_0 \sim p_0, \qquad Z \sim \mathcal{N}(0, I_d),$$

where X_0 and Z are independent. More generally, for any initial law μ, $P_t\mu$ denotes its OU-smoothed marginal at time t. In SDMs, the forward process is reversed from a large terminal time T, yielding

$$\mathrm{d}X_t^{\leftarrow} = \frac{1}{2}(X_t^{\leftarrow} + 2\nabla \log p_{T-t}(X_t^{\leftarrow}))\mathrm{d}t + \mathrm{d}W_t, \qquad X_0^{\leftarrow} \sim p_T, \tag{2.1}$$

where W_t is another standard Brownian motion. The term $\nabla \log p_t$ is the score function of p_t. Under mild conditions, the reverse process initialized at p_T coincides in distribution with the time-reversed forward process, and hence generates samples from p_0 (Anderson, 1982; Cattiaux et al., 2023). Since the score $\nabla \log p_t$ is unknown, it must be estimated from data.

We study score estimation over $t \in [t_0, T]$ with $t_0 > 0$. For later use, define

$$\rho_t := \sigma_t^2 = 1 - e^{-t}, \qquad \tau_0 := \frac{\rho_{t_0}}{a_{t_0}^2} = e^{t_0} - 1, \qquad \tau_T := \frac{\rho_T}{a_T^2} = e^T - 1.$$

2.2 Wasserstein Distribution Shifts

We model deployment shifts through Wasserstein neighborhoods of the reference distribution. Let $p_0 \in \mathcal{P}_2(\mathbb{R}^d)$ be the reference distribution. For $\varepsilon \geq 0$, define the perturbation class

$$\mathcal{Q}_\varepsilon(p_0) := \big\{q_0 \in \mathcal{P}_2(\mathbb{R}^d) : \mathsf{W}_2(q_0, p_0) \leq \varepsilon\big\}.$$

Here, ε quantifies the magnitude of the distribution shift. Neither p_0 nor q_0 is required to admit a density. For every $t > 0$, Gaussian OU smoothing maps any initial law in $\mathcal{P}_2(\mathbb{R}^d)$ to a strictly

positive smooth density. Hence, for $q_0 \in \mathcal{Q}_\varepsilon(p_0)$, the marginals $p_t = P_t p_0$ and $q_t = P_t q_0$ admit well-defined scores $\nabla \log p_t$ and $\nabla \log q_t$.

Under distribution shift, an estimator learned from p_0 is evaluated on the shifted marginal q_t and compared with the shifted score $\nabla \log q_t$, rather than the reference score $\nabla \log p_t$.

2.3 Shifted Score Risk

For a shifted distribution $q_0 \in \mathcal{Q}_\varepsilon(p_0)$, let $q_t = P_t q_0$. For a measurable score function $s : \mathbb{R}^d \times [t_0, T] \to \mathbb{R}^d$, define the shifted score loss

$$\mathcal{L}_{q_0}(s) := \frac{1}{T - t_0} \int_{t_0}^{T} \mathbb{E}_{X \sim q_t} \left[\|s(X,t) - \nabla \log q_t(X)\|_2^2 \right] \mathrm{d}t.$$

For a measurable, possibly random score estimator $\hat{s}$, define the robust shifted score risk

$$\mathcal{R}_{\varepsilon, p_0}(\hat{s}) := \sup_{q_0 \in \mathcal{Q}_\varepsilon(p_0)} \mathbb{E}[\mathcal{L}_{q_0}(\hat{s})], \tag{2.2}$$

where the expectation is taken over the reference training samples and any internal randomness used to construct $\hat{s}$. This risk measures worst-case score estimation error over all deployment distributions within Wasserstein distance ε of p_0. When $\hat{s}$ is learned from finitely many reference samples, it incorporates both the statistical difficulty of estimating the reference distribution and the additional uncertainty induced by distribution shift. Our goal is to characterize these two effects and their dependence on ε, t_0, and the sample size.

3 The Intrinsic Cost of Distribution Shift

We begin by isolating the effect of distribution shift from the statistical error of learning the reference distribution. The key observation is that, if an estimated initial law is propagated through the same OU semigroup as the data distribution, then the entire estimated score trajectory is generated by a single underlying density path. This allows the shifted score risk to be controlled through a divergence between the shifted and estimated marginals at one positive diffusion time.

3.1 A Robust Comparison Principle

Let $\hat{\mu}_0$ be a possibly random estimate of p_0 with finite second moment, and define

$$\hat{p}_t := P_t \hat{\mu}_0, \qquad \hat{s}(x,t) := \nabla \log \hat{p}_t(x).$$

The same estimated initial law therefore determines the entire score trajectory, rather than estimating the score independently at each diffusion time. For strictly positive densities p and r, define the order-two Rényi divergence

$$D_2(p\|r) := \log \int_{\mathbb{R}^d} \frac{p(x)^2}{r(x)} \, \mathrm{d}x.$$

The following proposition gives the basic comparison principle underlying our robustness analysis.

Proposition 3.1. *Suppose $p_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and $D_2(p_{t_0}\|\hat{p}_{t_0}) < \infty$ almost surely. Then, for every $q_0 \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathrm{KL}\left(q_{t_0}\|\hat{p}_{t_0}\right) \le \frac{a_{t_0}^2}{\rho_{t_0}} \mathsf{W}_2^2(q_0, p_0) + D_2\left(p_{t_0}\|\hat{p}_{t_0}\right) \tag{3.1}$$

almost surely. Consequently, for every $\varepsilon \ge 0$,

$$\mathcal{R}_{\varepsilon, p_0}(\hat{s}) \le \frac{2e^{-t_0}}{(T - t_0)(1 - e^{-t_0})} \varepsilon^2 + \frac{2}{T - t_0} \mathbb{E}[D_2\left(p_{t_0}\|\hat{p}_{t_0}\right)]. \tag{3.2}$$

Proposition 3.1 separates two sources of error in robust score estimation. The first term depends only on the discrepancy between the shifted and reference initial distributions and is quadratic in $\mathsf{W}_2(q_0, p_0)$, while the second measures the error in estimating the reference marginal p_{t_0}. The bound therefore controls the effect of distribution shift through an additive term of order ε^2. This reduces the finite-sample problem to controlling $D_2(p_{t_0}\|\hat{p}_{t_0})$, which we address in Section 4. Before turning to that construction, we show that the quadratic dependence on ε is unavoidable.

3.2 Minimax Optimality of the Quadratic Shift Cost

Proposition 3.1 gives a shift contribution of order $\varepsilon^2/[(T-t_0)(e^{t_0}-1)]$. We now show that this dependence is unavoidable: when $\tau_0 \asymp t_0$, every estimator incurs error of order $\varepsilon^2/[t_0(T-t_0)]$ for some distribution in the Wasserstein neighborhood.

The upper bound in Proposition 3.1 holds for arbitrary initial laws in $\mathcal{P}_2(\mathbb{R}^d)$. For the lower bound, it suffices to consider a compactly supported reference distribution. Let

$$p_0^{(1)}(u) := \frac{1}{2}\mathbf{1}_{[-1,1]}(u), \qquad p_0^{(d)}(x) := \prod_{j=1}^{d} p_0^{(1)}(x_j).$$

The following theorem gives a matching lower bound for the distribution shift term.

Theorem 3.2. *Fix $d \geq 1$. Let $P_t^{(d)}$ be the semigroup of the d-dimensional Ornstein–Uhlenbeck process. Suppose that $0 < t_0 \leq 1/(16\pi^2), 2t_0 \leq T$, and $0 < \varepsilon \leq \sqrt{t_0}$. Then,*

$$\inf_{\hat{s}} \mathcal{R}_{\varepsilon, p_0^{(d)}}(\hat{s}) \geq \frac{e^{-4/9}}{2592} \frac{\varepsilon^2}{t_0(T-t_0)}, \tag{3.3}$$

where the infimum is over all possibly randomized score estimators with measurable realizations.

This theorem shows that the quadratic dependence on the shift radius is intrinsic: even with unrestricted choice of estimator, uncertainty over $\mathcal{Q}_\varepsilon(p_0^{(d)})$ incurs a cost of order $\varepsilon^2/[t_0(T-t_0)]$. Since $\tau_0 \asymp t_0$ in the regime above, this matches the shift term in Proposition 3.1 up to universal constants. Thus, both the quadratic dependence on ε and its early-time scaling are optimal.

4 Finite-Sample Robust Score Estimation

Section 3 establishes the intrinsic quadratic cost of distribution shift. We now turn to the finite-sample setting, where p_0 is observed through n i.i.d. samples. Proposition 3.1 reduces robust score estimation to controlling the positive-time estimation error $D_2(p_{t_0}\|\hat{p}_{t_0})$. We construct an explicit estimator for which this term can be controlled sharply and establish the resulting minimax rate.

4.1 The Gaussian-Blanket Estimator

For the finite-sample analysis, we impose the following compact support assumption.

Assumption 1. *For some $D > 0$, $\mathrm{supp}(p_0) \subseteq \mathcal{B}(0, D)$.*

Assumption 1 is needed only for the finite-sample construction below. The oracle inequality in Proposition 3.1, and hence the quadratic dependence on the distribution shift established in Section 3, continues to hold over the larger class $\mathcal{P}_2(\mathbb{R}^d)$.

A direct empirical approximation of p_0 becomes strictly positive after OU smoothing, but positivity alone is insufficient for controlling the Rényi divergence required in our analysis, since the estimated density appears in the denominator and can be arbitrarily small in poorly sampled regions. We therefore introduce Gaussian blanket regularization, which provides a deterministic Gaussian lower envelope while preserving OU compatibility.

For $v > 0$, let $\varphi_v(x) := (2\pi v)^{-d/2}\exp(-\|x\|_2^2/(2v))$ denote the density of $\mathcal{N}(0, vI_d)$. Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_0$, fix $\vartheta > 0$ and define $\nu_0^\vartheta := \mathcal{N}(0, \vartheta^2 I_d)$. Let $v_t := \rho_t + a_t^2\vartheta^2$, and define the calibration factor

$$M_{\vartheta,t_0} := \exp\left(\frac{D^2}{2\vartheta^2}\right)\left(1 + \frac{\vartheta^2}{\tau_0}\right)^{d/2}. \tag{4.1}$$

The Gaussian-blanket estimator regularizes the empirical measure as

$$\hat{\mu}_{n,0}^\vartheta := \frac{1}{n + M_{\vartheta,t_0}} \sum_{i=1}^{n} \delta_{X_i} + \frac{M_{\vartheta,t_0}}{n + M_{\vartheta,t_0}} \nu_0^\vartheta .$$

Propagating this measure through the same OU semigroup gives

$$\hat{p}^{\vartheta}_{n,t}(x) := P_t \hat{\mu}^{\vartheta}_{n,0}(x) = \frac{\sum_{i=1}^{n} \varphi_{\rho_t}(x - a_t X_i) + M_{\vartheta,t_0} \varphi_{v_t}(x)}{n + M_{\vartheta,t_0}}. \tag{4.2}$$

The corresponding score estimator is

$$\hat{s}^{\vartheta}_{n}(x,t) := \nabla \log \hat{p}^{\vartheta}_{n,t}(x). \tag{4.3}$$

The choice of M_{ϑ,t_0} ensures that the Gaussian component provides uniform denominator control over all sample realizations allowed by Assumption 1. This is the key ingredient for bounding the Rényi-divergence term $D_2(p_{t_0} \| \hat{p}^{\vartheta}_{n,t_0})$.

4.2 Finite-Sample Upper Bound

We first quantify the approximation error of the Gaussian-blanket density.

Proposition 4.1. *Under Assumption 1, for every $t \in [t_0, T]$,*

$$\mathbb{E}\big[D_2\left(p_t \| \hat{p}^{\vartheta}_{n,t}\right)\big] \le \log\left(1 + \frac{M_{\vartheta,t_0} - 1}{n+1}\right). \tag{4.4}$$

Proposition 4.1 provides the finite-sample control of the estimation term in Proposition 3.1. Combining the two propositions yields the following robust score estimation guarantee.

Theorem 4.2. *Fix $d \ge 1$ and $D > 0$, and suppose Assumption 1 holds. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_0$. Set $\vartheta = D/\sqrt{d}$ and construct $\hat{s}^{\vartheta}_n$ according to (4.2)-(4.3). If $\tau_0 \le 1$, then, for every $\varepsilon \ge 0$,*

$$\mathcal{R}_{\varepsilon,p_0}(\hat{s}^{\vartheta}_n) \le \frac{C_{d,D}}{(n+1)(T-t_0)\tau_0^{d/2}} + \frac{2\varepsilon^2}{(T-t_0)\tau_0}, \tag{4.5}$$

where $C_{d,D} = 2e^{d/2}(1 + D^2/d)^{d/2}$.

For fixed d and D, since $\tau_0 \asymp t_0$ for bounded t_0, the two terms in Theorem 4.2 scale as $1/[n(T-t_0)t_0^{d/2}]$ and $\varepsilon^2/[t_0(T-t_0)]$, respectively. The first term corresponds to the finite-sample estimation error, while the second term captures the intrinsic cost of distribution shift. Notably, the estimator does not depend on the shift radius ε, and the coefficient of the shift term is independent of the dimension, sample size, and support radius.

Neural realizability. The Gaussian-blanket score also admits a constructive ReLU-based approximation. Under Assumption 1, Appendix G constructs a sample-dependent approximation that is uniform over $\mathbb{R}^d \times [t_0, T]$, with explicit width and depth bounds (see Proposition G.1). Moreover, Subsection G.5 shows that the resulting neural score inherits the robust finite-sample guarantee of Theorem 4.2, up to the prescribed neural approximation accuracy. In particular, choosing the approximation error sufficiently small preserves the robust minimax rate.

4.3 Minimax Optimality

We now show that the upper bound in Theorem 4.2 is minimax optimal. Define

$$\mathcal{P}_D := \{p_0 \in \mathcal{P}(\mathbb{R}^d) : \mathrm{supp}(p_0) \subseteq \mathcal{B}(0, D)\},$$

and the minimax robust risk

$$\mathfrak{R}^{\star}_{n,\varepsilon}(D; t_0, T) := \inf_{\hat{s}_n} \sup_{p_0 \in \mathcal{P}_D} \mathcal{R}_{\varepsilon,p_0}(\hat{s}_n), \tag{4.6}$$

where the infimum is over all measurable, possibly randomized score estimators based on n i.i.d. samples from p_0. When $\varepsilon = 0$, this reduces to the nominal time-averaged score estimation risk.

Theorem 3.2 already identifies the unavoidable contribution of distribution shift. It remains to characterize the statistical difficulty of estimating the reference score trajectory from finite data. Because the risk is integrated over diffusion time, a fixed-time lower bound does not immediately suffice: the hardest instance may depend on the diffusion time. The next theorem constructs a single family whose score separation persists over a nontrivial time interval.

Theorem 4.3. *Fix $d \geq 1$ and $D > 0$. Let $c_0, c_1 > 0$ be the constants given in Appendix C, which depend only on d and D. If $0 < t_0 \leq c_0$, and $2t_0 \leq T$, then, for every $n \geq 1$ and $\varepsilon \geq 0$,*

$$\mathfrak{R}^{\star}_{n,\varepsilon}(D; t_0, T) \geq \frac{c_1}{T - t_0}\left(1 \wedge \frac{1}{n\tau_0^{d/2}}\right). \tag{4.7}$$

Combining this result with the shift lower bound in Theorem 3.2 gives the full robust minimax rate.

Theorem 4.4. *Fix $d \geq 1$ and $D \geq \sqrt{d}$, and let c_0, c_1 be the constants in Theorem 4.3. If $0 < t_0 \leq \min\{c_0, 1/(16\pi^2)\}$, $2t_0 \leq T$, $n\tau_0^{d/2} \geq 1$, and $0 \leq \varepsilon \leq \sqrt{t_0}$, then*

$$c_{d,D}\left[\frac{1}{n(T-t_0)\tau_0^{d/2}} + \frac{\varepsilon^2}{t_0(T-t_0)}\right] \leq \mathfrak{R}^{\star}_{n,\varepsilon}(D; t_0, T) \leq C_{d,D}\left[\frac{1}{n(T-t_0)\tau_0^{d/2}} + \frac{\varepsilon^2}{t_0(T-t_0)}\right]. \tag{4.8}$$

Here, $C_{d,D} = 2e^{d/2}(1 + D^2/d)^{d/2}$, and $c_{d,D} := \frac{1}{2}\min\{c_1, e^{-4/9}/2592\}$.

This theorem shows that the upper bound in Theorem 4.2 is sharp, up to constants, in both the finite-sample and distribution-shift terms. Thus, the Gaussian-blanket estimator attains the robust minimax rate in the stated regime.

Relation to fixed-time score estimation. The statistical term agrees with existing fixed-time score-estimation rates. For fixed d and D, prior work gives, up to logarithmic factors, the fixed-time rate $n^{-1}\tau^{-d/2-1}$ (Zhang et al., 2024; Dou et al., 2024; Cai & Li, 2025). Reparameterizing OU time by $\tau = e^t - 1$, together with the corresponding score rescaling, integration over $[t_0, T]$ yields $1/[n(T-t_0)\tau_0^{d/2}]$, which matches the statistical term in Theorem 4.2.

Adaptation to the shift radius. The Gaussian-blanket estimator does not depend on ε. Under the conditions of Theorem 4.4, the same estimator is minimax optimal simultaneously for all $0 \leq \varepsilon \leq \sqrt{t_0}$, without radius-dependent tuning.

5 Adaptation to Intrinsic Dimension

The finite-sample rate in Section 4 depends on the ambient dimension d. We show that this dependence can be replaced by the intrinsic dimension when p_0 is supported on an unknown low-dimensional linear subspace. The shifted distribution is allowed to leave the source subspace, and the resulting estimator adapts to the intrinsic dimension without requiring exact subspace recovery.

We impose the following additional structure.

Assumption 2. *Let $k \in \{1, \ldots, d\}$ be known. There exists an unknown matrix $U \in \mathbb{R}^{d\times k}$ such that*

$$U^{\top}U = I_k, \qquad \mathrm{supp}(p_0) \subseteq \mathrm{range}(U).$$

Let $X \sim p_0$ and write $X = UZ$ with $Z \in \mathbb{R}^k$. We do not impose smoothness or non-degeneracy assumptions on the latent distribution. The assumption applies only to p_0. The perturbation class $\mathcal{Q}_\varepsilon(p_0)$ remains the full Wasserstein neighborhood in $\mathbb{R}^d$.

To estimate the unknown source subspace, we split the n i.i.d. samples from p_0 into two independent subsets of sizes $n_1 = \lceil n/2 \rceil$ and $n_2 = n - n_1$, with $n \geq 2$. Using the first subset, we construct the empirical subspace and its orthogonal projector

$$\hat{L} := \mathrm{span}(X_1, \ldots, X_{n_1}), \qquad \hat{\Pi} := \mathrm{Proj}_{\hat{L}},$$

and project the remaining samples as

$$Y_i := \hat{\Pi}X_{n_1+i}, \qquad 1 \leq i \leq n_2.$$

We then apply the Gaussian-blanket construction in the estimated subspace with intrinsic scale $\vartheta_k^2 = D^2/k$ and the calibration $M_{\vartheta_k,t_0} := e^{k/2}(1 + D^2/(k\tau_0))^{k/2}$. Specifically, define

$$\hat{\mu}^{\mathrm{sub}}_{n,0} := \frac{\sum_{i=1}^{n_2}\delta_{Y_i} + M_{\vartheta_k,t_0}\mathcal{N}(0, \vartheta_k^2\hat{\Pi})}{n_2 + M_{\vartheta_k,t_0}}, \qquad \hat{s}^{\mathrm{sub}}_n(x,t) := \nabla \log P_t\hat{\mu}^{\mathrm{sub}}_{n,0}(x).$$

The Gaussian component provides the required regularization, while the projection onto the estimated subspace enables adaptation from the ambient dimension d to the intrinsic dimension k.

The estimator does not require exact recovery of the source subspace. Instead, it suffices to control the source mass missed by the empirical span, leading to the following guarantee.

Theorem 5.1. *Under Assumptions 1 and 2, for every $n \geq 2$, $0 < t_0 < T < \infty$, and $\varepsilon \geq 0$,*

$$\mathcal{R}_{\varepsilon,p_0}(\hat{s}_n^{\mathrm{sub}}) \leq \frac{2}{T-t_0} \log\left(1 + \frac{M_{\vartheta_k,t_0} - 1}{n_2 + 1}\right) + \frac{2\varepsilon^2}{(T-t_0)\tau_0} + \frac{2kD^2}{(n_1+k)(T-t_0)\tau_0}. \tag{5.1}$$

In particular, if $k \geq 2$ and $\tau_0 \leq 1$, then

$$\mathcal{R}_{\varepsilon,p_0}(\hat{s}_n^{\mathrm{sub}}) \leq \frac{\tilde{C}_{k,D}}{n(T-t_0)\tau_0^{k/2}} + \frac{2\varepsilon^2}{(T-t_0)\tau_0}. \tag{5.2}$$

Here $\tilde{C}_{k,D} := 4[e^{k/2}(1 + D^2/k)^{k/2} + kD^2]$ is independent of the ambient dimension d.

The bound decomposes into three terms corresponding to estimation within the empirical subspace, distribution shift, and the missed source component. The last term is controlled by

$$\mathbb{E} \int \|(I_d - \hat{\Pi})x\|_2^2 \, p_0(\mathrm{d}x) \leq \frac{kD^2}{n_1 + k}.$$

For $k \geq 2$ and $\tau_0 \leq 1$, it is absorbed into the statistical term, yielding the simplified rate. Furthermore, the resulting rate replaces the ambient-dimensional dependence $\tau_0^{-d/2}$ in Theorem 4.2 by $\tau_0^{-k/2}$, while the distribution shift term remains unchanged. Thus, intrinsic structure reduces the statistical cost of learning the reference distribution but does not mitigate the uncertainty induced by distribution shift.

5.1 Minimax Optimality

The intrinsic-dimensional rate is also minimax optimal. Let $\mathcal{P}_{k,D}$ denote the subclass of $\mathcal{P}_D$ satisfying Assumption 2. Define the minimax robust score risk by

$$\mathfrak{R}^\star_{n,\varepsilon}(k, D; t_0, T) := \inf_{\hat{s}_n} \sup_{p_0 \in \mathcal{P}_{k,D}} \mathcal{R}_{\varepsilon,p_0}(\hat{s}_n).$$

By Corollary E.3[2], when $k \geq 2$,

$$\mathfrak{R}^\star_{n,\varepsilon}(k, D; t_0, T) \asymp \frac{1}{n(T-t_0)\tau_0^{k/2}} + \frac{\varepsilon^2}{t_0(T-t_0)},$$

with constants depending only on k and D, independently of the ambient dimension d.

6 From Robust Estimation to Robust Generation

We now translate the score estimation guarantees developed above into distributional guarantees for reverse-time sampling. Since the shifted data law q_0 may be singular, we evaluate the sampler at the positive forward time $t_0 > 0$ and measure its error relative to q_{t_0} in KL divergence.

6.1 Minimax Optimal Reverse Sampling

Under Assumption 1, take $\hat{s}_n = \hat{s}_n^\vartheta$ and construct the approximate reverse process by replacing the true score in (2.1) with $\hat{s}_n$. When initialized with the self-consistent terminal distribution $\nu_T = \hat{p}_{n,T}^\vartheta$, the reverse process exactly recovers the estimated OU trajectory, so its output distribution at time t_0 coincides with $\hat{p}_{n,t_0}^\vartheta$. Therefore, setting $\vartheta = D/\sqrt{d}$ and assuming $\tau_0 \leq 1$, combining Propositions 3.1 and 4.1 yields

$$\sup_{q_0 \in \mathcal{Q}_\varepsilon(p_0)} \mathbb{E}\big[\mathrm{KL}\big(q_{t_0} \| \hat{\mu}_{n,t_0}^\nu\big)\big] \leq \frac{C_{d,D}}{2(n+1)\tau_0^{d/2}} + \frac{\varepsilon^2}{\tau_0}, \tag{6.1}$$

[2]Due to space constraints, we defer the formal statement and proof to Appendix E.4.

where $\hat{\mu}^{\nu}_{n,t_0}$ denotes the sampler output distribution. Hence, the reverse sampler inherits the same decomposition as score estimation: a finite-sample statistical term and an intrinsic quadratic distribution-shift term. A corresponding bound under standard Gaussian initialization is given in Corollary F.4.

To characterize whether these two terms are unavoidable, define the minimax positive-time KL risk[3]

$$\mathfrak{K}^{\star}_{n,\varepsilon}(D;t_0) := \inf_{\hat{\mu}_n} \sup_{p_0\in\mathcal{P}_D} \sup_{q_0\in\mathcal{Q}_\varepsilon(p_0)} \mathbb{E}[\mathrm{KL}(q_{t_0}\|\hat{\mu}_n)].$$

The following theorem shows that both components are minimax optimal.

Theorem 6.1. *Fix $d \geq 1$ and $D > 0$. Let $C_{d,D}$ be as in Theorem 4.2. There exist constants $c_0, \tilde{c}_{d,D} > 0$, depending only on d and D, such that if $0 < t_0 \leq c_0, n\tau_0^{d/2} \geq 1, 0 \leq \varepsilon \leq \sqrt{\tau_0}$, then*

$$\tilde{c}_{d,D}\left[\frac{1}{n\tau_0^{d/2}} + \frac{\varepsilon^2}{\tau_0}\right] \leq \mathfrak{K}^{\star}_{n,\varepsilon}(D;t_0) \leq C_{d,D}\left[\frac{1}{n\tau_0^{d/2}} + \frac{\varepsilon^2}{\tau_0}\right]. \tag{6.2}$$

The lower-bound constants are specified in Appendices C and D.

Hence, the statistical and distribution shift costs identified at the score level persist in the generated distribution and are both minimax optimal.

6.2 Intrinsic-Dimensional Reverse Sampling

The intrinsic-dimensional estimator in Section 5 immediately yields corresponding generation guarantees. With $\hat{s}_n = \hat{s}^{\mathrm{sub}}_n$ and self-consistent initialization, the output distribution satisfies the same KL decomposition, with the ambient dimension replaced by the intrinsic dimension.[4]

For the class $\mathcal{P}_{k,D}$, define

$$\mathfrak{K}^{\star}_{n,\varepsilon}(k,D;t_0) := \inf_{\hat{\mu}_n} \sup_{p_0\in\mathcal{P}_{k,D}} \sup_{q_0\in\mathcal{Q}_\varepsilon(p_0)} \mathbb{E}[\mathrm{KL}(q_{t_0}\|\hat{\mu}_n)].$$

Then, by Corollary F.6[5], when $k \geq 2$,

$$\mathfrak{K}^{\star}_{n,\varepsilon}(k,D;t_0) \asymp \frac{1}{n\tau_0^{k/2}} + \frac{\varepsilon^2}{\tau_0},$$

with constants depending only on k and D, independently of the ambient dimension d. Thus, intrinsic structure reduces the statistical cost to the intrinsic dimension k, while the distribution-shift cost remains unchanged.

7 Discussion

This work develops a theory of robust score estimation and generation under distribution shift. We show that the robust error decomposes into a statistical estimation component and an intrinsic distribution-shift component. For the OU diffusion, the shift cost is quadratic in the Wasserstein radius and minimax optimal. We further construct an optimal finite-sample estimator, extend the analysis to low-dimensional source structures, and establish corresponding minimax guarantees for positive-time reverse sampling.

Several directions remain open. First, our analysis focuses on the OU process, which enables a sharp characterization of the costs of distribution shift; extending the framework to more general diffusion dynamics and noise schedules is a natural next step. Second, our intrinsic-dimensional results assume linear subspace structure, and extending them to nonlinear manifolds or other notions of intrinsic complexity remains open. Third, while we establish neural realizability, understanding when practical score-network training achieves the predicted robustness guarantees is an important question. Finally, incorporating discretization error would yield a more complete end-to-end theory of robust sampling.

[3] The infimum is over all measurable, possibly randomized estimators based on n i.i.d. samples from p_0.

[4] We defer the formal statement and proof to Appendix F.3.

[5] We defer the formal statement and proof to Appendix F.4.

AI Assistance Disclosure

In this work, we used generative AI tools only to assist with editing and polishing the writing, including improving the clarity, readability, and presentation of the manuscript. We have not used generative AI tools for research ideation, development of theoretical models, formulation or verification of mathematical claims, design or execution of experiments, generation of synthetic data, or drafting sections of the paper.

All AI-assisted edits were carefully reviewed by the authors for accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including all text, claims, and results presented in the manuscript.

References

Jason M. Altschuler and Sinho Chewi. Shifted composition I: Harnack and reverse transport inequalities. *IEEE Transactions on Information Theory*, 71(1):90–113, 2025. doi: 10.1109/TIT.2024.3475290.

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*, 2024.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Eliot Beyler and Francis Bach. Convergence of deterministic and stochastic diffusion-model samplers: A simple analysis in wasserstein distance. *arXiv preprint arXiv:2508.03210*, 2025.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising autoencoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Changxiao Cai and Gen Li. Minimax optimality of the probability flow ode for diffusion models. *arXiv preprint arXiv:2503.09583*, 2025.

Patrick Cattiaux, Giovanni Conforti, Ivan Gentil, and Christian Léonard. Time reversal of diffusion processes under a finite entropy condition. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 59(4):1844–1881, 2023.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pp. 4672–4712. PMLR, 2023a.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *International Conference on Learning Representations*, 2023b.

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024.

Xiuyuan Cheng, Yunqin Zhu, and Yao Xie. Generative models for decision-making under distributional shift. *arXiv preprint arXiv:2604.04342*, 2026.

Ziheng Cheng, Tianyu Xie, Shiyue Zhang, and Cheng Zhang. Provable sample-efficient transfer learning conditional diffusion models via representation learning. *arXiv preprint arXiv:2502.04491*, 2025.

Sinho Chewi, Alkis Kalavasis, Anay Mehrotra, and Omar Montasser. Ddpm score matching and distribution learning. *arXiv preprint arXiv:2504.05161*, 2025.

Frank Cole and Yulong Lu. Score-based generative models break the curse of dimensionality in learning a family of sub-gaussian probability distributions. *arXiv preprint arXiv:2402.08082*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Zehao Dou, Subhodh Kotekal, Zhehao Xu, and Harrison H Zhou. From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*, 2024.

Jianqing Fan, Yihong Gu, and Ximing Li. Optimal estimation of a factorizable density using diffusion models with relu neural networks. *arXiv preprint arXiv:2510.03994*, 2025.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*, 2024.

Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven adaptation to test-time corruption. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11786–11796. IEEE, 2023.

Xuefeng Gao, Hoang M Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of Machine Learning Research*, 26(43): 1–54, 2025.

Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models. *arXiv preprint arXiv:2404.18869*, 2024.

Shivam Gupta, Aditya Parulekar, Eric Price, and Zhiyang Xun. Improved sample complexity bounds for diffusion model training. *Advances in Neural Information Processing Systems*, 37:40976–41012, 2024.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7323–7334, 2023.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. *arXiv preprint arXiv:2401.15604*, 2024.

Miku Hirata, Aya Nemoto, and Hiroaki Yoshida. An integral representation of the relative entropy. *Entropy*, 14(8):1469–1477, 2012. doi: 10.3390/e14081469.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Julian Kleutgens, Claudio Battiloro, Lingkai Kong, Benjamin Grewe, Francesca Dominici, and Mauricio Tec. Guided transfer learning for discrete diffusion models. *arXiv preprint arXiv:2512.10877*, 2025.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. A sharp convergence theory for the probability flow odes of diffusion models. *arXiv preprint arXiv:2408.02320*, 2024.

Taehong Moon, Moonseok Choi, Gayoung Lee, Jung-Woo Ha, and Juho Lee. Fine-tuning diffusion models with limited data. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pp. 26517–26582. PMLR, 2023.

Yidong Ouyang, Liyan Xie, Hongyuan Zha, and Guang Cheng. Transfer learning for diffusion models. *Advances in Neural Information Processing Systems*, 37:136962–136989, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021b.

Zixing Song, Ziqiao Meng, and José Miguel Hernández-Lobato. Domain-adapted diffusion model for protac linker design through the lens of density ratio in chemical space. In *Forty-second International Conference on Machine Learning*, 2025.

Arthur Stéphanovitch, Eddie Aamari, and Clément Levrard. Generalization bounds for score-based generative models: a synthetic proof. *arXiv preprint arXiv:2507.04794*, 2025.

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics*, pp. 1648–1656. PMLR, 2024.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. doi: 10.1007/b13794.

Sergio Verdú. Mismatched estimation and relative entropy. *IEEE Transactions on Information Theory*, 56(8):3712–3720, 2010. doi: 10.1109/TIT.2010.2050800.

Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4230–4239, 2023.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *International Conference on Learning Representations*, 2022.

Konstantin Yakovlev and Nikita Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. *arXiv preprint arXiv:2502.13662*, 2025.

Konstantin Yakovlev, Anna Markovich, and Nikita Puchkin. Implicit score matching meets denoising score matching: improved rates of convergence and log-density hessian estimation. *arXiv preprint arXiv:2512.24378*, 2025.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017. doi: 10.1016/j.neunet.2017.07.002.

Hiroaki Yoshida. A dissipation of relative entropy by diffusion flows. *Entropy*, 19(1):9, 2017. doi: 10.3390/e19010009.

Yifeng Yu and Lu Yu. Advancing wasserstein convergence analysis of score-based models: Insights from discretization and second-order acceleration. *Advances in Neural Information Processing Systems*, 38:138411–138465, 2025.

Yifeng Yu and Lu Yu. Diffusion models with heavy-tailed targets: Score estimation and sampling guarantees. *arXiv preprint arXiv:2601.06715*, 2026a.

Yifeng Yu and Lu Yu. On the limits of latent reuse in diffusion models. *arXiv preprint arXiv:2605.13448*, 2026b.

Kaihong Zhang, Caitlyn H. Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. *Proceedings of Machine Learning Research*, 235:60134–60178, 2024. ISSN 2640-3498.

A Additional Related Work

A substantial literature studies the learnability of score functions in diffusion models. Early work established L^2-estimation guarantees in terms of hypothesis-class complexity (Block et al., 2020), while subsequent studies derived sharper sample-complexity bounds using expressive neural networks, bounded-parameter local approximations for sub-Gaussian targets, and neural tangent kernel techniques (Gupta et al., 2024; Cole & Lu, 2024; Han et al., 2024). Complementarily, Chewi et al. (2025) relate DDPM score estimation to classical distribution learning. More refined guarantees have been obtained under additional structural assumptions, including Gaussian-mixture models (Gatmiry et al., 2024; Chen et al., 2024), data supported on unknown linear subspaces (Chen et al., 2023a), and distributions concentrated on low-dimensional manifolds (Azangulov et al., 2024; Yakovlev & Puchkin, 2025; Yakovlev et al., 2025). A related line of work establishes minimax rates for score estimation and sampling under assumptions such as sub-Gaussian tails (Zhang et al., 2024; Stéphanovitch et al., 2025; Cai & Li, 2025; Dou et al., 2024; Fu et al., 2024), explicit density lower bounds (Oko et al., 2023; Fan et al., 2025), compact support (Stéphanovitch et al., 2025; Dou et al., 2024; Tang & Yang, 2024), and, more recently, heavy-tailed target distributions (Yu & Yu, 2026a).

Another line of work studies how diffusion models and their learned representations can be transferred or adapted across related domains (Cheng et al., 2025; Ouyang et al., 2024; Song et al., 2025; Kleutgens et al., 2025; Moon et al., 2022; Xie et al., 2023; Han et al., 2023). For example, some methods adapt a pretrained source-domain score through additional guidance (Ouyang et al., 2024; Kleutgens et al., 2025), while others establish sample-efficient transfer guarantees for conditional diffusion models through representation learning (Cheng et al., 2025). Complementing these adaptation-oriented approaches, Yu & Yu (2026b) characterize the reliability and limitations of reusing a frozen source-domain latent representation under distribution shift.

B Proofs for Robust Score Upper Bounds

We first establish two OU information estimates, then prove the robust oracle bound in Proposition 3.1 and the Gaussian-blanket upper bounds in Proposition 4.1 and Theorem 4.2. The corresponding statistical and shift lower bounds are proved in Appendices C and D, respectively.

B.1 Fixed-Time OU Gaussian-Channel Bound

The fixed-time bound below is a special case of a standard reverse transport inequality for Langevin semigroups; see, e.g., Altschuler & Chewi (2025, Corollary 3.11). We include a direct proof to make our OU normalization explicit.

Lemma B.1 (Fixed-time OU Gaussian-channel bound)**.** *Let p_0 and q_0 be Borel probability measures on $\mathbb{R}^d$ with finite second moments, and set $p_t = P_t p_0$ and $q_t = P_t q_0$. Recall that $a_t = e^{-t/2}$ and $\sigma_t^2 = 1 - e^{-t}$. Then, for every $t > 0$,*

$$\mathrm{KL}(q_t \| p_t) \le \frac{a_t^2}{2\sigma_t^2} \mathsf{W}_2^2(p_0, q_0) = \frac{e^{-t}}{2(1-e^{-t})} \mathsf{W}_2^2(p_0, q_0). \tag{B.1}$$

Proof. Fix $t > 0$ and let $\pi \in \Gamma(p_0, q_0)$ be an arbitrary coupling. The $\mathcal{P}_2$ assumption ensures that

$$\int \|y_0 - x_0\|_2^2 \, \pi(\mathrm{d}x_0, \mathrm{d}y_0) \le 2 \int \|x_0\|_2^2 \, p_0(\mathrm{d}x_0) + 2 \int \|y_0\|_2^2 \, q_0(\mathrm{d}y_0) < \infty.$$

The common OU transition kernel is $p_{t|0}(\cdot \mid x_0) := \mathcal{N}(a_t x_0, \sigma_t^2 I_d)$. Since the first and second marginals of π are p_0 and q_0, respectively, the two OU marginals admit the mixture representations

$$p_t = \int_{\mathbb{R}^d \times \mathbb{R}^d} p_{t|0}(\cdot \mid x_0) \, \pi(\mathrm{d}x_0, \mathrm{d}y_0), \qquad q_t = \int_{\mathbb{R}^d \times \mathbb{R}^d} p_{t|0}(\cdot \mid y_0) \, \pi(\mathrm{d}x_0, \mathrm{d}y_0).$$

We use the following mixture form of the joint convexity of relative entropy. Let λ be a probability measure and let (μ_z, ν_z) be a measurable pair of probability kernels indexed by z. Then

$$\mathrm{KL}\Big(\int \mu_z \,\lambda(\mathrm{d}z)\,\Big\|\,\int \nu_z \,\lambda(\mathrm{d}z)\Big) \le \int \mathrm{KL}(\mu_z\|\nu_z)\,\lambda(\mathrm{d}z).$$

For completeness, define joint probability measures by

$$\overline{\mu}(\mathrm{d}z,\mathrm{d}x) := \lambda(\mathrm{d}z)\mu_z(\mathrm{d}x), \qquad \overline{\nu}(\mathrm{d}z,\mathrm{d}x) := \lambda(\mathrm{d}z)\nu_z(\mathrm{d}x).$$

They have the same z-marginal λ, and hence the chain rule for relative entropy gives

$$\mathrm{KL}(\overline{\mu}\|\overline{\nu}) = \int \mathrm{KL}(\mu_z\|\nu_z)\,\lambda(\mathrm{d}z).$$

Their x-marginals are the two mixtures in the displayed inequality. Applying the data-processing inequality to the projection $(z,x)\mapsto x$ therefore proves the claim.

Apply this inequality with $z=(x_0,y_0)$, $\lambda=\pi$, $\mu_z = p_{t|0}(\cdot\mid y_0)$, and $\nu_z = p_{t|0}(\cdot\mid x_0)$. The preceding mixture representations yield

$$\mathrm{KL}(q_t\|p_t) \le \int_{\mathbb{R}^d\times\mathbb{R}^d} \mathrm{KL}\big(p_{t|0}(\cdot\mid y_0)\big\|p_{t|0}(\cdot\mid x_0)\big)\pi(\mathrm{d}x_0,\mathrm{d}y_0).$$

Since $t>0$, we have $\sigma_t^2>0$. The two transition laws inside the relative entropy are therefore nondegenerate Gaussians with the same covariance matrix $\sigma_t^2 I_d$, so

$$\mathrm{KL}\big(p_{t|0}(\cdot\mid y_0)\big\|p_{t|0}(\cdot\mid x_0)\big) = \mathrm{KL}\big(\mathcal{N}(a_t y_0,\sigma_t^2 I_d)\big\|\mathcal{N}(a_t x_0,\sigma_t^2 I_d)\big) = \frac{a_t^2}{2\sigma_t^2}\|y_0-x_0\|_2^2.$$

Consequently,

$$\mathrm{KL}(q_t\|p_t) \le \frac{a_t^2}{2\sigma_t^2}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|y_0-x_0\|_2^2\,\pi(\mathrm{d}x_0,\mathrm{d}y_0).$$

Taking the infimum over $\pi\in\Gamma(p_0,q_0)$ proves (B.1). □

B.2 Relative de Bruijn Identity

The relative de Bruijn identity with a moving reference under common Gaussian smoothing goes back to Verdú (2010). A direct heat-equation derivation and the corresponding exact integral representation were given by Hirata et al. (2012, Theorem 2.2). See also Yoshida (2017, Lemma 2 and Proposition 3) for the Gaussian identity and its extension to common Fokker–Planck diffusion flows under regularity assumptions. We record the precise positive-time OU formulation needed here.

Lemma B.2 (OU relative de Bruijn identity on $\mathcal{P}_2$)**.** *Let $p_0,q_0\in\mathcal{P}_2(\mathbb{R}^d)$, and set $p_t := P_t p_0$ and $q_t := P_t q_0$. For $t>0$, regard q_t and p_t as their strictly positive smooth densities, and define*

$$I(q_t\|p_t) := \int_{\mathbb{R}^d} q_t(x)\|\nabla\log q_t(x) - \nabla\log p_t(x)\|_2^2\,\mathrm{d}x.$$

Then $t\mapsto \mathrm{KL}(q_t\|p_t)$ is locally absolutely continuous on $(0,\infty)$, and

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(q_t\|p_t) = -\frac{1}{2}I(q_t\|p_t) \tag{B.2}$$

for almost every $t>0$. Equivalently, for every $0<t_-<t_+<\infty$,

$$\mathrm{KL}(q_{t_+}\|p_{t_+}) + \frac{1}{2}\int_{t_-}^{t_+} I(q_t\|p_t)\,\mathrm{d}t = \mathrm{KL}(q_{t_-}\|p_{t_-}). \tag{B.3}$$

Proof. Let $(\mathsf{H}_s)_{s\ge0}$ denote the heat semigroup $\mathsf{H}_s\eta := \eta * \mathcal{N}(0,sI_d)$, and set $\overline{p}_s := \mathsf{H}_s p_0$ and $\overline{q}_s := \mathsf{H}_s q_0$. We first verify the hypotheses of the moving-reference heat-flow identity in Hirata et al. (2012, Theorem 2.2). For every $s>0$, both $\overline{q}_s$ and $\overline{p}_s$ are strictly positive smooth densities;

in particular, $\overline{q}_s \ll \overline{p}_s$. For any $\eta \in \mathcal{P}_2(\mathbb{R}^d)$, if $Y \sim \eta$ and $Z \sim \mathcal{N}(0, I_d)$ are independent and $U_s := Y + \sqrt{s}Z$, the Gaussian score identity gives

$$\nabla \log(\mathsf{H}_s\eta)(U_s) = -s^{-1/2}\mathbb{E}[Z \mid U_s].$$

Consequently, conditional Jensen's inequality yields

$$\begin{aligned} J(\mathsf{H}_s\eta) := \int_{\mathbb{R}^d} \mathsf{H}_s\eta(x)\|\nabla \log(\mathsf{H}_s\eta)(x)\|_2^2 \,\mathrm{d}x &= \frac{1}{s}\mathbb{E}\Big[\|\mathbb{E}[Z \mid U_s]\|_2^2\Big] \\ &\le \frac{1}{s}\mathbb{E}\|Z\|_2^2 = \frac{d}{s}. \end{aligned} \tag{B.4}$$

It remains to check finiteness of the relative entropy. Given $s > 0$, put $u := \log(1+s)$, let $S_c(x) := cx$, and write $(S_c)_\#\lambda$ for the pushforward of λ under S_c. Since

$$P_u\eta = (S_{a_u})_\#\mathsf{H}_s\eta,$$

invariance of relative entropy under a common invertible dilation and Lemma B.1 imply

$$\begin{aligned} \mathrm{KL}(\overline{q}_s\|\overline{p}_s) &= \mathrm{KL}(P_uq_0\|P_up_0) \\ &\le \frac{a_u^2}{2\sigma_u^2}\mathsf{W}_2^2(p_0,q_0) = \frac{1}{2s}\mathsf{W}_2^2(p_0,q_0) < \infty. \end{aligned} \tag{B.5}$$

For each fixed $s > 0$, we now apply Hirata et al. (2012, Theorem 2.2) with $f = \overline{q}_s$ and $g = \overline{p}_s$. As noted above, both densities are strictly positive, and hence $\overline{q}_s \ll \overline{p}_s$. Applying (B.4) separately to $\eta = q_0$ and $\eta = p_0$ gives

$$J(\overline{q}_s) < \infty, \qquad J(\overline{p}_s) < \infty,$$

while (B.5) gives

$$\mathrm{KL}(\overline{q}_s\|\overline{p}_s) < \infty.$$

These are precisely the hypotheses of the cited theorem.

The integral representation in the cited theorem, together with the heat-semigroup property, gives

$$\mathrm{KL}(\overline{q}_s\|\overline{p}_s) = \frac{1}{2}\int_s^\infty I(\overline{q}_v\|\overline{p}_v)\,\mathrm{d}v.$$

It follows that $s \mapsto \mathrm{KL}(\overline{q}_s\|\overline{p}_s)$ is locally absolutely continuous and, for $0 < s_- < s_+ < \infty$,

$$\mathrm{KL}(\overline{q}_{s_+}\|\overline{p}_{s_+}) + \frac{1}{2}\int_{s_-}^{s_+} I(\overline{q}_s\|\overline{p}_s)\,\mathrm{d}s = \mathrm{KL}(\overline{q}_{s_-}\|\overline{p}_{s_-}). \tag{B.6}$$

We now transfer this identity to the OU clock. A common invertible dilation leaves relative entropy invariant and rescales relative Fisher information. In the present notation,

$$\begin{aligned} &\tau_t := \frac{\sigma_t^2}{a_t^2} = e^t - 1, \qquad q_t = (S_{a_t})_\#\overline{q}_{\tau_t}, \qquad p_t = (S_{a_t})_\#\overline{p}_{\tau_t}, \\ &\mathrm{KL}(q_t\|p_t) = \mathrm{KL}(\overline{q}_{\tau_t}\|\overline{p}_{\tau_t}), \qquad I(q_t\|p_t) = a_t^{-2}I(\overline{q}_{\tau_t}\|\overline{p}_{\tau_t}), \qquad \tau_t' = e^t = a_t^{-2}. \end{aligned}$$

Consequently, under the change of variables $s = \tau_t$, we have $\mathrm{d}s = \tau_t'\,\mathrm{d}t = a_t^{-2}\,\mathrm{d}t$, and

$$\int_{\tau_{t_-}}^{\tau_{t_+}} I(\overline{q}_s\|\overline{p}_s)\,\mathrm{d}s = \int_{t_-}^{t_+} I(q_t\|p_t)\,\mathrm{d}t.$$

Together with $\mathrm{KL}(\overline{q}_{\tau_{t_\pm}}\|\overline{p}_{\tau_{t_\pm}}) = \mathrm{KL}(q_{t_\pm}\|p_{t_\pm})$, substituting $s_\pm = \tau_{t_\pm}$ into (B.6) yields (B.3). The local absolute continuity and the almost-everywhere differential identity (B.2) follow from the same change of variables. □

B.3 Proof of Proposition 3.1

The proof uses the fixed-time Gaussian-channel estimate in Lemma B.1 and the $\mathcal{P}_2$ relative de Bruijn identity in Lemma B.2. The only additional ingredient is the following elementary change-of-reference inequality.

Lemma B.3 (Three-density KL inequality)**.** *Let q, p, r be strictly positive probability densities. If $\mathrm{KL}(q\|p) < \infty$ and $D_2(p\|r) < \infty$, then*

$$\mathrm{KL}(q\|r) \le 2\mathrm{KL}(q\|p) + D_2(p\|r). \tag{B.7}$$

Proof. For any bounded measurable function g, the entropy variational inequality applied to $g/2$ gives

$$\mathbb{E}_q g \le 2\mathrm{KL}(q\|p) + 2\log \mathbb{E}_p e^{g/2},$$
$$\mathbb{E}_p e^{g/2} = \mathbb{E}_r\left[\frac{p}{r}e^{g/2}\right] \le \left(\int \frac{p^2}{r}\right)^{1/2} (\mathbb{E}_r e^g)^{1/2}.$$

Here the second line rewrites the expectation under r and applies Cauchy–Schwarz. Therefore,

$$\mathbb{E}_q g - \log \mathbb{E}_r e^g \le 2\mathrm{KL}(q\|p) + D_2(p\|r).$$

Taking the supremum over bounded measurable g in the variational representation of $\mathrm{KL}(q\|r)$ proves (B.7). □

Proof of Proposition 3.1 Fix a realization of $\hat{\mu}_0$ and let $q_0 \in \mathcal{P}_2(\mathbb{R}^d)$. At time t_0, apply Lemma B.3 with $q = q_{t_0}$, $p = p_{t_0}$, and $r = \hat{p}_{t_0}$. All three densities are strictly positive. Lemma B.1 gives $2\mathrm{KL}(q_{t_0}\|p_{t_0}) \le a_{t_0}^2 \mathsf{W}_2^2(q_0, p_0)/\rho_{t_0}$, and $D_2(p_{t_0}\|\hat{p}_{t_0}) < \infty$ by assumption. Hence

$$\mathrm{KL}(q_{t_0}\|\hat{p}_{t_0}) \le \frac{a_{t_0}^2}{\rho_{t_0}}\mathsf{W}_2^2(q_0, p_0) + D_2(p_{t_0}\|\hat{p}_{t_0}),$$

which is (3.1).

Because $q_0, \hat{\mu}_0 \in \mathcal{P}_2(\mathbb{R}^d)$, Lemma B.2 applies directly to the two OU trajectories $q_t = P_t q_0$ and $\hat{p}_t = P_t \hat{\mu}_0$. It yields

$$\begin{aligned}\int_{t_0}^{T} \mathbb{E}_{q_t}\left[\|\hat{s}(X,t) - \nabla \log q_t(X)\|_2^2\right] \mathrm{d}t \\ &= 2[\mathrm{KL}(q_{t_0}\|\hat{p}_{t_0}) - \mathrm{KL}(q_T\|\hat{p}_T)] \\ &\le 2\mathrm{KL}(q_{t_0}\|\hat{p}_{t_0}).\end{aligned}$$

Combining the preceding display with (3.1), dividing by $T - t_0$, and using $a_{t_0}^2/\rho_{t_0} = 1/\tau_0$ gives, for every $q_0 \in \mathcal{Q}_\varepsilon(p_0)$,

$$\mathcal{L}_{q_0}(\hat{s}) \le \frac{2}{(T-t_0)\tau_0}\mathsf{W}_2^2(q_0, p_0) + \frac{2}{T-t_0}D_2(p_{t_0}\|\hat{p}_{t_0}),$$

for almost every realization of $\hat{\mu}_0$. Fixing q_0 and taking expectation over the randomness of $\hat{\mu}_0$ yields

$$\mathbb{E}\mathcal{L}_{q_0}(\hat{s}) \le \frac{2\varepsilon^2}{(T-t_0)\tau_0} + \frac{2}{T-t_0}\mathbb{E}D_2(p_{t_0}\|\hat{p}_{t_0}).$$

The right-hand side is independent of q_0. Taking the supremum over $q_0 \in \mathcal{Q}_\varepsilon(p_0)$ and using (2.2) proves (3.2). □

B.4 Proof of Proposition 4.1

The proof uses the following two ingredients. The first establishes the Gaussian domination property underlying the blanket construction, while the second is a general add-one bound for dominated kernels. Their proofs are deferred until after the short deduction of Proposition 4.1.

Lemma B.4 (Gaussian blanket domination). *For $t \geq t_0$, $y \in \mathcal{B}(0, D)$, and $x \in \mathbb{R}^d$, let $k_{t,y}(x) := \varphi_{\rho_t}(x - a_t y)$. Then*

$$k_{t,y}(x) \leq M_{\vartheta,t_0} \varphi_{v_t}(x). \tag{B.8}$$

More precisely,

$$\sup_{x \in \mathbb{R}^d} \frac{k_{t,y}(x)}{\varphi_{v_t}(x)} = \left(1 + \frac{\vartheta^2}{e^t - 1}\right)^{d/2} \exp\left(\frac{\|y\|_2^2}{2\vartheta^2}\right). \tag{B.9}$$

Lemma B.5 (Dominated-kernel add-one bound). *Let $(\mathsf{Y}, \mathcal{A})$ be a measurable space and let $n \geq 1$. Suppose $Y_1, \ldots, Y_n$ are i.i.d. copies of a random element Y. Let $k : \mathsf{Y} \times \mathbb{R}^d \to [0, \infty)$ be jointly measurable, write $k_y(x) := k(y, x)$, and suppose that k_y is a probability density on $\mathbb{R}^d$ for every $y \in \mathsf{Y}$. Let r be a strictly positive probability density. Assume that, for some $M \geq 1$,*

$$k_y(x) \leq M r(x) \qquad \text{for every } x \in \mathbb{R}^d \text{ and every } y \in \mathsf{Y}.$$

Define

$$p(x) := \mathbb{E}[k_Y(x)], \qquad \hat{p}_n(x) := \frac{\sum_{i=1}^n k_{Y_i}(x) + M r(x)}{n + M}.$$

Then

$$\mathbb{E} \int_{\mathbb{R}^d} \frac{p(x)^2}{\hat{p}_n(x)} \, \mathrm{d}x \leq \frac{n + M}{n + 1}. \tag{B.10}$$

Proof of Proposition 4.1 Fix $t \in [t_0, T]$. Take $\mathsf{Y} = \mathcal{B}(0, D)$, equipped with its Borel σ-field, and apply Lemma B.5 with $Y_i = X_i$, $k_y(x) = k_{t,y}(x) = \varphi_{\rho_t}(x - a_t y)$, $r(x) = \varphi_{v_t}(x)$, and $M = M_{\vartheta,t_0}$. The domination assumption follows from Lemma B.4, while $p(x) = \mathbb{E}[k_{t,X}(x)] = p_t(x)$ for a generic draw $X \sim p_0$, and the regularized density in the abstract lemma is exactly $\hat{p}^{\vartheta}_{n,t}$. Therefore,

$$\mathbb{E}_{X_{1:n}} \int_{\mathbb{R}^d} \frac{p_t(x)^2}{\hat{p}^{\vartheta}_{n,t}(x)} \, \mathrm{d}x \leq \frac{n + M_{\vartheta,t_0}}{n + 1}.$$

Finally, the concavity of the logarithm implies

$$\begin{aligned} \mathbb{E}_{X_{1:n}} D_2\big(p_t \big\| \hat{p}^{\vartheta}_{n,t}\big) &= \mathbb{E}_{X_{1:n}} \log \int \frac{p_t^2}{\hat{p}^{\vartheta}_{n,t}} \\ &\leq \log \mathbb{E}_{X_{1:n}} \int \frac{p_t^2}{\hat{p}^{\vartheta}_{n,t}} \leq \log \frac{n + M_{\vartheta,t_0}}{n + 1}, \end{aligned}$$

which is (4.4). □

Proof of Lemma B.4 Fix $t \geq t_0$ and $y \in \mathcal{B}(0, D)$. Recall that $k_{t,y}(x) = \varphi_{\rho_t}(x - a_t y)$ and $v_t = \rho_t + a_t^2 \vartheta^2$. For every $x \in \mathbb{R}^d$,

$$\log \frac{k_{t,y}(x)}{\varphi_{v_t}(x)} = \frac{d}{2} \log \frac{v_t}{\rho_t} - \frac{\|x - a_t y\|_2^2}{2\rho_t} + \frac{\|x\|_2^2}{2v_t}.$$

The last two terms form a strictly concave quadratic function of x. Since $v_t - \rho_t = a_t^2 \vartheta^2$, completing the square gives

$$-\frac{\|x - a_t y\|_2^2}{2\rho_t} + \frac{\|x\|_2^2}{2v_t} = \frac{\|y\|_2^2}{2\vartheta^2} - \frac{a_t^2 \vartheta^2}{2\rho_t v_t} \left\| x - \frac{v_t}{a_t \vartheta^2} y \right\|_2^2. \tag{B.11}$$

Taking the supremum over x in the preceding identity and simplifying the variance ratio gives

$$\sup_{x \in \mathbb{R}^d} \frac{k_{t,y}(x)}{\varphi_{v_t}(x)} = \left(\frac{v_t}{\rho_t}\right)^{d/2} \exp\left(\frac{\|y\|_2^2}{2\vartheta^2}\right),$$

$$\frac{v_t}{\rho_t} = 1 + \frac{a_t^2 \vartheta^2}{\rho_t} = 1 + \frac{\vartheta^2}{e^t - 1}.$$

These identities give (B.9). Since $t \mapsto (e^t - 1)^{-1}$ is decreasing and $\|y\|_2 \le D$, the right-hand side is bounded by M_{ϑ,t_0} from (4.1). This proves (B.8). □

Proof of Lemma B.5 Fix $x \in \mathbb{R}^d$ and write

$$U_i(x) := \frac{k_{Y_i}(x)}{Mr(x)} \in [0,1], \qquad \theta(x) := \mathbb{E}[U_i(x)] = \frac{p(x)}{Mr(x)} \in [0,1].$$

Then

$$\frac{p(x)^2}{\hat{p}_n(x)} = Mr(x)\theta(x)^2(n+M)\frac{1}{1+\sum_{i=1}^n U_i(x)}. \tag{B.12}$$

For every $c > 0$, the map $u \mapsto (c+u)^{-1}$ is convex on $[0,1]$. Hence, if $U \in [0,1]$ has mean θ, the chord inequality gives

$$\mathbb{E}\left[\frac{1}{c+U}\right] \le \frac{1-\theta}{c} + \frac{\theta}{c+1}.$$

Let $V_1, \dots, V_n$ be i.i.d. Bernoulli$(\theta(x))$ random variables, independent of $U_1(x), \dots, U_n(x)$. Conditional on the remaining summands, the preceding inequality shows that replacing $U_i(x)$ by V_i can only increase the expectation. Repeating this replacement successively for $i = 1, \dots, n$ yields

$$\mathbb{E}\frac{1}{1+\sum_{i=1}^n U_i(x)} \le \mathbb{E}\frac{1}{1+\sum_{i=1}^n V_i} = \mathbb{E}\frac{1}{1+B}, \qquad B := \sum_{i=1}^n V_i \sim \mathrm{Bin}(n, \theta(x)).$$

For $\theta \in (0,1]$,

$$\begin{aligned}\mathbb{E}\frac{1}{1+B} &= \int_0^1 \mathbb{E}[z^B]\,\mathrm{d}z = \int_0^1 (1-\theta+\theta z)^n\,\mathrm{d}z \\ &= \frac{1-(1-\theta)^{n+1}}{(n+1)\theta} \le \frac{1}{(n+1)\theta}.\end{aligned}$$

If $\theta(x) = 0$, then $p(x) = 0$ and the left-hand side of (B.12) vanishes. If $\theta(x) > 0$, substituting the preceding estimate into (B.12) yields

$$\mathbb{E}\frac{p(x)^2}{\hat{p}_n(x)} \le Mr(x)\theta(x)\frac{n+M}{n+1} = p(x)\frac{n+M}{n+1}.$$

Tonelli's theorem and $\int p = 1$ now give (B.10). □

We record the consequences of Lemma B.4 needed below. For every probability measure μ supported in $\mathcal{B}(0,D)$,

$$P_t\mu(x) = \int k_{t,y}(x)\,\mu(\mathrm{d}y) \le M_{\vartheta,t_0}\varphi_{v_t}(x), \qquad t \in [t_0, T]. \tag{B.13}$$

Moreover, by construction,

$$\hat{p}^{\vartheta}_{n,t}(x) \ge \frac{M_{\vartheta,t_0}}{n + M_{\vartheta,t_0}}\varphi_{v_t}(x).$$

Taking $\mu = p_0$ in (B.13) and combining the two displays gives

$$\frac{p_t(x)}{\hat{p}^{\vartheta}_{n,t}(x)} \le n + M_{\vartheta,t_0}, \qquad t \in [t_0, T],\ x \in \mathbb{R}^d. \tag{B.14}$$

Multiplying (B.14) by p_t and integrating gives

$$\int_{\mathbb{R}^d} \frac{p_t(x)^2}{\hat{p}^{\vartheta}_{n,t}(x)}\,\mathrm{d}x \le \left(n + M_{\vartheta,t_0}\right)\int_{\mathbb{R}^d} p_t(x)\,\mathrm{d}x = n + M_{\vartheta,t_0} < \infty.$$

Thus $D_2(p_t \| \hat{p}^{\vartheta}_{n,t}) < \infty$ whenever $X_1, \dots, X_n \in \mathcal{B}(0,D)$, and hence $p_0^{\otimes n}$-almost surely. This verifies the finiteness assumption in Proposition 3.1.

B.5 Proof of Theorem 4.2

Proof. Assumption 1 implies $p_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Since $X_i \in \mathcal{B}(0, D)$ almost surely, for $p_0^{\otimes n}$-almost every realization of the training sample, the blanket initial measure belongs to $\mathcal{P}_2(\mathbb{R}^d)$, since

$$\int_{\mathbb{R}^d} \|x\|_2^2 \, \hat{\mu}_{n,0}^{\vartheta}(\mathrm{d}x) = \frac{\sum_{i=1}^n \|X_i\|_2^2 + M_{\vartheta,t_0} d\vartheta^2}{n + M_{\vartheta,t_0}} < \infty.$$

Moreover, the pointwise bound (B.14) guarantees $D_2(p_{t_0} \| \hat{p}_{n,t_0}^{\vartheta}) < \infty$ for every such realization. Therefore, Proposition 3.1 applies to the random measure $\hat{\mu}_0 = \hat{\mu}_{n,0}^{\vartheta}$ and gives

$$\mathcal{R}_{\varepsilon,p_0}(\hat{s}_n^{\vartheta}) \le \frac{2\varepsilon^2}{(T-t_0)\tau_0} + \frac{2}{T-t_0} \mathbb{E}_{X_{1:n}} D_2\big(p_{t_0} \big\| \hat{p}_{n,t_0}^{\vartheta}\big).$$

Applying Proposition 4.1 at $t = t_0$, and then using $\log(1+u) \le u$ for $u \ge 0$, yields

$$\begin{aligned} \mathcal{R}_{\varepsilon,p_0}(\hat{s}_n^{\vartheta}) &\le \frac{2\varepsilon^2}{(T-t_0)\tau_0} + \frac{2}{T-t_0} \log\left(1 + \frac{M_{\vartheta,t_0} - 1}{n+1}\right) \\ &\le \frac{2\varepsilon^2}{(T-t_0)\tau_0} + \frac{2(M_{\vartheta,t_0} - 1)}{(n+1)(T-t_0)}. \end{aligned} \tag{B.15}$$

It remains to prove the explicit rate specialization. Set $\vartheta^2 = D^2/d$. Since $\tau_0 = e^{t_0} - 1$, we have

$$M_{\vartheta,t_0} = e^{d/2} \left(1 + \frac{D^2}{d\tau_0}\right)^{d/2}.$$

If $\tau_0 \le 1$, then

$$1 + \frac{D^2}{d\tau_0} \le \tau_0^{-1} \left(1 + \frac{D^2}{d}\right), \qquad M_{\vartheta,t_0} \le e^{d/2} \left(1 + \frac{D^2}{d}\right)^{d/2} \tau_0^{-d/2}.$$

Substituting this into (B.15) and using $M_{\vartheta,t_0} - 1 \le M_{\vartheta,t_0}$ proves (4.5) with the constant $C_{d,D}$ specified in Theorem 4.2. The order comparison following that theorem follows from $n + 1 \asymp n$ and $\tau_0 = e^{t_0} - 1 \asymp t_0$ whenever $\tau_0 \le 1$. □

C Proofs for Statistical Lower Bounds

This appendix proves the statistical lower bounds for time-integrated score estimation and positive-time KL estimation. The argument is self-contained apart from the standard Gilbert–Varshamov packing lemma and Fano's inequality, and its construction is adapted from the high-noise score-estimation lower bound of Dou et al. (2024). The essential additional ingredient for the integrated problem is that a single hard family remains score-separated throughout the heat-time window $[\tau_0, 2\tau_0]$. Subsections C.1– C.4 prove Theorem 4.3; Subsections C.5 and C.6 reuse the same hard family to prove the nominal positive-time KL lower bound. The robust minimax conclusions are completed in Appendix D by combining these statistical bounds with the corresponding shift lower bounds.

Throughout the appendix, ϕ_τ denotes the density of $\mathcal{N}(0, \tau I_d)$, and

$$(\mathsf{H}_\tau f)(x) := (\phi_\tau * f)(x)$$

denotes the heat semigroup. Thus $\partial_\tau \mathsf{H}_\tau f = \frac{1}{2} \Delta \mathsf{H}_\tau f$. The bump used below is fixed explicitly for each dimension, and every constant that enters the final lower bound is specified below.

C.1 Reduction to the Integrated Heat-Score Problem

The first reduction is immediate but important. Since $p_0 \in \mathcal{Q}_\varepsilon(p_0)$, the definition of the robust risk gives, for every estimator $\hat{s}_n$,

$$\mathcal{R}_{\varepsilon,p_0}(\hat{s}_n) \ge \mathbb{E}\left[\frac{1}{T-t_0} \int_{t_0}^{T} \mathbb{E}_{X \sim p_t}\big[\|\hat{s}_n(X,t) - \nabla \log p_t(X)\|_2^2\big] \, \mathrm{d}t\right]. \tag{C.1}$$

Here the expectation is over $X_{1:n} \sim p_0^{\otimes n}$ and any internal randomization used by the estimator. It therefore suffices to lower bound the nominal time-integrated score-estimation problem.

We next record the exact correspondence between the OU and variance-exploding parametrizations.

Lemma C.1 (Exact OU–heat loss identity)**.** *Let p_0 be any Borel probability measure and define*

$$a_t := e^{-t/2}, \qquad \tau_t := \frac{1-a_t^2}{a_t^2} = e^t - 1, \qquad g_\tau := p_0 * \phi_\tau, \qquad u_\tau := \nabla \log g_\tau.$$

Then

$$p_t(x) = a_t^{-d} g_{\tau_t}(x/a_t), \qquad \nabla \log p_t(x) = a_t^{-1} u_{\tau_t}(x/a_t). \tag{C.2}$$

For an arbitrary score estimator $\hat{s}_n(x,t)$, define

$$\tilde{s}_n(z,\tau) := a_t \hat{s}_n(a_t z, t), \qquad t = \log(1+\tau). \tag{C.3}$$

Then, for every $0 < t_a < t_b$,

$$\int_{t_a}^{t_b} \|\hat{s}_n(\cdot,t) - \nabla \log p_t\|^2_{L^2(p_t)} \,\mathrm{d}t = \int_{\tau_{t_a}}^{\tau_{t_b}} \|\tilde{s}_n(\cdot,\tau) - u_\tau\|^2_{L^2(g_\tau)} \,\mathrm{d}\tau. \tag{C.4}$$

Proof. The OU representation can be rewritten as $X_t = a_t X_0 + \sqrt{1-a_t^2}\, Z = a_t(X_0 + \sqrt{e^t - 1}\, Z)$, where $Z \sim \mathcal{N}(0, I_d)$ is independent of X_0. This proves the density scaling in (C.2); differentiating its logarithm gives the score scaling. After the change of variables $x = a_t z$,

$$\|\hat{s}_n(\cdot,t) - \nabla \log p_t\|^2_{L^2(p_t)} = a_t^{-2} \|\tilde{s}_n(\cdot,\tau_t) - u_{\tau_t}\|^2_{L^2(g_{\tau_t})}.$$

Finally, $\mathrm{d}\tau_t/\mathrm{d}t = e^t = a_t^{-2}$, equivalently $\mathrm{d}t = a_t^2 \,\mathrm{d}\tau$. The two scaling factors cancel exactly, proving (C.4). □

Set $\tau_0 = e^{t_0} - 1$ and $t_1 := \log(1 + 2\tau_0)$. Since $e^{2t_0} - (1 + 2\tau_0) = e^{2t_0} - 2e^{t_0} + 1 = (e^{t_0} - 1)^2 \geq 0$, we have $t_1 \leq 2t_0$. Moreover, $1 + 2\tau_0 > 1 + \tau_0 = e^{t_0}$, so $t_1 > t_0$. Under the assumption $2t_0 \leq T$, the nonempty interval $[t_0, t_1]$ is contained in $[t_0, T]$. Lemma C.1 therefore reduces the desired lower bound to the heat-time window $[\tau_0, 2\tau_0]$.

C.2 Construction of a Compactly Supported Hard Family

The calibration has two targets: the sample KL divergence must remain below a fixed fraction of the packing entropy, while the associated heat scores remain separated throughout $[\tau_0, 2\tau_0]$. We therefore use cells of width $\rho = K_0\sqrt{\tau_0}$, giving a number m of cells of order $\tau_0^{-d/2}$, and later choose $\eta^2 = \kappa(1 \wedge m/n)$. The exact admissible constants are fixed below.

Fix

$$\ell := \frac{D}{8\sqrt{d}}, \qquad Q := [-4\ell, 4\ell]^d, \qquad I := [-3\ell, 3\ell]^d.$$

The cube Q is contained in $\mathcal{B}(0, D/2)$, hence in $\mathcal{B}(0, D)$. Let $c_Q := |Q|^{-1} = (8\ell)^{-d}$ and $f_\circ := c_Q \mathbf{1}_Q$. To remove any auxiliary dependence from the construction, define

$$\psi(u) := \begin{cases} \exp\left(-\dfrac{1}{1-16u^2}\right), & |u| < 1/4, \\ 0, & |u| \geq 1/4, \end{cases} \qquad \Psi_d(x) := \prod_{k=1}^d \psi(x_k),$$

and fix the dimension-dependent bump

$$w = w_d := \frac{\partial_1 \Psi_d}{\|\partial_1 \Psi_d\|_\infty}.$$

Then $w_d \in C_c^\infty((-1/2, 1/2)^d)$, $w_d \not\equiv 0$, and

$$\int_{\mathbb{R}^d} w(x) \,\mathrm{d}x = 0, \qquad \|w\|_\infty = 1, \qquad A_1 := \int \|\nabla w\|_2^2 > 0. \tag{C.5}$$

We also write

$$A_0 := \int_{\mathbb{R}^d} w^2, \qquad A_2 := \int_{\mathbb{R}^d} (\Delta w)^2, \qquad L_w := \|\nabla w\|_\infty.$$

Thus $A_0, A_1, A_2, L_w \in (0, \infty)$ are fixed constants determined solely by d.

Set

$$K_0 := \max\left\{1, \sqrt{\frac{8A_2}{A_1}}\right\}, \qquad \rho := K_0\sqrt{\tau_0}. \tag{C.6}$$

In particular,

$$A_1 - \frac{2A_2}{K_0^2} \geq \frac{3A_1}{4}. \tag{C.7}$$

For the grid below, define $\{z_1, \ldots, z_m\}$ to be the lattice points

$$2\rho\mathbb{Z}^d \cap [-(\ell - \rho), \ell - \rho]^d. \tag{C.8}$$

Whenever $\rho \leq \ell/4$, the cubes $z_j + \rho[-1/2, 1/2]^d$, $1 \leq j \leq m$, are pairwise disjoint and

$$2^{-d}\left(\frac{\ell}{\rho}\right)^d \leq m \leq \left(\frac{\ell}{\rho}\right)^d. \tag{C.9}$$

Indeed, if $i \neq j$, then $|(z_i)_k - (z_j)_k| \geq 2\rho$ for some coordinate k. Since each displayed cube has coordinate half-width $\rho/2$, the two cubes are separated by at least ρ in that coordinate. Hence the cubes are pairwise disjoint; in particular, the rescaled bumps defined below have pairwise disjoint supports.

To count the lattice points, put $y := \ell/\rho \geq 4$. In each coordinate the admissible points are $2\rho k$ with $|k| \leq (y-1)/2$, so their number is $N = 2\lfloor (y-1)/2 \rfloor + 1$. The elementary bounds $u - 1 \leq \lfloor u \rfloor \leq u$ give $y - 2 \leq N \leq y$. Since $y \geq 4$, it follows that $y/2 \leq N \leq y$. The grid is the d-fold Cartesian product of this one-dimensional set, so $m = N^d$, which proves (C.9).

Define $w_{j,\rho}(x) := w((x - z_j)/\rho)$. For $b = (b_1, \ldots, b_m) \in \{0, 1\}^m$ and an amplitude $0 < \eta \leq 1/4$, set

$$f_b(x) := f_\circ(x)\left(1 + \eta \sum_{j=1}^m b_j w_{j,\rho}(x)\right). \tag{C.10}$$

Because the bump supports are disjoint, $|\sum_{j=1}^m b_j w_{j,\rho}| \leq 1$. Hence the factor in parentheses in (C.10) lies in $[3/4, 5/4]$. The mean-zero condition in (C.5) gives $\int f_b = 1$. Thus every f_b is a probability density supported in $Q \subset \mathcal{B}(0, D)$, with $3f_\circ/4 \leq f_b \leq 5f_\circ/4$ on Q. For $b, b' \in \{0, 1\}^m$, write $d_{\mathrm{Ham}}(b, b') := \sum_{j=1}^m \mathbf{1}_{\{b_j \neq b'_j\}}$.

The following lemma records the information bound for this family.

Lemma C.2 (Sample KL bound). *For every $b \in \{0, 1\}^m$,*

$$\mathrm{KL}(f_b^{\otimes n} \| f_\circ^{\otimes n}) \leq c_Q A_0 n \eta^2 m \rho^d.$$

Proof. Write $W_b := \sum_{j=1}^m b_j w_{j,\rho}$. Since $f_b \ll f_\circ$, tensorization and the standard inequality $\mathrm{KL}(P\|Q) \leq \chi^2(P\|Q)$ give

$$\begin{aligned}\mathrm{KL}(f_b^{\otimes n} \| f_\circ^{\otimes n}) &\leq n \int_Q \frac{(f_b - f_\circ)^2}{f_\circ}\,\mathrm{d}x = nc_Q\eta^2 \int_Q W_b^2\,\mathrm{d}x \\ &= nc_Q A_0 \eta^2 \rho^d \sum_{j=1}^m b_j \leq c_Q A_0 n \eta^2 m \rho^d.\end{aligned}$$

The second equality uses the pairwise disjoint bump supports and $b_j^2 = b_j$. □

C.3 Uniform Score Separation over the Heat-Time Window

For $\tau > 0$, let

$$g_{b,\tau} := \mathsf{H}_\tau f_b, \qquad s_{b,\tau} := \nabla \log g_{b,\tau}.$$

Define the dimensionally explicit envelope constants

$$\bar{\tau} := \frac{\ell^2}{2\log(4d)}, \qquad C_1 := 8\sqrt{\frac{d}{2\pi}}, \qquad C_2 := 4L_w. \tag{C.11}$$

We first establish common density and score envelopes on the fixed interior cube I.

Lemma C.3 (Interior density and score bounds)**.** *For every $0 < \tau \le \bar{\tau}$, every $b \in \{0,1\}^m$, every $x \in I$, and every $0 < \eta \le 1/4$, the constants in (C.11) satisfy*

$$\frac{c_Q}{4} \le g_{b,\tau}(x) \le \frac{5c_Q}{4}, \tag{C.12}$$

and

$$\|s_{b,\tau}(x)\|_2 \le C_1 \tau^{-1/2} \exp\left(-\frac{\ell^2}{2\tau}\right) + C_2 \frac{\eta}{\rho}. \tag{C.13}$$

Proof. For any bounded h and $Z \sim \mathcal{N}(0, I_d)$, $(\mathsf{H}_\tau h)(x) = \mathbb{E}[h(x + \sqrt{\tau}Z)]$. Since the Gaussian kernel is nonnegative and has unit mass, this representation gives $\|\mathsf{H}_\tau h\|_\infty \le \|h\|_\infty$; the same argument applies to vector-valued h with the Euclidean norm.

For $x \in I$, the distance in each coordinate from x to the boundary of Q is at least ℓ. A union bound and the Gaussian tail bound therefore give

$$(\mathsf{H}_\tau f_\circ)(x) = c_Q \mathbb{P}(x + \sqrt{\tau}Z \in Q) \ge c_Q \left[1 - 2d \exp\left(-\frac{\ell^2}{2\tau}\right)\right].$$

Since $\tau \le \bar{\tau}$ makes the bracket at least $1/2$, while $\|f_b - f_\circ\|_\infty \le \eta c_Q$, the contraction property yields $c_Q/2 - \eta c_Q \le g_{b,\tau}(x) \le c_Q + \eta c_Q$. As $\eta \le 1/4$, this proves (C.12).

The base convolution factorizes coordinatewise. Differentiating it explicitly gives, for each k,

$$|\partial_k \mathsf{H}_\tau f_\circ(x)| \le 2c_Q (2\pi\tau)^{-1/2} \exp\left(-\frac{\ell^2}{2\tau}\right).$$

Moreover, although $f_\circ$ is discontinuous at ∂Q, the difference $f_b - f_\circ$ is a smooth compactly supported sum of bumps lying strictly inside Q. The disjoint supports give $\|\nabla(f_b - f_\circ)\|_\infty \le c_Q \eta \rho^{-1} L_w$. Since H_τ commutes with derivatives, its contraction property then gives $\|\nabla \mathsf{H}_\tau (f_b - f_\circ)\|_\infty \le c_Q \eta \rho^{-1} L_w$. Using $s_{b,\tau} = [\nabla \mathsf{H}_\tau f_\circ + \nabla \mathsf{H}_\tau (f_b - f_\circ)]/g_{b,\tau}$ and (C.12),

$$\begin{aligned}\|s_{b,\tau}(x)\|_2 &\le \frac{4}{c_Q}\left[2c_Q\sqrt{\frac{d}{2\pi}}\,\tau^{-1/2}\exp\left(-\frac{\ell^2}{2\tau}\right) + c_Q\frac{\eta}{\rho}L_w\right] \\ &= C_1\tau^{-1/2}\exp\left(-\frac{\ell^2}{2\tau}\right) + C_2\frac{\eta}{\rho},\end{aligned}$$

which proves (C.13). □

We next state a standard off-diagonal heat-kernel estimate. A proof is included to make the localization step explicit.

Lemma C.4 (Off-diagonal gradient estimate)**.** *Let $A, B \subseteq \mathbb{R}^d$ be nonempty measurable sets, and write $\operatorname{dist}(A,B) := \inf_{x\in A,\, y\in B} \|x - y\|_2$. Suppose that $\operatorname{dist}(A,B) \ge r > 0$ and that $h \in L^2(\mathbb{R}^d)$ is supported in B. Then, for every $\tau > 0$,*

$$\|\mathbf{1}_A \nabla \mathsf{H}_\tau h\|_2 \le C_{\mathrm{off},d}\,\tau^{-1/2}\exp\left(-\frac{r^2}{16\tau}\right)\|h\|_2, \tag{C.14}$$

where $C_{\mathrm{off},d} := \sqrt{d}(4/3)^{d/4}$.

Proof. The integral kernel of $\nabla \mathsf{H}_\tau$ is $\nabla\phi_\tau(z) = -(z/\tau)\phi_\tau(z)$. For $x \in A$ and $y \in B$, $\|x-y\|_2 \geq r$. Let $Z \sim \mathcal{N}(0, I_d)$ and $R := r/\sqrt{\tau}$. The bound $\mathbf{1}_{\{\|Z\|_2 \geq R\}} \leq \exp((\|Z\|_2^2 - R^2)/16)$, followed by Cauchy–Schwarz, gives

$$\begin{aligned}\sup_{x\in A}\int_B \|\nabla\phi_\tau(x-y)\|_2\,\mathrm{d}y &\leq \int_{\|z\|_2\geq r}\frac{\|z\|_2}{\tau}\phi_\tau(z)\,\mathrm{d}z\\ &= \tau^{-1/2}\mathbb{E}\big[\|Z\|_2\mathbf{1}_{\{\|Z\|_2\geq R\}}\big]\\ &\leq \tau^{-1/2}e^{-R^2/16}(\mathbb{E}\|Z\|_2^2)^{1/2}(\mathbb{E}e^{\|Z\|_2^2/8})^{1/2}\\ &= C_{\mathrm{off},d}\tau^{-1/2}\exp\!\left(-\frac{r^2}{16\tau}\right),\end{aligned}$$

where the last line uses $\mathbb{E}\|Z\|_2^2 = d$, $\mathbb{E}e^{\|Z\|_2^2/8} = (4/3)^{d/2}$, and $R = r/\sqrt{\tau}$. The same bound holds with the supremum over $y \in B$ and the integral over $x \in A$. Thus both Schur integrals of the scalar majorant $\mathbf{1}_A(x)\mathbf{1}_B(y)\|\nabla\phi_\tau(x-y)\|_2$ are bounded by the right-hand coefficient in (C.14). Schur's test applied to this majorant proves (C.14). □

For $r > 0$, write $\log_+ r := \max\{0, \log r\}$. We now fix the remaining localization parameters by setting

$$R_{\mathrm{pers}} := \frac{4C_{\mathrm{off},d}^2 A_0 K_0^2}{A_1}, \qquad R_{\mathrm{sc}} := \frac{32A_0C_1^2K_0^2}{A_1}.$$

Define

$$\begin{aligned}\tau_\star &:= \min\left\{\frac{\ell^2}{16K_0^2}, \frac{\bar\tau}{2}, \frac{\ell^2}{16(1+\log_+ R_{\mathrm{pers}})}, \frac{\ell^2}{2(1+\log_+ R_{\mathrm{sc}})}\right\},\\ \eta_0 &:= \min\left\{\frac{1}{4}, \sqrt{\frac{A_1}{32A_0C_2^2}}\right\}.\end{aligned} \tag{C.15}$$

Both constants are positive and depend only on d, D. The four constraints defining $\tau_\star$, together with the displayed bound defining η_0, will be invoked below.

The key persistence estimate follows from the heat equation.

Lemma C.5 (Persistence of gradient energy)**.** *Let K_0 and $\tau_\star$ be given by (C.6) and (C.15). If $0 < \tau_0 \leq \tau_\star$, set $\rho = K_0\sqrt{\tau_0}$ and $h_{b,b'} := f_b - f_{b'}$. Then, for every $b, b' \in \{0,1\}^m$ and every $\tau \in [\tau_0, 2\tau_0]$,*

$$\int_I \|\nabla \mathsf{H}_\tau h_{b,b'}(x)\|_2^2\,\mathrm{d}x \geq \frac{c_Q^2 A_1}{2}\,\eta^2\rho^{d-2}d_{\mathrm{Ham}}(b,b').$$

Proof. Write $H = d_{\mathrm{Ham}}(b,b')$. The case $H = 0$ is immediate, so assume henceforth that $H \geq 1$. The disjoint supports of the bumps give the exact identities

$$\begin{aligned}\|\nabla h_{b,b'}\|_2^2 &= c_Q^2\eta^2 H\rho^{d-2}A_1,\\ \|\Delta h_{b,b'}\|_2^2 &= c_Q^2\eta^2 H\rho^{d-4}A_2.\end{aligned}$$

Since the bump supports lie strictly inside Q, $h_{b,b'} = c_Q\eta\sum_j(b_j - b'_j)w_{j,\rho} \in C_c^\infty$. Set $u_s := \mathsf{H}_s h_{b,b'}$ and $E(s) := \|\nabla u_s\|_2^2$. Since $\partial_s u_s = \frac{1}{2}\Delta u_s$, differentiation under the integral and integration by parts give

$$\begin{aligned}E'(s) &= 2\langle\nabla u_s, \nabla\partial_s u_s\rangle_{L^2} = \langle\nabla u_s, \nabla\Delta u_s\rangle_{L^2} = -\|\Delta u_s\|_2^2,\\ E(\tau) &= E(0) - \int_0^\tau \|\Delta \mathsf{H}_s h_{b,b'}\|_2^2\,\mathrm{d}s \geq E(0) - \tau\|\Delta h_{b,b'}\|_2^2.\end{aligned}$$

The last inequality uses $\Delta\mathsf{H}_s h_{b,b'} = \mathsf{H}_s\Delta h_{b,b'}$ and the L^2-contraction of H_s. Using the preceding identities, $\tau \leq 2\tau_0$, and $\rho^2 = K_0^2\tau_0$, we obtain

$$E(\tau) \geq c_Q^2\eta^2 H\rho^{d-2}\left(A_1 - \frac{2A_2}{K_0^2}\right).$$

By (C.7), the parenthesis is at least $3A_1/4$.

By (C.8),

$$\operatorname{supp}(w_{j,\rho}) \subset z_j + \rho[-1/2,1/2]^d \subset [-(\ell-\rho/2),\ell-\rho/2]^d.$$

The distance from this cube to I^c is $2\ell+\rho/2$, hence at least ℓ. Lemma C.4 therefore gives

$$\begin{aligned}\|\mathbf{1}_{I^c}\nabla \mathsf{H}_\tau h_{b,b'}\|_2^2 &\le C_{\mathrm{off},d}^2\tau^{-1}\exp\!\left(-\frac{\ell^2}{8\tau}\right)\|h_{b,b'}\|_2^2\\ &= C_{\mathrm{off},d}^2 c_Q^2\eta^2 H\rho^d\tau^{-1}\exp\!\left(-\frac{\ell^2}{8\tau}\right)A_0.\end{aligned}$$

Here $\rho^2/\tau \le K_0^2$ and $e^{-\ell^2/(8\tau)} \le e^{-\ell^2/(16\tau_0)}$. Moreover, the third constraint in (C.15) ensures that $\ell^2/(16\tau_0) \ge 1+\log_+ R_{\mathrm{pers}}$. Since $R_{\mathrm{pers}}e^{-(1+\log_+ R_{\mathrm{pers}})} \le 1$, it follows that $R_{\mathrm{pers}}e^{-\ell^2/(16\tau_0)} \le 1$. Consequently, the tail energy is at most $(c_Q^2A_1/4)\eta^2 H\rho^{d-2}$. Combining this estimate with the global lower bound yields

$$\begin{aligned}\int_I \|\nabla \mathsf{H}_\tau h_{b,b'}(x)\|_2^2\,\mathrm{d}x &= E(\tau) - \|\mathbf{1}_{I^c}\nabla \mathsf{H}_\tau h_{b,b'}\|_2^2\\ &\ge \left(\frac{3}{4}-\frac{1}{4}\right)c_Q^2A_1\eta^2H\rho^{d-2} = \frac{c_Q^2A_1}{2}\eta^2H\rho^{d-2},\end{aligned}$$

which proves the claim. □

We can now pass from the derivative of the smoothed density to the self-normalized score.

Lemma C.6 (Uniform score separation)**.** *Let $\tau_\star$ and η_0 be given by (C.15). If $0<\tau_0\le\tau_\star$, $0<\eta\le\eta_0$, and $\rho = K_0\sqrt{\tau_0}$, then, for every $b,b'\in\{0,1\}^m$ and every $\tau\in[\tau_0,2\tau_0]$,*

$$\int_I \|s_{b,\tau}(x)-s_{b',\tau}(x)\|_2^2\,\mathrm{d}x \ge \frac{2A_1}{25}\eta^2\rho^{d-2}d_{\mathrm{Ham}}(b,b').$$

Proof. Because $\tau\le 2\tau_0\le 2\tau_\star\le\bar\tau$, Lemma C.3 applies throughout the argument. Set $u := g_{b,\tau}$, $v := g_{b',\tau}$, and $\delta_\tau := u-v = \mathsf{H}_\tau(f_b-f_{b'})$. On I, the identity $\nabla\log u-\nabla\log v = (\nabla\delta_\tau - \delta_\tau\nabla\log v)/u$ and the upper density bound in (C.12) imply

$$\begin{aligned}\int_I \|s_{b,\tau}-s_{b',\tau}\|_2^2 &\ge \frac{16}{25c_Q^2}\int_I\|\nabla\delta_\tau-\delta_\tau s_{b',\tau}\|_2^2\\ &\ge \frac{8}{25c_Q^2}\|\nabla\delta_\tau\|_{L^2(I)}^2 - \frac{16}{25c_Q^2}\|\delta_\tau s_{b',\tau}\|_{L^2(I)}^2.\end{aligned}$$

Here we used $\|a-b\|_2^2\ge\frac12\|a\|_2^2-\|b\|_2^2$.

Writing $H = d_{\mathrm{Ham}}(b,b')$, the L^2-contraction of the heat semigroup gives $\|\delta_\tau\|_2^2 \le \|f_b-f_{b'}\|_2^2 = c_Q^2\eta^2H\rho^dA_0$. By Lemma C.3, the pointwise score bound, $(a+b)^2\le 2(a^2+b^2)$, and $\|\delta_\tau\|_{L^2(I)}\le\|\delta_\tau\|_2$,

$$\begin{aligned}\|\delta_\tau s_{b',\tau}\|_{L^2(I)}^2 &\le 2\left[C_1^2\tau^{-1}\exp\!\left(-\frac{\ell^2}{\tau}\right)+C_2^2\frac{\eta^2}{\rho^2}\right]\|\delta_\tau\|_2^2\\ &\le 2c_Q^2A_0\eta^2H\rho^{d-2}\left[C_1^2\frac{\rho^2}{\tau}\exp\!\left(-\frac{\ell^2}{\tau}\right)+C_2^2\eta^2\right].\end{aligned}$$

For $\tau\in[\tau_0,2\tau_0]$ and $\rho^2 = K_0^2\tau_0$, the first term in brackets is at most $C_1^2K_0^2\exp(-\ell^2/(2\tau_0))$. The fourth constraint in (C.15) and the definition of R_{sc} imply that this quantity is at most $A_1/(32A_0)$. The definition of η_0 gives the same upper bound for $C_2^2\eta^2$. It follows that $(16/(25c_Q^2))\|\delta_\tau s_{b',\tau}\|_{L^2(I)}^2 \le (2A_1/25)\eta^2H\rho^{d-2}$. On the other hand, Lemma C.5 shows that the positive term in the decomposition above is at least $(4A_1/25)\eta^2H\rho^{d-2}$. Subtracting proves the claim. □

Integrating the separation over the full window gives the scale needed by Fano's method.

Corollary C.7 (Integrated packing separation)**.** *Under the assumptions of Lemma C.6, suppose that*

$$d_{\mathrm{Ham}}(b, b') \geq \frac{m}{8}.$$

Then

$$\int_{\tau_0}^{2\tau_0} \int_I \|s_{b,\tau}(x) - s_{b',\tau}(x)\|_2^2 \,\mathrm{d}x \,\mathrm{d}\tau \geq c_{\mathrm{int}} \eta^2, \qquad c_{\mathrm{int}} := \frac{A_1 \ell^d}{100\, 2^d K_0^2}.$$

Proof. Integrate the bound in Lemma C.6 over $[\tau_0, 2\tau_0]$ and use $d_{\mathrm{Ham}}(b, b') \geq m/8$. Since this interval has length τ_0, we obtain

$$\int_{\tau_0}^{2\tau_0} \int_I \|s_{b,\tau} - s_{b',\tau}\|_2^2 \,\mathrm{d}x \,\mathrm{d}\tau \geq \frac{A_1}{100} \eta^2 \tau_0 m \rho^{d-2}.$$

Since $\rho^2 = K_0^2 \tau_0$ and $m\rho^d \geq 2^{-d}\ell^d$ by (C.9), the right-hand side is at least $c_{\mathrm{int}}\eta^2$, as claimed. □

C.4 Fano Reduction and Proof of Theorem 4.3

We use the Gilbert–Varshamov packing bound (Tsybakov, 2009, Lemma 2.9): for every $m \geq 8$, there exists a set $\mathcal{B} \subseteq \{0,1\}^m$, containing the zero vector, such that

$$\log |\mathcal{B}| \geq \frac{\log 2}{8} m, \qquad \min_{b \neq b' \in \mathcal{B}} d_{\mathrm{Ham}}(b, b') \geq \frac{m}{8}. \tag{C.16}$$

Below we use Fano's inequality (Tsybakov, 2009, Corollary 2.6).

For clarity, we collect the remaining constants before applying the two results. Set $\kappa := \min\{\eta_0^2, 8^d \log 2/(128 A_0)\}$, and define

$$\begin{aligned} c_0 &:= \log\left(1 + \min\left\{1,\ \tau_\star,\ \left(\frac{\ell}{K_0}\right)^2 \left(\frac{1}{16\, 2^d}\right)^{2/d}\right\}\right), \\ c_1 &:= \frac{7 c_Q c_{\mathrm{int}} \kappa}{256} \min\left\{1,\ 2^{-d}\left(\frac{\ell}{K_0}\right)^d\right\}. \end{aligned} \tag{C.17}$$

The constants κ, c_0, c_1 are strictly positive and depend only on d, D. The displayed c_0, c_1 are admissible choices in Theorem 4.3.

Proof of Theorem 4.3 Assume $0 < t_0 \leq c_0$. By (C.17), $\tau_0 = e^{t_0} - 1$ is bounded above by each term in the minimum defining c_0. In particular, $\tau_0 \leq \tau_\star$, and the first constraint in (C.15) gives $\rho = K_0\sqrt{\tau_0} \leq \ell/4$. Moreover, (C.9) and the other term in that minimum give $m \geq 2^{-d}(\ell/(K_0\sqrt{\tau_0}))^d \geq 16$, so the Gilbert–Varshamov bound applies. Let $\mathcal{B}$ be the packing in (C.16). Choose

$$\eta^2 := \kappa\Big(1 \wedge \frac{m}{n}\Big),$$

where κ is defined above; in particular, $0 < \eta \leq \eta_0$. By Lemma C.2, (C.9), and $m\rho^d \leq \ell^d$,

$$\max_{b \in \mathcal{B}} \mathrm{KL}(f_b^{\otimes n} \| f_\circ^{\otimes n}) \leq \frac{A_0}{8^d} n\eta^2 \leq \frac{A_0}{8^d} \kappa m \leq \frac{(\log 2) m}{128} \leq \frac{1}{16} \log |\mathcal{B}|. \tag{C.18}$$

Consider an arbitrary measurable, possibly randomized, vector-field estimator $\tilde{s}_n(x, \tau)$ based on the sample. Represent the estimator's internal randomization by a seed $U \sim \lambda$, independent of the sample. For $b \in \mathcal{B}$, let

$$\mathbb{P}_b := (f_b(x)\,\mathrm{d}x)^{\otimes n} \otimes \lambda, \qquad \mathbb{E}_b[\,\cdot\,] := \int (\,\cdot\,)\,\mathrm{d}\mathbb{P}_b.$$

Thus $\mathbb{P}_b$ is the augmented observation law under hypothesis b, and

$$\mathrm{KL}\big(f_b^{\otimes n} \otimes \lambda \,\|\, f_\circ^{\otimes n} \otimes \lambda\big) = \mathrm{KL}(f_b^{\otimes n} \| f_\circ^{\otimes n}).$$

Thus the calibration in (C.18) also applies to randomized estimators; all probabilities $\mathbb{P}_b$ and expectations $\mathbb{E}_b$ below include the seed U. Introduce the common semimetric

$$d_{\mathcal{I}}(r, r') := \left(\int_{\tau_0}^{2\tau_0} \int_I \|r(x,\tau) - r'(x,\tau)\|_2^2 \, \mathrm{d}x \, \mathrm{d}\tau \right)^{1/2}.$$

Let $\hat{b}$ be a nearest-neighbour decoder of $\tilde{s}_n$ in the finite family $\{s_{b,\cdot} : b \in \mathcal{B}\}$ with respect to $d_{\mathcal{I}}$, using a fixed deterministic rule to break ties. Since $\mathcal{B}$ is finite and each extended-real-valued map $d_{\mathcal{I}}(\tilde{s}_n, s_{b,\cdot})$, $b \in \mathcal{B}$, is measurable, so is $\hat{b}$. By Lemma C.3, for every $b \in \mathcal{B}$,

$$\int_{\tau_0}^{2\tau_0} \|\tilde{s}_n(\cdot,\tau) - s_{b,\tau}\|^2_{L^2(g_{b,\tau})} \, \mathrm{d}\tau \geq \frac{c_Q}{4} d^2_{\mathcal{I}}(\tilde{s}_n, s_{b,\cdot}). \tag{C.19}$$

If the distance on the right is infinite with positive $\mathbb{P}_b$-probability for some b, then the corresponding expected risk is infinite and the desired lower bound is immediate. Hence we may assume that $d_{\mathcal{I}}(\tilde{s}_n, s_{b,\cdot}) < \infty$ $\mathbb{P}_b$-almost surely for every $b \in \mathcal{B}$.

By Corollary C.7, the common separation satisfies

$$d^2_{\mathcal{I}}(s_{b,\cdot}, s_{b',\cdot}) \geq c_{\mathrm{int}} \eta^2 \qquad (b \neq b').$$

If $\hat{b} \neq b$, the triangle inequality implies

$$\sqrt{c_{\mathrm{int}}}\, \eta \leq d_{\mathcal{I}}(s_{\hat{b},\cdot}, s_{b,\cdot}) \leq d_{\mathcal{I}}(s_{\hat{b},\cdot}, \tilde{s}_n) + d_{\mathcal{I}}(\tilde{s}_n, s_{b,\cdot}) \leq 2 d_{\mathcal{I}}(\tilde{s}_n, s_{b,\cdot}),$$

where the last inequality follows from the nearest-neighbour property. Therefore $d^2_{\mathcal{I}}(\tilde{s}_n, s_{b,\cdot}) \geq (c_{\mathrm{int}}/4)\eta^2$. Since $m \geq 16$, (C.16) gives $\log|\mathcal{B}| \geq (\log 2)m/8 \geq 2\log 2$. Fano's inequality and (C.18) therefore give

$$\inf_{\hat{b}} \sup_{b \in \mathcal{B}} \mathbb{P}_b(\hat{b} \neq b) \geq 1 - \frac{1}{16} - \frac{\log 2}{\log|\mathcal{B}|} \geq \frac{7}{16}. \tag{C.20}$$

Combining the decoding bound above with (C.19) gives, under hypothesis b,

$$\int_{\tau_0}^{2\tau_0} \|\tilde{s}_n(\cdot,\tau) - s_{b,\tau}\|^2_{L^2(g_{b,\tau})} \, \mathrm{d}\tau \geq \frac{c_Q c_{\mathrm{int}}}{16} \eta^2 \mathbf{1}_{\{\hat{b} \neq b\}}.$$

Taking $\mathbb{E}_b$, then the supremum over b, and using (C.20) yields

$$\sup_{b \in \mathcal{B}} \mathbb{E}_b \int_{\tau_0}^{2\tau_0} \|\tilde{s}_n(\cdot,\tau) - s_{b,\tau}\|^2_{L^2(g_{b,\tau})} \, \mathrm{d}\tau \geq \frac{7 c_Q c_{\mathrm{int}}}{256} \eta^2.$$

The lower half of (C.9) and $\rho = K_0\sqrt{\tau_0}$ give $m \geq 2^{-d}(\ell/K_0)^d \tau_0^{-d/2}$. Consequently, using the elementary inequality $1 \wedge (uv) \geq (1 \wedge u)(1 \wedge v)$ for $u, v > 0$,

$$\eta^2 = \kappa\Big(1 \wedge \frac{m}{n}\Big) \geq \kappa \min\left\{1, 2^{-d}\left(\frac{\ell}{K_0}\right)^d\right\}\left(1 \wedge \frac{1}{n\tau_0^{d/2}}\right).$$

Now take an arbitrary OU score estimator $\hat{s}_n$ and transform it into $\tilde{s}_n$ by (C.3). Since every f_b, $b \in \mathcal{B}$, belongs to $\mathcal{P}_D$, restricting first the supremum to this finite family and then the time integral to $[t_0, t_1]$ gives, by the exact identity (C.4),

$$\begin{aligned}
&\sup_{p_0 \in \mathcal{P}_D} \mathbb{E}_{p_0^{\otimes n} \otimes \lambda} \int_{t_0}^{T} \|\hat{s}_n(\cdot,t) - \nabla \log p_t\|^2_{L^2(p_t)} \mathrm{d}t \\
&\quad \geq \sup_{b \in \mathcal{B}} \mathbb{E}_b \int_{\tau_0}^{2\tau_0} \|\tilde{s}_n(\cdot,\tau) - s_{b,\tau}\|^2_{L^2(g_{b,\tau})} \mathrm{d}\tau \\
&\quad \geq c_1 \left(1 \wedge \frac{1}{n\tau_0^{d/2}}\right).
\end{aligned}$$

Combining this display with (C.1), dividing by $T - t_0$, and then taking the infimum over $\hat{s}_n$ proves (4.7).

In the sample-rich regime $n\tau_0^{d/2} \geq 1$, the minimum equals its second argument. □

C.5 Fixed-Time Density Separation at the Heat Scale

We now reuse the family from Subsection C.2 for the fixed-time KL problem. Unlike the integrated score lower bound, the present argument does not require localization to the interior cube I or persistence over an interval of heat times. It suffices to show that the L^2 distance between the smoothed densities remains visible at the single time τ_0.

For probability measures μ and ν, define the squared Hellinger distance by

$$\mathsf{h}^2(\mu,\nu) := \int \left(\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}} - \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}}\right)^2 \mathrm{d}\lambda, \qquad \lambda := \mu + \nu.$$

This definition is independent of the dominating measure, is invariant under a common bimeasurable bijection, and satisfies

$$\mathrm{KL}(\mu\|\nu) \geq \mathsf{h}^2(\mu,\nu). \tag{C.21}$$

We identify a probability density with its induced measure whenever no confusion can arise. To verify (C.21), the inequality is immediate if $\mu \not\ll \nu$. Otherwise, with

$$A(\mu,\nu) := \int \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}}\,\mathrm{d}\lambda,$$

Jensen's inequality gives $\mathrm{KL}(\mu\|\nu) \geq -2\log A(\mu,\nu)$. Now use $-\log x \geq 1-x$ and $\mathsf{h}^2(\mu,\nu) = 2[1-A(\mu,\nu)]$.

Recall the notation $g_{b,\tau} := \mathsf{H}_\tau f_b$ from Subsection C.3.

Lemma C.8 (Fixed-time L^2 persistence). *Let K_0 be defined by (C.6), set $\rho = K_0\sqrt{\tau_0}$, and assume that $\rho \leq \ell/4$ and $0 < \eta \leq 1/4$. Then, for every $b, b' \in \{0,1\}^m$,*

$$\|g_{b,\tau_0} - g_{b',\tau_0}\|_2^2 \geq \frac{c_Q^2 A_0}{2}\,\eta^2\rho^d d_{\mathrm{Ham}}(b,b'). \tag{C.22}$$

Proof. Set $r := f_b - f_{b'}$ and $H := d_{\mathrm{Ham}}(b,b')$. The disjoint bump supports give

$$\|r\|_2^2 = c_Q^2\eta^2 H\rho^d A_0, \qquad \|\nabla r\|_2^2 = c_Q^2\eta^2 H\rho^{d-2}A_1.$$

Since $r \in C_c^\infty(\mathbb{R}^d)$, the function $F(s) := \|\mathsf{H}_s r\|_2^2$ satisfies

$$F'(s) = 2\left\langle \mathsf{H}_s r, \frac{1}{2}\Delta\mathsf{H}_s r\right\rangle_{L^2} = -\|\nabla\mathsf{H}_s r\|_2^2.$$

The heat semigroup commutes with derivatives and is an L^2 contraction, so

$$\|\mathsf{H}_{\tau_0} r\|_2^2 \geq \|r\|_2^2 - \tau_0\|\nabla r\|_2^2 = c_Q^2\eta^2 H\rho^d\left(A_0 - \frac{A_1}{K_0^2}\right). \tag{C.23}$$

Integration by parts and Cauchy–Schwarz give $A_1 = -\int w\Delta w \leq \sqrt{A_0 A_2}$. Moreover, (C.6) gives

$$\frac{A_1}{K_0^2} \leq \frac{A_1^2}{8A_2} \leq \frac{A_0}{8}.$$

Thus the parenthesis in (C.23) is at least $7A_0/8$, which implies (C.22). □

Corollary C.9 (Fixed-time Hellinger packing). *Suppose that $\rho \leq \ell/4$ and $d_{\mathrm{Ham}}(b,b') \geq m/8$. Then*

$$\mathsf{h}^2(g_{b,\tau_0}, g_{b',\tau_0}) \geq c_{\mathrm{H},d,D}\,\eta^2, \qquad c_{\mathrm{H},d,D} := \frac{c_Q A_0 \ell^d}{80\,2^d} > 0.$$

The same bound holds for the corresponding OU marginals $p_{b,t_0} := P_{t_0}(f_b(x)\,\mathrm{d}x)$.

Proof. The L^∞ contraction of the heat semigroup and $0 \leq f_b \leq 5c_Q/4$ give

$$0 \leq g_{b,\tau_0}(x) \leq \frac{5c_Q}{4} \qquad (x \in \mathbb{R}^d).$$

Therefore,

$$\mathsf{h}^2(g_{b,\tau_0}, g_{b',\tau_0}) = \int \frac{(g_{b,\tau_0} - g_{b',\tau_0})^2}{(\sqrt{g_{b,\tau_0}} + \sqrt{g_{b',\tau_0}})^2}\,\mathrm{d}x \geq \frac{1}{5c_Q}\|g_{b,\tau_0} - g_{b',\tau_0}\|_2^2.$$

Lemma C.8, the Hamming separation, and (C.9) yield

$$\mathsf{h}^2(g_{b,\tau_0}, g_{b',\tau_0}) \geq \frac{c_Q A_0}{10}\eta^2\rho^d d_{\mathrm{Ham}}(b,b') \geq \frac{c_Q A_0}{80}\eta^2 m\rho^d \geq \frac{c_Q A_0 \ell^d}{80\,2^d}\eta^2.$$

Finally, with $a_{t_0} = e^{-t_0/2}$, let $S_{a_{t_0}}(x) := a_{t_0}x$. Then (C.2) gives $p_{b,t_0} = (S_{a_{t_0}})_{\#}(g_{b,\tau_0}(x)\,\mathrm{d}x)$. The invariance of Hellinger distance under this common invertible dilation proves the last claim. □

C.6 Fixed-Time KL Lower Bound and Nominal Minimax Rate

We first record the stronger lower bound that includes the sample-poor truncation.

With $c_{\mathrm{H},d,D}$ defined in Corollary C.9, κ as in Subsection C.4, and ℓ, K_0 as in Subsection C.2, set

$$c_{d,D}^{\mathrm{KL}} := \frac{7c_{\mathrm{H},d,D}\kappa}{64}\min\left\{1,\, 2^{-d}\left(\frac{\ell}{K_0}\right)^d\right\}. \tag{C.24}$$

This constant is strictly positive and depends only on d, D.

Proposition C.10 (Fixed-time statistical KL lower bound)**.** *Let c_0 be defined by (C.17), and let $c_{d,D}^{\mathrm{KL}}$ be defined by (C.24). Then, for every $0 < t_0 \leq c_0$ and $n \geq 1$,*

$$\inf_{\hat{\mu}_n}\sup_{p_0\in\mathcal{P}_D}\mathbb{E}\,\mathrm{KL}(p_{t_0}\|\hat{\mu}_n) \geq c_{d,D}^{\mathrm{KL}}\left(1 \wedge \frac{1}{n\tau_0^{d/2}}\right), \tag{C.25}$$

where the infimum is over all measurable, possibly randomized, probability-measure estimators based on n i.i.d. observations from p_0; the expectation includes both sample and estimator randomness.

Proof. Assume $0 < t_0 \leq c_0$. By (C.17), $\rho = K_0\sqrt{\tau_0} \leq \ell/4$ and $m \geq 16$. Let $\mathcal{B} \subset \{0,1\}^m$ be the Gilbert–Varshamov packing in (C.16), and choose

$$\eta^2 := \kappa\Big(1 \wedge \frac{m}{n}\Big),$$

with κ as above. The calculation leading to (C.18) applies verbatim and gives

$$\max_{b\in\mathcal{B}}\mathrm{KL}(f_b^{\otimes n}\|f_\circ^{\otimes n}) \leq \frac{1}{16}\log|\mathcal{B}|. \tag{C.26}$$

Consider an arbitrary measurable, possibly randomized, probability-measure estimator $\hat{\mu}_n$ based on the sample. Represent its internal randomization by a seed $U \sim \lambda$, independent of the sample, and define

$$\mathbb{P}_b := (f_b(x)\,\mathrm{d}x)^{\otimes n} \otimes \lambda, \qquad b \in \mathcal{B}.$$

Let $\mathbb{E}_b$ denote expectation under $\mathbb{P}_b$. Decode among the OU marginals by

$$\hat{b} \in \arg\min_{c\in\mathcal{B}} \mathsf{h}(\hat{\mu}_n, p_{c,t_0}),$$

using a fixed deterministic rule to break ties. Equip $\mathcal{P}(\mathbb{R}^d)$ with the Borel σ-field induced by weak convergence. Since Hellinger distance is lower semicontinuous, each map $\mu \mapsto \mathsf{h}(\mu, p_{c,t_0})$ is Borel measurable. Because $\mathcal{B}$ is finite, the nearest-neighbour decoder with deterministic tie-breaking is measurable. Corollary C.9 gives, for distinct packing elements,

$$\mathsf{h}^2(p_{b,t_0}, p_{b',t_0}) \geq c_{\mathrm{H},d,D}\eta^2.$$

If $\hat{b} \neq b$, the triangle inequality and the nearest-neighbour property imply

$$\sqrt{c_{\mathrm{H},d,D}}\,\eta \leq \mathsf{h}(p_{b,t_0}, p_{\hat{b},t_0}) \leq 2\mathsf{h}(p_{b,t_0}, \hat{\mu}_n).$$

Consequently, (C.21) yields

$$\mathrm{KL}(p_{b,t_0}\|\hat{\mu}_n) \geq \frac{c_{\mathrm{H},d,D}}{4}\eta^2 \mathbf{1}_{\{\hat{b}\neq b\}}. \tag{C.27}$$

The Fano calculation in (C.20) applies without change because (C.26) is the same information calibration. Taking expectations in (C.27) therefore gives

$$\sup_{b\in\mathcal{B}} \mathbb{E}_b \,\mathrm{KL}(p_{b,t_0}\|\hat{\mu}_n) \geq \frac{7c_{\mathrm{H},d,D}}{64}\eta^2. \tag{C.28}$$

By (C.9), $\rho = K_0\sqrt{\tau_0}$, and $1\wedge(uv) \geq (1\wedge u)(1\wedge v)$ for $u, v > 0$,

$$\eta^2 \geq \kappa \min\left\{1, 2^{-d}\left(\frac{\ell}{K_0}\right)^d\right\}\left(1\wedge\frac{1}{n\tau_0^{d/2}}\right). \tag{C.29}$$

Every f_b is the density of a member of $\mathcal{P}_D$. Combining (C.28) and (C.29), restricting the supremum to this finite family, and then taking the infimum over $\hat{\mu}_n$, with $c_{d,D}^{\mathrm{KL}}$ as in (C.24), proves (C.25). □

Theorem C.11 (Nominal positive-time KL minimax rate)**.** *Fix $d \geq 1$ and $D > 0$. Let c_0 and $c_{d,D}^{\mathrm{KL}}$ be defined by (C.17) and (C.24), respectively. If $0 < t_0 \leq c_0$ and $n\tau_0^{d/2} \geq 1$, then*

$$\frac{c_{d,D}^{\mathrm{KL}}}{n\tau_0^{d/2}} \leq \mathfrak{K}_{n,0}^{\star}(D;t_0) \leq \frac{C_{d,D}}{2n\tau_0^{d/2}}. \tag{C.30}$$

Proof. The lower bound follows from Proposition C.10; under $n\tau_0^{d/2} \geq 1$, the minimum in (C.25) equals its second argument. The upper bound follows from (6.1) with $\varepsilon = 0$, followed by $n+1 \geq n$. This proves (C.30). □

D Proofs for Shift Lower Bounds and Minimax Optimality

This appendix proves the shift lower bounds and completes the robust score and KL minimax results. Subsections D.1– D.3 develop the one-dimensional oscillatory construction through Wasserstein admissibility, two-point score separation, and persistence under OU smoothing. Subsection D.4 lifts the construction to arbitrary fixed dimension and proves Theorems 3.2 and 4.4. Finally, Subsection D.5 proves the positive-time KL shift lower bound by a two-point Gaussian argument and combines it with the nominal KL lower bound from Appendix C to prove Theorem 6.1.

D.1 One-Dimensional Construction and Lower-Bound Statement

Throughout the one-dimensional argument, write $p_0 := p_0^{(1)}$.

Consider the one-dimensional Ornstein–Uhlenbeck process $\mathrm{d}X_t = -\frac{1}{2}X_t\,\mathrm{d}t + \mathrm{d}W_t$. Its explicit solution is

$$X_t = a_t X_0 + \sigma_t Z, \qquad a_t = e^{-t/2}, \qquad \sigma_t^2 = 1 - e^{-t}, \qquad Z \sim \mathcal{N}(0,1).$$

We use below the elementary bound $\sigma_t^2 \leq t$.

Let the reference density be $p_0(x) := \frac{1}{2}\mathbf{1}_{[-1,1]}(x)$. For an integer $k \geq 1$, set $\lambda_k := \pi k$, and $h_k(x) := \cos(\lambda_k x)$. Since

$$\int_{-1}^{1} h_k(x)\,\mathrm{d}x = \int_{-1}^{1} \cos(\pi k x)\,\mathrm{d}x = \frac{2\sin(\pi k)}{\pi k} = 0,$$

the following perturbations preserve total mass. For $0 < \alpha \leq 1/2$, define

$$q_0^{\pm}(x) := \frac{1}{2}\big(1 \pm \alpha h_k(x)\big)\mathbf{1}_{[-1,1]}(x).$$

Because $|h_k| \le 1$ and $\alpha \le 1/2$, these are probability densities, and $\mathrm{supp}(q_0^{\pm}) = \mathrm{supp}(p_0) = [-1,1]$. Thus the alternatives are compactly supported.

For $t > 0$, let

$$K_t(x,y) := \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\frac{(x-a_t y)^2}{2\sigma_t^2}\right)$$

denote the OU transition density. The smoothed reference density is

$$p_t(x) = \frac{1}{2}\int_{-1}^{1} K_t(x,y)\,\mathrm{d}y.$$

Similarly, $q_t^{\pm}(x) = p_t(x) \pm \alpha r_t(x)$, where $r_t(x) := \frac{1}{2}\int_{-1}^{1} h_k(y)K_t(x,y)\,\mathrm{d}y$. Define $\eta_t(x) := r_t(x)/p_t(x)$. Then

$$q_t^{\pm}(x) = p_t(x)\big(1 \pm \alpha\eta_t(x)\big).$$

Under the joint law induced by $X_0 \sim p_0$ and the OU transition, $\eta_t(x) = \mathbb{E}[h_k(X_0) \mid X_t = x]$, so $|\eta_t(x)| \le 1$ and $1 \pm \alpha\eta_t(x) \ge 1 - \alpha \ge 1/2$. Thus $q_t^{\pm}$ are strictly positive for every $t > 0$, and their scores are well defined on all of $\mathbb{R}$.

The preceding construction underlies the following one-dimensional lower bound, whose proof is given after the auxiliary estimates.

Proposition D.1. *Let $p_0(x) = \frac{1}{2}\mathbf{1}_{[-1,1]}(x)$, and let P_t denote the semigroup of the one-dimensional Ornstein–Uhlenbeck process $\mathrm{d}X_t = -\frac{1}{2}X_t\,\mathrm{d}t + \mathrm{d}W_t$. Fix $0 < t_0 \le 1/(16\pi^2)$, $2t_0 \le T$, and $0 < \varepsilon \le \sqrt{t_0}$. For each $q_0 \in \mathcal{Q}_\varepsilon(p_0)$, set $q_t = P_t q_0$. Then*

$$\inf_{\hat{s}} \sup_{q_0 \in \mathcal{Q}_\varepsilon(p_0)} \mathbb{E}[\mathcal{L}_{q_0}(\hat{s})] \ge \frac{e^{-4/9}}{2592}\,\frac{\varepsilon^2}{t_0(T-t_0)}. \tag{D.1}$$

The infimum is over all possibly random score estimators whose realizations are measurable maps from $\mathbb{R} \times [t_0, T]$ to $\mathbb{R}$, and the expectation is over all randomness used to construct $\hat{s}$.

D.2 Auxiliary Estimates

We next establish the three estimates used in the proof of Proposition D.1. The Wasserstein estimate verifies admissibility of the alternatives, the two-point inequality reduces the score risk to the energy of η_t, and the persistence estimate controls this energy over the early-time window.

D.2.1 Wasserstein Size of the Perturbation

Lemma D.2. *For every $k \ge 1$ and every $0 < \alpha \le 1/2$,*

$$\mathsf{W}_2(p_0, q_0^{\pm}) \le \frac{\alpha}{\lambda_k}.$$

Proof. We prove the claim for q_0^{+}, and the proof for q_0^{-} is identical after reversing the sign. For $s \in [0,1]$, define the interpolating density and its flux by

$$\rho_s(x) := \frac{1}{2}\big(1 + s\alpha\cos(\lambda_k x)\big)\mathbf{1}_{[-1,1]}(x),$$
$$\rho_s(x)v_s(x) := -\frac{\alpha}{2\lambda_k}\sin(\lambda_k x)\mathbf{1}_{[-1,1]}(x).$$

Since $\lambda_k = \pi k$, $\sin(\lambda_k) = \sin(\pi k) = 0$. Thus the flux vanishes at the boundary points ± 1. Moreover,

$$\partial_s \rho_s(x) = \frac{\alpha}{2}\cos(\lambda_k x), \qquad \partial_x(\rho_s v_s)(x) = -\frac{\alpha}{2}\cos(\lambda_k x).$$

Thus $\partial_s\rho_s + \partial_x(\rho_s v_s) = 0$ on $[-1,1]$. After extending both ρ_s and $\rho_s v_s$ by zero outside $[-1,1]$, the same continuity equation holds on $\mathbb{R}$ in the distributional sense because the flux vanishes at ± 1. By the Benamou–Brenier formula,

$$\mathsf{W}_2^2(p_0, q_0^{+}) \le \int_0^1 \int_{-1}^1 |v_s(x)|^2 \rho_s(x)\,\mathrm{d}x\,\mathrm{d}s = \int_0^1 \int_{-1}^1 \frac{\alpha^2}{4\lambda_k^2}\,\frac{\sin^2(\lambda_k x)}{\rho_s(x)}\,\mathrm{d}x\,\mathrm{d}s.$$

Since $\alpha \le 1/2$, $\rho_s(x) \ge (1-\alpha)/2 \ge 1/4$ for all $s \in [0,1]$ and $x \in [-1,1]$. Therefore

$$\mathsf{W}_2^2(p_0, q_0^+) \le \frac{\alpha^2}{\lambda_k^2} \int_{-1}^{1} \sin^2(\lambda_k x)\,\mathrm{d}x = \frac{\alpha^2}{\lambda_k^2}.$$

Taking square roots proves the claim. □

D.2.2 Two-Point Score Separation

Lemma D.3. *Assume $0 < \alpha \le 1/2$. For every measurable $s : \mathbb{R}\times[t_0,T] \to \mathbb{R}$ and every $t \in [t_0,T]$, define*

$$R_\xi(t,s) := \mathbb{E}_{q_t^\xi}\Big[|s(X,t) - \partial_x \log q_t^\xi(X)|^2\Big], \qquad \xi \in \{+,-\}.$$

The two risks satisfy

$$R_+(t,s) + R_-(t,s) \ge 2\alpha^2 \int_{\mathbb{R}} |\partial_x \eta_t(x)|^2 p_t(x)\,\mathrm{d}x.$$

Proof. For any $\lambda, \mu > 0$ and $a, b, z \in \mathbb{R}$, $\lambda|z-a|^2 + \mu|z-b|^2 \ge \lambda\mu|a-b|^2/(\lambda+\mu)$. Apply this pointwise with $\lambda = q_t^+(x)$, $\mu = q_t^-(x)$, $a = \partial_x \log q_t^+(x)$, $b = \partial_x \log q_t^-(x)$, and $z = s(x,t)$. This gives

$$R_+(t,s) + R_-(t,s) \ge \int_{\mathbb{R}} \frac{q_t^+(x) q_t^-(x)}{q_t^+(x) + q_t^-(x)} \big|\partial_x \log q_t^+(x) - \partial_x \log q_t^-(x)\big|^2\,\mathrm{d}x.$$

Using $q_t^\pm = p_t(1 \pm \alpha\eta_t)$, we have $\partial_x \log q_t^\pm = \partial_x \log p_t + \partial_x \log(1 \pm \alpha\eta_t)$. Moreover,

$$\partial_x \log q_t^+ - \partial_x \log q_t^- = \frac{2\alpha\,\partial_x \eta_t}{1-\alpha^2\eta_t^2},$$

$$\frac{q_t^+ q_t^-}{q_t^+ + q_t^-} = \frac{1}{2} p_t(1-\alpha^2\eta_t^2).$$

Substitution yields

$$R_+(t,s) + R_-(t,s) \ge 2\alpha^2 \int_{\mathbb{R}} \frac{|\partial_x \eta_t(x)|^2}{1-\alpha^2\eta_t(x)^2} p_t(x)\,\mathrm{d}x.$$

Since $|\eta_t| \le 1$ and $\alpha \le 1/2$, we have $(1-\alpha^2\eta_t^2)^{-1} \ge 1$, which proves the result. □

D.2.3 Persistence of One Oscillatory Mode

We next show that an oscillatory mode remains visible during the early-time window whenever its frequency is below the critical scale $t_0^{-1/2}$.

Set $u := x/a_t$ and $\tau_t := \sigma_t^2/a_t^2 = e^t - 1$, so $\sqrt{\tau_t} = \sigma_t/a_t$. After the change of variables $x = a_t u$, define

$$p_t(x) = \frac{1}{2a_t} A_t(u), \qquad A_t(u) := \int_{-1}^{1} \frac{1}{\sqrt{2\pi\tau_t}} \exp\left(-\frac{(u-y)^2}{2\tau_t}\right)\mathrm{d}y,$$

$$r_t(x) = \frac{1}{2a_t} B_t(u), \qquad B_t(u) := \int_{-1}^{1} \cos(\lambda_k y) \frac{1}{\sqrt{2\pi\tau_t}} \exp\left(-\frac{(u-y)^2}{2\tau_t}\right)\mathrm{d}y,$$

$$\eta_t(x) = \frac{B_t(u)}{A_t(u)}.$$

Primes below denote derivatives with respect to u.

Lemma D.4. *Assume $k \ge 1$ and that the small-time and frequency–time conditions*

$$0 < t \le \frac{1}{64}, \qquad \lambda_k^2 t \le 1$$

hold. Then

$$\int_{\mathbb{R}} |\partial_x \eta_t(x)|^2 p_t(x)\,\mathrm{d}x \ge \frac{1}{32}\lambda_k^2 \exp\big(-2\lambda_k^2\sigma_t^2\big). \tag{D.2}$$

Proof. We work on the central interval $|u| \le 1/2$.

Step 1: Bounds for the truncated Gaussian mass. For $0 < t \le 1/64$, the elementary inequality $e^t \le (1-t)^{-1}$ gives

$$\tau_t = e^t - 1 \le \frac{t}{1-t} \le \frac{1}{63} < \frac{1}{8\log 4}.$$

Thus $\tau_t \le 1$. For $|u| \le 1/2$, the Gaussian mass omitted from $A_t(u)$ outside $[-1,1]$ lies at distance at least $1/(2\sqrt{\tau_t})$ in standard normal units. Hence, the one-dimensional Gaussian tail bound gives

$$\frac{1}{2} \le A_t(u) \le 1, \qquad |u| \le 1/2, \tag{D.3}$$

and

$$|A_t'(u)| \le \frac{2}{\sqrt{2\pi}} \tau_t^{-1/2} \exp\left(-\frac{1}{8\tau_t}\right), \qquad |u| \le 1/2. \tag{D.4}$$

Step 2: Comparison with the full Gaussian convolution. Let

$$\mathcal{B}_t(u) := \int_{\mathbb{R}} \cos(\lambda_k y) \frac{1}{\sqrt{2\pi\tau_t}} \exp\left(-\frac{(u-y)^2}{2\tau_t}\right) \mathrm{d}y.$$

The Gaussian convolution of a cosine can be computed explicitly, as $\mathcal{B}_t(u) = e^{-\lambda_k^2\tau_t/2}\cos(\lambda_k u)$, and hence $\mathcal{B}_t'(u) = -\lambda_k e^{-\lambda_k^2\tau_t/2}\sin(\lambda_k u)$. The difference between B_t and $\mathcal{B}_t$ comes only from the Gaussian tails outside $[-1,1]$. For $|u| \le 1/2$, those tails are at distance at least $1/(2\sqrt{\tau_t})$ in standard normal units. Using $|\cos(\lambda_k y)| \le 1$, differentiating the Gaussian kernel, and $\tau_t \le 1$, the same boundary-tail estimate gives

$$|B_t'(u) - \mathcal{B}_t'(u)| \le \frac{2}{\sqrt{2\pi}} \tau_t^{-1} \exp\left(-\frac{1}{8\tau_t}\right), \qquad |u| \le 1/2. \tag{D.5}$$

Step 3: Lower bound for the derivative of B_t/A_t. Since

$$\partial_u\left(\frac{B_t}{A_t}\right) = \frac{B_t'}{A_t} - \frac{B_t A_t'}{A_t^2} = \frac{\mathcal{B}_t'}{A_t} + \frac{B_t' - \mathcal{B}_t'}{A_t} - \frac{B_t A_t'}{A_t^2},$$

we obtain

$$\left|\partial_u\left(\frac{B_t}{A_t}\right)\right| \ge \frac{|\mathcal{B}_t'|}{A_t} - \frac{|B_t' - \mathcal{B}_t'|}{A_t} - \frac{|B_t||A_t'|}{A_t^2}.$$

Because $A_t \le 1$, we have $|\mathcal{B}_t'|/A_t \ge |\mathcal{B}_t'|$. Moreover, $|B_t| \le A_t$, and (D.3), (D.4), and (D.5) imply

$$\frac{|B_t' - \mathcal{B}_t'|}{A_t} + \frac{|B_t||A_t'|}{A_t^2} \le 4\tau_t^{-1} \exp\left(-\frac{1}{8\tau_t}\right).$$

Now define $\delta_t := 4\tau_t^{-1}\exp(-1/(8\tau_t))$. For $|u| \le 1/2$, it holds that

$$\left|\partial_u\left(\frac{B_t}{A_t}\right)(u)\right| \ge \left(\lambda_k e^{-\lambda_k^2\tau_t/2}|\sin(\lambda_k u)| - \delta_t\right)_+.$$

For $r, s \ge 0$, the elementary inequality $(r-s)_+^2 \ge \frac{1}{2}r^2 - s^2$ holds. Hence

$$\int_{-1/2}^{1/2} \left|\partial_u\left(\frac{B_t}{A_t}\right)(u)\right|^2 \mathrm{d}u \ge \frac{1}{2}\lambda_k^2 e^{-\lambda_k^2\tau_t} \int_{-1/2}^{1/2} \sin^2(\lambda_k u)\,\mathrm{d}u - \delta_t^2.$$

Since $\lambda_k = \pi k$, $\int_{-1/2}^{1/2}\sin^2(\lambda_k u)\,\mathrm{d}u = 1/2$. Moreover, since $0 < t \le 1/64 < \log 2$ and $\lambda_k^2 t \le 1$, it follows that

$$\lambda_k^2\tau_t = \lambda_k^2(e^t - 1) \le 2\lambda_k^2 t \le 2.$$

Since $k \ge 1$,

$$\lambda_k^2 e^{-\lambda_k^2\tau_t} \ge \lambda_k^2 e^{-2} \ge \pi^2 e^{-2}.$$

On the other hand, set $g(y) := 16y^{-2}\exp(-1/(4y))$. This function is increasing on $(0, 1/8]$, since

$$\frac{g'(y)}{g(y)} = \frac{1-8y}{4y^2} \ge 0.$$

As $\tau_t \le 1/63 < 1/8$, we obtain the explicit bound

$$\delta_t^2 = g(\tau_t) \le g(1/63) = 16 \cdot 63^2 e^{-63/4} < \frac{\pi^2}{8} e^{-2}.$$

Consequently,

$$\delta_t^2 \le \frac{\pi^2}{8} e^{-2} \le \frac{1}{8} \lambda_k^2 e^{-\lambda_k^2 \tau_t}$$

uniformly over all $k \ge 1$ satisfying $\lambda_k^2 t \le 1$. Therefore,

$$\int_{-1/2}^{1/2} \left| \partial_u \left(\frac{B_t}{A_t} \right)(u) \right|^2 \mathrm{d}u \ge \frac{1}{8} \lambda_k^2 e^{-\lambda_k^2 \tau_t}.$$

Step 4: Return to the original spatial variable. Since $x = a_t u$, we have

$$\partial_x \eta_t(a_t u) = a_t^{-1} \partial_u \left(\frac{B_t}{A_t} \right)(u), \qquad p_t(a_t u) a_t \, \mathrm{d}u = \frac{1}{2} A_t(u) \, \mathrm{d}u.$$

Using $A_t(u) \ge 1/2$ and $a_t^{-2} \ge 1$, we get

$$\begin{aligned} \int_{\mathbb{R}} |\partial_x \eta_t(x)|^2 p_t(x) \, \mathrm{d}x &\ge \int_{-1/2}^{1/2} |\partial_x \eta_t(a_t u)|^2 p_t(a_t u) a_t \, \mathrm{d}u \\ &\ge \frac{1}{4} \int_{-1/2}^{1/2} \left| \partial_u \left(\frac{B_t}{A_t} \right)(u) \right|^2 \mathrm{d}u \ge \frac{1}{32} \lambda_k^2 e^{-\lambda_k^2 \tau_t}. \end{aligned}$$

Finally, since $0 < t \le 1/64$, $\tau_t = e^t \sigma_t^2 \le 2\sigma_t^2$. Thus,

$$\int_{\mathbb{R}} |\partial_x \eta_t(x)|^2 p_t(x) \, \mathrm{d}x \ge \frac{1}{32} \lambda_k^2 e^{-\lambda_k^2 \tau_t} \ge \frac{1}{32} \lambda_k^2 e^{-2\lambda_k^2 \sigma_t^2},$$

which proves the claimed lower bound (D.2). □

D.3 Proof of Proposition D.1

Proof. Since $\pi^2 > 8$, the assumption $t_0 \le 1/(16\pi^2)$ implies $t_0 < \frac{1}{128}$. We select the largest integer frequency that is uniformly compatible with the amplitude constraint under $\varepsilon \le \sqrt{t_0}$. Let

$$\kappa_0 := \frac{1}{2\pi\sqrt{t_0}}, \qquad k := \lfloor \kappa_0 \rfloor, \qquad \lambda_k = \pi k.$$

The assumption on t_0 gives $\kappa_0 \ge 2$. For every $\kappa_0 \ge 2$, $2\kappa_0/3 \le \lfloor \kappa_0 \rfloor \le \kappa_0$. Indeed, this is immediate when $2 \le \kappa_0 < 3$, while for $\kappa_0 \ge 3$, $\lfloor \kappa_0 \rfloor \ge \kappa_0 - 1 \ge 2\kappa_0/3$. Therefore, with $y := \lambda_k^2 t_0 = k^2/(4\kappa_0^2)$, we have

$$\frac{1}{9} \le y \le \frac{1}{4}. \tag{D.6}$$

Set $\alpha := \varepsilon \lambda_k$. By (D.6) and $\varepsilon \le \sqrt{t_0}$,

$$0 < \alpha \le \sqrt{t_0} \lambda_k = \sqrt{y} \le \frac{1}{2}.$$

By Lemma D.2,

$$\mathsf{W}_2(p_0, q_0^{\pm}) \le \frac{\alpha}{\lambda_k} = \varepsilon.$$

Thus $q_0^{\pm}$ are valid probability densities supported on $[-1, 1]$ and satisfy $q_0^{\pm} \in \mathcal{Q}_\varepsilon(p_0)$. First fix an arbitrary deterministic measurable score field s. For $t \in [t_0, 2t_0]$, we have $t \le 2t_0 < 1/64$, $\lambda_k^2 t \le 2y \le 1/2 < 1$, and $2\lambda_k^2 \sigma_t^2 \le 4y$. Hence Lemma D.4 and Lemma D.3 give

$$\int_{\mathbb{R}} |\partial_x \eta_t(x)|^2 p_t(x) \, \mathrm{d}x \ge \frac{1}{32} \lambda_k^2 e^{-4y},$$

$$R_+(t,s) + R_-(t,s) \geq \frac{1}{16}\alpha^2\lambda_k^2 e^{-4y}.$$

Therefore

$$\begin{aligned}\frac{1}{2}\Big[\mathcal{L}_{q_0^+}(s) + \mathcal{L}_{q_0^-}(s)\Big] &= \frac{1}{2(T-t_0)}\int_{t_0}^{T}\big(R_+(t,s) + R_-(t,s)\big)\,\mathrm{d}t \\ &\geq \frac{1}{2(T-t_0)}\int_{t_0}^{2t_0}\big(R_+(t,s) + R_-(t,s)\big)\,\mathrm{d}t \\ &\geq \frac{1}{32}\frac{\alpha^2\lambda_k^2 t_0}{T-t_0}e^{-4y} \\ &= \frac{\varepsilon^2}{32t_0(T-t_0)}y^2e^{-4y}.\end{aligned}$$

The map $y \mapsto y^2e^{-4y}$ is increasing on $[1/9, 1/4]$, since $\frac{\mathrm{d}}{\mathrm{d}y}(y^2e^{-4y}) = 2y(1-2y)e^{-4y} \geq 0$. Using (D.6), we conclude that

$$\frac{1}{2}\Big[\mathcal{L}_{q_0^+}(s) + \mathcal{L}_{q_0^-}(s)\Big] \geq \frac{e^{-4/9}}{2592}\frac{\varepsilon^2}{t_0(T-t_0)}.$$

Now let $\hat{s}$ be any possibly random score estimator. Applying the preceding inequality to each realization and then taking expectation gives

$$\frac{1}{2}\sum_{\xi\in\{+,-\}}\mathbb{E}\Big[\mathcal{L}_{q_0^\xi}(\hat{s})\Big] \geq \frac{e^{-4/9}}{2592}\frac{\varepsilon^2}{t_0(T-t_0)}.$$

Since both alternatives belong to $\mathcal{Q}_\varepsilon(p_0)$,

$$\sup_{q_0\in\mathcal{Q}_\varepsilon(p_0)}\mathbb{E}[\mathcal{L}_{q_0}(\hat{s})] \geq \max_{\xi\in\{+,-\}}\mathbb{E}\Big[\mathcal{L}_{q_0^\xi}(\hat{s})\Big] \geq \frac{1}{2}\sum_{\xi\in\{+,-\}}\mathbb{E}\Big[\mathcal{L}_{q_0^\xi}(\hat{s})\Big].$$

Taking the infimum over $\hat{s}$ proves (D.1). □

D.4 Proofs of Theorems 3.2 and 4.4

We finally lift the preceding one-dimensional construction to arbitrary fixed dimension through a product embedding.

Proof of Theorem 3.2 The case $d = 1$ is exactly Proposition D.1. We therefore assume $d \geq 2$.

Let $q_{0,1}^\pm$ be the one-dimensional alternatives constructed in the proof of Proposition D.1. Thus they use the same integer k, frequency $\lambda_k = \pi k$, and amplitude $\alpha = \varepsilon\lambda_k$. With $y := \lambda_k^2 t_0$, they satisfy $1/9 \leq y \leq 1/4$ and $0 < \alpha \leq 1/2$. Define the d-dimensional embeddings by

$$q_0^{\pm,(d)}(x) := q_{0,1}^\pm(x_1)\prod_{j=2}^{d}p_0^{(1)}(x_j).$$

Then $\mathrm{supp}(q_0^{\pm,(d)}) = \mathrm{supp}(p_0^{(d)}) = [-1,1]^d \subseteq \mathcal{B}(0,\sqrt{d})$.

We next verify the Wasserstein constraint. Let π be a coupling of $q_{0,1}^\pm$ and $p_0^{(1)}$. Let $Y_2,\dots,Y_d$ be independent with common law $p_0^{(1)}$, independent of $(U,V) \sim \pi$. Then $X = (U, Y_2, \dots, Y_d)$ and $Y = (V, Y_2, \dots, Y_d)$ form a coupling of $q_0^{\pm,(d)}$ and $p_0^{(d)}$, with $\|X - Y\|_2^2 = |U - V|^2$. Taking the infimum over all one-dimensional couplings π gives

$$\mathsf{W}_2\big(q_0^{\pm,(d)}, p_0^{(d)}\big) \leq \mathsf{W}_2\big(q_{0,1}^\pm, p_0^{(1)}\big) \leq \varepsilon.$$

The preceding Wasserstein bound therefore gives $q_0^{\pm,(d)} \in \mathcal{Q}_\varepsilon(p_0^{(d)})$. Let $q_{t,1}^\pm$ and $p_t^{(1)}$ denote the one-dimensional OU evolutions of $q_{0,1}^\pm$ and $p_0^{(1)}$, respectively. Since the d-dimensional OU semigroup factorizes across coordinates, set

$$q_t^{\pm,(d)} := P_t^{(d)}q_0^{\pm,(d)}, \qquad p_t^{(d)} := P_t^{(d)}p_0^{(d)},$$

$$q_t^{\pm,(d)}(x) = q_{t,1}^{\pm}(x_1)\prod_{j=2}^{d} p_t^{(1)}(x_j), \qquad p_t^{(d)}(x) = \prod_{j=1}^{d} p_t^{(1)}(x_j).$$

For $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, write $x_{2:d} := (x_2, \dots, x_d) \in \mathbb{R}^{d-1}$ and set $p_{t,\perp}(x_{2:d}) := \prod_{j=2}^{d} p_t^{(1)}(x_j)$. Then $p_{t,\perp}$ is a probability density on $\mathbb{R}^{d-1}$, and $q_t^{\pm,(d)}(x) = q_{t,1}^{\pm}(x_1)p_{t,\perp}(x_{2:d})$. The logarithm of this product factorizes additively, and therefore

$$\nabla \log q_t^{\pm,(d)}(x) = \Big(\partial_x \log q_{t,1}^{\pm}(x_1), \partial_x \log p_t^{(1)}(x_2), \dots, \partial_x \log p_t^{(1)}(x_d)\Big).$$

Thus the score separation occurs only in the first coordinate:

$$\nabla \log q_t^{+,(d)}(x) - \nabla \log q_t^{-,(d)}(x) = \big(\partial_x \log q_{t,1}^{+}(x_1) - \partial_x \log q_{t,1}^{-}(x_1), 0, \dots, 0\big).$$

For a deterministic measurable vector-valued score field s, define

$$R_\xi^{(d)}(t,s) := \mathbb{E}_{q_t^{\xi,(d)}}\left[\left\|s(X,t) - \nabla \log q_t^{\xi,(d)}(X)\right\|_2^2\right], \qquad \xi \in \{+,-\}.$$

The Hilbert-space version of the same two-point inequality used in Lemma D.3 gives

$$R_+^{(d)}(t,s) + R_-^{(d)}(t,s) \geq \int_{\mathbb{R}^d} \frac{q_t^{+,(d)}(x)q_t^{-,(d)}(x)}{q_t^{+,(d)}(x) + q_t^{-,(d)}(x)} \left\|\nabla \log q_t^{+,(d)}(x) - \nabla \log q_t^{-,(d)}(x)\right\|_2^2 \mathrm{d}x.$$

The overlap factor satisfies

$$\frac{q_t^{+,(d)}(x)q_t^{-,(d)}(x)}{q_t^{+,(d)}(x) + q_t^{-,(d)}(x)} = \frac{q_{t,1}^{+}(x_1)q_{t,1}^{-}(x_1)}{q_{t,1}^{+}(x_1) + q_{t,1}^{-}(x_1)} p_{t,\perp}(x_{2:d}).$$

Since $p_{t,\perp}$ integrates to one, the resulting d-dimensional lower bound reduces exactly to the one-dimensional two-point separation:

$$R_+^{(d)}(t,s) + R_-^{(d)}(t,s) \geq 2\alpha^2 \int_{\mathbb{R}} |\partial_x \eta_t(x_1)|^2 p_t^{(1)}(x_1)\,\mathrm{d}x_1.$$

For $t \in [t_0, 2t_0]$, the same bounds as in the one-dimensional proof give

$$t \leq 2t_0 < \frac{1}{64}, \qquad \lambda_k^2 t \leq 2y \leq \frac{1}{2} < 1, \qquad 2\lambda_k^2\sigma_t^2 \leq 4y.$$

Hence Lemma D.4 and the preceding two-point bound give

$$\int_{\mathbb{R}} |\partial_x \eta_t(x_1)|^2 p_t^{(1)}(x_1)\,\mathrm{d}x_1 \geq \frac{1}{32}\lambda_k^2 e^{-4y},$$
$$R_+^{(d)}(t,s) + R_-^{(d)}(t,s) \geq \frac{1}{16}\alpha^2\lambda_k^2 e^{-4y}.$$

Therefore, for every deterministic measurable vector-valued score field s,

$$\begin{aligned}
\frac{1}{2}\Big[\mathcal{L}_{q_0^{+,(d)}}(s) + \mathcal{L}_{q_0^{-,(d)}}(s)\Big] &= \frac{1}{2(T-t_0)}\int_{t_0}^{T}\big(R_+^{(d)}(t,s) + R_-^{(d)}(t,s)\big)\,\mathrm{d}t \\
&\geq \frac{1}{2(T-t_0)}\int_{t_0}^{2t_0}\big(R_+^{(d)}(t,s) + R_-^{(d)}(t,s)\big)\,\mathrm{d}t \\
&\geq \frac{1}{32}\frac{\alpha^2\lambda_k^2 t_0}{T-t_0}e^{-4y} \\
&= \frac{\varepsilon^2}{32t_0(T-t_0)}y^2e^{-4y} \geq \frac{e^{-4/9}}{2592}\frac{\varepsilon^2}{t_0(T-t_0)}.
\end{aligned}$$

The last step uses the monotonicity of y^2e^{-4y} on $[1/9, 1/4]$, as in the one-dimensional proof. The numerical constant $e^{-4/9}/2592$ is independent of t_0, ε, T, d, and the estimator s. Now let $\hat{s}$ be

any possibly random score estimator. Taking expectation in the preceding inequality and using $q_0^{\pm,(d)} \in \mathcal{Q}_\varepsilon(p_0^{(d)})$ gives

$$\sup_{q_0 \in \mathcal{Q}_\varepsilon(p_0^{(d)})} \mathbb{E}[\mathcal{L}_{q_0}(\hat{s})] \geq \frac{1}{2} \sum_{\xi \in \{+,-\}} \mathbb{E}\Big[\mathcal{L}_{q_0^{\xi,(d)}}(\hat{s})\Big] \geq \frac{e^{-4/9}}{2592} \frac{\varepsilon^2}{t_0(T-t_0)}.$$

Taking the infimum over $\hat{s}$ proves (3.3). □

Proof of Theorem 4.4 The assumed bound on t_0 implies $\tau_0 \leq 1$. Applying (4.5) with $\vartheta^2 = D^2/d$, and using $n+1 \geq n$, $\tau_0 \geq t_0$, and $C_{d,D} \geq 2$, gives the stated bound uniformly over $p_0 \in \mathcal{P}_D$. Taking the supremum over p_0 and then the infimum over estimators proves the minimax upper bound.

For the lower bound, set $A := 1/[n(T-t_0)\tau_0^{d/2}]$ and $B := \varepsilon^2/[t_0(T-t_0)]$. Theorem 4.3 and the sample-rich condition give $\mathfrak{R}^\star_{n,\varepsilon}(D; t_0, T) \geq c_1 A$. Because $D \geq \sqrt{d}$, the reference law $p_0^{(d)}$ belongs to $\mathcal{P}_D$. Theorem 3.2 therefore gives $\mathfrak{R}^\star_{n,\varepsilon}(D; t_0, T) \geq (e^{-4/9}/2592)B$ for $\varepsilon > 0$. For $\varepsilon = 0$, this bound holds trivially because $B = 0$. Consequently,

$$\mathfrak{R}^\star_{n,\varepsilon}(D; t_0, T) \geq (c_1 A) \vee \left(\frac{e^{-4/9}}{2592} B\right) \geq \frac{1}{2} \min\left\{c_1, \frac{e^{-4/9}}{2592}\right\}(A+B) = c_{d,D}(A+B).$$

This is the lower bound in (4.8). □

D.5 Proof of Theorem 6.1

Proposition D.5 (Intrinsic positive-time KL shift cost)**.** *Fix $d \geq 1$ and $D > 0$. For every $n \geq 1$, $t_0 > 0$, and $0 \leq \varepsilon \leq \sqrt{\tau_0}$,*

$$\mathfrak{K}^\star_{n,\varepsilon}(D; t_0) \geq \frac{e^{-1}}{\pi} \frac{\varepsilon^2}{\tau_0}. \tag{D.7}$$

Proof. The case $\varepsilon = 0$ is immediate, so assume $0 < \varepsilon \leq \sqrt{\tau_0}$. Fix an arbitrary measurable, possibly randomized probability-measure estimator $\hat{\mu}_n$ based on n reference samples. Take $p_0 := \delta_0 \in \mathcal{P}_D$ and $q_0^\pm := \delta_{\pm \varepsilon e_1}$, where e_1 is the first standard basis vector. Then $\mathsf{W}_2(q_0^\pm, p_0) = \varepsilon$, hence $q_0^\pm \in \mathcal{Q}_\varepsilon(p_0)$. Under p_0, all reference samples vanish almost surely, so the observation law is identical for the two target alternatives and the only randomness in $\hat{\mu}_n$ is internal. Moreover, $q_{t_0}^\pm = \mathcal{N}(\pm a_{t_0}\varepsilon e_1, \rho_{t_0} I_d)$. Let Φ and ϕ be the standard Gaussian distribution function and density, and set $x := a_{t_0}\varepsilon/\sqrt{\rho_{t_0}} = \varepsilon/\sqrt{\tau_0} \in (0, 1]$. Let $d_{\mathrm{TV}}(\mu, \nu) := \sup_{A \text{ Borel}} |\mu(A) - \nu(A)|$ denote the total variation distance. Taking the half-space $A := \{z \in \mathbb{R}^d : z_1 \geq 0\}$ gives

$$\begin{aligned} d_{\mathrm{TV}}(q_{t_0}^+, q_{t_0}^-) &\geq |q_{t_0}^+(A) - q_{t_0}^-(A)| = 2\Phi(x) - 1 \\ &= 2\int_0^x \phi(u)\,\mathrm{d}u \geq 2\phi(1)x = 2\phi(1)\frac{\varepsilon}{\sqrt{\tau_0}}. \end{aligned} \tag{D.8}$$

For any probability measure μ, Pinsker's inequality and the triangle inequality for total variation imply

$$\begin{aligned} \mathrm{KL}(q_{t_0}^+ \| \mu) + \mathrm{KL}(q_{t_0}^- \| \mu) &\geq 2 d_{\mathrm{TV}}(q_{t_0}^+, \mu)^2 + 2 d_{\mathrm{TV}}(q_{t_0}^-, \mu)^2 \\ &\geq \big[d_{\mathrm{TV}}(q_{t_0}^+, \mu) + d_{\mathrm{TV}}(q_{t_0}^-, \mu)\big]^2 \geq d_{\mathrm{TV}}(q_{t_0}^+, q_{t_0}^-)^2, \end{aligned}$$

with the inequality understood trivially if either KL divergence is infinite. Applying this pointwise to each realization of $\hat{\mu}_n$, taking expectation over its internal randomness, and using (D.8) yields

$$\frac{1}{2} \sum_{\xi \in \{+,-\}} \mathbb{E}\Big[\mathrm{KL}(q_{t_0}^\xi \| \hat{\mu}_n)\Big] \geq \frac{1}{2} d_{\mathrm{TV}}(q_{t_0}^+, q_{t_0}^-)^2 \geq 2\phi(1)^2 \frac{\varepsilon^2}{\tau_0} = \frac{e^{-1}}{\pi} \frac{\varepsilon^2}{\tau_0}.$$

The supremum over $q_0 \in \mathcal{Q}_\varepsilon(p_0)$ is at least the maximum, hence the average, of these two risks. Taking the infimum over $\hat{\mu}_n$ proves (D.7). □

Proof of Theorem 6.1 Take c_0 and $c_{d,D}^{\mathrm{KL}}$ as in Theorem C.11, and set

$$\tilde{c}_{d,D} := \frac{1}{2}\min\left\{c_{d,D}^{\mathrm{KL}}, \frac{e^{-1}}{\pi}\right\}.$$

Set $A := 1/(n\tau_0^{d/2})$ and $B := \varepsilon^2/\tau_0$. Since $p_0 \in \mathcal{Q}_\varepsilon(p_0)$, the robust minimax risk dominates its nominal counterpart. Theorem C.11 and the sample-rich condition therefore give $\mathfrak{K}^\star_{n,\varepsilon}(D;t_0) \geq c_{d,D}^{\mathrm{KL}} A$. Proposition D.5 also gives $\mathfrak{K}^\star_{n,\varepsilon}(D;t_0) \geq (e^{-1}/\pi)B$. Consequently,

$$\mathfrak{K}^\star_{n,\varepsilon}(D;t_0) \geq \left(c_{d,D}^{\mathrm{KL}} A\right) \vee \left(\frac{e^{-1}}{\pi} B\right) \geq \frac{1}{2}\min\left\{c_{d,D}^{\mathrm{KL}}, \frac{e^{-1}}{\pi}\right\}(A+B) = \tilde{c}_{d,D}(A+B).$$

For the upper bound, use the self-consistent Gaussian-blanket output $\hat{\mu}_n(\mathrm{d}x) = \hat{p}_{n,t_0}^{D/\sqrt{d}}(x)\,\mathrm{d}x$. Since $t_0 \leq c_0 \leq \log 2$, one has $\tau_0 \leq 1$, and (6.1) gives, uniformly over $p_0 \in \mathcal{P}_D$ and $q_0 \in \mathcal{Q}_\varepsilon(p_0)$, $\mathbb{E}[\mathrm{KL}(q_{t_0}\|\hat{\mu}_n)] \leq B + C_{d,D}/[2(n+1)\tau_0^{d/2}] \leq C_{d,D}(A+B)/2$, where the last step uses $n+1 \geq n$ and $C_{d,D} \geq 2$. Taking the relevant suprema and then the infimum over estimators proves the upper bound in (6.2). □

E Proofs for Adaptation to Intrinsic Dimension

This appendix proves the results of Section 5. The upper bound combines two ingredients: a bound on the source energy missed by the empirical span, and an OU comparison that separates this error from estimation within the span. Conditional Gaussian-blanket estimates then give Theorem 5.1. The final subsection establishes the intrinsic-dimensional minimax rate for score estimation. The corresponding positive-time KL results are given in Appendix F.

Throughout this appendix, write $\tau_t := \rho_t/a_t^2 = e^t - 1$ for $t > 0$, retaining $\tau_0 = e^{t_0} - 1$ for its value at t_0.

E.1 Missed Energy of the Empirical Span

Lemma E.1 (Missed source energy)**.** *Let $X_1, X_2, \ldots$ be i.i.d. samples from the law of a random vector $X \in \mathbb{R}^d$ satisfying $\|X\|_2 \leq D$ almost surely for some $D > 0$. Suppose that the second-moment matrix $C := \mathbb{E}[XX^\top]$ has rank at most $k \geq 1$. If Π_j is the orthogonal projector onto $\mathrm{span}(X_1, \ldots, X_j)$, with $\Pi_0 = 0$, then, for every integer $m \geq 0$,*

$$\mathbb{E}\,\mathrm{tr}[C(I_d - \Pi_m)] \leq \frac{kD^2}{m+k}. \tag{E.1}$$

Proof. Set $Q_j := I_d - \Pi_j$, $e_j := \mathrm{tr}(CQ_j)$, and $b_j := \mathbb{E}e_j$. Since $Q_j^\top = Q_j^2 = Q_j$,

$$e_j = \mathrm{tr}(Q_j C Q_j) = \int \|Q_j x\|_2^2 \, P_X(\mathrm{d}x),$$

where P_X denotes the law of X. Thus e_j is the source energy outside the observed span. The next observation X_{j+1} is independent of $X_1, \ldots, X_j$. Set $z := Q_j X_{j+1}$. On $\{z \neq 0\}$, the rank-one projector update and $\|z\|_2 \leq \|X_{j+1}\|_2 \leq D$ give

$$\Pi_{j+1} = \Pi_j + \frac{zz^\top}{\|z\|_2^2}, \qquad e_j - e_{j+1} = \frac{z^\top C z}{\|z\|_2^2} \geq \frac{z^\top C z}{D^2}.$$

On $\{z = 0\}$, the span is unchanged and the same inequality holds with both sides equal to zero. Since $\mathbb{E}[zz^\top \mid X_1, \ldots, X_j] = Q_j C Q_j$,

$$\mathbb{E}[e_{j+1} \mid X_1, \ldots, X_j] \leq e_j - \frac{\mathrm{tr}(CQ_jCQ_j)}{D^2}$$

$$= e_j - \frac{\operatorname{tr}[(Q_j C Q_j)^2]}{D^2} \le e_j - \frac{e_j^2}{kD^2}.$$

Indeed, $Q_j C Q_j$ is positive semidefinite and has rank $r \le k$. Writing $\lambda_1, \dots, \lambda_r$ for its nonzero eigenvalues, Cauchy–Schwarz gives

$$e_j^2 = \left(\sum_{i=1}^{r} \lambda_i\right)^2 \le r \sum_{i=1}^{r} \lambda_i^2 \le k \operatorname{tr}[(Q_j C Q_j)^2].$$

Taking expectations and using Jensen's inequality gives

$$b_{j+1} \le b_j - \frac{b_j^2}{kD^2}. \tag{E.2}$$

If $b_m = 0$, the claim is immediate. Otherwise, (E.2) gives $b_j \ge b_m > 0$ for $0 \le j \le m$, so, for $j < m$,

$$\frac{1}{b_{j+1}} \ge \frac{1}{b_j(1 - b_j/(kD^2))} \ge \frac{1}{b_j} + \frac{1}{kD^2}.$$

Since $b_0 \le D^2$, iteration yields

$$b_m \le \frac{1}{b_0^{-1} + m/(kD^2)} \le \frac{kD^2}{m+k}.$$

□

Under Assumptions 1 and 2, $C = U\mathbb{E}[ZZ^\top]U^\top$ has rank at most k, even when the latent second-moment matrix is singular. Thus the lemma bounds the missed energy without a lower bound on the nonzero eigenvalues.

E.2 An Orthogonal OU Comparison

The projected source law is supported on the empirical span, whereas the target need not be. The following comparison separates error within the span from the energy in its orthogonal complement.

Lemma E.2 (Orthogonal KL comparison)**.** *Fix $t_0 > 0$. Let Π be an orthogonal projector on $\mathbb{R}^d$ and let $p_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Define*

$$p_0^\Pi := \Pi_\# p_0, \qquad e_\Pi := \int \|(I_d - \Pi)x\|_2^2 \, p_0(\mathrm{d}x).$$

Let $\hat{\mu}_0 \in \mathcal{P}_2(\mathbb{R}^d)$ be supported on $\operatorname{range}(\Pi)$, and set $\hat{p}_t := P_t \hat{\mu}_0$. If $D_2(P_{t_0} p_0^\Pi \| \hat{p}_{t_0}) < \infty$, then, for every $q_0 \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathrm{KL}(q_{t_0} \| \hat{p}_{t_0}) \le D_2\big(P_{t_0} p_0^\Pi \big\| \hat{p}_{t_0}\big) + \frac{\mathsf{W}_2^2(q_0, p_0) + e_\Pi}{\tau_0}, \tag{E.3}$$

where $q_t := P_t q_0$.

Proof. Set $Q := I_d - \Pi$, $p_t^\Pi := P_t p_0^\Pi$, and $q_t^\Pi := P_t(\Pi_\# q_0)$. Here projection is applied to the initial law before the full d-dimensional OU evolution; in general, $q_t^\Pi \ne \Pi_\# q_t$. Let $r := \operatorname{rank}(\Pi)$ and choose orthonormal coordinates $x = (z, w) \in \mathbb{R}^r \times \mathbb{R}^{d-r}$ adapted to $\operatorname{range}(\Pi) \oplus \operatorname{range}(Q)$. Write $\varphi_v^{(j)}$ for the density of $\mathcal{N}(0, vI_j)$, with $\varphi_v^{(0)} \equiv 1$. Let $q_{\|,t}$ and $\hat{p}_{\|,t}$ denote the densities of the z-marginals of q_t and $\hat{p}_t$, respectively. Projection of the initial law preserves its z-component, so q_t^Π and q_t have the same z-marginal. Since $\Pi_\# q_0$ and $\hat{\mu}_0$ are supported on $\operatorname{range}(\Pi)$, their w-components under the OU evolution consist solely of Gaussian noise, independent of their z-components. Hence

$$q_t^\Pi(z, w) = q_{\|,t}(z) \varphi_{\rho_t}^{(d-r)}(w), \qquad \hat{p}_t(z, w) = \hat{p}_{\|,t}(z) \varphi_{\rho_t}^{(d-r)}(w).$$

The density q_t itself need not factor. All KL divergences below are finite by Lemma B.1, since the initial laws have finite second moments. The log density ratio $\log(q_t^\Pi / \hat{p}_t)$ depends only on z, and thus

$$\int_{\mathbb{R}^d} q_t(x) \log \frac{q_t^\Pi(x)}{\hat{p}_t(x)} \, \mathrm{d}x = \int_{\mathbb{R}^r} q_{\|,t}(z) \log \frac{q_{\|,t}(z)}{\hat{p}_{\|,t}(z)} \, \mathrm{d}z = \int_{\mathbb{R}^d} q_t^\Pi(x) \log \frac{q_t^\Pi(x)}{\hat{p}_t(x)} \, \mathrm{d}x.$$

Decomposing $\log(q_t/\hat{p}_t)$ through q_t^{Π} therefore gives

$$\mathrm{KL}(q_t\|\hat{p}_t) = \mathrm{KL}(q_t\|q_t^{\Pi}) + \mathrm{KL}(q_t^{\Pi}\|\hat{p}_t). \tag{E.4}$$

Let (X, Y) be an optimal coupling of (p_0, q_0) and put $\delta := Y - X$, so that $\mathbb{E}\|\delta\|_2^2 = \mathsf{W}_2^2(p_0, q_0)$. Applying Lemma B.1 with the couplings $(Y, \Pi Y)$ and $(\Pi Y, \Pi X)$ gives

$$\mathrm{KL}(q_{t_0}\|q_{t_0}^{\Pi}) \le \frac{\mathbb{E}\|QY\|_2^2}{2\tau_0}, \qquad \mathrm{KL}(q_{t_0}^{\Pi}\|p_{t_0}^{\Pi}) \le \frac{\mathbb{E}\|\Pi\delta\|_2^2}{2\tau_0}.$$

Applying Lemma B.3 to the second term in (E.4) at $t = t_0$, with intermediate density $p_{t_0}^{\Pi}$, and using the preceding bounds yields

$$\mathrm{KL}(q_{t_0}\|\hat{p}_{t_0}) \le D_2(p_{t_0}^{\Pi}\|\hat{p}_{t_0}) + \frac{\mathbb{E}\|\Pi\delta\|_2^2 + \frac{1}{2}\mathbb{E}\|QY\|_2^2}{\tau_0}.$$

Since $QY = QX + Q\delta$ and $\mathbb{E}\|QX\|_2^2 = e_{\Pi}$,

$$\frac{1}{2}\mathbb{E}\|QY\|_2^2 \le e_{\Pi} + \mathbb{E}\|Q\delta\|_2^2.$$

Together with the orthogonal decomposition $\mathbb{E}\|\Pi\delta\|_2^2 + \mathbb{E}\|Q\delta\|_2^2 = \mathsf{W}_2^2(p_0, q_0)$, this proves (E.3). □

E.3 Proof of Theorem 5.1

Proof. Let $\mathcal{F}_1 := \sigma(X_1, \dots, X_{n_1})$ be the σ-field generated by the pilot sample, and write $M := M_{\vartheta_k, t_0}$. Set $\hat{r} := \mathrm{rank}(\hat{\Pi}) = \dim \hat{L}$. For $t > 0$, write $\hat{p}_{n,t}^{\mathrm{sub}} := P_t\hat{\mu}_{n,0}^{\mathrm{sub}}$. The empirical projector $\hat{\Pi}$ is measurable, since it can be constructed by applying Gram–Schmidt to the pilot sample with zero residuals omitted. Hence the projected observations Y_i are also measurable. The maps $y \mapsto \delta_y$ and $\Sigma \mapsto \mathcal{N}(0, \Sigma)$ are continuous for weak convergence, including singular covariance matrices. Their finite mixture $\hat{\mu}_{n,0}^{\mathrm{sub}}$ is therefore measurable for the Borel σ-field of the weak topology. Almost surely, $\|Y_i\|_2 \le D$, $\hat{r} \le k$, and

$$\int \|x\|_2^2 \, \hat{\mu}_{n,0}^{\mathrm{sub}}(\mathrm{d}x) = \frac{\sum_{i=1}^{n_2} \|Y_i\|_2^2 + M\vartheta_k^2\hat{r}}{n_2 + M} \le D^2. \tag{E.5}$$

Since $Y_i \in \mathrm{range}(\hat{\Pi})$ and $\mathcal{N}(0, \vartheta_k^2\hat{\Pi})$ is supported on the same subspace, so is their mixture $\hat{\mu}_{n,0}^{\mathrm{sub}}$, as required by Lemma E.2.

Step 1: Conditional Rényi bound. Condition on $\mathcal{F}_1$. Then $\hat{\Pi}$ is fixed and the Y_i are independent with law $p_0^{\hat{\Pi}} := \hat{\Pi}_{\#}p_0$, supported on $\mathcal{B}(0, D) \cap \mathrm{range}(\hat{\Pi})$. Set

$$\Sigma_t := \rho_t I_d + a_t^2\vartheta_k^2\hat{\Pi}, \qquad k_{t,y}(x) := \varphi_{\rho_t I_d}(x - a_t y),$$

where φ_{Σ} denotes the density of $\mathcal{N}(0, \Sigma)$ for positive definite Σ. Then

$$\hat{p}_{n,t}^{\mathrm{sub}}(x) = \frac{\sum_{i=1}^{n_2} k_{t,Y_i}(x) + M\varphi_{\Sigma_t}(x)}{n_2 + M}. \tag{E.6}$$

This formula also shows that the score is jointly measurable in the observations and (x, t) for $t > 0$. For $y \in \mathrm{range}(\hat{\Pi})$, the Gaussian factors of $k_{t,y}$ and φ_{Σ_t} on $\mathrm{range}(\hat{\Pi})^{\perp}$ coincide and cancel in the density ratio. Repeating the Gaussian-ratio calculation of Lemma B.4 in the $\hat{r}$ tangential directions gives, for $y \in \mathcal{B}(0, D) \cap \mathrm{range}(\hat{\Pi})$ and $t \ge t_0$,

$$\sup_{x \in \mathbb{R}^d} \frac{k_{t,y}(x)}{\varphi_{\Sigma_t}(x)} = \exp\left(\frac{\|y\|_2^2}{2\vartheta_k^2}\right)\left(1 + \frac{\vartheta_k^2}{\tau_t}\right)^{\hat{r}/2} \le M. \tag{E.7}$$

For $\hat{r} = 0$, the only possible center is $y = 0$ and the ratio equals one. The last inequality uses $\|y\|_2 \le D$, $\hat{r} \le k$, and $\tau_t \ge \tau_0$.

Conditional on $\mathcal{F}_1$, the mean kernel density is $\mathbb{E}[k_{t,Y_i}(x) \mid \mathcal{F}_1] = (P_t p_0^{\hat{\Pi}})(x)$. By (E.6) and (E.7), applying Lemma B.5 with $n = n_2$, kernels $k_{t,y}$, reference density φ_{Σ_t}, and domination constant M yields

$$\mathbb{E}\left[\int \frac{(P_t p_0^{\hat{\Pi}})(x)^2}{\hat{p}_{n,t}^{\mathrm{sub}}(x)}\,\mathrm{d}x \;\middle|\; \mathcal{F}_1\right] \le \frac{n_2 + M}{n_2 + 1}.$$

Conditional Jensen's inequality for the logarithm yields

$$\mathbb{E}\Big[D_2(P_t p_0^{\hat{\Pi}} \| \hat{p}_{n,t}^{\mathrm{sub}}) \;\Big|\; \mathcal{F}_1\Big] \le \log\left(1 + \frac{M-1}{n_2+1}\right). \tag{E.8}$$

The finiteness condition in Lemma E.2 holds for every such sample, since (E.7) and (E.6) give

$$P_t p_0^{\hat{\Pi}} \le M\varphi_{\Sigma_t}, \qquad \hat{p}_{n,t}^{\mathrm{sub}} \ge \frac{M\varphi_{\Sigma_t}}{n_2 + M}, \qquad \int \frac{(P_t p_0^{\hat{\Pi}})^2}{\hat{p}_{n,t}^{\mathrm{sub}}} \le n_2 + M.$$

Step 2: KL and score-risk bounds. Fix $q_0 \in \mathcal{Q}_\varepsilon(p_0)$ and define

$$e_{n_1} := \int \|(I_d - \hat{\Pi})x\|_2^2\, p_0(\mathrm{d}x) = \mathrm{tr}[C(I_d - \hat{\Pi})], \qquad C := \mathbb{E}_{p_0}[XX^\top].$$

Taking conditional expectation in Lemma E.2 and using (E.8) at $t = t_0$ gives

$$\mathbb{E}\big[\mathrm{KL}\big(q_{t_0} \big\| \hat{p}_{n,t_0}^{\mathrm{sub}}\big) \mid \mathcal{F}_1\big] \le \log\left(1 + \frac{M-1}{n_2+1}\right) + \frac{\varepsilon^2 + e_{n_1}}{\tau_0}.$$

Since $\mathrm{rank}(C) \le k$, (E.1) gives $\mathbb{E}e_{n_1} \le kD^2/(n_1 + k)$. Averaging over the pilot sample therefore yields

$$\mathbb{E}\big[\mathrm{KL}\big(q_{t_0} \big\| \hat{p}_{n,t_0}^{\mathrm{sub}}\big)\big] \le \log\left(1 + \frac{M-1}{n_2+1}\right) + \frac{\varepsilon^2}{\tau_0} + \frac{kD^2}{(n_1+k)\tau_0}. \tag{E.9}$$

For each sample realization, q_0 and $\hat{\mu}_{n,0}^{\mathrm{sub}}$ belong to $\mathcal{P}_2(\mathbb{R}^d)$, so Lemma B.2 gives

$$\mathcal{L}_{q_0}(\hat{s}_n^{\mathrm{sub}}) = \frac{2}{T - t_0}\big[\mathrm{KL}\big(q_{t_0} \big\| \hat{p}_{n,t_0}^{\mathrm{sub}}\big) - \mathrm{KL}\big(q_T \big\| \hat{p}_{n,T}^{\mathrm{sub}}\big)\big] \le \frac{2}{T - t_0}\mathrm{KL}\big(q_{t_0} \big\| \hat{p}_{n,t_0}^{\mathrm{sub}}\big).$$

Taking expectations and using (E.9) gives a bound independent of $q_0 \in \mathcal{Q}_\varepsilon(p_0)$. Taking the supremum over this class proves (5.1).

Finally, if $k \ge 2$ and $\tau_0 \le 1$, then $M \le e^{k/2}(1 + D^2/k)^{k/2}\tau_0^{-k/2}$ and $\tau_0^{-1} \le \tau_0^{-k/2}$. Applying these bounds, $n_2 + 1 \ge n/2$, $n_1 + k \ge n/2$, and $\log(1+u) \le u$ to (5.1) gives (5.2) with the stated constant. □

E.4 Intrinsic-Dimensional Score Minimax Rate

We use the source class $\mathcal{P}_{k,D}$ and minimax risk $\mathfrak{R}^\star_{n,\varepsilon}(k, D; t_0, T)$ defined in Subsection 5.1.

Corollary E.3. *Fix $k \ge 2$ and $D \ge \sqrt{k}$, and let $d \ge k$ be arbitrary. Let $c_0, c_{k,D} > 0$ be the lower-bound constants in Theorem 4.4 with the dimension set to k. If $n \ge 2$,*

$$0 < t_0 \le \min\{c_0, 1/(16\pi^2)\}, \qquad 2t_0 \le T, \qquad n\tau_0^{k/2} \ge 1, \qquad 0 \le \varepsilon \le \sqrt{t_0},$$

then

$$\begin{aligned} c_{k,D}\left[\frac{1}{n(T-t_0)\tau_0^{k/2}} + \frac{\varepsilon^2}{t_0(T-t_0)}\right] &\le \mathfrak{R}^\star_{n,\varepsilon}(k, D; t_0, T) \\ &\le \tilde{C}_{k,D}\left[\frac{1}{n(T-t_0)\tau_0^{k/2}} + \frac{\varepsilon^2}{t_0(T-t_0)}\right]. \end{aligned} \tag{E.10}$$

Here $\tilde{C}_{k,D}$ is the constant in Theorem 5.1.

Proof. The conditions imply $t_0 \le 1/(16\pi^2) < \log 2$, so $\tau_0 \le 1$. The upper bound follows uniformly over the source class from Theorem 5.1, using $\tau_0 \ge t_0$ and $\tilde{C}_{k,D} \ge 2$.

To transfer the k-dimensional lower bound, we construct from any ambient estimator a latent estimator with no greater risk. Fix a linear isometry $U : \mathbb{R}^k \to \mathbb{R}^d$, and let the columns of $V \in \mathbb{R}^{d\times(d-k)}$ form an orthonormal basis of $\mathrm{range}(U)^\perp$. For latent laws $\bar{p}_0, \bar{q}_0$, set $p_0 := U_\# \bar{p}_0$ and $q_0 := U_\# \bar{q}_0$. Then

$$\mathsf{W}_2(p_0, q_0) = \mathsf{W}_2(\bar{p}_0, \bar{q}_0). \tag{E.11}$$

Indeed, one inequality follows by pushing a latent coupling forward, and the other by applying $U^\top$ to an ambient coupling, whose marginals are supported on $\mathrm{range}(U)$.

At positive time, with $\bar{q}_t := P_t^{(k)} \bar{q}_0$,

$$\begin{aligned} q_t(Uz + Vw) &= \bar{q}_t(z)\varphi_{\rho_t}^{(d-k)}(w), \\ U^\top \nabla \log q_t(Uz + Vw) &= \nabla \log \bar{q}_t(z). \end{aligned} \tag{E.12}$$

Consider an arbitrary measurable, possibly randomized ambient estimator $\hat{s}_n$, and represent its internal randomization by a seed ω independent of the training sample. We write $\hat{s}_n(x, t; x_1, \dots, x_n, \omega)$ to make the sample and seed dependence explicit. Given latent observations $Z_1, \dots, Z_n$, feed $UZ_1, \dots, UZ_n$ to the estimator. To match the orthogonal OU noise, introduce an additional seed $G \sim \mathcal{N}(0, I_{d-k})$, independent of the sample and ω. Define

$$\tilde{s}_n(z, t) := U^\top \hat{s}_n(Uz + \sqrt{\rho_t} VG, t; UZ_1, \dots, UZ_n, \omega). \tag{E.13}$$

The same G is used for all (z, t), and composition of the measurable maps in (E.13) defines a measurable random score field. Its construction does not depend on $\bar{p}_0$ or $\bar{q}_0$. When $d = k$, the orthogonal component and the extra seed are absent.

Fix the training sample and ω. Since $\sqrt{\rho_t} G \sim \mathcal{N}(0, \rho_t I_{d-k})$ and $[U, V]$ is orthogonal, (E.12) gives

$$\begin{aligned} \mathbb{E}_G \int_{\mathbb{R}^k} \|\tilde{s}_n(z, t) - \nabla \log \bar{q}_t(z)\|_2^2 \bar{q}_t(z)\, \mathrm{d}z &= \int_{\mathbb{R}^d} \|U^\top(\hat{s}_n(x, t) - \nabla \log q_t(x))\|_2^2 q_t(x)\, \mathrm{d}x \\ &\le \int_{\mathbb{R}^d} \|\hat{s}_n(x, t) - \nabla \log q_t(x)\|_2^2 q_t(x)\, \mathrm{d}x. \end{aligned}$$

The inequality uses contraction under orthogonal projection. Integrating in t, dividing by $T - t_0$, and averaging over the sample and ω yields

$$\mathbb{E}\mathcal{L}_{\bar{q}_0}^{(k)}(\tilde{s}_n) \le \mathbb{E}\mathcal{L}_{U_\#\bar{q}_0}^{(d)}(\hat{s}_n). \tag{E.14}$$

Here $\mathcal{L}^{(j)}$ denotes the score loss in dimension j, and the expectation on the left also includes G. All integrands are nonnegative, so Tonelli's theorem applies even if a risk is infinite.

Every latent source law supported on $\mathcal{B}(0, D) \subset \mathbb{R}^k$ embeds into $\mathcal{P}_{k,D}$. Fix such a law $\bar{p}_0$ and set $p_0 = U_\# \bar{p}_0$. By (E.11), every target in its latent Wasserstein ball $\mathcal{Q}_\varepsilon^{(k)}(\bar{p}_0)$ embeds into $\mathcal{Q}_\varepsilon(p_0)$. The training sample and both seeds have the same joint law for all these targets. Restricting the target supremum to embedded laws and using (E.14) gives

$$\sup_{q_0 \in \mathcal{Q}_\varepsilon(p_0)} \mathbb{E}\mathcal{L}_{q_0}^{(d)}(\hat{s}_n) \ge \sup_{\bar{q}_0 \in \mathcal{Q}_\varepsilon^{(k)}(\bar{p}_0)} \mathbb{E}\mathcal{L}_{U_\#\bar{q}_0}^{(d)}(\hat{s}_n) \ge \sup_{\bar{q}_0 \in \mathcal{Q}_\varepsilon^{(k)}(\bar{p}_0)} \mathbb{E}\mathcal{L}_{\bar{q}_0}^{(k)}(\tilde{s}_n).$$

Now take the supremum over latent sources and the infimum over ambient estimators. Since the induced estimators form a subset of all admissible randomized latent estimators, the unknown-subspace minimax risk is at least the k-dimensional minimax risk in (4.6). The lower bound in Theorem 4.4, with $d = k$, now proves (E.10), with the same constants c_0 and $c_{k,D}$. □

F Proofs for Robust Reverse Sampling

Subsection F.1 establishes the KL comparison with general initialization. Subsection F.2 gives the standard Gaussian initialization bound. Subsection F.3 states and proves Theorem F.5 for the unknown-subspace estimator, and Subsection F.4 establishes its intrinsic-dimensional KL minimax rate. The robust KL minimax result in Theorem 6.1 is proved in Appendix D.5.

F.1 KL Comparison with General Initialization

Given a possibly randomized score estimator $\hat{s}_n$ and a terminal law ν_T, consider the approximate reverse process

$$\begin{aligned} \mathrm{d}\hat{X}_t^{\leftarrow} &= \frac{1}{2}\Big(\hat{X}_t^{\leftarrow} + 2\hat{s}_n(\hat{X}_t^{\leftarrow}, T-t)\Big)\mathrm{d}t + \mathrm{d}W_t, \\ \hat{X}_0^{\leftarrow} &\sim \nu_T, \qquad 0 \le t \le T - t_0. \end{aligned} \tag{F.1}$$

Conditional on the training sample and any randomness used to construct $\hat{s}_n$ and ν_T, denote the law of $\hat{X}^{\leftarrow}_{T-t_0}$ by $\hat{\mu}^{\nu}_{n,t_0}$.

Lemma F.1 (Conditional score-to-KL bridge). *Fix $q_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Conditional on the training sample and the randomness used to construct $\hat{s}_n$ and ν_T, suppose that the exact shifted reverse equation and (F.1) admit non-explosive weak solutions that are unique in law and whose drifts are jointly locally bounded in space and reverse time. Then, in the extended-real sense,*

$$\mathrm{KL}\big(q_{t_0}\big\|\hat{\mu}^{\nu}_{n,t_0}\big) \le \mathrm{KL}(q_T\|\nu_T) + \frac{1}{2}\int_{t_0}^{T} \|\hat{s}_n(\cdot,t) - \nabla \log q_t\|^2_{L^2(q_t)}\,\mathrm{d}t. \tag{F.2}$$

Proof. Condition on the training sample and the randomness used to construct $\hat{s}_n$ and ν_T, and set $S := T - t_0$. Let $\mathbf{Q}$ be the path law of the exact shifted reverse process initialized from q_T and driven by $\nabla \log q_{T-u}$, and let $\hat{\mathbf{Q}}$ be the path law of (F.1). By the standard time-reversal formula, the time-u marginal of $\mathbf{Q}$ is q_{T-u}; see Anderson (1982) and Cattiaux et al. (2023, Theorem 4.9).

If either term on the right of (F.2) is infinite, the claim is immediate. Otherwise $q_T \ll \nu_T$. The standard localized Girsanov argument on path space applies after disintegrating with respect to the initial coordinate. For a related localization argument in diffusion sampling, see Chen et al. (2023b). Indeed, let X_u be the canonical coordinate, set $\tau_R := \inf\{u \in [0,S] : \|X_u\|_2 \ge R\} \wedge S$, and define $X^{(R)}_u := X_{u\wedge\tau_R}$. Let $\mathbf{Q}^{(R)}$ and $\hat{\mathbf{Q}}^{(R)}$ denote the laws of $X^{(R)}$ under $\mathbf{Q}$ and $\hat{\mathbf{Q}}$, respectively. The stopped drift difference is bounded, so Novikov's condition and the entropy chain rule give

$$\mathrm{KL}(\mathbf{Q}^{(R)}\|\hat{\mathbf{Q}}^{(R)}) \le \mathrm{KL}(q_T\|\nu_T) + \frac{1}{2}\mathbb{E}_{\mathbf{Q}}\int_0^{\tau_R} \|\nabla \log q_{T-u}(X_u) - \hat{s}_n(X_u, T-u)\|_2^2\,\mathrm{d}u.$$

Non-explosion implies that $\tau_R \uparrow S$ almost surely under both path laws, and hence $\mathbf{Q}^{(R)} \Rightarrow \mathbf{Q}$ and $\hat{\mathbf{Q}}^{(R)} \Rightarrow \hat{\mathbf{Q}}$. Therefore, the lower semicontinuity of relative entropy and monotone convergence yield

$$\mathrm{KL}(\mathbf{Q}\|\hat{\mathbf{Q}}) \le \mathrm{KL}(q_T\|\nu_T) + \frac{1}{2}\mathbb{E}_{\mathbf{Q}}\int_0^{S} \|\nabla \log q_{T-u}(X_u) - \hat{s}_n(X_u, T-u)\|_2^2\,\mathrm{d}u.$$

Applying data processing under the endpoint map, using that the time-u marginal of $\mathbf{Q}$ is q_{T-u}, and changing variables $t = T - u$ prove (F.2). □

Proposition F.2 (OU-compatible KL comparison). *Suppose Assumption 1 holds and fix $\vartheta > 0$. Take $\hat{s}_n = \hat{s}^{\vartheta}_n$ in (F.1). Then the reverse equation has a unique non-explosive strong solution. For every $q_0 \in \mathcal{Q}_\varepsilon(p_0)$ and every possibly data-dependent terminal law ν_T, in the extended-real sense,*

$$\mathrm{KL}\big(q_{t_0}\big\|\hat{\mu}^{\nu}_{n,t_0}\big) \le \mathrm{KL}\big(q_{t_0}\big\|\hat{p}^{\vartheta}_{n,t_0}\big) + \mathrm{KL}(q_T\|\nu_T) - \mathrm{KL}\big(q_T\big\|\hat{p}^{\vartheta}_{n,T}\big) \tag{F.3}$$

almost surely.

Proof. Condition on the training sample. Let $\gamma_d := \mathcal{N}(0, I_d)$ denote the stationary OU law. Since $q_0 \in \mathcal{P}_2(\mathbb{R}^d)$, Lemma B.1, applied with $(p_0, q_0) = (\gamma_d, q_0)$, gives $\mathrm{KL}(q_{t_0}\|\gamma_d) < \infty$. The forward OU path laws on $[t_0, T]$ initialized from q_{t_0} and γ_d have the same conditional path kernel given their initial coordinate. Hence the entropy chain rule shows that their path-space relative entropy equals $\mathrm{KL}(q_{t_0}\|\gamma_d)$. The finite-entropy time-reversal theorem therefore realizes the time reversal of the former path law as a non-explosive weak solution of the exact shifted reverse equation; see Anderson (1982) and Cattiaux et al. (2023, Theorem 4.9). Moreover, the map $(x,t) \mapsto q_t(x)$ is strictly positive and smooth on $\mathbb{R}^d \times (0,\infty)$. Thus the exact reverse drift is jointly locally bounded in space and reverse time and locally Lipschitz in space. Together with the non-explosive time-reversed construction, local Lipschitzness gives uniqueness in law up to the terminal reverse time.

Conditional on the training sample, $\hat{p}^{\vartheta}_{n,t}$ is a finite mixture of nondegenerate Gaussians and is therefore strictly positive and smooth. The empirical-component scores are $-(x - a_t X_i)/\rho_t$, while the blanket-component score is $-x/v_t$. Since $\|X_i\|_2 \le D$, $\rho_t \ge \rho_{t_0}$, and $v_t \ge \rho_{t_0}$, their convex combination satisfies

$$\|\nabla \log \hat{p}^{\vartheta}_{n,t}(x)\|_2 \le \frac{\|x\|_2 + D}{\rho_{t_0}}, \qquad t \in [t_0, T].$$

Consequently, the estimated reverse drift is jointly locally bounded, locally Lipschitz in space, and of linear growth uniformly in reverse time. The reverse equation therefore has a unique non-explosive strong solution, and both reverse drifts satisfy the conditions of Lemma F.1.

Finally, $q_0, \hat{\mu}^{\vartheta}_{n,0} \in \mathcal{P}_2(\mathbb{R}^d)$, so Lemma B.2 gives

$$\frac{1}{2}\int_{t_0}^{T} \|\nabla \log q_t - \nabla \log \hat{p}^{\vartheta}_{n,t}\|^2_{L^2(q_t)}\,\mathrm{d}t = \mathrm{KL}(q_{t_0}\|\hat{p}^{\vartheta}_{n,t_0}) - \mathrm{KL}(q_T\|\hat{p}^{\vartheta}_{n,T}).$$

Substituting this identity into (F.2) proves (F.3). □

F.2 Standard Gaussian Initialization

Throughout this subsection, let $\gamma_d := \mathcal{N}(0, I_d)$.

Lemma F.3 (Standard Gaussian terminal mismatch). *Under Assumption 1, for every $T > 0$ and $q_0 \in \mathcal{Q}_\varepsilon(p_0)$,*

$$\mathrm{KL}(q_T\|\gamma_d) \le \frac{\varepsilon^2}{\tau_T} + \frac{e^{-T}D^2}{1+e^{-T}} - \frac{d}{2}\log(1-e^{-2T}). \tag{F.4}$$

Proof. Lemma B.3, applied with $(q, p, r) = (q_T, p_T, \gamma_d)$, gives

$$\mathrm{KL}(q_T\|\gamma_d) \le 2\mathrm{KL}(q_T\|p_T) + D_2(p_T\|\gamma_d).$$

By Lemma B.1,

$$2\mathrm{KL}(q_T\|p_T) \le \frac{\varepsilon^2}{\tau_T}.$$

It remains to control $D_2(p_T\|\gamma_d)$. Write $a = a_T = e^{-T/2}$ and $\rho = \rho_T = 1 - e^{-T}$, and let $k_y(x) := \varphi_\rho(x - ay)$. A direct completion of the square gives

$$\int_{\mathbb{R}^d} \frac{k_y(x)^2}{\varphi_1(x)}\,\mathrm{d}x = (1-a^4)^{-d/2}\exp\left(\frac{a^2\|y\|_2^2}{1+a^2}\right).$$

Since $p_T(x) = \mathbb{E}_{Y\sim p_0} k_Y(x)$, Jensen's inequality gives

$$p_T(x)^2 \le \mathbb{E}_{Y\sim p_0} k_Y(x)^2.$$

Consequently, Tonelli's theorem and $\|Y\|_2 \le D$ yield

$$D_2(p_T\|\gamma_d) \le \log \mathbb{E}_{Y\sim p_0}\int \frac{k_Y(x)^2}{\varphi_1(x)}\,\mathrm{d}x \le \frac{a^2D^2}{1+a^2} - \frac{d}{2}\log(1-a^4).$$

Combining the two estimates and substituting $a^2 = e^{-T}$ proves (F.4). □

Corollary F.4 (Standard Gaussian initialization). *Under Assumption 1, set $\vartheta = D/\sqrt{d}$, take $\hat{s}_n = \hat{s}^{\vartheta}_n$ in (F.1), and initialize the reverse process from $\nu_T = \gamma_d$. If $\tau_0 \le 1$ and $T \ge \log 2$, then*

$$\sup_{q_0\in\mathcal{Q}_\varepsilon(p_0)} \mathbb{E}\,\mathrm{KL}\big(q_{t_0}\big\|\hat{\mu}^{\gamma}_{n,t_0}\big) \le \frac{C_{d,D}}{2(n+1)\tau_0^{d/2}} + \frac{\varepsilon^2}{\tau_0} + (D^2 + 2\varepsilon^2)e^{-T} + de^{-2T}. \tag{F.5}$$

Proof. With $\vartheta = D/\sqrt{d}$, taking $\nu_T = \gamma_d$ in (F.3) and dropping its final nonpositive term gives

$$\mathrm{KL}\big(q_{t_0}\big\|\hat{\mu}^{\gamma}_{n,t_0}\big) \le \mathrm{KL}\big(q_{t_0}\big\|\hat{p}^{\vartheta}_{n,t_0}\big) + \mathrm{KL}(q_T\|\gamma_d).$$

Taking expectations and the supremum, (6.1) and Lemma F.3 yield

$$\sup_{q_0\in\mathcal{Q}_\varepsilon(p_0)} \mathbb{E}\,\mathrm{KL}\big(q_{t_0}\big\|\hat{\mu}^{\gamma}_{n,t_0}\big) \le \frac{C_{d,D}}{2(n+1)\tau_0^{d/2}} + \frac{\varepsilon^2}{\tau_0} + \frac{\varepsilon^2}{\tau_T} + \frac{e^{-T}D^2}{1+e^{-T}} - \frac{d}{2}\log(1-e^{-2T}).$$

Since $T \ge \log 2$,

$$\frac{1}{\tau_T} \le 2e^{-T}, \qquad \frac{e^{-T}D^2}{1+e^{-T}} \le D^2 e^{-T}, \qquad -\frac{d}{2}\log(1-e^{-2T}) \le de^{-2T},$$

where the last inequality uses $-\log(1-x) \le 2x$ for $0 \le x \le 1/2$. Substitution proves (F.5). □

F.3 Intrinsic-Dimensional KL Upper Bound

Theorem F.5. *Under the assumptions and construction of Theorem 5.1, suppose that $k \ge 2$ and $\tau_0 \le 1$. Then*

$$\sup_{q_0\in\mathcal{Q}_\varepsilon(p_0)} \mathbb{E}\big[\mathrm{KL}\left(q_{t_0}\big\|\hat{p}^{\mathrm{sub}}_{n,t_0}\right)\big] \le \frac{\tilde{C}_{k,D}}{2n\tau_0^{k/2}} + \frac{\varepsilon^2}{\tau_0}. \tag{F.6}$$

Here $\tilde{C}_{k,D}$ is the constant in Theorem 5.1.

Proof. Under the assumptions and construction of Theorem 5.1, taking the supremum over $q_0 \in \mathcal{Q}_\varepsilon(p_0)$ in (E.9) yields

$$\sup_{q_0\in\mathcal{Q}_\varepsilon(p_0)} \mathbb{E}\big[\mathrm{KL}\big(q_{t_0}\big\|\hat{p}^{\mathrm{sub}}_{n,t_0}\big)\big] \le \log\left(1 + \frac{M_{\vartheta_k,t_0}-1}{n_2+1}\right) + \frac{\varepsilon^2}{\tau_0} + \frac{kD^2}{(n_1+k)\tau_0}. \tag{F.7}$$

For $k \ge 2$ and $\tau_0 \le 1$, simplifying (F.7) as in the derivation of (5.2) gives (F.6).

We also verify that self-consistent initialization yields the conditional output law $\hat{p}^{\mathrm{sub}}_{n,t_0}$. Condition on the full sample and consider the reverse process (F.1) with $\hat{s}_n = \hat{s}^{\mathrm{sub}}_n$ and self-consistent initialization $\nu_T = \hat{p}^{\mathrm{sub}}_{n,T}$. By (E.6), the score is a convex combination of the component scores $-(x-a_tY_i)/\rho_t$ and $-\Sigma_t^{-1}x$, where $\Sigma_t = \rho_t I_d + a_t^2\vartheta_k^2\hat{\Pi}$. Since $\Sigma_t \succeq \rho_t I_d$ and $\|Y_i\|_2 \le D$,

$$\|\hat{s}^{\mathrm{sub}}_n(x,t)\|_2 \le \frac{\|x\|_2 + D}{\rho_{t_0}}, \qquad t \in [t_0, T].$$

The mixture density is strictly positive and jointly smooth in (x,t) for $t > 0$. Thus, on each compact spatial set, its score has bounded spatial derivatives and is locally Lipschitz, uniformly over $t \in [t_0, T]$. Together with the linear-growth bound, this gives a unique non-explosive strong solution of (F.1) with the stated score and initialization.

To identify the output law, start a forward OU process from $\hat{\mu}^{\mathrm{sub}}_{n,0}$ and restrict it to $[t_0, T]$. The moment bound (E.5) and Lemma B.1 give $\mathrm{KL}(\hat{p}^{\mathrm{sub}}_{n,t_0}\|\mathcal{N}(0, I_d)) < \infty$. On $[t_0, T]$, this process and the stationary OU process have the same conditional path law given their position at t_0. By the entropy chain rule, their relative entropy on path space equals this finite initial KL divergence. By the finite-entropy time-reversal theorem (Cattiaux et al., 2023, Theorem 4.9), the reversed OU process is a weak solution of the estimated reverse equation with marginals $\hat{p}^{\mathrm{sub}}_{n,T-t}$. By uniqueness in law, the strong solution has the same marginals. Its output law at reverse time $T - t_0$ is therefore $\hat{p}^{\mathrm{sub}}_{n,t_0}$. □

F.4 Intrinsic-Dimensional KL Minimax Rate

For the same source class $\mathcal{P}_{k,D}$, define the minimax positive-time KL risk by

$$\mathfrak{K}^{\star}_{n,\varepsilon}(k, D; t_0) := \inf_{\hat{\mu}_n} \sup_{p_0\in\mathcal{P}_{k,D}} \sup_{q_0\in\mathcal{Q}_\varepsilon(p_0)} \mathbb{E}[\mathrm{KL}(q_{t_0}\|\hat{\mu}_n)],$$

with the same estimator class as in Subsection 6.1.

Corollary F.6. *Fix $k \geq 2$ and $D > 0$, and let $d \geq k$ be arbitrary. Let $c_0 \in (0, \log 2]$ and $\tilde{c}_{k,D} > 0$ be as in Theorem 6.1 with dimension k. If $n \geq 2$ and*

$$0 < t_0 \leq c_0, \qquad n\tau_0^{k/2} \geq 1, \qquad 0 \leq \varepsilon \leq \sqrt{\tau_0},$$

then

$$\tilde{c}_{k,D}\left[\frac{1}{n\tau_0^{k/2}} + \frac{\varepsilon^2}{\tau_0}\right] \leq \mathfrak{K}^\star_{n,\varepsilon}(k, D; t_0) \leq \frac{\tilde{C}_{k,D}}{2n\tau_0^{k/2}} + \frac{\varepsilon^2}{\tau_0}.$$

Here $\tilde{C}_{k,D}$ is the constant in Theorem 5.1.

Proof. The upper bound follows from Theorem F.5, since $\tau_0 \leq 1$. For the lower bound, use the embedding and notation from the proof of Corollary E.3. Given any ambient distribution estimator $\hat{\mu}_n$, apply it to the embedded observations and define $\tilde{\mu}_n := (U^\top)_\# \hat{\mu}_n$, retaining any internal randomization. Since $(U^\top)_\# q_{t_0} = \bar{q}_{t_0}$, the data processing inequality gives

$$\mathrm{KL}(q_{t_0} \| \hat{\mu}_n) \geq \mathrm{KL}\big((U^\top)_\# q_{t_0} \,\big\|\, (U^\top)_\# \hat{\mu}_n\big) = \mathrm{KL}(\bar{q}_{t_0} \| \tilde{\mu}_n).$$

The same reduction bounds this minimax risk below by its k-dimensional counterpart. The lower bound follows from Theorem 6.1 with dimension k. □

G Neural Approximation of the Gaussian-Blanket Score

We approximate the isotropic Gaussian-blanket score from Subsection 4.1, uniformly over $\mathbb{R}^d \times [t_0, T]$. We separate the known score of the Gaussian blanket from the sample-dependent correction, retain the former exactly, and approximate the latter by a ReLU network. The construction combines spatial localization with elementary ReLU arithmetic and gives explicit width, depth, and sparsity bounds, together with a time-integrated bound on the compatibility defect, uniformly over samples in the support ball. The final subsection transfers this approximation guarantee to robust score risk via Corollary G.4.

G.1 Uniform Approximation Guarantee

Fix

$$\Theta = (d, D, \vartheta, t_0, T), \qquad d \geq 1, \quad D, \vartheta > 0, \quad 0 < t_0 < T < \infty,$$

and let $X_1, \ldots, X_n \in \mathcal{B}(0, D)$, with $n \geq 1$. Write $X_{1:n} := (X_1, \ldots, X_n)$ and initially regard the sample as deterministic. Retain a_t, ρ_t, τ_0, and the Gaussian density φ_v from Subsection 4.1, and write

$$v_t = \rho_t + a_t^2 \vartheta^2, \qquad M := M_{\vartheta, t_0} = \exp\left(\frac{D^2}{2\vartheta^2}\right)\left(1 + \frac{\vartheta^2}{\tau_0}\right)^{d/2}.$$

Recall the density estimator (4.2),

$$\hat{p}^\vartheta_{n,t}(x) = \frac{M\varphi_{v_t}(x) + \sum_{i=1}^n \varphi_{\rho_t}(x - a_t X_i)}{n + M}.$$

Since the Gaussian blanket has score $\nabla_x \log \varphi_{v_t}(x) = -x/v_t$, we only need to approximate the correction

$$G_{n,t}(x) := \nabla_x \log \frac{\hat{p}^\vartheta_{n,t}(x)}{\varphi_{v_t}(x)} = \nabla_x \log \hat{p}^\vartheta_{n,t}(x) + \frac{x}{v_t}.$$

Lemma G.2 shows that this correction decays at spatial infinity, allowing its uniform approximation to be reduced to a bounded region.

We use feedforward ReLU networks with layer dimensions $m_0, \ldots, m_L$ and affine maps $\mathcal{A}_j(z) = A_j z + b_j$, where $A_j \in \mathbb{R}^{m_j \times m_{j-1}}$ and $b_j \in \mathbb{R}^{m_j}$. The network has the form

$$\Phi = \mathcal{A}_L \circ \mathrm{ReLU} \circ \cdots \circ \mathrm{ReLU} \circ \mathcal{A}_1,$$

with $\mathrm{ReLU}(z) = z_+$ applied componentwise between consecutive affine maps. Its width, depth, and sparsity are

$$W(\Phi) := \max_{1\le j<L} m_j, \qquad L(\Phi) := L, \qquad S(\Phi) := \sum_{j=1}^{L} \big(\|A_j\|_0 + \|b_j\|_0\big),$$

where $\|\cdot\|_0$ counts nonzero entries. Depth includes the output affine layer; for a purely affine network ($L = 1$), set $W(\Phi) = 0$. Weights and biases may be arbitrary real numbers. For the network approximating $G_{n,t}$, $m_0 = d + 1$ and $m_L = d$; the complexity bounds below concern this correction network.

All vector norms are Euclidean, and all logarithms are natural unless a base is indicated. The $L^\infty(\mathbb{R}^d)$ norm denotes the essential supremum with respect to Lebesgue measure; for continuous vector fields, $\|f\|_{L^\infty(\mathbb{R}^d)} = \sup_{x\in\mathbb{R}^d} \|f(x)\|_2$. Constants denoted by C_Θ are finite and depend only on Θ; their values may change from line to line.

Proposition G.1 (Uniform neural score approximation)**.** *For every $n \ge 1$, $X_{1:n} \in \mathcal{B}(0, D)^n$, and $\eta \in (0, 1/2)$, there exists a ReLU network $\Phi_{X_{1:n},\eta} : \mathbb{R}^{d+1} \to \mathbb{R}^d$ such that*

$$\sup_{\substack{x\in\mathbb{R}^d \\ t\in[t_0,T]}} \|\Phi_{X_{1:n},\eta}(x, t) - G_{n,t}(x)\|_2 \le \eta. \tag{G.1}$$

Writing $\Phi = \Phi_{X_{1:n},\eta}$ and $\ell_{n,\eta} = 1 + \log((n + 1)/\eta)$, we have

$$W(\Phi) \le C(n + d), \quad L(\Phi) \le C_\Theta \ell_{n,\eta}^2, \quad S(\Phi) \le C_\Theta \big[nd + (n + d)\ell_{n,\eta}^2\big], \tag{G.2}$$

where C is absolute. For fixed Θ, n, η, the networks admit a common architecture and set of allowed parameter positions, with parameters continuous in $X_{1:n}$.

G.2 Ratio Representation and Localization

To obtain a representation with a denominator bounded below by one, we compare each sample kernel with the Gaussian blanket. Define

$$R_{i,t}(x) := \frac{\varphi_{\rho_t}(x - a_t X_i)}{M\varphi_{v_t}(x)}, \qquad U_{n,t}(x) := \sum_{i=1}^{n} R_{i,t}(x).$$

By Lemma B.4, for every $x \in \mathbb{R}^d$ and $t \in [t_0, T]$,

$$0 < R_{i,t}(x) \le 1 \quad (i = 1, \dots, n), \qquad 1 \le 1 + U_{n,t}(x) \le n + 1. \tag{G.3}$$

Factoring the density gives

$$\hat{p}^{\vartheta}_{n,t} = \frac{M\varphi_{v_t}}{n + M}(1 + U_{n,t}), \qquad G_{n,t} = \nabla \log(1 + U_{n,t}) = \frac{\nabla U_{n,t}}{1 + U_{n,t}}.$$

Thus the Gaussian density cancels from the score correction, leaving a denominator bounded below by one. To compute the numerator, expand the Gaussian log ratio as

$$\log R_{i,t}(x) = c_t - \frac{\lambda_t}{2}\|x\|_2^2 + \gamma_t x^\top X_i - \beta_t \|X_i\|_2^2, \tag{G.4}$$

where the four time-dependent coefficients are

$$c_t := \frac{d}{2}\log\frac{v_t}{\rho_t} - \log M, \qquad \lambda_t := \rho_t^{-1} - v_t^{-1} = \frac{a_t^2\vartheta^2}{\rho_t v_t},$$
$$\gamma_t := \frac{a_t}{\rho_t}, \qquad \beta_t := \frac{a_t^2}{2\rho_t}.$$

Differentiation yields $\nabla R_{i,t}(x) = R_{i,t}(x)(\gamma_t X_i - \lambda_t x)$. Introducing the weighted sample sum $V_{n,t}(x) := \sum_{i=1}^n R_{i,t}(x) X_i$, we therefore obtain

$$G_{n,t}(x) = \frac{\gamma_t V_{n,t}(x) - \lambda_t x U_{n,t}(x)}{1 + U_{n,t}(x)}. \tag{G.5}$$

The next lemma shows that the correction is uniformly small outside a ball, providing the localization needed for the network construction.

Lemma G.2 (Localization of the score correction)**.** *For every $\eta \in (0, 1/2)$, there is a radius $R = R_{n,\eta} \geq 1$, independent of the sample locations, such that*

$$R \leq C_\Theta \sqrt{1 + \log \frac{n+1}{\eta}}, \qquad \sup_{\substack{\|x\|_2 \geq R \\ t \in [t_0, T]}} \|G_{n,t}(x)\|_2 \leq \frac{\eta}{4}. \tag{G.6}$$

Proof. Applying the completed-square identity (B.11) with $y = X_i$ gives

$$R_{i,t}(x) = R_{i,t}(z_{i,t}) \exp\left(-\frac{\lambda_t}{2}\|x - z_{i,t}\|_2^2\right), \qquad z_{i,t} = \frac{v_t}{a_t \vartheta^2} X_i.$$

Here $0 < R_{i,t}(z_{i,t}) \leq 1$ by (G.3), and differentiation gives

$$\nabla R_{i,t}(x) = -\lambda_t (x - z_{i,t}) R_{i,t}(x).$$

For the tail estimate, set

$$\lambda_- := \min_{t \in [t_0,T]} \lambda_t, \qquad C_* := D \max_{t \in [t_0,T]} \frac{v_t}{a_t \vartheta^2}, \qquad B_* := \max\left\{1, \sqrt{\frac{2}{e} \max_{t \in [t_0,T]} \lambda_t}\right\}.$$

These constants depend only on Θ, and $\lambda_- > 0$ because λ_t is continuous and strictly positive on $[t_0, T]$. Set $u = \sqrt{\lambda_t}\|x - z_{i,t}\|_2$. The inequality $ue^{-u^2/2} \leq \sqrt{2/e}\, e^{-u^2/4}$ for $u \geq 0$ then gives

$$\begin{aligned} \|\nabla R_{i,t}(x)\|_2 &= \lambda_t \|x - z_{i,t}\|_2 R_{i,t}(z_{i,t}) e^{-\lambda_t \|x - z_{i,t}\|_2^2/2} \\ &\leq \sqrt{\lambda_t}\, u e^{-u^2/2} \leq \sqrt{\frac{2\lambda_t}{e}}\, e^{-u^2/4} \\ &\leq B_*\, e^{-\lambda_- \|x - z_{i,t}\|_2^2/4}. \end{aligned}$$

Using $G_{n,t} = \nabla U_{n,t}/(1 + U_{n,t})$ and $1 + U_{n,t} \geq 1$, the triangle inequality and $\|z_{i,t}\|_2 \leq C_*$ give, for $\|x\|_2 \geq C_*$,

$$\|G_{n,t}(x)\|_2 \leq n B_*\, e^{-\lambda_- (\|x\|_2 - C_*)^2/4}. \tag{G.7}$$

Choose

$$R_{n,\eta} := \max\left\{1,\ C_* + \sqrt{\frac{4}{\lambda_-} \log \frac{4nB_*}{\eta}}\right\}. \tag{G.8}$$

Substitution into (G.7) proves the tail bound in (G.6). Since $\log(4nB_*/\eta) \leq (1 + \log(4B_*))\ell_{n,\eta}$ and $\ell_{n,\eta} \geq 1$, the radius bound in this lemma holds with

$$C_\Theta = 1 + C_* + 2\sqrt{\frac{1 + \log(4B_*)}{\lambda_-}},$$

which depends only on Θ. □

G.3 Elementary ReLU Arithmetic

For $a < b$, the clipping map

$$\mathrm{clip}_{[a,b]}(u) = a + (u - a)_+ - (u - b)_+$$

is represented exactly by a ReLU network of constant width and depth. Clipping cannot increase the distance to a point in $[a, b]$.

We use multiplication, negative-exponential, and reciprocal networks to approximate the score correction in (G.5). The logarithm, evaluated on a fixed positive interval, is used only to construct the time coefficient c_t. The square and multiplication modules build on Yarotsky (2017, Propositions 2–3). The remaining modules are constructed from these building blocks using repeated squaring and elementary series expansions.

Lemma G.3 (Elementary ReLU arithmetic modules)**.** *Fix $0 < \xi < 1/2$. The following ReLU networks exist with width at most an absolute constant C. In each item, L and S denote the depth and sparsity of the stated network.*

(i) Multiplication: for $B \geq 1$, $\mathrm{Mult}_{B,\xi} : \mathbb{R}^2 \to \mathbb{R}$ satisfies

$$\sup_{|u|,|v|\leq B} |\mathrm{Mult}_{B,\xi}(u,v) - uv| \leq \xi, \qquad L, S \leq C(1 + \log B + \log \xi^{-1}).$$

(ii) Negative exponential: for $B \geq 1$, $\mathrm{Exp}_{B,\xi} : \mathbb{R} \to [0,1]$ satisfies

$$\sup_{z\in[0,B]} |\mathrm{Exp}_{B,\xi}(z) - e^{-z}| \leq \xi, \qquad L, S \leq C(1 + \log B + \log \xi^{-1})^2.$$

(iii) Reciprocal: for $B \geq 1$, $\mathrm{Rec}_{B,\xi} : \mathbb{R} \to [0,1]$ satisfies

$$\sup_{z\in[1,B]} |\mathrm{Rec}_{B,\xi}(z) - z^{-1}| \leq \xi, \qquad L, S \leq C(1 + \log B + \log \xi^{-1})^2.$$

(iv) Logarithm: for $0 < b < B < \infty$, $\mathrm{Log}_{b,B,\xi} : \mathbb{R} \to \mathbb{R}$ satisfies

$$\sup_{z\in[b,B]} |\mathrm{Log}_{b,B,\xi}(z) - \log z| \leq \xi, \qquad L, S \leq C\frac{B}{b}\left(1 + \log\frac{B}{b} + \log \xi^{-1}\right)^2.$$

Proof. *(i) Multiplication.* We first construct a square network. Let S_m be the piecewise-linear interpolant of u^2 on the dyadic grid $\{k2^{-m} : 0 \leq k \leq 2^m\}$, so $S_0(u) = u$ on $[0,1]$. When refining from level $m-1$ to level m, the values at old nodes remain unchanged. At the midpoint of each coarse interval $[a,b]$, the old interpolant exceeds u^2 by

$$S_{m-1}\left(\frac{a+b}{2}\right) - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{4} = 4^{-m}.$$

To implement this correction, use the tent map and its iterates

$$g(u) = 2u_+ - 4(u - 1/2)_+ + 2(u-1)_+, \qquad g_j = g^{\circ j}.$$

The function g_m is zero at the old nodes, one at the new midpoints, and linear between consecutive nodes of the refined grid. Hence

$$S_m = S_{m-1} - 4^{-m} g_m, \qquad S_m(u) = u - \sum_{j=1}^{m} 4^{-j} g_j(u).$$

On each refined interval $[a,b]$, the interpolation error is $(u-a)(b-u)$, so

$$\sup_{u\in[0,1]} |S_m(u) - u^2| \leq 4^{-m-1}.$$

Keeping the current iterate g_j and the partial sum gives constant width and $O(m+1)$ depth and sparsity. For later use, denote by $\mathrm{Sq}_{\xi_*} : \mathbb{R} \to [0,1]$ the network obtained by choosing the smallest integer $m \geq 0$ with $4^{-m-1} \leq \xi_*$ and clipping S_m to $[0,1]$. For $0 < \xi_* < 1/2$, it approximates u^2 on $[0,1]$ with error at most ξ_* and depth and sparsity $O(1 + \log \xi_*^{-1})$.

For $|u|, |v| \leq B$, use

$$uv = B^2\left[\left(\frac{|u+v|}{2B}\right)^2 - \left(\frac{|u-v|}{2B}\right)^2\right].$$

The absolute values are exact ReLU operations. Replacing the two squares by parallel copies of $\mathrm{Sq}_{\xi/(2B^2)}$ gives error at most ξ and the claimed complexity.

(ii) Negative exponential. We approximate e^{-z} on $[0,B]$ by $(1 - z/N)^N$ and implement this power by repeated squaring. Let $N = 2^k$ be the smallest power of two with

$$N \geq \max\{2B, 2B^2/\xi\}.$$

For $0 \le w \le 1/2$, the logarithm series gives

$$0 \le -\log(1-w) - w = \sum_{r=2}^{\infty} \frac{w^r}{r} \le \frac{w^2}{2(1-w)} \le w^2.$$

Taking $w = z/N$ yields $-z - z^2/N \le N\log(1-z/N) \le -z$. Exponentiating and using $1-e^{-r} \le r$ for $r \ge 0$, we obtain

$$0 \le e^{-z} - (1-z/N)^N \le e^{-z}(1-e^{-z^2/N}) \le B^2/N \le \xi/2. \tag{G.9}$$

It remains to approximate the power by a network. The exact recursion $y_0 = 1 - z/N$, $y_{j+1} = y_j^2$ stays in $[0,1]$ and ends at $y_k = (1-z/N)^N$. With $\xi_* = \xi/(2N)$, use the square network from part (i) to define

$$\tilde{y}_0 = \mathrm{clip}_{[0,1]}(y_0), \qquad \tilde{y}_{j+1} = \mathrm{Sq}_{\xi_*}(\tilde{y}_j), \qquad 0 \le j < k.$$

All approximate iterates also lie in $[0,1]$, and the initial clipping is inactive for $z \in [0,B]$. Since $|u^2 - v^2| \le 2|u-v|$ on $[0,1]$, the square approximation gives

$$|\tilde{y}_{j+1} - y_{j+1}| \le 2|\tilde{y}_j - y_j| + \xi_*, \qquad |\tilde{y}_k - y_k| \le (N-1)\xi_* \le \xi/2.$$

Define $\mathrm{Exp}_{B,\xi}(z) := \tilde{y}_k$. Together with (G.9), the last estimate proves the required error bound. The initial clipping and the range of Sq_{ξ_*} ensure that the output lies in $[0,1]$ for every real input. There are $k = O(1 + \log B + \log \xi^{-1})$ squaring stages, each with depth and sparsity $O(1 + \log \xi_*^{-1})$ of the same order. Only the current iterate is needed, so the width remains bounded. This proves (ii).

(iii) Reciprocal. For $z \in [1,B]$, set $a = 1 - z/B$. The geometric series gives

$$z^{-1} = \frac{1}{B} \sum_{r=0}^{\infty} a^r.$$

We approximate this series by partial sums whose number of terms doubles at each stage. Specifically, set

$$p_j = a^{2^j}, \qquad h_j = \frac{1}{B} \sum_{r=0}^{2^j - 1} a^r = \frac{1 - a^{2^j}}{z}.$$

These quantities satisfy the recursion

$$p_0 = a, \quad h_0 = B^{-1}, \qquad p_{j+1} = p_j^2, \quad h_{j+1} = h_j(1 + p_j).$$

Since $z \ge 1$ and $0 \le a \le 1 - 1/B$, we have $p_j, h_j \in [0,1]$ and the truncation error satisfies

$$|h_j - z^{-1}| = \frac{a^{2^j}}{z} \le (1 - 1/B)^{2^j} \le e^{-2^j/B}. \tag{G.10}$$

Choose

$$k = \left\lceil \log_2 \left(B \log \frac{2}{\xi} \right) \right\rceil, \qquad \xi_* = \frac{\xi}{2 \cdot 3^k}.$$

Starting from $\tilde{p}_0 = \mathrm{clip}_{[0,1]}(a)$ and $\tilde{h}_0 = B^{-1}$, define

$$\tilde{p}_{j+1} = \mathrm{Sq}_{\xi_*}(\tilde{p}_j), \qquad \tilde{h}_{j+1} = \mathrm{clip}_{[0,1]}\big(\mathrm{Mult}_{2,\xi_*}(\tilde{h}_j, 1 + \tilde{p}_j)\big).$$

Both updates use the preceding pair of iterates; initial clipping is inactive for $z \in [1,B]$. Define $\mathrm{Rec}_{B,\xi}(z) := \tilde{h}_k(z)$. For fixed $z \in [1,B]$, set $e_j := \max\{|\tilde{p}_j - p_j|, |\tilde{h}_j - h_j|\}$, with the dependence on z suppressed. Then $e_0 = 0$ and

$$e_{j+1} \le 3e_j + \xi_*, \qquad e_k \le \frac{3^k - 1}{2}\xi_* \le \xi/4.$$

Indeed, the square map is 2-Lipschitz on $[0,1]$, while

$$|\tilde{h}(1+\tilde{p}) - h(1+p)| \le 2|\tilde{h} - h| + |\tilde{p} - p|$$

for $p, \tilde{p}, h, \tilde{h} \in [0,1]$. The choice of k makes (G.10) at most $\xi/2$. Hence the total error is less than ξ. The output is clipped to $[0,1]$, and only a fixed number of scalars is carried between stages. Since

$$k \le C(1 + \log B + \log \xi^{-1}), \qquad 1 + \log \xi_*^{-1} \le C(1 + \log B + \log \xi^{-1}),$$

part (i) proves the claimed complexity.

(iv) Logarithm. Let $u = 1 - z/B$ and $q = 1 - b/B \in (0,1)$. For $z \in [b, B]$,

$$\log z = \log B - \sum_{j=1}^{\infty} \frac{u^j}{j}, \qquad 0 \le u \le q.$$

Set $\kappa := B/b > 1$ and choose

$$m = \left\lceil \kappa \log \frac{2\kappa}{\xi} \right\rceil.$$

Since $q = 1 - \kappa^{-1} \le e^{-1/\kappa}$, the omitted tail satisfies

$$\sum_{j=m+1}^{\infty} \frac{u^j}{j} \le \frac{q^{m+1}}{1-q} \le \kappa e^{-(m+1)/\kappa} \le \xi/2.$$

Approximate the powers recursively, starting at $\tilde{v}_0 = 1$, by

$$\tilde{v}_j = \mathrm{clip}_{[0,1]}\big(\mathrm{Mult}_{1,\xi_*}(\tilde{v}_{j-1}, u)\big), \qquad \xi_* = \frac{\xi}{2m}.$$

Since $u, \tilde{v}_{j-1} \in [0,1]$ and clipping cannot increase the distance to u^j, the multiplication error gives

$$|\tilde{v}_j - u^j| \le \xi_* + u|\tilde{v}_{j-1} - u^{j-1}| \le \xi_* + |\tilde{v}_{j-1} - u^{j-1}|.$$

Starting from $\tilde{v}_0 = 1$, induction therefore gives $|\tilde{v}_j - u^j| \le j\xi_*$. Define the output network by

$$\mathrm{Log}_{b,B,\xi}(z) := \log B - \sum_{j=1}^{m} \frac{\tilde{v}_j(z)}{j}.$$

Its arithmetic error relative to the exact truncated series is at most $\sum_{j=1}^{m} j\xi_*/j = m\xi_* = \xi/2$. Adding the truncation error gives total error at most ξ. Only u, the current power, and the running sum need to be stored. Thus the width is bounded. Since $m \le C\kappa(1 + \log\kappa + \log\xi^{-1})$, the depth and sparsity satisfy

$$L, S \le Cm(1 + \log(m/\xi)) \le C\kappa(1 + \log\kappa + \log\xi^{-1})^2,$$

where C is universal. □

G.4 Proof of the Uniform Approximation Guarantee

Proof of Proposition G.1 The network implements the ratio formula (G.5) through the following computations:

$$\begin{aligned}(x,t) &\longmapsto \big(\|x\|_2^2, (x^\top X_i)_{i=1}^n, c_t, \lambda_t, \gamma_t, \beta_t\big) \\ &\longmapsto (R_{i,t}(x))_{i=1}^n \longmapsto \big(U_{n,t}(x), V_{n,t}(x)\big) \longmapsto G_{n,t}(x).\end{aligned}$$

The spatial input and time coefficients are carried forward for use in the final quotient. We approximate these operations on a cube, constructing each ratio approximation $\tilde{R}_i$ to take values in $[0,1]$. The resulting denominator $1 + \sum_{i=1}^n \tilde{R}_i$ then lies in $[1, n+1]$, allowing us to apply the reciprocal approximation. The tail estimate then extends the approximation to the whole space.

Step 1: Choose the internal accuracy and approximate the time coefficients. Choose $R = R_{n,\eta}$ from (G.8). Throughout the local construction, $(x,t) \in [-R,R]^d \times [t_0, T]$. Let

$$\begin{aligned}K &:= 1 + \max_{t\in[t_0,T]} \max\{|c_t|, \lambda_t, \gamma_t, \beta_t\}, \qquad Q := 1 + \sqrt{d}\,R + D, \\ \xi &:= \frac{\eta}{96(n+1)K^2Q^3}, \qquad B_{\mathrm{op}} := 8K(n+1)Q^2.\end{aligned} \tag{G.11}$$

Here K bounds the time coefficients and Q controls the spatial scales. We use ξ as the internal accuracy and B_{op} as a common input bound for the subsequent multiplication modules; the choice of ξ accounts for the accumulated error in Step 3. In particular, $K, Q \geq 1$, $Q \geq \sqrt{d}$, $D + \|x\|_2 \leq Q$, and $0 < \xi < 1/2$.

We construct a constant-width network with outputs $\tilde{c}_t, \tilde{\lambda}_t, \tilde{\gamma}_t, \tilde{\beta}_t$ in $[-K, K]$ such that

$$\sup_{t \in [t_0, T]} \max\{|\tilde{c}_t - c_t|, |\tilde{\lambda}_t - \lambda_t|, |\tilde{\gamma}_t - \gamma_t|, |\tilde{\beta}_t - \beta_t|\} \leq \xi. \tag{G.12}$$

Its depth and sparsity are $O_\Theta((1 + \log \xi^{-1})^2)$. To see this, set

$$b := \frac{1}{2} \min\{\rho_{t_0}, 1, \vartheta^2\} > 0, \qquad B_0 := 2 \max\{1, \vartheta^2\}.$$

Since $\rho_t \in [\rho_{t_0}, 1]$ and $v_t = (1 - e^{-t}) + \vartheta^2 e^{-t}$ is a convex combination of 1 and ϑ^2, both lie in $[b, B_0]$, an interval determined only by Θ.

Take an auxiliary accuracy $\xi_0 \in (0, 1/2)$. Apply $\mathrm{Exp}_{\max\{1,T\},\xi_0}$ to t and $\mathrm{Exp}_{\max\{1,T/2\},\xi_0}$ to $t/2$, and clip the outputs to $[e^{-T}, e^{-t_0}]$ and $[e^{-T/2}, e^{-t_0/2}]$, respectively. The resulting $\tilde{u}_t$ and $\tilde{a}_t$ approximate e^{-t} and a_t with error at most ξ_0. Define

$$\tilde{\rho}_t = 1 - \tilde{u}_t, \qquad \tilde{v}_t = 1 + (\vartheta^2 - 1)\tilde{u}_t.$$

They remain in $[b, B_0]$ and satisfy $|\tilde{\rho}_t - \rho_t| \leq \xi_0$ and $|\tilde{v}_t - v_t| \leq |\vartheta^2 - 1|\xi_0$.

Using the reciprocal module in Lemma G.3, set

$$\tilde{J}_{\rho,t} := b^{-1} \mathrm{Rec}_{B_0/b, b\xi_0}(\tilde{\rho}_t / b), \qquad \tilde{J}_{v,t} := b^{-1} \mathrm{Rec}_{B_0/b, b\xi_0}(\tilde{v}_t / b).$$

The inputs lie in $[1, B_0/b]$, and $b\xi_0 < 1/2$ because $b \leq 1/2$. These outputs lie in $[0, b^{-1}]$ and approximate $\tilde{\rho}_t^{-1}$ and $\tilde{v}_t^{-1}$ with error at most ξ_0. To include the input error, use the b^{-2}-Lipschitz property of the reciprocal on $[b, B_0]$. For example,

$$\begin{aligned} |\tilde{J}_{\rho,t} - \rho_t^{-1}| &\leq |\tilde{J}_{\rho,t} - \tilde{\rho}_t^{-1}| + |\tilde{\rho}_t^{-1} - \rho_t^{-1}| \\ &\leq \xi_0 + b^{-2}|\tilde{\rho}_t - \rho_t| \leq (1 + b^{-2})\xi_0. \end{aligned}$$

The estimate for $\tilde{J}_{v,t}$ follows in the same way.

We now form the four coefficient approximations before final clipping:

$$\begin{aligned} c_{\xi_0}(t) &= \tfrac{d}{2}\big[\mathrm{Log}_{b,B_0,\xi_0}(\tilde{v}_t) - \mathrm{Log}_{b,B_0,\xi_0}(\tilde{\rho}_t)\big] - \log M, \\ \lambda_{\xi_0}(t) &= \tilde{J}_{\rho,t} - \tilde{J}_{v,t}, \\ \gamma_{\xi_0}(t) &= \mathrm{Mult}_{b^{-1},\xi_0}(\tilde{a}_t, \tilde{J}_{\rho,t}), \\ \beta_{\xi_0}(t) &= \tfrac{1}{2}\,\mathrm{Mult}_{b^{-1},\xi_0}(\tilde{u}_t, \tilde{J}_{\rho,t}). \end{aligned}$$

The multiplication inputs lie in $[0, b^{-1}]$, since $\tilde{a}_t, \tilde{u}_t \in [0, 1]$ and $b^{-1} \geq 2$. The logarithm is b^{-1}-Lipschitz on $[b, B_0]$, so its input errors are controlled in the same way. The remaining operations are affine maps and products of bounded scalars. Thus each of the four errors is at most $C_{\mathrm{time}}\xi_0$ for a constant $C_{\mathrm{time}} \geq 1$ depending only on Θ. Choosing $\xi_0 = \xi / C_{\mathrm{time}}$ and clipping each output to $[-K, K]$ proves (G.12). There are a fixed number of scalar modules, so the arithmetic lemma also gives the claimed depth and sparsity.

Step 2: Approximate the ratios $R_{i,t}$ on the cube. For each coordinate, the square network constructed in the proof of Lemma G.3 gives $R^2\,\mathrm{Sq}_{\xi/(dR^2)}(|x_j|/R)$. This output lies in $[0, R^2]$ and approximates x_j^2 with error at most ξ/d on the cube. Summing over j, we obtain $\tilde{Z}(x)$ with

$$|\tilde{Z}(x) - \|x\|_2^2| \leq \xi, \qquad 0 \leq \tilde{Z}(x) \leq dR^2 \leq Q^2. \tag{G.13}$$

For each i, the inner product $x^\top X_i$ is an exact affine function of x, and $\|X_i\|_2^2$ is a fixed sample-dependent coefficient. Both $|x^\top X_i|$ and $\|X_i\|_2^2$ are at most Q^2. Unless stated otherwise, all subsequent products use $\mathrm{Mult}_{B_{\mathrm{op}},\xi}$. Write $\ell_i(x,t) := \log R_{i,t}(x)$ and approximate it by

$$\tilde{\ell}_i(x,t) := \tilde{c}_t - \frac{1}{2}\,\mathrm{Mult}_{B_{\mathrm{op}},\xi}(\tilde{\lambda}_t, \tilde{Z}(x)) + \mathrm{Mult}_{B_{\mathrm{op}},\xi}(\tilde{\gamma}_t, x^\top X_i) - \tilde{\beta}_t \|X_i\|_2^2. \tag{G.14}$$

The last term is exact multiplication by a fixed coefficient and hence an affine operation. Equations (G.12) and (G.13) give

$$|\tilde{\ell}_i - \ell_i| \le \xi + \tfrac{1}{2}(1+K+Q^2)\xi + (1+Q^2)\xi + Q^2\xi \le 6KQ^2\xi. \tag{G.15}$$

Also, $\ell_i \le 0$ by (G.3), while (G.4) gives $|\ell_i| \le 4KQ^2$. Set $H := 4KQ^2$ and let

$$\tilde{\ell}_i^{\mathrm{cl}} := \mathrm{clip}_{[-H,0]}(\tilde{\ell}_i), \qquad \tilde{R}_i := \mathrm{Exp}_{H,\xi}(-\tilde{\ell}_i^{\mathrm{cl}}).$$

Clipping does not increase the error in (G.15). Since $-\tilde{\ell}_i^{\mathrm{cl}}, -\ell_i \in [0, H]$ and $z \mapsto e^{-z}$ is 1-Lipschitz on $[0,\infty)$, adding the exponential-network error ξ gives

$$0 \le \tilde{R}_i \le 1, \qquad |\tilde{R}_i - R_{i,t}(x)| \le \xi + 6KQ^2\xi \le e_R, \qquad e_R := 7KQ^2\xi.$$

The last inequality uses $K, Q \ge 1$. All n ratio modules are evaluated in parallel.

Step 3: Assemble the score correction and bound its error. For this step, suppress (x,t) in the notation and write $U = U_{n,t}(x)$ and $V = V_{n,t}(x)$. Form the exact affine aggregates

$$\tilde{U} = \sum_{i=1}^{n} \tilde{R}_i, \qquad \tilde{V} = \sum_{i=1}^{n} \tilde{R}_i X_i.$$

Then

$$\begin{aligned} 0 \le \tilde{U} \le n, \qquad & \|\tilde{V}\|_2 \le nD, \\ |\tilde{U} - U| \le ne_R, \qquad & \|\tilde{V} - V\|_2 \le nDe_R. \end{aligned} \tag{G.16}$$

Let N be the exact numerator and $\overline{N}$ its approximation using exact products:

$$N := \gamma_t V - \lambda_t x U, \qquad \overline{N} := \tilde{\gamma}_t \tilde{V} - \tilde{\lambda}_t x \tilde{U}.$$

Using $U \le n$, $\|V\|_2 \le nD$ and $D + \|x\|_2 \le Q$, we have

$$\begin{aligned} \|\overline{N} - N\|_2 &\le K\|\tilde{V} - V\|_2 + \xi\|V\|_2 + K\|x\|_2\,|\tilde{U} - U| + \xi\|x\|_2 U \\ &\le nQ(Ke_R + \xi) \le 8nK^2Q^3\xi. \end{aligned} \tag{G.17}$$

Replacing these products by multiplication networks gives the implemented numerator $\tilde{N}$:

$$\begin{aligned} \tilde{w} &= \mathrm{Mult}_{B_{\mathrm{op}},\xi}(\tilde{\lambda}_t, \tilde{U}), \\ \tilde{N}_j &= \mathrm{Mult}_{B_{\mathrm{op}},\xi}(\tilde{\gamma}_t, \tilde{V}_j) - \mathrm{Mult}_{B_{\mathrm{op}},\xi}(\tilde{w}, x_j), \qquad 1 \le j \le d. \end{aligned}$$

All multiplication inputs are in range: $K, n, nD, Q \le B_{\mathrm{op}}$, and the first product gives $|\tilde{w}| \le Kn + \xi \le B_{\mathrm{op}}$.

The two final products contribute at most 2ξ, and the error in $\tilde{w}$ contributes at most $|x_j|\xi$. Hence

$$\begin{aligned} |\tilde{N}_j - \overline{N}_j| &\le \xi + \xi + |x_j|\xi = (2 + |x_j|)\xi, \\ \|\tilde{N} - \overline{N}\|_2 &\le (2\sqrt{d} + \|x\|_2)\xi \le 3Q\xi. \end{aligned}$$

Combining this with (G.17) gives

$$\|\tilde{N} - N\|_2 \le 11nK^2Q^3\xi, \qquad \|\tilde{N}\|_2 \le 4KnQ. \tag{G.18}$$

For the last bound, use $\|\overline{N}\|_2 \le KnQ$ and $\xi < 1$.

The denominator also improves the bound on the exact correction. Since $\|V\|_2 \le DU$, we have

$$\|G_{n,t}(x)\|_2 = \frac{\|N\|_2}{1+U} \le K(D + \|x\|_2)\frac{U}{1+U} \le KQ.$$

To compare the two quotients, we use the identity

$$\frac{\tilde{N}}{1+\tilde{U}} - G_{n,t}(x) = \frac{\tilde{N} - N + (U - \tilde{U})G_{n,t}(x)}{1+\tilde{U}}.$$

The denominator error is thus multiplied by a correction bounded by KQ, avoiding an additional factor of n. Since $1+\tilde{U} \geq 1$, equations (G.16) and (G.18) give

$$\begin{aligned}\left\| \frac{\tilde{N}}{1+\tilde{U}} - G_{n,t}(x) \right\|_2 &\leq \|\tilde{N} - N\|_2 + KQ|\tilde{U} - U| \\ &\leq 11nK^2Q^3\xi + KQne_R \leq 18nK^2Q^3\xi.\end{aligned} \tag{G.19}$$

To implement the division, use

$$\tilde{J} := \mathrm{Rec}_{n+1,\xi}(1+\tilde{U}), \qquad \Psi_j(x,t) := \mathrm{Mult}_{B_{\mathrm{op}},\xi}(\tilde{J}, \tilde{N}_j), \quad 1 \leq j \leq d.$$

These modules are evaluated within their approximation domains: $1+\tilde{U} \in [1, n+1]$, $\tilde{J} \in [0,1]$, and $|\tilde{N}_j| \leq 4KnQ \leq B_{\mathrm{op}}$. In particular, $|\tilde{J} - (1+\tilde{U})^{-1}| \leq \xi$, so

$$\left\| \Psi(x,t) - \frac{\tilde{N}}{1+\tilde{U}} \right\|_2 \leq \sqrt{d}\,\xi + \|\tilde{N}\|_2\xi \leq Q\xi + 4KnQ\xi.$$

Combining this with (G.19) yields

$$\begin{aligned}\|\Psi(x,t) - G_{n,t}(x)\|_2 &\leq 18nK^2Q^3\xi + Q\xi + 4KnQ\xi \\ &\leq 24nK^2Q^3\xi = \frac{\eta n}{4(n+1)} \leq \frac{\eta}{4}.\end{aligned} \tag{G.20}$$

Every approximation above is applied directly to the explicit score correction; no approximation of a density is differentiated.

Step 4: Extend to the whole space and verify complexity and measurability. Let $\mathcal{C}_R$ clip each spatial coordinate to $[-R, R]$ and set $\Phi_{X_{1:n},\eta}(x,t) := \Psi(\mathcal{C}_R(x), t)$. On the cube, (G.20) gives error at most $\eta/4$. Outside the cube, one clipped coordinate has magnitude R, so both $\|x\|_2$ and $\|\mathcal{C}_R(x)\|_2$ are at least R. By (G.6),

$$\begin{aligned}\|\Phi_{X_{1:n},\eta}(x,t) - G_{n,t}(x)\|_2 &\leq \|\Psi(\mathcal{C}_R(x),t) - G_{n,t}(\mathcal{C}_R(x))\|_2 \\ &\quad + \|G_{n,t}(\mathcal{C}_R(x))\|_2 + \|G_{n,t}(x)\|_2 \leq \frac{3\eta}{4}.\end{aligned}$$

Thus (G.1) holds on the whole space.

We now bound the width, depth, and sparsity. From (G.6) and (G.11),

$$Q \leq C_\Theta\sqrt{\ell_{n,\eta}}, \qquad 1 + \log H + \log B_{\mathrm{op}} + \log \xi^{-1} \leq C_\Theta \ell_{n,\eta}.$$

For the square modules, the underlying accuracy $\xi/(dR^2)$ gives $1 + \log(dR^2/\xi) \leq C_\Theta\ell_{n,\eta}$. The square, log-ratio, and final multiplication modules therefore have depth $O_\Theta(\ell_{n,\eta})$, while the time-coefficient, exponential, and reciprocal modules have depth $O_\Theta(\ell^2_{n,\eta})$. Each scalar arithmetic module has constant width, so its sparsity is bounded by a constant times its depth. With $O(n+d)$ parallel branches and only a fixed number of serial stages, these bounds give the width and depth estimates in (G.2).

For sparsity, the two sample-dependent affine transformations

$$x \longmapsto (x^\top X_i)_{i=1}^n, \qquad (r_i)_{i=1}^n \longmapsto \sum_{i=1}^n r_i X_i$$

use $O(nd)$ nonzero parameters in total. The coefficients $\|X_i\|_2^2$ in (G.14) require another $O(n)$ parameters. To connect modules, store a signed scalar z as $(z_+, (-z)_+)$ and recover it by subtraction. Passing m such scalars through k layers costs $O(mk)$ parameters and $O(m)$ width, without multiplying consecutive dense affine matrices. The arithmetic modules and the channels needed to store intermediate values and align branch depths thus use $O_\Theta((n+d)\ell^2_{n,\eta})$ parameters in total. The coordinatewise input clipping adds $O(d)$ parameters at constant depth. Consequently,

$$S(\Phi) \leq Cnd + C_\Theta(n+d)\ell^2_{n,\eta},$$

which completes (G.2).

For fixed Θ, n, η, the ranges, clipping thresholds, accuracies, and integer depths are independent of the sample locations. Hence the architecture and allowed parameter positions can be fixed across samples. The parameters depend continuously on $X_{1:n}$ through the entries of X_i and $\|X_i\|_2^2$, so the network output is jointly continuous in the sample and its inputs. The same holds for $G_{n,t}$ by the smoothness and positivity of the underlying Gaussian mixture. Hence their difference is jointly continuous, and continuity in x gives

$$\sup_{x\in\mathbb{R}^d} \|\Phi_{X_{1:n},\eta}(x,t) - G_{n,t}(x)\|_2 = \sup_{x\in\mathbb{Q}^d} \|\Phi_{X_{1:n},\eta}(x,t) - G_{n,t}(x)\|_2.$$

The right-hand side is a countable supremum of continuous functions of the sample and time, and is therefore measurable. □

G.5 Transfer to Robust Score Risk

We now relate the neural approximation guarantee above to the robust score risk. The following result shows that the robustness guarantee for an OU-compatible score is stable under a uniformly controlled approximation.

Corollary G.4. *Under the assumptions of Proposition 3.1, consider a measurable, possibly random estimator*

$$s_\theta(x,t) = \nabla \log \hat{p}_t(x) + r_\theta(x,t), \qquad \hat{p}_t = P_t\hat{\mu}_0.$$

Suppose that, for some $\delta \geq 0$,

$$\mathbb{E}\left[\frac{1}{T-t_0}\int_{t_0}^{T} \|r_\theta(\cdot,t)\|^2_{L^\infty(\mathbb{R}^d)}\,\mathrm{d}t\right] \leq \delta, \tag{G.21}$$

Then, for every $\varepsilon \geq 0$,

$$\mathcal{R}_{\varepsilon,p_0}(s_\theta) \leq \frac{4\varepsilon^2}{(T-t_0)\tau_0} + \frac{4}{T-t_0}\,\mathbb{E}[D_2\left(p_{t_0}\|\hat{p}_{t_0}\right)] + 2\delta. \tag{G.22}$$

Proof. Fix $\varepsilon \geq 0$ and an arbitrary deterministic $q_0 \in \mathcal{Q}_\varepsilon(p_0)$, and write $q_t = P_t q_0$ and $\hat{s}(x,t) = \nabla \log \hat{p}_t(x)$. The fields $\hat{s}$ and r_θ may depend on the same training sample; all expectations below are over their joint randomness.

For each estimator realization and almost every $t \in [t_0, T]$, the density q_t has unit mass, so

$$\int_{\mathbb{R}^d} \|r_\theta(x,t)\|_2^2 q_t(x)\,\mathrm{d}x \leq \|r_\theta(\cdot,t)\|^2_{L^\infty(\mathbb{R}^d)}.$$

Consequently, Tonelli's theorem and (G.21) give

$$\mathbb{E}\left[\frac{1}{T-t_0}\int_{t_0}^{T}\int_{\mathbb{R}^d} \|r_\theta(x,t)\|_2^2 q_t(x)\,\mathrm{d}x\,\mathrm{d}t\right] \leq \mathbb{E}\left[\frac{1}{T-t_0}\int_{t_0}^{T} \|r_\theta(\cdot,t)\|^2_{L^\infty(\mathbb{R}^d)}\,\mathrm{d}t\right] \leq \delta. \tag{G.23}$$

This bound is uniform over the choice of q_0.

Next, the decomposition $s_\theta = \hat{s} + r_\theta$ and the inequality $\|u+v\|_2^2 \leq 2\|u\|_2^2 + 2\|v\|_2^2$ imply, for each realization,

$$\|s_\theta(x,t) - \nabla \log q_t(x)\|_2^2 \leq 2\|\hat{s}(x,t) - \nabla \log q_t(x)\|_2^2 + 2\|r_\theta(x,t)\|_2^2.$$

Integrating against $q_t(x)\,\mathrm{d}x\,\mathrm{d}t/(T-t_0)$, taking expectation, and using (G.23) yields

$$\mathbb{E}[\mathcal{L}_{q_0}(s_\theta)] \leq 2\mathbb{E}[\mathcal{L}_{q_0}(\hat{s})] + 2\delta.$$

Taking the supremum over $q_0 \in \mathcal{Q}_\varepsilon(p_0)$, outside the expectation as in (2.2), therefore gives

$$\mathcal{R}_{\varepsilon,p_0}(s_\theta) \leq 2\mathcal{R}_{\varepsilon,p_0}(\hat{s}) + 2\delta.$$

Finally, the oracle bound (3.2) and $\tau_0 = (1 - e^{-t_0})/e^{-t_0}$ give

$$\mathcal{R}_{\varepsilon,p_0}(s_\theta) \le \frac{4\varepsilon^2}{(T-t_0)\tau_0} + \frac{4}{T-t_0}\,\mathbb{E}[D_2(p_{t_0}\|\hat{p}_{t_0})] + 2\delta,$$

which is (G.22). □

We next apply Corollary G.4 to the neural approximation constructed in Proposition G.1. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_0$, with p_0 satisfying Assumption 1. Recall that the Gaussian-blanket score can be written as

$$\hat{s}_n^{\vartheta}(x,t) = -\frac{x}{v_t} + G_{n,t}(x),$$

where the Gaussian term $-x/v_t$ is retained exactly and only the sample-dependent correction $G_{n,t}$ is approximated by a ReLU network.

For every $\eta \in (0, 1/2)$, Proposition G.1 constructs a sample-dependent ReLU network $\Phi_{X_{1:n},\eta}$ satisfying

$$\sup_{x\in\mathbb{R}^d,\ t\in[t_0,T]} \|\Phi_{X_{1:n},\eta}(x,t) - G_{n,t}(x)\|_2 \le \eta.$$

Define the corresponding neural score estimator by

$$\hat{s}_{n,\eta}^{\mathrm{NN}}(x,t) := -\frac{x}{v_t} + \Phi_{X_{1:n},\eta}(x,t).$$

Then

$$\sup_{x\in\mathbb{R}^d,\ t\in[t_0,T]} \left\|\hat{s}_{n,\eta}^{\mathrm{NN}}(x,t) - \hat{s}_n^{\vartheta}(x,t)\right\|_2 \le \eta.$$

Hence (G.21) holds with $\delta = \eta^2$, and the proof of Corollary G.4 gives

$$\mathcal{R}_{\varepsilon,p_0}\left(\hat{s}_{n,\eta}^{\mathrm{NN}}\right) \le 2\mathcal{R}_{\varepsilon,p_0}\left(\hat{s}_n^{\vartheta}\right) + 2\eta^2.$$

Applying Minkowski's inequality before taking the supremum over q_0 gives the sharper bound

$$\sqrt{\mathcal{R}_{\varepsilon,p_0}\left(\hat{s}_{n,\eta}^{\mathrm{NN}}\right)} \le \sqrt{\mathcal{R}_{\varepsilon,p_0}\left(\hat{s}_n^{\vartheta}\right)} + \eta.$$

With $\vartheta = D/\sqrt{d}$ and $\tau_0 \le 1$, Theorem 4.2 yields

$$\sqrt{\mathcal{R}_{\varepsilon,p_0}\left(\hat{s}_{n,\eta}^{\mathrm{NN}}\right)} \le \left[\frac{C_{d,D}}{(n+1)(T-t_0)\tau_0^{d/2}} + \frac{2\varepsilon^2}{(T-t_0)\tau_0}\right]^{1/2} + \eta.$$

Thus, the neural approximation preserves the Gaussian-blanket robustness rate, up to a constant factor, provided that η^2 is bounded by the sum of the statistical and distribution-shift terms.

Scope of the neural representation. The approximation result is constructive: the network parameters are specified from the observed sample, and the width, depth, and sparsity are controlled explicitly in Proposition G.1. The result is a realizability statement and does not imply that a particular score-matching training procedure recovers the constructed parameters. Moreover, the approximation bounds are stated for fixed $0 < t_0 < T < \infty$; their constants may deteriorate as $t_0 \to 0$ or $T \to \infty$.

H Numerical Illustration

We provide a controlled numerical study illustrating the finite-sample and distribution-shift effects predicted by our theory. We consider a full-rank three-dimensional example, which permits accurate numerical evaluation of both the time-averaged score risk and the positive-time KL error while retaining a nontrivial distribution shift setting.

Source distribution and shift family. We set $d = 3$ and $D = 1.5$, and consider the four support points

$$\mu_1 = \frac{D}{\sqrt{3}}(1,1,1), \qquad \mu_2 = \frac{D}{\sqrt{3}}(1,-1,-1),$$
$$\mu_3 = \frac{D}{\sqrt{3}}(-1,1,-1), \qquad \mu_4 = \frac{D}{\sqrt{3}}(-1,-1,1).$$

The reference distribution is

$$p_0 = 0.45\delta_{\mu_1} + 0.30\delta_{\mu_2} + 0.15\delta_{\mu_3} + 0.10\delta_{\mu_4}.$$

The support points are noncoplanar and lie in $\mathcal{B}(0, D)$, so p_0 has full-rank covariance.

To generate distribution shifts, let $\phi = (1 + \sqrt{5})/2$ and let $v_1, \dots, v_{12}$ be the normalized vectors obtained from

$$(0, \pm 1, \pm\phi), \qquad (\pm 1, \pm\phi, 0), \qquad (\pm\phi, 0, \pm 1).$$

For $k = 1, \dots, 12$, define the translated distributions

$$q^{\varepsilon}_{0,k} := (x \mapsto x + \varepsilon v_k)_{\#} p_0, \qquad q^{\varepsilon}_{t,k} := P_t q^{\varepsilon}_{0,k}.$$

The translation coupling gives $\mathsf{W}_2(p_0, q^{\varepsilon}_{0,k}) \le \varepsilon$, while the difference of means gives the reverse inequality; hence

$$\mathsf{W}_2(p_0, q^{\varepsilon}_{0,k}) = \varepsilon.$$

We evaluate the worst-case error over these twelve translation directions:

$$R^{\mathrm{tr}}_{12}(n, \varepsilon) := \max_{1 \le k \le 12} \mathbb{E}_{X_{1:n}} \mathcal{L}_{q^{\varepsilon}_{0,k}}(\hat{s}^{\vartheta}_n),$$
$$K_{12}(n, \varepsilon) := \max_{1 \le k \le 12} \mathbb{E}_{X_{1:n}} \mathrm{KL}\big(q^{\varepsilon}_{t_0,k} \big\| \hat{p}^{\vartheta}_{n,t_0}\big),$$

where the expectation over the source samples is taken before maximizing over the translation directions.

Estimator and theoretical bounds. We take

$$t_0 = 0.05, \qquad T = 1,$$

and consider 23 logarithmically spaced shift radii in $[10^{-5}, 0.2]$. The Gaussian-blanket estimator follows Section 4, with

$$\vartheta = \frac{D}{\sqrt{d}}, \qquad M = M_{\vartheta,t_0} = \exp\left(\frac{D^2}{2\vartheta^2}\right)\left(1 + \frac{\vartheta^2}{\tau_0}\right)^{d/2}.$$

For each source-sample realization, we generate

$$(N_1, \dots, N_4) \sim \mathrm{Multinomial}(n; (0.45, 0.30, 0.15, 0.10)),$$

which exactly represents the empirical measure for the finite-support source. The resulting density estimator is

$$\hat{p}^{\vartheta}_{n,t}(x) = \frac{\sum_{j=1}^{4} N_j \varphi_{\rho_t}(x - a_t\mu_j) + M\varphi_{\rho_t + a_t^2\vartheta^2}(x)}{n + M}.$$

The theoretical comparison curves are

$$U_K(n, \varepsilon) := \log\left(1 + \frac{M-1}{n+1}\right) + \frac{\varepsilon^2}{\tau_0},$$

and

$$U_S(n, \varepsilon) := \frac{2}{T - t_0} U_K(n, \varepsilon),$$

as implied by Propositions 3.1 and 4.1.

Numerical evaluation. For the main experiment, we use

$$n \in \{5 \times 10^3,\ 5 \times 10^5,\ 5 \times 10^7\}.$$

For each n, we average over 12 independent source-sample replicates, reusing the same replicates across all shift radii and translation directions.

Since the Gaussian-blanket score is OU-compatible, the time-averaged score risk is evaluated through the endpoint identity

$$\mathcal{L}_{q^{\varepsilon}_{0,k}}(\hat{s}^{\vartheta}_n) = \frac{2}{T - t_0}\left[\mathrm{KL}\big(q^{\varepsilon}_{t_0,k} \big\| \hat{p}^{\vartheta}_{n,t_0}\big) - \mathrm{KL}\big(q^{\varepsilon}_{T,k} \big\| \hat{p}^{\vartheta}_{n,T}\big)\right],$$

which follows from Lemma B.2. The KL terms at t_0 and T are computed using component-stratified order-32 tensor Gauss–Hermite quadrature. Under self-consistent initialization, the reverse sampler output at forward time t_0 is exactly $\hat{p}^{\vartheta}_{n,t_0}$, allowing the KL error to be evaluated directly without simulating the reverse SDE.

Figure 1 reports the resulting time-averaged score risk and positive-time KL error. Increasing the source sample size moves both quantities toward their known-source counterparts, reflecting the reduction in finite-sample estimation error. As the shift radius increases, the distribution-shift contribution becomes dominant, with the observed behavior consistent with the quadratic dependence on ε predicted by Theorem 4.2 and (6.1). Across the displayed configurations, the empirical errors remain below the corresponding theoretical upper bounds.

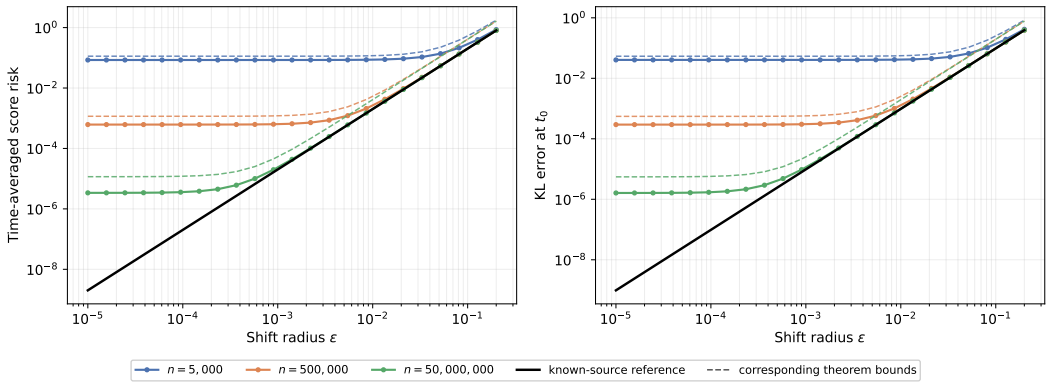


Figure 1: Time-averaged score risk (left) and positive-time KL error (right) in the controlled three-dimensional experiment. Increasing the source sample size reduces the finite-sample contribution, while the distribution-shift contribution dominates at larger shift radii. Dashed curves show the corresponding theoretical upper bounds.

The black curves denote known-source references, using $\nabla \log p_t$ for the score risk and p_{t_0} for the KL error. The shaded regions are pointwise 95% paired-bootstrap intervals over the 12 source-sample replicates, with the maximization over translation directions recomputed in each of 4,000 bootstrap resamples.

Reference-domain finite-sample behavior. To isolate the finite-sample contribution, Figure 2 reports the score risk at $\varepsilon = 0$ for seven sample sizes between 5×10^3 and 5×10^9. The risk is evaluated using the same endpoint identity and shows approximately n^{-1} decay over the displayed range. This experiment is intended only as a finite-sample illustration and does not probe the worst-case $\tau_0^{-d/2}$ dependence or the full minimax rate.

Quadrature validation. As a numerical stability check, increasing the endpoint quadrature order from 32 to 48 changes the mean score risks by at most 0.11% over the configurations examined.

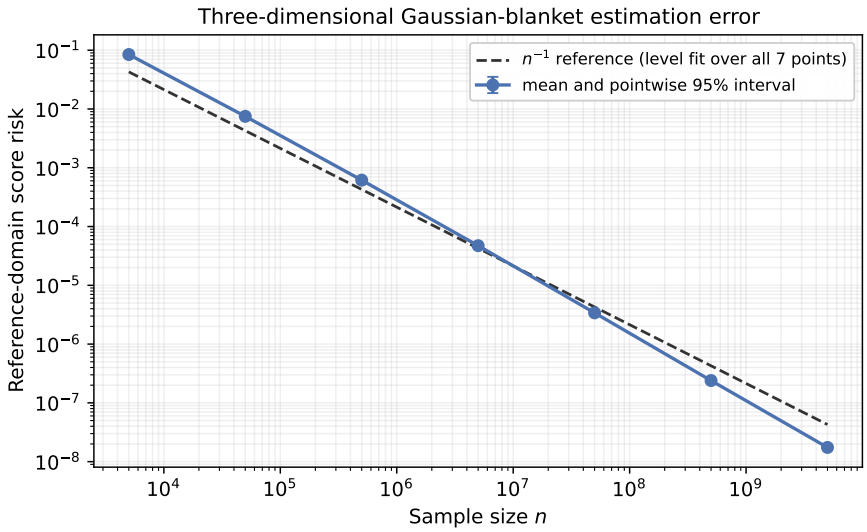


Figure 2: Reference-domain Gaussian-blanket score risk as a function of source sample size. Points are averaged over 256 independent multinomial-count replicates; error bars indicate ± 1.96 standard errors, and the dashed line shows an n^{-1} reference.