PREDICTION TARGETS IN DIFFUSION MODELS: GEOMETRY, CAPACITY, AND SHORTCUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Although x-, ε-, and v-prediction yield the same score function at optimum, their prediction targets can have markedly different geometric structures. Empirical studies have reported substantial performance gaps between these parameterizations in both pixel and representation spaces. We study this discrepancy through the interaction between prediction-target geometry and limited output rank. For data with low-dimensional structure, the clean predictor follows the underlying signal structure, whereas noise and velocity predictors additionally carry ambient components inherited from the noisy input. These additional components can create ambient approximation floors for ε- and v-prediction, even when the same output rank is sufficient for x-prediction. This geometric characterization also suggests a simple time-dependent shortcut $c(t)x_t$ that directly carries the ambient component, leaving the core network to represent a scaled conditional clean signal. With the appropriate coefficient, the shortcut removes the resulting ambient approximation floor. Controlled experiments validate the predicted approximation scaling and evaluate the learned shortcut on nonlinear synthetic data. On ImageNet 256×256 with JiT-B/16, the shortcut reduces v-prediction FID from 96.53 to 7.83, compared with 8.62 for x-prediction.

1 INTRODUCTION

Diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020) have achieved strong results in image and video generation (Dhariwal & Nichol, 2021; Rombach et al., 2022; Podell et al., 2023; Ho et al., 2022; Blattmann et al., 2023; Saharia et al., 2022) by learning to reverse a noise-corruption process. The reverse process can be learned by predicting the clean signal x, the noise ε, or the velocity v (Section 2). Although these parameterizations yield the same score function at optimum, they can differ substantially in practice. Recent work highlights the benefits of x-prediction in high-dimensional pixel space (Li & He, 2025; Lu et al., 2026), fixed VAE latent space (Fu et al., 2026), and pretrained representation space (Zhang et al., 2026). These observations raise a central question: why do theoretically equivalent parameterizations perform differently in practice?

We investigate this question through the interaction between prediction-target geometry and finite model capacity. Through synthetic and image-generation experiments, Li & He (2025) show that x-prediction can remain effective with limited model capacity in high-dimensional spaces, while noise and velocity prediction deteriorate. They interpret these results through the manifold hypothesis: clean data have low-dimensional structure, whereas noise and velocity contain components outside that structure.

To make this interaction concrete, we focus on output rank as a tractable aspect of model capacity and analyze how it constrains approximation and learning for different targets. For subspace data, the optimal clean predictor stays within the signal structure, whereas noise and velocity predictors retain ambient components. We further show that this geometric separation extends to the one-step Mean Flow framework (Geng et al., 2025; Lu et al., 2026) (Appendix E). In the Gaussian setting, this separation can be quantified explicitly: we characterize the target-dependent approximation errors and the convergence of gradient descent. We also extend the geometric analysis to smooth manifold unions and Gaussian mixtures, and establish local convergence of a one-gate x-prediction model for data drawn from a two-component Gaussian mixture.

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.

This geometric distinction motivates a time-dependent scalar shortcut $c_\phi(t)x_t$ that carries the ambient component directly, reducing the representational burden on the core network. For Gaussian data and Gaussian mixtures, we establish conditions under which an appropriate coefficient eliminates the approximation floor (Section 4). This motivates a practical variant that learns the coefficient jointly with the core, which we evaluate on synthetic data and image generation.

Our experiments evaluate target-dependent approximation errors and learning behavior. Gaussian experiments confirm that an output rank sufficient for x-prediction can leave approximation floors for ε- and v-prediction (Figure 1). On the Swiss roll, x-prediction recovers the spiral while direct v-prediction fails within the training budget under an output bottleneck (Figure 2a); without this bottleneck, v-prediction still recovers more slowly (Figure 2b). The learned shortcut improves v-prediction in both settings and benefits both ε- and v-prediction on Gaussian mixtures (Figure 4). On ImageNet 256×256 with JiT-B/16, it reduces v-prediction FID from 96.53 to 7.83, compared with 8.62 for x-prediction (Table 1).

Related Work. Diffusion theory has studied sampling complexity (De Bortoli, 2022; Chen et al., 2023a; Benton et al., 2024; Li & Jiao, 2025) and score estimation (Oko et al., 2023; Chen et al., 2023b; Wang et al., 2024; Cui et al., 2024). Under the manifold hypothesis, these guarantees can depend on intrinsic rather than ambient dimension when the model is sufficiently expressive (Azangulov et al., 2024; Tang & Yang, 2024; Huang et al., 2026). These results motivate examining how low-dimensional data structure interacts with the prediction target when model capacity is limited.

One approach to this question is to compare targets in tractable model classes. Jin & Wang (2026) analyze $u_k = kx - (1-k)\varepsilon$ using a single time-shared linear predictor. Their time-integrated MSE includes both the Bayes risk and approximation error from sharing the predictor across noise levels. Our analysis isolates excess risk under output constraints, relates it to score error, and covers both Gaussian and nonlinear Gaussian-mixture targets.

Beyond comparing prediction targets, network preconditioning changes how the network represents them to improve training. EDM (Karras et al., 2022) combines a noise-dependent skip connection with input and output scaling to normalize signal magnitudes and limit amplification of network errors. EDM2 further improves training dynamics through magnitude-preserving network designs (Karras et al., 2024). Our shortcut uses the same residual form as EDM, but derives its coefficient from target geometry. We quantify how this choice affects approximation under limited model output rank and when it eliminates the ambient floor. Appendix D.2 compares the two coefficient choices under the same output constraint.

Contributions.

- **Target geometry, approximation, and learning.** We characterize target geometry in subspaces and smooth manifold unions. Under limited output rank, we quantify Gaussian approximation errors and analyze gradient-descent convergence, and establish local convergence of a one-gate x-predictor for a two-component Gaussian mixture.
- **Geometry-motivated shortcuts.** We identify a scalar shortcut that separates the ambient component from the signal-dependent core. Relative to direct ε- and v-prediction, it removes output-rank-induced approximation floors under our Gaussian and Gaussian-mixture settings, and can accelerate training and improve generation quality even without an output-width bottleneck.
- **Empirical evidence.** We validate target-dependent errors on Gaussian data, compare learning with and without output-width bottlenecks on Swiss rolls, and demonstrate the benefits of learned shortcuts on synthetic data and ImageNet.

2 Setup and Prediction Targets

Forward and reverse processes. Let $x_0 \sim p_{\mathrm{data}}$ in $\mathbb{R}^D$, where D is the ambient dimension, and let $\varepsilon \sim \mathcal{N}(0, I_D)$ be independent of x_0. We consider the following stochastic differential equation (SDE):

$$dx_t = f(x_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}W_t, \tag{1}$$

where W_t is standard D-dimensional Brownian motion and $f(x,t) = \mathbf{f}(t)x$ with scalar $\mathbf{f}(t)$. Its marginal transition is

$$x_t = a_t x_0 + b_t \varepsilon. \tag{2}$$

The coefficients $a_t > 0$ and $b_t \geq 0$ are determined by the SDE; standard VP and VE schedules appear in Appendix H. Writing p_t for the law of x_t at time t, the reverse-time SDE is (Anderson, 1982; Song et al., 2021)

$$dx_t = \big[f(x_t,t) - g(t)^2 \nabla \log p_t(x_t)\big]\,\mathrm{d}t + g(t)\,\mathrm{d}\bar{W}_t, \tag{3}$$

where t runs from T toward 0 and $\bar{W}_t$ is Brownian motion in reverse time. In practice, the unknown score $s_t = \nabla \log p_t$ is replaced by a learned estimator $\hat{s}_t$, and samples are generated by simulating the resulting reverse process. This estimator can be obtained from any of the three prediction targets introduced below.

Prediction targets. We consider $y_x = x_0$ (Karras et al., 2022), $y_\varepsilon = \varepsilon$ (Ho et al., 2020), and $y_v = v_t$ (Salimans & Ho, 2022), using the interpolation-velocity convention

$$v_t := a_t' x_0 + b_t' \varepsilon. \tag{4}$$

Primes denote derivatives with respect to t. For each $\tau \in \{x, \varepsilon, v\}$, the training objective is

$$\min_\theta \; \mathbb{E}_{t,x_0,\varepsilon}\Big[w_t \big\| f_\theta^{(\tau)}(x_t,t) - y_\tau \big\|^2\Big], \tag{5}$$

where t is sampled from the training-time distribution and w_t is a nonnegative time weight. We also denote the three model outputs by x_θ, ε_θ, and v_θ, respectively. We assume $b_t > 0$ at the noise levels analyzed and $\Delta_t := a_t b_t' - b_t a_t' \neq 0$ whenever converting v to a score.

Optimal predictors and excess risk. We assume x_0 has finite second moment. At each fixed t, the unrestricted optimal predictor is the conditional mean $f_t^{\star,(\tau)}(x) = \mathbb{E}[y_\tau \mid x_t = x]$, also written $x_t^\star$, $\varepsilon_t^\star$, and $v_t^\star$. The prediction MSE decomposes as (Lemma G.1)

$$\mathbb{E}\big\| f_\theta^{(\tau)}(x_t,t) - y_\tau \big\|^2 = \mathbb{E}\big\| f_\theta^{(\tau)}(x_t,t) - f_t^{\star,(\tau)}(x_t) \big\|^2 + \mathbb{E}\big\| f_t^{\star,(\tau)}(x_t) - y_\tau \big\|^2. \tag{6}$$

The first term is the excess risk studied below. The second is the Bayes risk, which captures the intrinsic uncertainty of denoising and is independent of θ.

Score equivalence and score error. The three optimal predictors determine the same score (Lemma G.2):

$$\nabla \log p_t(x_t) = \frac{1}{b_t^2}\big(a_t x_t^\star(x_t) - x_t\big) = -\frac{1}{b_t}\varepsilon_t^\star(x_t) = -\frac{a_t v_t^\star(x_t) - a_t' x_t}{b_t \Delta_t}. \tag{7}$$

Substituting a model output into the corresponding expression gives

$$\hat{s}_t^{(\tau)} - s_t = \lambda_t^{(\tau)}\big(f_\theta^{(\tau)} - f_t^{\star,(\tau)}\big),$$

where $\lambda_t^{(x)} = a_t/b_t^2$, $\lambda_t^{(\varepsilon)} = -1/b_t$, and $\lambda_t^{(v)} = -a_t/(b_t\Delta_t)$. Thus, score MSE equals excess risk multiplied by $(\lambda_t^{(\tau)})^2$. This relation connects the output-range approximation errors below to the reverse-SDE KL bounds.

3 Target Geometry, Approximation, and Learning

The score identities in Section 2 reveal a geometric difference between prediction targets. We first characterize this difference for subspace data, then use Gaussian data to quantify its effect on approximation and learning under limited model output rank.

3.1 Target Geometry in General Subspaces

We first study prediction-target geometry for data supported on a low-dimensional affine subspace.

Assumption 1 (General subspace prior)**.** *The clean data lie in an r-dimensional affine subspace $\mathcal{S} \subset \mathbb{R}^D$ and admit the representation $x_0 = \mu + Uz$, where $\mu \in \mathbb{R}^D$, $U \in \mathbb{R}^{D\times r}$ has orthonormal columns, and $z \in \mathbb{R}^r$ is drawn from an arbitrary distribution with $\mathbb{E}\|z\|^2 < \infty$.*

Theorem 1 (Geometric separation in general subspaces)**.** *Under Assumption 1, fix $t > 0$ with $b_t > 0$. Let $\mathcal{P}_{\mathcal{S}^\perp} = I_D - UU^\top$ denote the projection onto the directions orthogonal to the affine support. Then*

$$\mathcal{P}_{\mathcal{S}^\perp}\big(x^\star(x_t) - \mu\big) = 0, \qquad \mathcal{P}_{\mathcal{S}^\perp}\varepsilon^\star(x_t) = \mathcal{P}_{\mathcal{S}^\perp}\varepsilon.$$

Consequently,

$$\mathbb{E}\|\mathcal{P}_{\mathcal{S}^\perp}\varepsilon^\star(x_t)\|^2 = D - r,$$
$$\mathbb{E}\|\mathcal{P}_{\mathcal{S}^\perp}(v^\star(x_t) - a_t'\mu)\|^2 = (b_t')^2(D - r).$$

Theorem 1 identifies a geometric asymmetry between prediction targets: the x-predictor stays within the affine support, whereas the noise and velocity predictors retain orthogonal components with energy $D - r$ and $(b_t')^2(D - r)$, respectively. This distinction holds for any distribution satisfying the subspace assumption and creates different representational demands: an output range sufficient for the clean predictor may exclude ambient directions needed by the other targets.

The same geometric distinction extends beyond affine subspaces. For data supported near unions of low-dimensional smooth manifolds, the clean predictor need not lie on the data support, so exact subspace containment is replaced by a bound on its normal displacement relative to a local reference tangent plane. Under the geometric regularity conditions in Appendix C, the normal displacement energy of the x-predictor remains bounded, whereas the noise and velocity predictors can retain normal energy that scales with the ambient dimension D.

Our theoretical results extend to the one-step generation framework of Mean Flow (Geng et al., 2025; Lu et al., 2026). Under the affine-subspace assumption, we establish an analogous separation of the Mean Flow prediction targets into signal and ambient components. The precise statement and proof are given in Appendix E.

3.2 Gaussian Distribution

The preceding subspace analysis identifies ambient components in noise and velocity prediction. We now consider Gaussian data to quantify how this geometry affects approximation under limited model output rank and what errors remain after training.

Assumption 2 (Low-rank Gaussian prior)**.** *The clean data follow $x_0 \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma \succeq 0$ and $\mathrm{rank}(\Sigma) = r \ll D$. Write $\Sigma = U\Lambda_r U^\top$, where $U \in \mathbb{R}^{D\times r}$ has orthonormal columns and $\Lambda_r \in \mathbb{R}^{r\times r}$ is positive definite. Equivalently, $x_0 = \mu + U\Lambda_r^{1/2}z$ with $z \sim \mathcal{N}(0, I_r)$, so the data have an r-dimensional affine support.*

Optimal predictors. Fix t with $a_t, b_t > 0$ and finite a_t', b_t'. The noisy marginal is $\mathcal{N}(\mu_t, \Sigma_t)$, where $\mu_t = a_t\mu$ and $\Sigma_t = a_t^2\Sigma + b_t^2 I_D \succ 0$. With $B_t = a_t' a_t \Sigma + b_t' b_t I_D$, the three optimal predictors are affine (Lemma G.3):

$$\begin{aligned} x^\star(x_t) &= \mu + a_t\Sigma\Sigma_t^{-1}(x_t - \mu_t), \\ \varepsilon^\star(x_t) &= b_t\Sigma_t^{-1}(x_t - \mu_t), \\ v^\star(x_t) &= a_t'\mu + B_t\Sigma_t^{-1}(x_t - \mu_t). \end{aligned} \tag{8}$$

Consistent with Theorem 1, the x-predictor lies in the affine support and hence in the signal span $\mathcal{M} = \mathrm{span}(\mathrm{range}(\Sigma) \cup \{\mu\})$ of dimension $d_{\mathrm{int}} = \dim(\mathcal{M})$. We include the mean in this span because the model below has a linear output range.

Approximation and learning under limited output rank. This constraint can be viewed as an idealized model of a hidden-width bottleneck in DiTs, where the hidden width is smaller than the output token dimension, limiting the performance of v-prediction (Zheng et al., 2025). To isolate

this constraint, for each target $\tau \in \{x, \varepsilon, v\}$ we consider the affine model $A_\tau \Theta_\tau \psi(x_t)$, where $A_\tau \in \mathbb{R}^{D\times m}$, $\Theta_\tau \in \mathbb{R}^{m\times(D+1)}$, and $\psi(x) = (x^\top, 1)^\top$. Both factors are trainable, and the output rank is at most m. Denote the corresponding optimal predictor by $f_t^{\star,(\tau)}$. To separate model approximation from intrinsic denoising uncertainty, we measure the excess risk

$$\mathcal{R}_{\tau,t}(A, \Theta) = \mathbb{E}\|A\Theta\psi(x_t) - f_t^{\star,(\tau)}(x_t)\|^2. \tag{9}$$

By (6), this objective has the same population gradients as the unweighted target MSE.

We train each target separately at the fixed time t, jointly updating A_τ and Θ_τ by population gradient descent. Write $\mathcal{R}_{\tau,t}^{(k)} = \mathcal{R}_{\tau,t}(A_\tau^{(k)}, \Theta_\tau^{(k)})$ for the risk after k iterations. The next theorem compares the best attainable errors and shows when training attains them.

Theorem 2 (Target-dependent approximation and learning)**.** *Under Assumption 2, fix t with $a_t, b_t > 0$ and finite a_t', b_t', and suppose $1 \le d_{\mathrm{int}} \le m \le D$. For each $\tau \in \{x, \varepsilon, v\}$, initialize $(A_\tau^{(0)}, \Theta_\tau^{(0)})$ from an absolutely continuous distribution with bounded support $\mathcal{B}_\tau$. Train both factors by population gradient descent with a constant stepsize $0 < \eta_\tau \le \eta_{\mathcal{B},\tau}$, where the sufficient bound $\eta_{\mathcal{B},\tau} > 0$ depends on $\mathcal{B}_\tau$, the noisy input distribution, and the target. Its explicit form and the gradient descent updates are given in Appendix A.1.*

For each target $\tau \in \{x, \varepsilon, v\}$, the excess risk at iteration k admits the decomposition

$$\mathcal{R}_{\tau,t}^{(k)} = \mathcal{A}_{\tau,t}^{\mathrm{opt}}(m) + \mathcal{E}_{\tau,t}^{(k)}, \qquad \mathcal{A}_{\tau,t}^{\mathrm{opt}}(m) := \inf_{A,\Theta} \mathcal{R}_{\tau,t}(A, \Theta),$$

where $\mathcal{A}_{\tau,t}^{\mathrm{opt}}(m)$ is the approximation error, i.e., the smallest excess risk achievable by the rank-constrained model $A\Theta\psi(x_t)$, and $\mathcal{E}_{\tau,t}^{(k)} \ge 0$ is the optimization error.

Almost surely, the factor iterates converge to finite limits $(A_\tau^{(\infty)}, \Theta_\tau^{(\infty)})$ that attain a global minimum, and

$$\mathcal{E}_{\tau,t}^{(k)} \to 0 \qquad \text{as } k \to \infty.$$

The approximation errors satisfy

$$\mathcal{A}_{x,t}^{\mathrm{opt}}(m) = 0,$$
$$(D - r - m)_+ \le \mathcal{A}_{\varepsilon,t}^{\mathrm{opt}}(m) \le D - m,$$
$$(b_t')^2(D - r - m)_+ \le \mathcal{A}_{v,t}^{\mathrm{opt}}(m) \le (b_t')^2(D - m),$$

where $(u)_+ = \max\{u, 0\}$. Consequently, x-prediction recovers the optimal clean predictor:

$$A_x^{(\infty)}\Theta_x^{(\infty)}\psi(x_t) = x^\star(x_t) \quad \text{almost surely}, \qquad \mathcal{M} \subseteq \mathrm{range}(A_x^{(\infty)}).$$

In contrast, noise and velocity prediction can retain nonzero approximation errors even after their optimization errors vanish. When $m = D$, all three approximation errors are zero.

The proof is given in Appendix A.1.

From prediction error to sampling. The score identities in Section 2 relate prediction error to score error. The following corollary uses this relation to bound the sampling error in terms of the approximation error $\mathcal{A}_{\tau,t}^{\mathrm{opt}}(m)$ and the optimization error $\mathcal{E}_{\tau,t}^{(k)}$.

Corollary 1 (Score-induced KL bound)**.** *Consider the forward SDE of Section 2 on $[\delta, T]$, where $0 < \delta < T$ and $a_t, b_t > 0$. At each fixed time, use the separately trained predictors of Theorem 2. For the evaluated iteration k, assume that $\mathbf{f}$, g, and the induced affine score coefficients are bounded and piecewise continuous on this interval, with $\Delta_t \neq 0$ for velocity prediction. Then the learned reverse SDE, initialized from p_T, satisfies*

$$\mathrm{KL}(p_\delta \| \hat{p}_\delta^{(\tau,k)}) \le \frac{1}{2}\int_\delta^T g(t)^2 |\lambda_t^{(\tau)}|^2 \left[\mathcal{A}_{\tau,t}^{\mathrm{opt}}(m) + \mathcal{E}_{\tau,t}^{(k)}\right] dt, \tag{10}$$

where $\lambda_t^{(x)} = a_t/b_t^2$, $\lambda_t^{(\varepsilon)} = -1/b_t$, and $\lambda_t^{(v)} = -a_t/(b_t\Delta_t)$.

For x-prediction, $\mathcal{A}_{x,t}^{\mathrm{opt}}(m) = 0$, so the bound contains only the optimization error. For ε- and v-prediction, the approximation errors are bounded above by $D - m$ and $(b_t')^2(D - m)$, respectively. Together with the lower bounds in Theorem 2, these approximation errors grow linearly with the ambient dimension at fixed r and m. The proof is in Appendix A.2.

3.3 MIXTURES OF LOW-RANK GAUSSIANS

The Gaussian analysis connects target geometry to approximation and learning under limited output rank. We now extend this comparison to Gaussian mixtures, whose optimal predictors are generally nonlinear. We first characterize the signal and ambient noise components for a mixture with a shared covariance subspace. We then specialize to two components with shared covariance to construct an exact x-predictor and analyze its local learning dynamics.

Target Geometry in Gaussian Mixtures. We assume the target distribution is a mixture of low-rank Gaussians as specified in Assumption 3.

Assumption 3 (Mixture of low-rank Gaussians)**.** *The clean-data distribution is a mixture of low-rank Gaussians (MoLRG),*

$$p_0 = \sum_{k=1}^{M} \pi_k \mathcal{N}(\mu_k, \Sigma_k), \qquad \pi_k \geq 0, \qquad \sum_{k=1}^{M} \pi_k = 1.$$

Here, $\mu_k \in \mathbb{R}^D$, and all component covariances share an r-dimensional subspace, with $r \ll D$:

$$\Sigma_k = U \Lambda_k U^\top, \qquad U^\top U = I_r, \qquad \Lambda_k \succ 0, \quad k \in [M],$$

where $U \in \mathbb{R}^{D \times r}$ and $\Lambda_k \in \mathbb{R}^{r \times r}$.

Let $\bar{\mu}_0 = \sum_{k=1}^{M} \pi_k \mu_k$ and $\bar{\Sigma} = \sum_{k=1}^{M} \pi_k \Sigma_k$ denote the mixture mean and average within-component covariance. At time t, component k has mean $a_t \mu_k$ and covariance $\Sigma_{k,t} = a_t^2 \Sigma_k + b_t^2 I_D$. Write $\bar{\mu}_t = a_t \bar{\mu}_0$ and $\bar{\Sigma}_t = \sum_{k=1}^{M} \pi_k \Sigma_{k,t} = a_t^2 \bar{\Sigma} + b_t^2 I_D$. The posterior component weights are

$$w_k(x) = \frac{\pi_k \mathcal{N}(x; a_t \mu_k, \Sigma_{k,t})}{\sum_{\ell=1}^{M} \pi_\ell \mathcal{N}(x; a_t \mu_\ell, \Sigma_{\ell,t})}.$$

They determine the nonlinear correction

$$\Psi_t(x) = \sum_{k=1}^{M} w_k(x) \left[\bar{\Sigma}_t^{-1}(x - \bar{\mu}_t) - \Sigma_{k,t}^{-1}(x - a_t \mu_k) \right].$$

The score therefore decomposes into a reference Gaussian term and a mixture correction (Lemma G.4):

$$\nabla \log p_t(x) = -\bar{\Sigma}_t^{-1}(x - \bar{\mu}_t) + \Psi_t(x). \tag{11}$$

Combining this decomposition with the score identities in Section 2 yields the following geometric separation; the explicit predictor formulas are derived in Appendix B.

Theorem 3 (Dimensional separation for MoLRGs)**.** *Under Assumption 3, fix t with $a_t, b_t > 0$ and finite a_t', b_t', and define the centered signal subspace*

$$\mathcal{S}_0 = \operatorname{span}\big(\operatorname{range}(\bar{\Sigma}) \cup \{\mu_k - \bar{\mu}_0\}_{k=1}^{M}\big).$$

The optimal clean predictor satisfies $x^\star(x_t) \in \bar{\mu}_0 + \mathcal{S}_0$, where $\dim(\mathcal{S}_0) \leq r + M - 1$. Writing $P_\perp = P_{\mathcal{S}_0^\perp}$ for the orthogonal projection onto $\mathcal{S}_0^\perp$, we have

$$P_\perp(x^\star(x_t) - \bar{\mu}_0) = 0, \quad P_\perp \varepsilon^\star(x_t) = \frac{P_\perp(x_t - \bar{\mu}_t)}{b_t}, \quad P_\perp(v^\star(x_t) - a_t' \bar{\mu}_0) = \frac{b_t'}{b_t} P_\perp(x_t - \bar{\mu}_t).$$

Thus, x-prediction has affine image dimension at most $r + M - 1$, a bound independent of the ambient dimension D. In contrast, ε- and v-prediction can retain ambient noise components in $\mathcal{S}_0^\perp$, whose dimension $D - \dim(\mathcal{S}_0)$ grows with D.

The proof is in Appendix B. Although the mixture correction makes the targets nonlinear, the clean predictor remains in a low-dimensional affine subspace. The orthogonal noise and velocity components have the same form as in Theorem 1. To examine what this distinction implies for a nonlinear model with limited output rank, we next consider a mixture whose clean predictor admits an explicit one-gate representation.

Exact Representation and Local Learning. Consider the balanced two-component mixture with shared covariance

$$p_0 = \tfrac{1}{2}\mathcal{N}(\mu_1, \Sigma) + \tfrac{1}{2}\mathcal{N}(\mu_2, \Sigma), \qquad \Sigma = U\Lambda U^\top, \quad \Lambda \succ 0.$$

At the same fixed time t, write $\bar{\mu}_0 = (\mu_1 + \mu_2)/2$, $d_0 = (\mu_1 - \mu_2)/2$, $\bar{\mu}_t = a_t\bar{\mu}_0$, $d_t = a_t d_0$, and $\Sigma_t = a_t^2\Sigma + b_t^2 I_D$. The posterior weights reduce to a single hyperbolic-tangent gate,

$$\phi_t(x_t) = \tanh\big((\Sigma_t^{-1}d_t)^\top(x_t - \bar{\mu}_t)\big).$$

As in the Gaussian model, the linear output range must contain the mean as well as the centered signal directions. We therefore use

$$\mathcal{S} = \mathrm{span}(\mathrm{range}(\Sigma) \cup \{\mu_1, \mu_2\}) = \mathrm{span}(\mathcal{S}_0 \cup \{\bar{\mu}_0\}).$$

This is the counterpart of the signal span $\mathcal{M}$ in Section 3.2. Choose $\dim(\mathcal{S}) \le m \le D$ and fix a column-orthonormal $Q \in \mathbb{R}^{D\times m}$ with $\mathcal{S} \subseteq \mathrm{range}(Q)$. Consider the one-gate model

$$f_\theta(x) = Q\big(Wx + u\tanh(\omega^\top x + \beta) + c\big), \tag{12}$$

where $W \in \mathbb{R}^{m\times D}$, $u, c \in \mathbb{R}^m$, $\omega \in \mathbb{R}^D$, and $\beta \in \mathbb{R}$ are trainable. The model can match the gate ϕ_t and exactly represent the clean predictor; the construction of $\theta_\star$ is given in Appendix B. We train $\theta = (W, u, c, \omega, \beta)$ while keeping Q fixed.

Theorem 4 (Local convergence of the one-gate x-predictor)**.** *In the balanced two-component setting above, suppose $d_t \neq 0$ and define $\mathcal{R}_x(\theta) = \mathbb{E}\|f_\theta(x_t) - x^\star(x_t)\|^2$. At the exact teacher $\theta_\star$ in* (27)*, the Hessian $\nabla^2\mathcal{R}_x(\theta_\star)$ is positive definite. Consequently, there exist constants $r_{\mathrm{loc}}, \mu_{\mathrm{loc}}, L_{\mathrm{loc}} > 0$ such that, if $\|\theta^{(0)} - \theta_\star\| \le r_{\mathrm{loc}}$ and $0 < \eta \le 1/L_{\mathrm{loc}}$, population gradient descent satisfies*

$$\mathcal{R}_x(\theta^{(K)}) \le (1 - \eta\mu_{\mathrm{loc}})^K \mathcal{R}_x(\theta^{(0)}).$$

The proof and local constants are in Appendix B. This result establishes geometric convergence near an exact x-predictor for a nonlinear core. As in the Gaussian case, under the output range constraint, ε- and v-prediction incur approximation floors of order $\Theta(D - m)$.

4 Target-Matched Shortcuts

The preceding analysis quantifies how target geometry affects approximation under limited output rank. The same geometry motivates a shortcut that carries the explicit noisy-input component directly, leaving the signal-dependent part for the core. For a core network f_θ of arbitrary architecture, define

$$f^{\mathrm{shortcut}}_{\theta,\phi}(x_t, t) = f_\theta(x_t, t) + c_\phi(t)x_t, \tag{13}$$

where $c_\phi(t)$ is a scalar coefficient. To see what this shortcut represents, condition the target definitions on x_t:

$$\varepsilon^\star(x_t) = \frac{1}{b_t}x_t - \frac{a_t}{b_t}x^\star(x_t), \quad v^\star(x_t) = \frac{b_t'}{b_t}x_t - \frac{\Delta_t}{b_t}x^\star(x_t). \tag{14}$$

Thus, the coefficients $\gamma_\varepsilon(t) = 1/b_t$ and $\gamma_v(t) = b_t'/b_t$ capture the explicit noisy-input component, while the core represents the signal-dependent part. These identities hold independently of the data distribution and the core architecture. By separating the ambient component from the signal-dependent part, the shortcut can remove approximation floors induced by output-rank constraints and can also improve learning and generation even without an output-width bottleneck (Figure 2b and Table 1). For the Gaussian model of Section 3.2, the former effect admits an exact characterization.

Theorem 5 (Shortcut removes the Gaussian approximation floor)**.** *Under Assumption 2, fix t with $a_t, b_t > 0$ and finite a_t', b_t', and suppose $d_{\mathrm{int}} \le m \le D$. Consider the affine core from Section 3.2 with a scalar shortcut,*

$$f_{A,\Theta,c}(x) = A\Theta\psi(x) + cx, \qquad A \in \mathbb{R}^{D\times m}, \quad \Theta \in \mathbb{R}^{m\times(D+1)}, \quad c \in \mathbb{R}. \tag{15}$$

For $\tau \in \varepsilon, v$, define $\mathcal{R}^{\mathrm{sc}}_{\tau,c}(A, \Theta) = \mathbb{E}|f_{A,\Theta,c}(x_t) - f_t^{\star,(\tau)}(x_t)|^2$. For every scalar c, the optimal approximation error satisfies

$$\inf_{A,\Theta} \mathcal{R}^{\mathrm{sc}}_{\tau,c}(A, \Theta) \le b_t^2\big(c - \gamma_\tau(t)\big)^2(D - m). \tag{16}$$

In particular, $c = \gamma_\tau(t)$ gives zero approximation error. Consequently, $\inf_{A,\Theta,c} \mathcal{R}^{\mathrm{sc}}_{\tau,c}(A, \Theta) = 0$.

The proof is given in Appendix D. The construction also extends to a balanced two-component Gaussian mixture: the shortcut carries the ambient component, while a one-gate core represents the remaining signal-dependent component exactly, provided that its output subspace contains the signal span. Together, they recover the optimal predictor; see Appendix D.1. Relatedly, the construction takes a residual form similar to EDM preconditioning (Karras et al., 2022), but its coefficient is derived from the target geometry to represent the ambient component directly. Appendix D.2 compares the two coefficient choices under the same output constraint.

5 Experiments

We test approximation errors and shortcut benefits under output constraints, then examine learning and generation beyond output-width bottlenecks. Detailed settings are provided in Appendix I.

Gaussian data. We use Gaussian data with signal dimension $d_{\text{int}} = 8$. For each target, gradient descent optimizes A_τ and Θ_τ. We vary D at $m = 8$ and m at $D = 64$ to test Theorem 2, evaluating excess risk against the Bayes predictor on 20,000 samples per seed over 10 seeds (Appendix I.1).

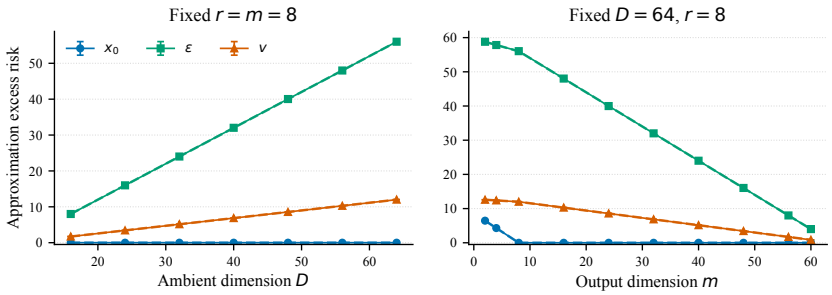


Figure 1: Gaussian excess risks after gradient descent (mean and standard deviation over 10 seeds). Left: varying D with $m = d_{\text{int}} = 8$. Right: varying m with $D = 64$ and $d_{\text{int}} = 8$.

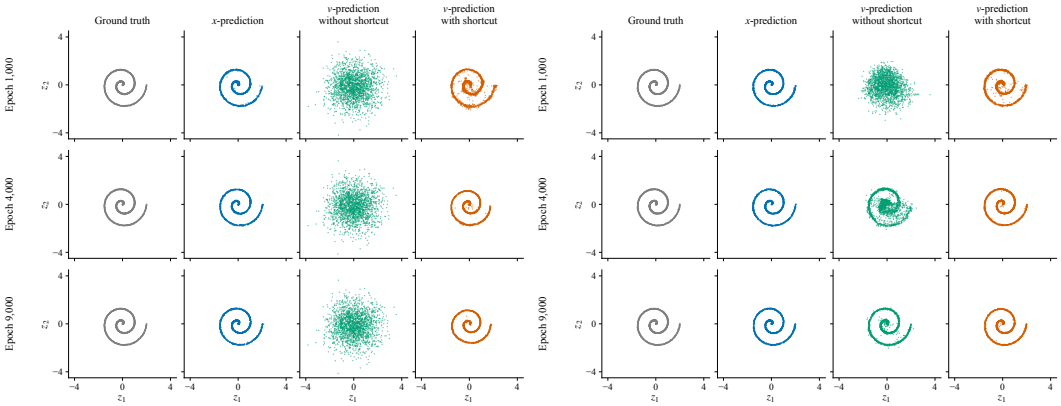
Figure 1 supports the predicted separation: x-error becomes negligible at $m \geq 8$, while noise and velocity errors increase with D and decrease with m. For $m < 8$, x-prediction also incurs error because the output rank is insufficient to represent the full signal subspace. Appendix D.3 checks the floor-removing shortcut of Theorem 5. Under a fixed output constraint, learned shortcuts also improve ε- and v-prediction sample quality on Gaussian mixtures (Appendix F.2).

Swiss roll. On Swiss-roll data (Li & He, 2025), we compare x-prediction, direct v-prediction, and the learned v-shortcut (13) with a five-layer, 256-wide ReLU MLP core. The data satisfy Theorem 1. All methods train for 9000 epochs. They share data, core initialization, and training random inputs within each seed. Sampling uses 250 Euler steps with shared initial noise; full settings appear in Appendix I.2.

With an output-width bottleneck ($D = 512$), the 256-wide core cannot represent all ambient components of the velocity target. The shortcut provides a direct path for these components, allowing the core to focus on the signal-dependent part, as motivated by (14). Figure 2a shows that both x-prediction and shortcut-assisted v-prediction recover the spiral structure, whereas direct v-prediction remains diffuse even at the final checkpoint. These results support the benefit of bypassing the core's output constraint.

At $D = 32$, the 256-wide output head no longer imposes a rank bottleneck. All methods eventually recover the spiral, but shortcut-assisted v-prediction recovers the structure substantially earlier than direct v-prediction (Figure 2b). This suggests that separating the noisy-input component from the signal-dependent part can facilitate learning even when the core has sufficient output width.

ImageNet. We compare x-prediction and direct/shortcut-assisted v-prediction using JiT-B/16 on pixel-space ImageNet 256×256 with a common 200-epoch budget, following Li & He (2025)



(a) $D = 512$: with an output-width bottleneck. (b) $D = 32$: without an output-width bottleneck.

Figure 2: Swiss-roll samples projected back onto the signal plane (seed 0). Rows: epochs 1000, 4000, 9000. Columns: ground truth, x-prediction, direct v-prediction, shortcut-assisted v-prediction.

(Appendix I.5). The shortcut adds $c_\phi(t)x_t$ to the core output. Table 1 reports FID: the shortcut improves v-prediction from 96.53 to 7.83, versus 8.62 for x-prediction.

Table 1: ImageNet 256×256 FID with JiT-B/16 after 200 epochs.

Parameterization	FID ↓
x-prediction	8.62
Direct v-prediction	96.53
Shortcut-assisted v-prediction	**7.83**

6 Conclusion and Limitations

Conclusion. We characterize geometric differences between prediction targets: x-prediction follows the signal structure, whereas noise and velocity prediction can retain ambient components. We further extend this characterization to the one-step generation framework of Mean Flow. To understand how these differences affect approximation and learning, we analyze models with limited output rank, using Gaussian data and Gaussian mixtures as tractable settings. The target decomposition motivates a scalar shortcut that separates the ambient component from the signal-dependent part. We show that this shortcut removes output-rank-induced approximation floors under the stated conditions, while experiments further demonstrate improved learning and generation even without an output-width bottleneck.

Limitations. Output rank captures only one aspect of model capacity, and extending the analysis to broader architectural constraints would provide a more complete understanding of target-dependent learning. The mixture learning result is local and specific to a one-gate model for a two-component Gaussian mixture; extending it to more general initializations and mixtures remains open. Our experiments cover a limited set of architectures and training configurations, and broader evaluations are needed to assess the roles of architecture, loss weighting, and optimization. Finally, while the shortcut also improves learning and generation without an output-width bottleneck, the mechanism underlying this benefit remains to be theoretically characterized.

AI USE STATEMENT

Generative AI tools were used for language polishing and writing guidance, including improvements to grammar, clarity, organization, and presentation. They also assisted with preparing and refining experimental visualizations, including figure layouts, labels, and captions, and with checking mathematical proofs for algebraic errors, consistency of notation and assumptions, and potential gaps in logical arguments.

REFERENCES

Brian DO 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.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=r5njV3BsuD.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pp. 4735–4763. PMLR, 2023a.

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, 2023b.

Hugo Cui, Florent Krzakala, Eric Vanden-Eijnden, and Lenka Zdeborova. Analysis of learning a flow-based generative model from limited sample complexity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ndCJeysCPe.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

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

Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Funing Fu, Tenghui Wang, Guanyu Zhou, Junyong Cen, and Qichao Zhu. Jlt: Clean-latent prediction in latent diffusion transformers. *arXiv preprint arXiv:2605.27102*, 2026.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean Flows for One-step Generative Modeling, May 2025.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Zhihan Huang, Yuting Wei, and Yuxin Chen. Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *Mathematics of Operations Research*, 2026.

Qing Jin and Chaoyang Wang. Revisiting diffusion model predictions through dimensionality. *arXiv preprint arXiv:2601.21419*, 2026.

Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184. IEEE, 2024.

Gen Li and Yuchen Jiao. Improved convergence rate for diffusion probabilistic models. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows, 2026. URL https://arxiv.org/abs/2601.22158.

Gabin Maxime Nguegnang, Holger Rauhut, and Ulrich Terstiege. Convergence of gradient descent for learning linear neural networks. *Advances in Continuous and Discrete Models*, 2024(1):23, 2024.

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.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

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.

Yuqing Wang, Ye He, and Molei Tao. Evaluating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 37:19307–19352, 2024.

Le Zhang, Ning Mang, and Aishwarya Agrawal. Rit: Vanilla diffusion transformers suffice in representation space. *arXiv preprint arXiv:2605.21981*, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

A Proof of Low-Rank Gaussian Case

We first establish the invertibility of the affine feature matrix, then prove the Gaussian approximation and learning results and their score-induced KL corollary.

Lemma A.1 (Positive definiteness of the linear Gram matrix)**.** *At every analyzed noise level with* $b_t > 0$*,*

$$G_t \succ 0.$$

Proof. Since $b_t > 0$, we have $\Sigma_t \succ 0$, and therefore X_t has a strictly positive density on $\mathbb{R}^D$. Suppose that for some $(a, \beta) \in \mathbb{R}^D \times \mathbb{R}$,

$$\mathbb{E}\big[(a^\top X_t + \beta)^2\big] = 0.$$

Then $a^\top X_t + \beta = 0$ almost surely. Since X_t has positive density on all of $\mathbb{R}^D$, this implies

$$a^\top x + \beta = 0 \qquad \text{for all } x \in \mathbb{R}^D.$$

Hence $a = 0$ and $\beta = 0$. Therefore, $G_t \succ 0$. □

A.1 Approximation and Learning under Limited Output Rank

This subsection provides the approximation and learning analysis for Section 3.2. We first characterize the optimal approximation errors in Theorem 2, then prove gradient-descent convergence and the prediction properties after convergence.

Optimal Predictors and Target-Dependent Approximation Errors

We use $\mathcal{M}$, d_{int}, and ψ as defined in Section 3.2. Fix $t > 0$ with $a_t, b_t > 0$, and let $G_t = \mathbb{E}[\psi(X_t)\psi(X_t)^\top] \succ 0$ by Lemma A.1. Under Assumption 2, each optimal predictor is affine and admits a unique representation

$$f_t^{\star,(\tau)}(x) = W_t^{\star,(\tau)}\psi(x), \qquad \tau \in \{x, \varepsilon, v\}.$$

For the clean target, write $W_t^\star = W_t^{\star,(x)}$. At this fixed time, abbreviate the excess risk in (9) as

$$\mathcal{R}_\tau(A, \Theta) := \mathcal{R}_{\tau,t}(A, \Theta) = \|(A\Theta - W_t^{\star,(\tau)})G_t^{1/2}\|_F^2.$$

The optimal approximation error is governed by the target's uncentered second moment,

$$H_{\tau,t} := W_t^{\star,(\tau)} G_t (W_t^{\star,(\tau)})^\top = \mathbb{E}\Big[f_t^{\star,(\tau)}(X_t) f_t^{\star,(\tau)}(X_t)^\top\Big],$$

which accounts for both its mean and covariance. The predictor formulas in (8) give

$$\begin{aligned} H_{x,t} &= \mu\mu^\top + a_t^2 \Sigma \Sigma_t^{-1} \Sigma, \\ H_{\varepsilon,t} &= b_t^2 \Sigma_t^{-1}, \\ H_{v,t} &= (a_t')^2 \mu\mu^\top + B_t \Sigma_t^{-1} B_t. \end{aligned} \tag{17}$$

Both terms in the expression for $H_{x,t}$ in (17) are positive semidefinite. Since $a_t > 0$ and $\Sigma_t \succ 0$, the second term has range $\mathrm{range}(\Sigma)$. Consequently,

$$\mathrm{range}(H_{x,t}) = \mathcal{M}, \qquad \mathrm{rank}(H_{x,t}) = d_{\mathrm{int}}. \tag{18}$$

The following propositions characterize the optimal output range and quantify the approximation floor when the output range is constrained to have rank at most m.

Proposition 1 (Spectral-tail characterization of the optimal output range)**.** *Under Assumption 2, fix t with $a_t, b_t > 0$. For $1 \le m \le D$ and each $\tau \in \{x, \varepsilon, v\}$,*

$$\boxed{\mathcal{A}^{\mathrm{opt}}_{\tau,t}(m) := \inf_{A,\Theta} \mathcal{R}_\tau(A,\Theta) = \sum_{j>m} \lambda_j(H_{\tau,t}),}$$

where eigenvalues are ordered nonincreasingly. In particular, $\mathrm{rank}(H_{x,t}) = d_{\mathrm{int}}$*, so* $\mathcal{A}^{\mathrm{opt}}_{x,t}(m) = 0$ *whenever* $m \ge d_{\mathrm{int}}$.

The following bounds quantify the ambient contribution when the range is also optimized.

Proposition 2 (Ambient burden after optimizing the output range)**.** *Under the conditions of Proposition 1, set $S = \mathrm{range}(\Sigma)$ and $r = \mathrm{rank}(\Sigma)$. For $d_{\mathrm{int}} \le m \le D$,*

$$\boxed{\begin{aligned}(D-r-m)_+ \le \mathcal{A}^{\mathrm{opt}}_{\varepsilon,t}(m) \le D-m,\\ (b'_t)^2(D-r-m)_+ \le \mathcal{A}^{\mathrm{opt}}_{v,t}(m) \le (b'_t)^2(D-m).\end{aligned}}$$

Here $(u)_+ = \max\{u, 0\}$. The lower bounds also hold for $m < d_{\mathrm{int}}$. A positive v lower bound requires $(b'_t)^2 > 0$ and $D - r - m > 0$; if $(b'_t)^2 = 0$, then $\mathcal{A}^{\mathrm{opt}}_{v,t}(m) = 0$ for $m \ge d_{\mathrm{int}}$.

Proof of Proposition 1. For a fixed output subspace with orthogonal projector P, each $Pf^{\star,(\tau)}_t$ is representable by an affine head in that subspace. Orthogonal decomposition therefore gives the profiled risk $\mathrm{Tr}((I_D - P)H_{\tau,t})$. Every subspace of dimension at most m can be represented by $A \in \mathbb{R}^{D\times m}$. The Ky Fan variational characterization gives

$$\min_{\substack{P=P^\top=P^2\\ \mathrm{rank}(P)\le m}} \mathrm{Tr}((I_D - P)H_{\tau,t}) = \sum_{j>m} \lambda_j(H_{\tau,t}).$$

The x-claim follows from (18) at the chosen time. □

Proof of Proposition 2. Put $P_\perp = P_{S^\perp}$ and $B_t = a'_t a_t \Sigma + b'_t b_t I_D$. Equation (8) gives

$$P_\perp \varepsilon^\star_t(X_t) = P_\perp \varepsilon, \qquad P_\perp(v^\star_t(X_t) - a'_t\mu) = b'_t P_\perp \varepsilon.$$

To obtain operator inequalities, note that Σ_t and B_t preserve $S \oplus S^\perp$. Since $\mathbb{E}\varepsilon^\star_t = 0$ and $\mathbb{E}v^\star_t = a'_t\mu$,

$$\begin{aligned} H_{\varepsilon,t} &= b_t^2 \Sigma_t^{-1} \succeq P_\perp,\\ \mathrm{Cov}(v^\star_t(X_t)) &= B_t\Sigma_t^{-1}B_t = (b'_t)^2 P_\perp + C_t, \qquad C_t \succeq 0,\\ H_{v,t} &= \mathrm{Cov}(v^\star_t(X_t)) + (a'_t)^2\mu\mu^\top \succeq (b'_t)^2 P_\perp, \end{aligned}$$

where C_t is supported on S. Thus the uncentered mean term cannot invalidate the lower bound, even when $\mu \notin S$. For every orthogonal projector P of rank at most m,

$$\mathrm{Tr}((I_D - P)P_\perp) = D - r - \mathrm{Tr}(PP_\perp) \ge (D - r - m)_+.$$

Taking traces against $I_D - P \succeq 0$ and minimizing proves the lower bounds. For $m \ge d_{\mathrm{int}}$, choose an orthogonal projector P of rank m whose range contains $\mathcal{M}$. The affine predictors $P\varepsilon^\star$ and $Pv^\star$ are admissible rank-m candidates. Since $(I_D - P)x_0 = (I_D - P)x^\star(x_t) = 0$, the target identities give

$$(I_D - P)\varepsilon^\star(x_t) = (I_D - P)\varepsilon, \qquad (I_D - P)v^\star(x_t) = b'_t(I_D - P)\varepsilon.$$

Their errors are $D - m$ and $(b'_t)^2(D - m)$, respectively. □

When $r \ll D - m$, the optimized ε risk remains $\Theta(D-m)$, and the optimized v risk is $\Theta((b'_t)^2(D - m))$. At a fixed time with $b'_t \ne 0$ independent of D, the latter is also $\Theta(D - m)$.

Gradient-Descent Convergence and Prediction after Convergence

Let $E^{(k)}_\tau = A^{(k)}_\tau \Theta^{(k)}_\tau - W^{\star,(\tau)}_t$. Simultaneous gradient descent updates both factors using the same iterate:

$$\begin{aligned} A^{(k+1)}_\tau &= A^{(k)}_\tau - 2\eta_\tau E^{(k)}_\tau G_t (\Theta^{(k)}_\tau)^\top,\\ \Theta^{(k+1)}_\tau &= \Theta^{(k)}_\tau - 2\eta_\tau (A^{(k)}_\tau)^\top E^{(k)}_\tau G_t. \end{aligned} \tag{19}$$

Initialization and stepsize. The convergence guarantee requires a sufficiently small constant stepsize, chosen before the initial factors are sampled. For each target τ, let $\mathcal{B}_\tau$ be the bounded initialization support from Theorem 2 and choose a finite upper bound on the initial risk, $S_\tau \geq \sup_{(A,\Theta)\in\mathcal{B}_\tau} \mathcal{R}_{\tau,t}(A,\Theta)$. For the stepsize calculation, set $\mathsf{X} = G_t^{1/2}$ and $\mathsf{Y}_\tau = W_t^{\star,(\tau)}\mathsf{X}$. Define

$$\begin{aligned} \delta_{\mathcal{B},\tau} &= 2 \sup_{(A,\Theta)\in\mathcal{B}_\tau} \|A^\top A - \Theta\Theta^\top\|_2 + 1, \\ K_{\mathcal{B},\tau} &= \frac{\sqrt{S_\tau} + \|\mathsf{Y}_\tau\|_2}{\sigma_{\min}(\mathsf{X})} + 9\delta_{\mathcal{B},\tau}, \qquad (20) \\ C_\tau &= 4\mathrm{e}K_{\mathcal{B},\tau}\|\mathsf{X}\|_2^2 + 2\sqrt{\mathrm{e}}\|\mathsf{Y}_\tau\mathsf{X}^\top\|_2, \\ \eta_{\mathcal{B},\tau} &= \frac{\delta_{\mathcal{B},\tau}}{4S_\tau + \delta_{\mathcal{B},\tau}C_\tau}. \qquad (21) \end{aligned}$$

Here e is the base of the natural logarithm. Denote the smallest and largest eigenvalues of G_t by $\lambda_{\min}^t$ and $\lambda_{\max}^t$. All quantities are determined by the initialization support and the fixed regression problem.

Regression formulation. We prove Theorem 2 using the simultaneous updates (19). For the clean target, abbreviate $A = A_x$, $\Theta = \Theta_x$, $\mathcal{R}_x^{(k)} = \mathcal{R}_{x,t}^{(k)}$, $P^{(k)} = A^{(k)}(A^{(k)})^\dagger$, and $W_t^\star = W_t^{\star,(x)}$. Let $U_\mathcal{M}$ be an orthonormal basis of $\mathcal{M}$. Recall that $G_t = \mathbb{E}[\psi(X_t)\psi(X_t)^\top] \succ 0$. Define the regression matrices

$$\mathsf{X} = G_t^{1/2} \in \mathbb{R}^{(D+1)\times(D+1)}, \qquad \mathsf{Y} = W_t^\star G_t^{1/2} \in \mathbb{R}^{D\times(D+1)}.$$

Then X is invertible and

$$\mathsf{Y}\mathsf{Y}^\top = H_{x,t}, \qquad \mathrm{rank}(\mathsf{Y}) = d_{\mathrm{int}}.$$

The objective becomes

$$\mathcal{R}_x(A,\Theta) = \|A\Theta\mathsf{X} - \mathsf{Y}\|_F^2.$$

Thus (19) is gradient descent for a two-layer linear network. The global recovery result follows from Nguegnang et al. (2024, Theorem 2.6(b)). The uniform factor bounds used below are

$$\|A^{(k)}\|_2^2 \leq K_{\mathcal{B},x}, \qquad \|\Theta^{(k)}\|_2^2 \leq K_{\mathcal{B},x}. \qquad (22)$$

Proof of Theorem 2. For each target, set $\mathsf{Y}_\tau = W_t^{\star,(\tau)}\mathsf{X}$. Then

$$\mathcal{R}_{\tau,t}(A,\Theta) = \|A\Theta\mathsf{X} - \mathsf{Y}_\tau\|_F^2, \qquad \mathsf{Y}_\tau\mathsf{Y}_\tau^\top = H_{\tau,t}.$$

Apply Nguegnang et al. (2024, Theorem 2.6(b) and Proposition 3.4) with depth 2, $W_1 = \Theta$, $W_2 = A$, $\alpha = 1/2$, and $\delta_{\mathcal{B},\tau}$ from (20). The assumptions hold because X is invertible and the initial imbalance and loss are uniformly bounded on $\mathcal{B}_\tau$. Accounting for their loss normalization gives the stepsize bound (21) and the factor bounds (22), with $K_{\mathcal{B},\tau}$ for each target. Absolute continuity of the initialization then ensures almost-sure convergence to finite global minimizers for all three targets. Propositions 1 and 2 give the limiting approximation errors and their bounds.

For the clean target, $\mathrm{rank}(H_{x,t}) = d_{\mathrm{int}} \leq m$, so its limiting risk is zero. Since $\mathrm{range}(W_t^\star) = \mathcal{M}$ and $m \geq d_{\mathrm{int}}$, there is a column-orthonormal $Q_\star$ whose range contains $\mathcal{M}$. The factors $A = Q_\star$ and $\Theta = Q_\star^\top W_t^\star$ attain zero risk. Thus the global minimum is zero. Since $G_t \succ 0$, this implies

$$A^{(\infty)}\Theta^{(\infty)} = W_t^\star, \qquad \mathcal{R}_x^{(k)} \longrightarrow 0.$$

Let $\gamma = \lambda_{d_{\mathrm{int}}}(H_{x,t}) > 0$. For any factors, the orthogonal projector $P = AA^\dagger$ satisfies

$$\begin{aligned} \mathcal{R}_x(A,\Theta) &\geq \|(I_D - P)\mathsf{Y}\|_F^2 \\ &= \mathrm{Tr}\big((I_D - P)H_{x,t}\big) \geq \gamma\|(I_D - P)U_\mathcal{M}\|_2^2. \end{aligned} \qquad (23)$$

Since $\mathcal{R}_x^{(k)} \to 0$, this proves the containment limit. The limiting product identity also gives $\mathcal{M} \subseteq \mathrm{range}(A^{(\infty)})$, with equality when $m = d_{\mathrm{int}}$.

For noise prediction, $H_{\varepsilon,t} = b_t^2\Sigma_t^{-1} \succ 0$, so its spectral tail is strictly positive for every $m < D$ and zero for $m = D$. For velocity prediction, the lower bound in Proposition 2 is strictly positive when $b_t' \neq 0$ and $D - r - m > 0$. If $b_t' = 0$, then $v^\star = a_t' x^\star$, which is exactly representable for $m \geq d_{\mathrm{int}}$. At $m = D$, every affine target is exactly representable. Convergence to global minimizers transfers these approximation conclusions to the trained factor limits. □

A.2 Proof of the Gaussian KL bound

We now give Proposition 3, which is used in the proof of Corollary 1 to link the score estimation error to the KL divergence between the generated and true distributions.

Proposition 3. *Fix $0 < \delta < T$. Let p_δ be the true marginal at time δ, and let $\hat{p}_\delta$ be the marginal induced by the reverse-time SDE driven by a learned score $\hat{s}_t$ on $[\delta, T]$. Suppose*

$$dX_t = f(X_t, t)\,dt + g(t)\,dW_t.$$

Under Assumption 2, let $f(x,t) = \mathbf{f}(t)x$ and let $\hat{s}_t$ be affine, with $a_t, b_t > 0$ on $[\delta, T]$. Assume that $\mathbf{f}$, g, and the coefficients of $\hat{s}_t$ are bounded and piecewise continuous, and that both reverse SDEs start from p_T. Then

$$\mathrm{KL}(p_\delta \| \hat{p}_\delta) \le \frac{1}{2}\int_\delta^T g(t)^2\, \mathbb{E}_{X_t\sim p_t}\left\|\hat{s}_t(X_t) - s_t(X_t)\right\|^2 dt, \tag{24}$$

where $s_t = \nabla \log p_t$.

Proof of Proposition 3. Following the entropy argument of Chen et al. (2023a, Lemma C.1 and Proposition C.3), set $u = T - t$, $q_u = p_{T-u}$, and let $\hat{q}_u$ denote the learned marginal. Writing $h_u = g(T-u)$, the two reverse drifts satisfy

$$F_u - \hat{F}_u = h_u^2\big(s_{T-u} - \hat{s}_{T-u}\big), \qquad q_0 = \hat{q}_0 = p_T.$$

Let

$$z_u = \nabla \log(q_u/\hat{q}_u).$$

The Fokker–Planck equations yield, for almost every u,

$$\begin{aligned}
\frac{d}{du}\mathrm{KL}(q_u\|\hat{q}_u) &= -\frac{h_u^2}{2}\mathbb{E}_{q_u}\|z_u\|^2 + \mathbb{E}_{q_u}\langle F_u - \hat{F}_u, z_u\rangle \\
&= -\frac{h_u^2}{2}\mathbb{E}_{q_u}\|z_u\|^2 + h_u^2\mathbb{E}_{q_u}\langle s_{T-u} - \hat{s}_{T-u}, z_u\rangle \\
&= \frac{h_u^2}{2}\mathbb{E}_{q_u}\left[\|s_{T-u} - \hat{s}_{T-u}\|^2 - \|z_u - (s_{T-u} - \hat{s}_{T-u})\|^2\right] \\
&\le \frac{h_u^2}{2}\mathbb{E}_{q_u}\|s_{T-u} - \hat{s}_{T-u}\|^2.
\end{aligned}$$

Integrating over $u \in [0, T-\delta]$, using $\mathrm{KL}(q_0\|\hat{q}_0) = 0$, and changing variables back to $t = T - u$ gives (24). □

Now we proceed to prove Corollary 1.

Proof of Corollary 1. For the evaluated family of affine predictors, the score identity gives

$$\mathbb{E}_{X_t\sim p_t}\|\hat{s}_t^{(\tau,k)}(X_t) - s_t(X_t)\|^2 = |\lambda_t^{(\tau)}|^2\mathcal{R}_{\tau,t}^{(k)} = |\lambda_t^{(\tau)}|^2\left[\mathcal{A}_{\tau,t}^{\mathrm{opt}}(m) + \mathcal{E}_{\tau,t}^{(k)}\right].$$

Apply Proposition 3 and substitute $\lambda_t^{(x)} = a_t/b_t^2$, $\lambda_t^{(\varepsilon)} = -1/b_t$, and $\lambda_t^{(v)} = -a_t/(b_t\Delta_t)$. The clean approximation term vanishes by Theorem 2, yielding all three integrands in the corollary. No head-only contraction bound is used. □

B Proofs of Mixtures of Low-Rank Gaussians

This appendix proves the mixture results in Section 3.3. We now prove Theorem 3.

Proof of Theorem 3. We first derive explicit expressions for the optimal predictors. Substituting the score decomposition (11) into (7) gives

$$\varepsilon^\star(x_t) = b_t\bar{\Sigma}_t^{-1}(x_t - \bar{\mu}_t) - b_t\Psi_t(x_t). \tag{25}$$

For the clean predictor, the same identities give

$$x^\star(x_t) = \frac{x_t}{a_t} - \frac{b_t^2}{a_t}\bar{\Sigma}_t^{-1}(x_t - \bar{\mu}_t) + \frac{b_t^2}{a_t}\Psi_t(x_t).$$

Using $\bar{\mu}_t = a_t\bar{\mu}_0$ and $I_D - b_t^2\bar{\Sigma}_t^{-1} = a_t^2\bar{\Sigma}\bar{\Sigma}_t^{-1}$, we obtain

$$x^\star(x_t) = \bar{\mu}_0 + a_t\bar{\Sigma}\bar{\Sigma}_t^{-1}(x_t - \bar{\mu}_t) + \frac{b_t^2}{a_t}\Psi_t(x_t). \tag{26}$$

The velocity predictor follows from $v^\star(x_t) = a_t'x^\star(x_t) + b_t'\varepsilon^\star(x_t)$. We use these expressions to establish the geometric separation stated in the theorem.

For geometric containment, write $\Sigma_{k,t}^{-1} = b_t^{-2}(I_D - a_t^2\Sigma_k\Sigma_{k,t}^{-1})$ and set $\Delta_{k,t} = (a_t^2/b_t^2)\Sigma_k\Sigma_{k,t}^{-1}$ and $\bar{\Delta}_t = (a_t^2/b_t^2)\bar{\Sigma}\bar{\Sigma}_t^{-1}$. Both operators have range in $\mathrm{range}(U) \subseteq \mathcal{S}_0$, where U is the shared basis from Assumption 3. Substituting these identities into Ψ_t yields

$$\begin{aligned}
\Psi_t(x) &= \sum_{k=1}^M w_k(x)\left[\left(\frac{1}{b_t^2}I - \bar{\Delta}_t\right)(x - \bar{\mu}_t) - \left(\frac{1}{b_t^2}I - \Delta_{k,t}\right)(x - a_t\mu_k)\right] \\
&= \sum_{k=1}^M w_k(x)\left[\frac{1}{b_t^2}(x - \bar{\mu}_t - x + a_t\mu_k) - \bar{\Delta}_t(x - \bar{\mu}_t) + \Delta_{k,t}(x - a_t\mu_k)\right] \\
&= \frac{a_t}{b_t^2}\sum_{k=1}^M w_k(x)(\mu_k - \bar{\mu}_0) + \sum_{k=1}^M w_k(x)\left[\Delta_{k,t}(x - a_t\mu_k) - \bar{\Delta}_t(x - \bar{\mu}_t)\right],
\end{aligned}$$

where we used $\bar{\mu}_t = a_t\bar{\mu}_0$. The first term is a linear combination of the centered mean shifts $\mu_k - \bar{\mu}_0$, and the remaining terms lie in $\mathrm{range}(U)$. Hence $\Psi_t(x) \in \mathcal{S}_0$ and $x^\star(x) - \bar{\mu}_0 \in \mathcal{S}_0$. Since the M centered mean shifts have one linear dependence, $\dim(\mathcal{S}_0) \le r + M - 1$.

It remains to characterize the orthogonal components of the noise-based targets. Each mixture component is supported in the affine space $\bar{\mu}_0 + \mathcal{S}_0$, so $P_{\mathcal{S}_0^\perp}(X_0 - \bar{\mu}_0) = 0$ almost surely. Projecting $X_t - a_t\bar{\mu}_0 = a_t(X_0 - \bar{\mu}_0) + b_t\varepsilon$ and conditioning on $X_t = x$ gives

$$P_{\mathcal{S}_0^\perp}\varepsilon^\star(x) = \frac{1}{b_t}P_{\mathcal{S}_0^\perp}(x - a_t\bar{\mu}_0).$$

Since x ranges over $\mathbb{R}^D$ for $t > 0$, the projected image is exactly $\mathcal{S}_0^\perp$. Finally, $v^\star = a_t'x^\star + b_t'\varepsilon^\star$ and $P_{\mathcal{S}_0^\perp}(x^\star - \bar{\mu}_0) = 0$, so

$$P_{\mathcal{S}_0^\perp}(v^\star(x) - a_t'\bar{\mu}_0) = \frac{b_t'}{b_t}P_{\mathcal{S}_0^\perp}(x - a_t\bar{\mu}_0).$$

For $b_t' \neq 0$, its projected image is also $\mathcal{S}_0^\perp$. □

For the special case $M = 2$ with shared covariance, the score further reduces to an explicit $\tanh$ form.

Lemma B.1 (Score Function of Two-Component MoLRG)**.** *Let the data distribution be*

$$p_0 = \frac{1}{2}\mathcal{N}(\mu_1, \Sigma) + \frac{1}{2}\mathcal{N}(\mu_2, \Sigma).$$

Under the Gaussian diffusion process

$$x_t = a_t x_0 + b_t\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_D),$$

the score function $\nabla \log p_t(x)$ at time t reduces to

$$\nabla \log p_t(x) = -\Sigma_t^{-1}(x - \bar{\mu}_t) + \Sigma_t^{-1}d_t\,\tanh\big(\langle\Sigma_t^{-1}d_t,\ x - \bar{\mu}_t\rangle\big),$$

where

$$\Sigma_t = a_t^2\Sigma + b_t^2 I_D, \qquad \bar{\mu}_t = a_t\frac{\mu_1 + \mu_2}{2}, \qquad d_t = a_t\frac{\mu_1 - \mu_2}{2}.$$

Proof of Lemma B.1. By Gaussian convolution, the distribution at time t is

$$p_t(x) = \frac{1}{2}\mathcal{N}(x; a_t\mu_1, \Sigma_t) + \frac{1}{2}\mathcal{N}(x; a_t\mu_2, \Sigma_t),$$

with $\Sigma_t = a_t^2\Sigma + b_t^2 I_D$. Let

$$\phi_1(x) \coloneqq \mathcal{N}(x; a_t\mu_1, \Sigma_t), \qquad \phi_2(x) \coloneqq \mathcal{N}(x; a_t\mu_2, \Sigma_t).$$

Then

$$p_t(x) = \frac{1}{2}\phi_1(x) + \frac{1}{2}\phi_2(x),$$

and hence

$$\nabla \log p_t(x) = \frac{\phi_1(x)\nabla \log \phi_1(x) + \phi_2(x)\nabla \log \phi_2(x)}{\phi_1(x) + \phi_2(x)}.$$

Using the Gaussian score identity

$$\nabla \log \phi_i(x) = -\Sigma_t^{-1}(x - a_t\mu_i), \qquad i = 1, 2,$$

we obtain

$$\nabla \log p_t(x) = \frac{\phi_1(x)\big[-\Sigma_t^{-1}(x - a_t\mu_1)\big] + \phi_2(x)\big[-\Sigma_t^{-1}(x - a_t\mu_2)\big]}{\phi_1(x) + \phi_2(x)}.$$

Now write

$$a_t\mu_1 = \bar{\mu}_t + d_t, \qquad a_t\mu_2 = \bar{\mu}_t - d_t.$$

Substituting these identities yields

$$\begin{aligned}\nabla \log p_t(x) &= \frac{\phi_1(x)\big[-\Sigma_t^{-1}(x - \bar{\mu}_t - d_t)\big] + \phi_2(x)\big[-\Sigma_t^{-1}(x - \bar{\mu}_t + d_t)\big]}{\phi_1(x) + \phi_2(x)} \\ &= -\Sigma_t^{-1}(x - \bar{\mu}_t) + \Sigma_t^{-1}d_t\,\frac{\phi_1(x) - \phi_2(x)}{\phi_1(x) + \phi_2(x)}.\end{aligned}$$

It remains to simplify the ratio term. We compute

$$\begin{aligned}\log\frac{\phi_1(x)}{\phi_2(x)} &= -\frac{1}{2}\|x - \bar{\mu}_t - d_t\|^2_{\Sigma_t^{-1}} + \frac{1}{2}\|x - \bar{\mu}_t + d_t\|^2_{\Sigma_t^{-1}} \\ &= 2\left\langle \Sigma_t^{-1}d_t,\ x - \bar{\mu}_t\right\rangle.\end{aligned}$$

Let

$$z \coloneqq \left\langle \Sigma_t^{-1}d_t,\ x - \bar{\mu}_t\right\rangle.$$

Then $\phi_1(x)/\phi_2(x) = e^{2z}$, and therefore

$$\frac{\phi_1(x) - \phi_2(x)}{\phi_1(x) + \phi_2(x)} = \frac{e^{2z} - 1}{e^{2z} + 1} = \tanh(z).$$

Substituting this back gives

$$\nabla \log p_t(x) = -\Sigma_t^{-1}(x - \bar{\mu}_t) + \Sigma_t^{-1}d_t\,\tanh\big(\left\langle \Sigma_t^{-1}d_t,\ x - \bar{\mu}_t\right\rangle\big),$$

which proves the claim. □

Exact teacher construction. In the balanced two-component setting of Section 3.3, Lemma G.5 gives

$$x^\star(x) = \bar{\mu}_0 + a_t\Sigma\Sigma_t^{-1}(x - \bar{\mu}_t) + \frac{b_t^2}{a_t}\Sigma_t^{-1}d_t\,\phi_t(x), \qquad \phi_t(x) = \tanh\big((\Sigma_t^{-1}d_t)^\top(x - \bar{\mu}_t)\big).$$

For the fixed Q with $\mathcal{S} \subseteq \mathrm{range}(Q)$, choose the parameters in (12) as

$$\begin{aligned}\omega_\star &= \Sigma_t^{-1}d_t, & \beta_\star &= -\omega_\star^\top\bar{\mu}_t, \\ W_\star &= a_tQ^\top\Sigma\Sigma_t^{-1}, & u_\star &= \frac{b_t^2}{a_t}Q^\top\omega_\star, \\ c_\star &= Q^\top\bar{\mu}_0 - W_\star\bar{\mu}_t.\end{aligned} \tag{27}$$

The gate matches ϕ_t, and the affine and gate coefficients of $x^\star$ lie in $\mathcal{S}$, so $f_{\theta_\star}(x) = x^\star(x)$ for all x. When $d_t \neq 0$, both $\omega_\star$ and $u_\star$ are nonzero, as required for the local convergence argument below.

We give Lemma B.2 to show that the Jacobian Gram matrix at the teacher point is positive definite, which is a key ingredient in the proof of Theorem 4.

Lemma B.2 (Jacobian nondegeneracy for the trainable one-gate model)**.** *In the two-component setting, assume $a_t, b_t > 0$ and $d_t \neq 0$. Let*

$$\omega_\star \coloneqq \Sigma_t^{-1} d_t, \qquad \beta_\star \coloneqq -\omega_\star^\top \bar{\mu}_t.$$

Define

$$\phi_\star(x) \coloneqq \tanh(\omega_\star^\top x + \beta_\star), \qquad \rho_\star(x) \coloneqq \operatorname{sech}^2(\omega_\star^\top x + \beta_\star).$$

Then the scalar feature family

$$1, \qquad x, \qquad \phi_\star(x), \qquad \rho_\star(x), \qquad \rho_\star(x)x$$

is linearly independent in $L^2(p_t)$: if

$$\mathbb{E}\Big[\big(a^\top X_t + \alpha\phi_\star(X_t) + \beta + \rho_\star(X_t)(c^\top X_t + \delta)\big)^2\Big] = 0,$$

for some $(a, \alpha, \beta, c, \delta) \in \mathbb{R}^D \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^D \times \mathbb{R}$, then

$$a = 0, \qquad \alpha = 0, \qquad \beta = 0, \qquad c = 0, \qquad \delta = 0.$$

Consequently, if $u_\star \neq 0$, then the Jacobian Gram at $\theta_\star$ is positive definite. More precisely, for every nonzero perturbation

$$\Delta\theta = (\Delta W, \Delta u, \Delta c, \Delta\omega, \Delta\beta),$$

we have

$$\mathbb{E}\Big[\big\|Df_{\theta_\star}(X_t)[\Delta\theta]\big\|^2\Big] > 0, \tag{28}$$

where

$$Df_{\theta_\star}(x)[\Delta\theta] = Q\Big(\Delta W x + \Delta u\,\phi_\star(x) + \Delta c + u_\star\,\rho_\star(x)\,(\Delta\omega^\top x + \Delta\beta)\Big).$$

Equivalently, the self-adjoint Jacobian Gram operator

$$\mathcal{H}_\star \coloneqq \mathbb{E}\big[J_{\theta_\star}(X_t)^\top J_{\theta_\star}(X_t)\big]$$

satisfies

$$\mathcal{H}_\star \succ 0.$$

Proof. Since $b_t > 0$, the diffused distribution p_t has a strictly positive density on $\mathbb{R}^D$. Therefore, if a continuous function vanishes p_t-almost surely, then it must vanish identically on $\mathbb{R}^D$.

Step 1: linear independence of $1, x, \phi_\star, \rho_\star, \rho_\star x$. Assume

$$\mathbb{E}\Big[\big(a^\top X_t + \alpha\phi_\star(X_t) + \beta + \rho_\star(X_t)(c^\top X_t + \delta)\big)^2\Big] = 0.$$

Then the function

$$h(x) \coloneqq a^\top x + \alpha\tanh(\omega_\star^\top x + \beta_\star) + \beta + \operatorname{sech}^2(\omega_\star^\top x + \beta_\star)(c^\top x + \delta)$$

satisfies

$$h(x) = 0 \qquad \text{for all } x \in \mathbb{R}^D.$$

We first show that both a and c must be parallel to $\omega_\star$. Write

$$a = a_\parallel + a_\perp, \qquad c = c_\parallel + c_\perp,$$

where $a_\parallel, c_\parallel \in \operatorname{span}\{\omega_\star\}$ and $a_\perp, c_\perp \perp \omega_\star$.

Fix any $q \perp \omega_\star$, and consider the affine line

$$x = x_0 + sq, \qquad s \in \mathbb{R}.$$

Along this line,

$$\omega_\star^\top x + \beta_\star = \omega_\star^\top x_0 + \beta_\star$$

is constant, so both

$$\tanh(\omega_\star^\top x + \beta_\star), \qquad \operatorname{sech}^2(\omega_\star^\top x + \beta_\star)$$

remain constant as functions of s. Hence $h(x_0 + sq)$ is affine in s, namely

$$h(x_0 + sq) = C_0 + s\Big(a^\top q + \rho_0\, c^\top q\Big),$$

where

$$\rho_0 \coloneqq \operatorname{sech}^2(\omega_\star^\top x_0 + \beta_\star).$$

Since $h \equiv 0$, the slope must vanish:

$$a^\top q + \rho_0\, c^\top q = 0.$$

Now x_0 may be chosen so that $\omega_\star^\top x_0 + \beta_\star$ takes any prescribed value, hence ρ_0 can take infinitely many values in $(0, 1]$. Therefore the identity above holds for at least two distinct values of ρ_0, which forces

$$a^\top q = 0, \qquad c^\top q = 0.$$

Since this is true for every $q \perp \omega_\star$, we conclude that

$$a_\perp = 0, \qquad c_\perp = 0.$$

Thus there exist scalars $\kappa, \lambda \in \mathbb{R}$ such that

$$a = \kappa\omega_\star, \qquad c = \lambda\omega_\star.$$

Substituting this back into h, choose

$$x = \frac{s - \beta_\star}{\|\omega_\star\|^2}\omega_\star,$$

so that

$$\omega_\star^\top x + \beta_\star = s.$$

Then $h(x) = 0$ becomes

$$\kappa(s - \beta_\star) + \alpha\tanh(s) + \beta + \operatorname{sech}^2(s)\big(\lambda(s - \beta_\star) + \delta\big) = 0 \qquad \text{for all } s \in \mathbb{R}.$$

Equivalently,

$$A + Bs + C\tanh(s) + D\operatorname{sech}^2(s) + E\, s\operatorname{sech}^2(s) = 0 \qquad \text{for all } s \in \mathbb{R},$$

for suitable constants A, B, C, D, E.

We now identify these coefficients. Since $\operatorname{sech}^2(s) \to 0$ as $s \to \pm\infty$, the term Bs is the only unbounded one, so necessarily

$$B = 0.$$

Taking $s \to +\infty$ and $s \to -\infty$, and using $\tanh(s) \to \pm 1$, we obtain

$$A + C = 0, \qquad A - C = 0,$$

hence

$$A = 0, \qquad C = 0.$$

The identity then reduces to

$$D\operatorname{sech}^2(s) + E\, s\operatorname{sech}^2(s) = 0 \qquad \text{for all } s \in \mathbb{R}.$$

Since $\operatorname{sech}^2(s) > 0$ for all s, this implies

$$D + Es = 0 \qquad \text{for all } s \in \mathbb{R},$$

and therefore

$$D = 0, \qquad E = 0.$$

Tracing back the definitions of A, B, C, D, E, we conclude that

$$a = 0, \qquad \alpha = 0, \qquad \beta = 0, \qquad c = 0, \qquad \delta = 0.$$

This proves the claimed linear independence.

Step 2: positive definiteness of the Jacobian Gram. Suppose

$$\mathbb{E}\Big[\big\|Df_{\theta_\star}(X_t)[\Delta\theta]\big\|^2\Big] = 0,$$

for some perturbation

$$\Delta\theta = (\Delta W, \Delta u, \Delta c, \Delta\omega, \Delta\beta).$$

Since Q has orthonormal columns, this is equivalent to

$$\mathbb{E}\Big[\|g(X_t)\|^2\Big] = 0,$$

where

$$g(x) \coloneqq \Delta W x + \Delta u\, \phi_\star(x) + \Delta c + u_\star\, \rho_\star(x)(\Delta\omega^\top x + \Delta\beta).$$

Hence $g(x) = 0$ for all $x \in \mathbb{R}^D$.

Let $P_{u_\star^\perp}$ denote the orthogonal projection onto the subspace orthogonal to $u_\star$. Since the last term in $g(x)$ is always parallel to $u_\star$, applying $P_{u_\star^\perp}$ gives

$$P_{u_\star^\perp}\Delta W x + P_{u_\star^\perp}\Delta u\, \phi_\star(x) + P_{u_\star^\perp}\Delta c = 0 \qquad \text{for all } x.$$

Now fix any $z \in \mathbb{R}^m$ with $z \perp u_\star$, and take inner products with z. We obtain

$$(z^\top \Delta W)x + (z^\top \Delta u)\phi_\star(x) + z^\top \Delta c = 0 \qquad \text{for all } x.$$

By the linear independence proved in Step 1, this implies

$$z^\top \Delta W = 0, \qquad z^\top \Delta u = 0, \qquad z^\top \Delta c = 0 \qquad \text{for all } z \perp u_\star.$$

Therefore $\Delta W, \Delta u, \Delta c$ all take values in $\operatorname{span}\{u_\star\}$. Equivalently, there exist

$$\tilde{a} \in \mathbb{R}^D, \qquad \tilde{\alpha}, \tilde{\beta} \in \mathbb{R}$$

such that

$$\Delta W = u_\star \tilde{a}^\top, \qquad \Delta u = \tilde{\alpha}\, u_\star, \qquad \Delta c = \tilde{\beta}\, u_\star.$$

Substituting these expressions into $g(x) = 0$ and factoring out $u_\star \neq 0$, we get

$$\tilde{a}^\top x + \tilde{\alpha}\, \phi_\star(x) + \tilde{\beta} + \rho_\star(x)(\Delta\omega^\top x + \Delta\beta) = 0 \qquad \text{for all } x.$$

Applying Step 1 again, now with

$$a = \tilde{a}, \qquad \alpha = \tilde{\alpha}, \qquad \beta = \tilde{\beta}, \qquad c = \Delta\omega, \qquad \delta = \Delta\beta,$$

yields

$$\tilde{a} = 0, \qquad \tilde{\alpha} = 0, \qquad \tilde{\beta} = 0, \qquad \Delta\omega = 0, \qquad \Delta\beta = 0.$$

Hence

$$\Delta W = 0, \qquad \Delta u = 0, \qquad \Delta c = 0, \qquad \Delta\omega = 0, \qquad \Delta\beta = 0.$$

Therefore $\Delta\theta = 0$, proving

$$\mathbb{E}\Big[\big\|Df_{\theta_\star}(X_t)[\Delta\theta]\big\|^2\Big] > 0 \qquad \text{for every } \Delta\theta \neq 0.$$

Equivalently,

$$\mathcal{H}_\star \succ 0.$$

□

Proof of Theorem 4. Write $F \coloneqq \mathcal{R}_x$. We prove the result in the following steps.

Step 1: Hessian at the exact teacher point. By the explicit construction in (27), the teacher network exactly computes the target, yielding $f_{\theta_\star}(x) = x^\star(x)$ for all x. Consequently, the residual evaluated at the teacher is identically zero.

$$r_\star(x) \coloneqq f_{\theta_\star}(x) - x^\star(x) = 0.$$

The gradient of the risk relies on this residual, evaluating directly to zero for every perturbation $\Delta\theta$.

$$\nabla F(\theta_\star)[\Delta\theta] = 2\,\mathbb{E}\langle r_\star(X_t), Df_{\theta_\star}(X_t)[\Delta\theta]\rangle = 0.$$

Thus, $\theta_\star$ is a critical point. The second directional derivative involves both the Jacobian and the Hessian of the network mapping.

$$\nabla^2 F(\theta_\star)[\Delta\theta, \Delta\theta] = 2\,\mathbb{E}\Big[\big\|Df_{\theta_\star}(X_t)[\Delta\theta]\big\|^2\Big] + 2\,\mathbb{E}\langle r_\star(X_t), D^2 f_{\theta_\star}(X_t)[\Delta\theta, \Delta\theta]\rangle.$$

Because $r_\star(x) = 0$, the second term vanishes entirely.

$$\nabla^2 F(\theta_\star)[\Delta\theta, \Delta\theta] = 2\,\mathbb{E}\Big[\big\|Df_{\theta_\star}(X_t)[\Delta\theta]\big\|^2\Big].$$

By Lemma B.2, the Jacobian Gram matrix is strictly positive definite, establishing $\nabla^2 F(\theta_\star) \succ 0$. We denote the teacher Hessian as $H_\star \coloneqq \nabla^2 F(\theta_\star)$, ensuring its minimum eigenvalue $\lambda_\star$ is strictly positive. Because the network exactly matches the target, the global infimum of the objective evaluates to zero.

$$F(\theta_\star) = \mathbb{E}\|f_{\theta_\star}(X_t) - x^\star(X_t)\|^2 = 0.$$

Step 2: Local strong convexity and smoothness. Let $\lambda_\star > 0$ and $\Lambda_\star$ be the smallest and largest eigenvalues of $H_\star$, and define

$$L_H \coloneqq \sup_{\|\theta - \theta_\star\| \le 1} \|\nabla^3 \mathcal{R}_x(\theta)\|_{\mathrm{op}}.$$

This constant is finite: the parameter ball is bounded, derivatives of $\tanh$ are bounded, and the Gaussian mixture has all polynomial moments. We instantiate the local neighborhood radius and the convexity constants as below.

$$r_{\mathrm{loc}} \coloneqq \min\left\{1, \frac{\lambda_\star}{2(1 + L_H)}\right\}, \qquad \mu_{\mathrm{loc}} \coloneqq \frac{\lambda_\star}{2}, \qquad L_{\mathrm{loc}} \coloneqq \Lambda_\star + \frac{\lambda_\star}{2}.$$

We consider an arbitrary parameter θ within the closed local ball $B(\theta_\star, r_{\mathrm{loc}})$. Because $r_{\mathrm{loc}} \le 1$, the parameter lies strictly within the unit neighborhood where the Hessian-Lipschitz constant L_H is well-defined. By applying the mean value theorem to the Hessian operator, we bound the spectral deviation between the perturbed Hessian $\nabla^2 F(\theta)$ and the teacher Hessian $H_\star$ using the parameter distance.

$$\|\nabla^2 F(\theta) - H_\star\|_{\mathrm{op}} \le L_H \|\theta - \theta_\star\|.$$

By the explicit construction of the local radius, the parameter distance is bounded by r_{loc}. Substituting this radius bound into the Lipschitz inequality directly yields a uniform spectral bound on the Hessian deviation.

$$L_H \|\theta - \theta_\star\| \le L_H r_{\mathrm{loc}} \le L_H \frac{\lambda_\star}{2(1 + L_H)} < \frac{\lambda_\star}{2}.$$

We then apply Weyl's inequality to bound the minimum eigenvalue

$$\lambda_{\min}(\nabla^2 F(\theta)) \ge \lambda_{\min}(H_\star) - \|\nabla^2 F(\theta) - H_\star\|_{\mathrm{op}} \ge \lambda_\star - \frac{\lambda_\star}{2} = \mu_{\mathrm{loc}}.$$

Similarly, the maximum eigenvalue is bounded by

$$\lambda_{\max}(\nabla^2 F(\theta)) \le \lambda_{\max}(H_\star) + \|\nabla^2 F(\theta) - H_\star\|_{\mathrm{op}} \le \Lambda_\star + \frac{\lambda_\star}{2} = L_{\mathrm{loc}}.$$

Thus we obtain

$$\mu_{\mathrm{loc}} I \preceq \nabla^2 F(\theta) \preceq L_{\mathrm{loc}} I \qquad \text{for all } \theta \in B(\theta_\star, r_{\mathrm{loc}}).$$

Step 3: Invariance of the local ball. We define the parameter error sequence as $e^{(K)} \coloneqq \theta^{(K)} - \theta_\star$. Because $\nabla F(\theta_\star) = 0$, the gradient descent iteration can be expressed as a difference of gradients.

$$e^{(K+1)} = e^{(K)} - \eta\big(\nabla F(\theta^{(K)}) - \nabla F(\theta_\star)\big).$$

Assume the current iterate lies within the local ball $\theta^{(K)} \in B(\theta_\star, r_{\mathrm{loc}})$. By the fundamental theorem of calculus, we express the gradient difference via the integrated Hessian.

$$\nabla F(\theta^{(K)}) - \nabla F(\theta_\star) = A_K e^{(K)}, \qquad A_K \coloneqq \int_0^1 \nabla^2 F\big(\theta_\star + s e^{(K)}\big)\, ds.$$

Because the entire line segment lies within the local convex ball, the integrated matrix is strictly bounded.

$$\mu_{\mathrm{loc}} I \preceq A_K \preceq L_{\mathrm{loc}} I.$$

This allows us to rewrite the iteration as $e^{(K+1)} = (I - \eta A_K) e^{(K)}$. For any step size $0 < \eta \leq 1/L_{\mathrm{loc}}$, all eigenvalues of $I - \eta A_K$ reside in the interval $[0, 1 - \eta\mu_{\mathrm{loc}}]$. This forces a strict contraction.

$$\|e^{(K+1)}\| \leq (1 - \eta\mu_{\mathrm{loc}})\|e^{(K)}\| \leq \|e^{(K)}\|.$$

Starting from a valid initialization $\|e^{(0)}\| \leq r_{\mathrm{loc}}$, mathematical induction guarantees that all iterates remain confined to the local neighborhood.

$$\theta^{(K)} \in B(\theta_\star, r_{\mathrm{loc}}) \qquad \text{for all iterations } K \geq 0.$$

Step 4: Linear convergence. Because every iterate remains within the smooth local ball, the standard descent lemma applies.

$$F(\theta^{(K+1)}) \leq F(\theta^{(K)}) - \eta\|\nabla F(\theta^{(K)})\|^2 + \frac{L_{\mathrm{loc}}}{2}\eta^2\|\nabla F(\theta^{(K)})\|^2.$$

Applying the step size upper bound $\eta \leq 1/L_{\mathrm{loc}}$ simplifies this inequality.

$$F(\theta^{(K+1)}) \leq F(\theta^{(K)}) - \frac{\eta}{2}\|\nabla F(\theta^{(K)})\|^2.$$

Local strong convexity ensures that the Polyak-Łojasiewicz (PL) inequality holds on this ball.

$$\|\nabla F(\theta)\|^2 \geq 2\mu_{\mathrm{loc}}\big(F(\theta) - F(\theta_\star)\big).$$

Substituting the PL inequality into the descent bound yields the linear decay of the excess risk.

$$F(\theta^{(K+1)}) - F(\theta_\star) \leq (1 - \eta\mu_{\mathrm{loc}})\big(F(\theta^{(K)}) - F(\theta_\star)\big).$$

Furthermore, using local strong convexity and smoothness, we bound the excess risk by the parameter distance for any $\theta \in B(\theta_\star, r_{\mathrm{loc}})$.

$$\frac{\mu_{\mathrm{loc}}}{2}\|\theta - \theta_\star\|^2 \leq F(\theta) - F(\theta_\star) \leq \frac{L_{\mathrm{loc}}}{2}\|\theta - \theta_\star\|^2.$$

Applying the left inequality to $\theta^{(K)}$, the right inequality to $\theta^{(0)}$, and incorporating the risk contraction yields the final parameter convergence bound.

$$\|\theta^{(K)} - \theta_\star\|^2 \leq \frac{L_{\mathrm{loc}}}{\mu_{\mathrm{loc}}}(1 - \eta\mu_{\mathrm{loc}})^K\|\theta^{(0)} - \theta_\star\|^2.$$

The identical mathematical argument applies directly to the sign-flipped teacher tuple, as the odd parity of the hyperbolic tangent strictly preserves the network output.

$$(-u_\star)\tanh(-\omega_\star^\top x - \beta_\star) = u_\star \tanh(\omega_\star^\top x + \beta_\star).$$

□

C Subspace Data Case

We first prove the geometric separation for affine-subspace data, then extend it to the case of a union of smooth manifolds.

C.1 AFFINE SUBSPACES

Proof of Theorem 1. **The clean predictor remains in the affine support.** Under Assumption 1,

$$x^\star(x_t) = \mathbb{E}[x_0 \mid x_t] = \mu + U\mathbb{E}[z \mid x_t].$$

Thus $x^\star(x_t) - \mu$ lies in $\mathrm{range}(U)$, the direction space of $\mathcal{S}$. Since $\mathcal{P}_{\mathcal{S}^\perp} U = 0$,

$$\mathcal{P}_{\mathcal{S}^\perp}(x^\star(x_t) - \mu) = 0.$$

The other predictors retain the orthogonal noise. Conditioning the forward relation on x_t gives $x_t = a_t x^\star(x_t) + b_t \varepsilon^\star(x_t)$. Projecting this identity and the original forward relation yields

$$\mathcal{P}_{\mathcal{S}^\perp}(x_t - a_t\mu) = b_t \mathcal{P}_{\mathcal{S}^\perp}\varepsilon^\star(x_t) = b_t \mathcal{P}_{\mathcal{S}^\perp}\varepsilon. \tag{29}$$

Since $b_t > 0$, the two projected noise terms agree. The velocity identity $v^\star(x_t) = a_t' x^\star(x_t) + b_t' \varepsilon^\star(x_t)$ then gives

$$\mathcal{P}_{\mathcal{S}^\perp}(v^\star(x_t) - a_t'\mu) = b_t' \mathcal{P}_{\mathcal{S}^\perp}\varepsilon.$$

Finally, $\mathcal{P}_{\mathcal{S}^\perp}$ is a fixed rank-$(D-r)$ orthogonal projector and $\mathbb{E}[\varepsilon\varepsilon^\top] = I_D$, so

$$\mathbb{E}\|\mathcal{P}_{\mathcal{S}^\perp}\varepsilon\|^2 = \mathrm{Tr}(\mathcal{P}_{\mathcal{S}^\perp}) = D - r.$$

The two target energies are therefore $D - r$ and $(b_t')^2(D-r)$. □

C.2 SETTING FOR SMOOTH MANIFOLD UNIONS

We extend the affine-subspace analysis to data near a finite union of smooth manifolds. Conditional averaging need not preserve a curved manifold, so the clean predictor may lie outside the data support. We therefore measure its normal displacement from a local affine tangent plane and compare it with the normal components of the noise and velocity predictors.

Assumption 4 (Bounded manifold patches)**.** *Let $C \in [M]$ be a latent component label with $\mathbb{P}(C = k) = \pi_k > 0$. Conditional on $C = k$, the clean data satisfy*

$$x_0 = \phi_k(z_k) + \xi_k, \qquad z_k \in K_k, \quad \|z_k\| \le R_z, \quad \|\xi_k\| \le R_\xi \quad \textit{almost surely}.$$

Here $K_k \subset \mathbb{R}^{r_k}$ is a nonempty compact convex set, $1 \le r_k < D$, and $\phi_k : U_k \to \mathbb{R}^D$ is a C^2 embedding on an open neighborhood U_k of K_k. Write $\mathcal{M}_k = \phi_k(U_k)$ for the corresponding manifold branch. Assume that the curvature of each embedding is bounded: for every $z \in K_k$ and $u, v \in \mathbb{R}^{r_k}$,

$$\|D^2\phi_k(z)[u,v]\| \le \bar{L}\|u\|\|v\|.$$

Define the diameter of the union of the manifold patches by

$$\Delta_{\max} = \max_{j,k} \sup_{z \in K_j,\, z' \in K_k} \|\phi_j(z) - \phi_k(z')\|.$$

The latent coordinate z_k describes variation within a branch, while ξ_k allows bounded deviations from the manifold. The constants $\bar{L}$ and $\Delta_{\max}$ control local curvature and the overall spatial extent of the patches, respectively. Convexity of K_k ensures that posterior mean coordinates and the line segments used in Taylor expansions remain in the same patch.

Global tangent complexity. To control how much ambient noise can align with an observation-dependent tangent space, collect all possible unit tangent directions in

$$\mathcal{A} = \bigcup_{k=1}^{M} \bigcup_{z \in K_k} \left(\mathrm{range}\, D\phi_k(z) \cap \mathbb{S}^{D-1}\right), \qquad \Gamma_{\tan} = \mathbb{E}\left[\sup_{u \in \mathcal{A}} |\langle \varepsilon, u\rangle|^2\right],$$

where $\varepsilon \sim \mathcal{N}(0, I_D)$. The embedding and compactness assumptions make $\mathcal{A}$ compact, so the supremum is measurable and $\Gamma_{\tan} \le D$. This complexity accounts for variation of the tangent spaces across both locations and branches. The finite-dimensional bounds below retain this dependence explicitly. For dimension scaling, $\Gamma_{\tan} = o(D)$ ensures that most of the ambient noise energy remains in the normal space.

Prediction targets and local reference geometry. Fix t with $a_t, b_t > 0$ and finite a'_t, b'_t, and let

$$x_t = a_t x_0 + b_t \varepsilon, \qquad v_t = a'_t x_0 + b'_t \varepsilon,$$

where ε is independent of all clean-data variables. As in the affine-subspace case, write

$$x^\star(x) = \mathbb{E}[x_0 \mid x_t = x], \qquad \varepsilon^\star(x) = \mathbb{E}[\varepsilon \mid x_t = x], \qquad v^\star(x) = \mathbb{E}[v_t \mid x_t = x].$$

Thus $v^\star = a'_t x^\star + b'_t \varepsilon^\star$; as before, $\Delta_t = a_t b'_t - a'_t b_t$.

For an observation x, choose the branch with the largest posterior probability and evaluate its embedding at the posterior mean latent coordinate:

$$w_k(x) = \mathbb{P}(C = k \mid x_t = x), \qquad \hat{k}(x) \in \operatorname*{arg\,max}_{k \in [M]} w_k(x),$$
$$\bar{z}_{k,t}(x) = \mathbb{E}[z_k \mid x_t = x, C = k], \quad \tilde{x}_t(x) = \phi_{\hat{k}(x)}\big(\bar{z}_{\hat{k}(x),t}(x)\big).$$

Use a fixed deterministic rule to break ties. Since $b_t > 0$, every branch has a strictly positive observation density, and compactness and convexity give $\bar{z}_{k,t}(x) \in K_k$ almost surely. The point $\tilde{x}_t(x)$ supplies a local reference; it need not be a nearest-point projection of either the observation or the clean predictor. Define the tangent and normal spaces at this reference point by

$$T_t(x) = \operatorname{range} D\phi_{\hat{k}(x)}\big(\bar{z}_{\hat{k}(x),t}(x)\big), \qquad N_t(x) = T_t(x)^\perp.$$

The corresponding affine tangent plane is $\tilde{x}_t(x) + T_t(x)$. Below, $\mathcal{P}_{T_t}$ and $\mathcal{P}_{N_t}$ denote the orthogonal projectors evaluated at the random observation x_t. Because $T_t(x_t) \cap \mathbb{S}^{D-1} \subseteq \mathcal{A}$,

$$\mathbb{E}\|\mathcal{P}_{T_t}\varepsilon\|^2 \le \Gamma_{\tan}, \qquad \mathbb{E}\|\mathcal{P}_{N_t}\varepsilon\|^2 \ge D - \Gamma_{\tan}.$$

Quantities controlling normal displacement. The normal displacement depends on structural noise, curvature within the selected branch, and uncertainty about the branch label. To track the latter two effects, define

$$\omega_t(x) = w_{\hat{k}(x)}(x), \qquad \eta_t = \mathbb{E}[1 - \omega_t(x_t)] = \mathbb{P}(C \ne \hat{k}(x_t)),$$

and, for $j \in \{2, 4\}$,

$$m_{j,t}(x) = \mathbb{E}\Big[\|z_{\hat{k}(x)} - \bar{z}_{\hat{k}(x),t}(x)\|^j \,\Big|\, x_t = x, C = \hat{k}(x)\Big].$$

Here η_t is the probability of selecting the wrong branch, and $m_{j,t}$ measures posterior spread within the selected branch. Combine these quantities into the geometric bound $\mathcal{G}_t \ge 0$:

$$\begin{aligned} \mathcal{G}_t^2 &= 2R_\xi^2 + \frac{\bar{L}^2}{2}\mathbb{E}[\omega_t(x_t) m_{4,t}(x_t)] + 2\Delta_{\max}^2 \eta_t \\ &\le 2R_\xi^2 + 8\bar{L}^2 R_z^4 + 2\Delta_{\max}^2 \eta_t. \end{aligned} \tag{30}$$

The inequality uses $m_{4,t} \le 16 R_z^4$. In particular, dimension-independent R_ξ, $\bar{L}$, R_z, and $\Delta_{\max}$ suffice for $\mathcal{G}_t^2 = O(1)$.

Finally, distinguish the normal displacement of the clean predictor from the remaining normal uncertainty in the clean data:

$$\mathcal{B}_t = \mathbb{E}\|\mathcal{P}_{N_t}(x^\star(x_t) - \tilde{x}_t(x_t))\|^2,$$
$$\mathcal{V}_t = \mathbb{E}\|\mathcal{P}_{N_t}(x_0 - x^\star(x_t))\|^2.$$

The next theorem bounds $\mathcal{B}_t + \mathcal{V}_t$ by $\mathcal{G}_t^2$ and uses this control, together with $\Gamma_{\tan}$, to compare the three prediction targets.

C.3 Normal-energy bounds and dimension scaling

Theorem 6 (Normal separation near manifold unions)**.** *Under Assumption 4 and the fixed-noise-level conditions above, the following statements hold. Pointwise statements are understood for p_t-almost every observation.*

(i) Clean-target geometry.

$$\|\mathcal{P}_{N_t(x)}(x^\star(x) - \tilde{x}_t(x))\| \le R_\xi + \frac{\bar{L}}{2}\omega_t(x)m_{2,t}(x) + (1-\omega_t(x))\Delta_{\max}. \tag{31}$$

Moreover,

$$\mathbb{E}\|\mathcal{P}_{N_t}(x_0 - \tilde{x}_t(x_t))\|^2 = \mathcal{B}_t + \mathcal{V}_t \le \mathcal{G}_t^2. \tag{32}$$

(ii) Noise-target energy. *The exact identity is*

$$\mathbb{E}\|\mathcal{P}_{N_t}\varepsilon^\star(x_t)\|^2 = \mathbb{E}\|\mathcal{P}_{N_t}\varepsilon\|^2 - \frac{a_t^2}{b_t^2}\mathcal{V}_t. \tag{33}$$

With

$$\ell_t = \left(D - \Gamma_{\tan} - \frac{a_t^2}{b_t^2}\mathcal{G}_t^2\right)_+,$$

it follows that

$$\ell_t \le \mathbb{E}\|\mathcal{P}_{N_t}\varepsilon^\star(x_t)\|^2 \le D. \tag{34}$$

(iii) Velocity-target energy. *Let*

$$\mathcal{E}_{v,t}^{\mathrm{loc}} = \mathbb{E}\|\mathcal{P}_{N_t}(v^\star(x_t) - a_t'\tilde{x}_t(x_t))\|^2.$$

Then

$$\begin{aligned}\left(|b_t'|\sqrt{\ell_t} - |a_t'|\mathcal{G}_t\right)_+^2 &\le \mathcal{E}_{v,t}^{\mathrm{loc}} \\ &\le \left(|b_t'|\sqrt{D} + |a_t'|\mathcal{G}_t\right)^2.\end{aligned} \tag{35}$$

Corollary 2 (Dimension scaling at a fixed noise level)**.** *Consider a sequence of models satisfying Theorem 6, indexed by D. Assume a_t, b_t, a_t', b_t' are fixed independently of D, with $b_t > 0$. If $\mathcal{G}_t^2 = O(1)$ and $\Gamma_{\tan} = o(D)$, then*

$$\mathcal{B}_t = O(1), \qquad \frac{1}{D}\mathbb{E}\|\mathcal{P}_{N_t}\varepsilon^\star(x_t)\|^2 \longrightarrow 1, \qquad \frac{1}{D}\mathcal{E}_{v,t}^{\mathrm{loc}} \longrightarrow (b_t')^2.$$

In particular, the noise predictor has $\Theta(D)$ normal energy, as does the locally centered velocity predictor when $b_t' \neq 0$.

Proof of Theorem 6. **Step 1: Control the normal displacement.** Fix x and let $k = \hat{k}(x)$, $\bar{z} = \bar{z}_{k,t}(x)$, and $\tilde{x} = \phi_k(\bar{z})$. Since K_k is convex, Taylor's theorem along the segment from $\bar{z}$ to $z \in K_k$ gives

$$\phi_k(z) = \tilde{x} + D\phi_k(\bar{z})(z - \bar{z}) + R_k(z;x), \qquad \|R_k(z;x)\| \le \frac{\bar{L}}{2}\|z - \bar{z}\|^2.$$

The projector $\mathcal{P}_{N_t(x)}$ annihilates the linear term. On the branch $C = k$,

$$\|\mathcal{P}_{N_t(x)}(x_0 - \tilde{x})\| \le R_\xi + \frac{\bar{L}}{2}\|z_k - \bar{z}\|^2.$$

On a branch $C = j \neq k$, both $\phi_j(z_j)$ and $\tilde{x}$ belong to $\bigcup_\ell \phi_\ell(K_\ell)$, so

$$\|\mathcal{P}_{N_t(x)}(x_0 - \tilde{x})\| \le R_\xi + \Delta_{\max}.$$

Conditional Jensen's inequality and averaging over the branches prove (31).

Applying $\|u+v\|^2 \le 2\|u\|^2 + 2\|v\|^2$ separately to the perturbation and geometric displacement on each branch yields

$$\begin{aligned}&\mathbb{E}\big[\|\mathcal{P}_{N_t}(x_0 - \tilde{x}_t(x_t))\|^2 \mid x_t = x\big] \\ &\quad\le 2R_\xi^2 + \frac{\bar{L}^2}{2}\omega_t(x)m_{4,t}(x) + 2\Delta_{\max}^2(1-\omega_t(x)).\end{aligned}$$

Taking expectations gives an upper bound of $\mathcal{G}_t^2$. Now set

$$\rho = \mathcal{P}_{N_t}(x^\star(x_t) - \tilde{x}_t(x_t)), \qquad R = \mathcal{P}_{N_t}(x_0 - x^\star(x_t)).$$

Both ρ and $\mathcal{P}_{N_t}$ are x_t-measurable, and

$$\mathbb{E}[R \mid x_t] = \mathcal{P}_{N_t}\mathbb{E}[x_0 - x^\star(x_t) \mid x_t] = 0.$$

Consequently $\mathbb{E}\langle \rho, R\rangle = 0$. Expanding $\mathbb{E}\|\rho + R\|^2$ proves (32).

Step 2: Establish the noise identity and bounds. The x_t-measurability of $\mathcal{P}_{N_t}$ implies

$$\mathbb{E}[\mathcal{P}_{N_t}\varepsilon \mid x_t] = \mathcal{P}_{N_t}\varepsilon^\star(x_t).$$

Conditional orthogonality therefore gives

$$\begin{aligned}\mathbb{E}\|\mathcal{P}_{N_t}\varepsilon\|^2 &= \mathbb{E}\|\mathcal{P}_{N_t}\varepsilon^\star(x_t)\|^2 \\ &\quad + \mathbb{E}\|\mathcal{P}_{N_t}(\varepsilon - \varepsilon^\star(x_t))\|^2.\end{aligned}$$

Subtracting the conditional-mean forward identity from the forward model shows that

$$\varepsilon - \varepsilon^\star(x_t) = -\frac{a_t}{b_t}(x_0 - x^\star(x_t)).$$

This proves (33).

Pointwise,

$$\begin{aligned}\|\mathcal{P}_{T_t}\varepsilon\|^2 &= \sup_{u \in T_t(x_t)\cap \mathbb{S}^{D-1}} |\langle \varepsilon, u\rangle|^2 \\ &\le \sup_{u \in \mathcal{A}} |\langle \varepsilon, u\rangle|^2.\end{aligned}$$

Since $\mathcal{P}_{N_t} = I_D - \mathcal{P}_{T_t}$ and these are orthogonal complementary projectors,

$$D - \Gamma_{\tan} \le \mathbb{E}\|\mathcal{P}_{N_t}\varepsilon\|^2 \le D.$$

Combining this inequality with (33), $\mathcal{V}_t \le \mathcal{G}_t^2$, and nonnegativity proves (34). No independence between $\mathcal{P}_{N_t}$ and ε has been used.

Step 3: Transfer the bounds to velocity. The target identity gives

$$\mathcal{P}_{N_t}(v^\star(x_t) - a_t'\tilde{x}_t(x_t)) = a_t'\rho + b_t'\mathcal{P}_{N_t}\varepsilon^\star(x_t).$$

The triangle and reverse triangle inequalities in L^2, together with $\|\rho\|_{L^2} = \sqrt{\mathcal{B}_t} \le \mathcal{G}_t$ and (34), prove (35). □

Proof of Corollary 2. By (32), $\mathcal{B}_t \le \mathcal{G}_t^2 = O(1)$. Moreover,

$$\frac{\ell_t}{D} = \left(1 - \frac{\Gamma_{\tan}}{D} - \frac{a_t^2\mathcal{G}_t^2}{b_t^2 D}\right)_+ \longrightarrow 1.$$

Divide (34) by D and apply the squeeze theorem. Likewise, dividing (35) by D and using $\mathcal{G}_t/\sqrt{D} \to 0$ proves the velocity limit, including $b_t' = 0$. □

Dimension scaling. The assumptions $\mathcal{G}_t^2 = O(1)$ and $\Gamma_{\tan} = o(D)$ are additional requirements. Dimension-independent $R_z, R_\xi, L, \Delta_{\max}$ suffice for the first. More generally, (30) permits growing $\Delta_{\max}$ provided $\Delta_{\max}^2\eta_t = O(1)$ and the other terms remain bounded. Small patch dimensions alone do not imply $\Gamma_{\tan} = o(D)$; the variation and number of tangent spaces must also be controlled. Lemma G.7 gives a sufficient condition for sublinear tangent complexity.

Shortcut consequence. The target identities imply

$$\begin{aligned}\varepsilon^\star(x_t) - \frac{x_t}{b_t} &= -\frac{a_t}{b_t}x^\star(x_t), \\ v^\star(x_t) - \frac{b_t'}{b_t}x_t &= -\frac{\Delta_t}{b_t}x^\star(x_t).\end{aligned}$$

Consequently, after the corresponding reference centering,

$$\begin{aligned}\mathbb{E}\left\|\mathcal{P}_{N_t}\left(\varepsilon^\star(x_t) - \frac{x_t}{b_t} + \frac{a_t}{b_t}\tilde{x}_t(x_t)\right)\right\|^2 &= \frac{a_t^2}{b_t^2}\mathcal{B}_t, \\ \mathbb{E}\left\|\mathcal{P}_{N_t}\left(v^\star(x_t) - \frac{b_t'}{b_t}x_t + \frac{\Delta_t}{b_t}\tilde{x}_t(x_t)\right)\right\|^2 &= \frac{\Delta_t^2}{b_t^2}\mathcal{B}_t.\end{aligned} \tag{36}$$

Both residual normal energies are controlled by $\mathcal{B}_t \le \mathcal{G}_t^2$.

D THEORETICAL ANALYSIS OF SHORTCUT PARAMETERIZATION

We first derive the residual targets for a general core, then specialize to the affine factorization used in Section 3.2.

Target identities. Conditioning $x_t = a_t x_0 + b_t \varepsilon$ on $x_t = x$ gives

$$\varepsilon^\star(x) = \frac{x}{b_t} - \frac{a_t}{b_t} x^\star(x). \tag{37}$$

Using $v^\star = a_t' x^\star + b_t' \varepsilon^\star$ yields

$$v^\star(x) = \frac{b_t'}{b_t} x - \frac{\Delta_t}{b_t} x^\star(x). \tag{38}$$

Define

$$\gamma_\varepsilon(t) = \frac{1}{b_t}, \qquad \gamma_v(t) = \frac{b_t'}{b_t}. \tag{39}$$

Then $f_t^{\star,(\tau)}(x) = \gamma_\tau(t) x + h_\tau^\star(x)$, where $h_\varepsilon^\star = -a_t x^\star / b_t$ and $h_v^\star = -\Delta_t x^\star / b_t$.

Proof of Theorem 5. For Gaussian data, the residual target is affine. Define its uncentered second moment by $H_{\tau,t}^{\mathrm{sc}}(c) = \mathbb{E}[r_c(x_t) r_c(x_t)^\top]$, where $r_c(x) = f_t^{\star,(\tau)}(x) - cx$. Write $f_t^{\star,(\tau)}(x) = W_t^{\star,(\tau)} \psi(x)$, $G_t = \mathbb{E}[\psi(x_t)\psi(x_t)^\top] \succ 0$, and $J = (I_D, 0) \in \mathbb{R}^{D \times (D+1)}$. Then

$$H_{\tau,t}^{\mathrm{sc}}(c) = (W_t^{\star,(\tau)} - cJ) G_t (W_t^{\star,(\tau)} - cJ)^\top.$$

As A and Θ vary, $A\Theta G_t^{1/2}$ ranges over all $D \times (D+1)$ matrices of rank at most m. The best rank-m approximation to $(W_t^{\star,(\tau)} - cJ) G_t^{1/2}$ therefore gives

$$\inf_{A,\Theta} \mathcal{R}_{\tau,c}^{\mathrm{sc}}(A, \Theta) = \sum_{j=m+1}^{D} \lambda_j \left(H_{\tau,t}^{\mathrm{sc}}(c) \right),$$

where eigenvalues are ordered nonincreasingly. We next prove the upper bound in (16) by an explicit construction.

For the upper bound, choose A with orthonormal columns whose span contains $\mathcal{M}$, and set $P = AA^\top$. Such a choice is possible because $m \geq d_{\mathrm{int}}$. The affine function $h_\tau^\star(x) + (\gamma_\tau(t) - c) Px$ takes values in $\mathrm{range}(A)$ and can thus be written as $A\Theta\psi(x)$. For this choice, the prediction residual is $(\gamma_\tau(t) - c)(I_D - P)x_t$. Since $(I_D - P)x_0 = 0$, its expected squared norm is

$$(\gamma_\tau(t) - c)^2 b_t^2 \mathbb{E}\|(I_D - P)\varepsilon\|^2 = b_t^2 (c - \gamma_\tau(t))^2 (D - m).$$

This gives the stated upper bound. At $c = \gamma_\tau(t)$ the residual vanishes, proving exact representation. □

Aligned cores for controlled comparisons. The construction above also gives the exact error when a rank-m matrix A is held fixed with $\mathcal{M} \subseteq \mathrm{range}(A)$. Let P be the orthogonal projector onto its range. Every core output lies in this range, so the orthogonal component of the residual is necessarily $(\gamma_\tau(t) - c)(I_D - P)x_t$. The affine core can match the remaining component exactly, giving

$$\inf_{\Theta} \mathcal{R}_{\tau,c}^{\mathrm{sc}}(A, \Theta) = b_t^2 (c - \gamma_\tau(t))^2 (D - m). \tag{40}$$

At $c = 0$, the errors are $D - m$ for noise and $(b_t')^2 (D - m)$ for velocity; the clean predictor is exactly representable because its image lies in $\mathcal{M}$. These identities describe the controlled coefficient comparison below. They are not asserted as equalities after optimizing over all A.

D.1 Exact Representation for a Two-Component Mixture

The same target identities apply to the nonlinear predictors of Section 3.3. Rescaling the output of the exact one-gate x-predictor gives the required residual core.

Corollary 3 (Exact shortcut representation for the two-component MoLRG)**.** *In the balanced two-component setting of Section 3.3 at $a_t, b_t > 0$, suppose that the output subspace of the one-gate model contains $\mathcal{S} = \mathrm{span}(\mathrm{range}(\Sigma) \cup \{\mu_1, \mu_2\})$. Augment* (12) *with a scalar shortcut sx. For $\tau \in \{\varepsilon, v\}$, choosing $s = \gamma_\tau(t)$ permits exact representation of $f_t^{\star,(\tau)}$.*

Proof of Corollary 3. The one-gate core exactly represents $x^\star$. Equations (37) and (38) express each target as $\gamma_\tau(t)x$ plus a scalar multiple of $x^\star(x)$. Rescaling the one-gate core output therefore represents the residual exactly. □

D.2 Comparison with EDM Preconditioning

The contribution of the target-matched shortcut is not the generic use of a skip connection. Residual parameterizations, EDM preconditioning (Karras et al., 2022), and our method all use the affine form

$$\widehat{y}_t(x) = c(t)x + f_\theta(x, t).$$

The distinction lies in the choice of $c(t)$. Our coefficient is derived from the exact target identities and removes the full-dimensional ambient component, leaving only the signal-dependent residual to the capacity-limited core.

For this coefficient comparison, fix a rank-m matrix A with $\mathcal{M} \subseteq \mathcal{L} = \mathrm{range}(A)$, and let the core range over all affine maps into $\mathcal{L}$. Consider an EDM x-head with nonzero output scaling

$$\widehat{x}^{\mathrm{EDM}}(x, t) = c_{\mathrm{skip}}(t)x + c_{\mathrm{out}}(t)F_\theta(x, t), \tag{41}$$

where $F_\theta(x, t)$ has output in $\mathcal{L}$. Converting (41) into the two noise-based coordinates gives

$$\widehat{\varepsilon}^{\mathrm{EDM}}(x, t) = \frac{1 - a_t c_{\mathrm{skip}}(t)}{b_t}x - \frac{a_t c_{\mathrm{out}}(t)}{b_t}F_\theta(x, t), \tag{42}$$

and

$$\widehat{v}^{\mathrm{EDM}}(x, t) = \frac{b_t' - \Delta_t c_{\mathrm{skip}}(t)}{b_t}x - \frac{\Delta_t c_{\mathrm{out}}(t)}{b_t}F_\theta(x, t). \tag{43}$$

Applying (40) to the effective shortcut coefficients gives the following exact errors for this aligned core class (with $\Delta_t \neq 0$ for velocity conversion):

$$\mathcal{A}_\varepsilon^{\mathrm{EDM}}(t) = a_t^2 c_{\mathrm{skip}}(t)^2(D - m), \qquad \mathcal{A}_v^{\mathrm{EDM}}(t) = \Delta_t^2 c_{\mathrm{skip}}(t)^2(D - m). \tag{44}$$

Thus, standard EDM preconditioning can reduce the ambient approximation error but generally does not eliminate it in the constrained output class. For example, under the additive-noise parameterization $a_t = 1$ and $b_t = \sigma_t$, EDM uses

$$c_{\mathrm{skip}}(t) = \frac{\sigma_{\mathrm{data}}^2}{\sigma_{\mathrm{data}}^2 + \sigma_t^2}.$$

At $\sigma_t = \sigma_{\mathrm{data}}$, this coefficient equals $1/2$, so (44) gives $\mathcal{A}_\varepsilon^{\mathrm{EDM}}(t) = (D - m)/4$. Thus, the converted EDM predictor retains a positive approximation error when $m < D$, whereas the target-matched coefficient $c_\varepsilon^\star(t) = 1/b_t$ eliminates it in this aligned core class. This comparison concerns representability only; EDM input scaling, loss weighting, and noise conditioning may provide separate optimization benefits.

D.3 Gaussian Verification

We verify this coefficient-level distinction in a Gaussian experiment with intrinsic rank $r = 4$ and output dimension $m = 8$. We vary $D - m$ and compare direct v-prediction, migrated EDM preconditioning, the theory-specified shortcut, and a learned scalar shortcut. Results are reported as mean $\pm$ standard deviation over five seeds.

Table 2: Gaussian v-prediction excess error under different shortcut parameterizations. EDM denotes the migrated EDM parameterization.

Ambient codimension $D-m$	Direct	Migrated EDM	Ours (theory)	Ours (learned)
8	7.993 ± 0.015	2.859 ± 0.004	0	≈ 0
24	23.996 ± 0.003	8.562 ± 0.012	0	≈ 0
56	56.035 ± 0.033	20.015 ± 0.015	0	≈ 0

Direct v-prediction exhibits the predicted linear scaling with $D-m$. Migrated EDM preconditioning reduces the error but retains a nonzero ambient-dimensional floor. Both the theory-specified and learned target-matched shortcuts instead reduce the error to numerical precision, and the learned coefficient converges to the target-identity coefficient.

Let $\mathcal{A}_{\tau,c}(t) = \inf_{A,\Theta} \mathcal{R}^{\mathrm{sc}}_{\tau,c}(A,\Theta)$. If the learned coefficient satisfies $|c(t) - \gamma_\tau(t)| \le \delta_c$, then (16) gives

$$\mathcal{A}_{\tau,c}(t) \le b_t^2 \delta_c^2 (D-m), \qquad \tau \in \{\varepsilon, v\}. \tag{45}$$

This upper bound is quadratic in the coefficient error and linear in the ambient codimension.

E Extension to the Mean Flow Framework

The geometric separation principles established in the main text are not limited to instantaneous diffusion models. They naturally extend to the recently proposed Mean Flow framework (Geng et al., 2025; Lu et al., 2026). Lu et al. (2026) empirically demonstrate that the x-prediction parameterization is more favorable than the v-prediction parameterization in Pixel Mean Flows. In this section, we provide a geometric analysis demonstrating that the x-prediction parameterization structurally resolves the dimensional penalty inherent in the Mean Flow velocity objective.

In the Mean Flow framework, the forward process interpolates linearly between the clean data and the standard normal noise.

$$x_t = (1-t)x_0 + t\varepsilon, \qquad t \in [0,1], \tag{46}$$

where $x_0 \sim p_{\mathrm{data}}$ and $\varepsilon \sim \mathcal{N}(0, I_D)$. The instantaneous velocity field is given by the conditional expectation $v^\star(x_t,t) = \mathbb{E}[\varepsilon - x_0 \mid x_t]$. However, the training target in Mean Flow is defined as the average velocity over an integration interval $[s,t]$ with $0 < s < t \le 1$.

$$u^\star(x_t,s,t) \coloneqq \frac{1}{t-s}\int_s^t v^\star(x_\tau,\tau)\,\mathrm{d}\tau. \tag{47}$$

Here x_τ is the solution of $\dot{x}_\tau = v^\star(x_\tau,\tau)$ ending at x_t, on an interval where this trajectory is defined. The x-prediction parameterization is formulated by defining the optimal x-target via an algebraic transformation of the average velocity.

$$x^\star(x_t,s,t) \coloneqq x_t - t \cdot u^\star(x_t,s,t). \tag{48}$$

To formalize the properties of these targets, we restate the general affine subspace prior within this specific interpolative setting.

Theorem 7 (Geometric analysis in Mean Flow)**.** *Let the data distribution reside in an r-dimensional affine subspace $\mathcal{S}$ defined by $x_0 = \mu + Uz$, as formulated in Assumption 1. Let $\mathcal{P}_{\mathcal{S}^\perp} \coloneqq I_D - UU^\top$ denote the orthogonal projection onto the ambient noise subspace. For any interval $[s,t]$ with $0 < s < t \le 1$ on which the trajectory is defined, the average velocity target $u^\star$ retains an ambient component, whereas the data-centric target $x^\star$ vanishes in the orthogonal noise subspace. Specifically,*

$$\mathcal{P}_{\mathcal{S}^\perp} u^\star(x_t,s,t) = \frac{1}{t}\mathcal{P}_{\mathcal{S}^\perp}(x_t - \mu), \qquad \mathcal{P}_{\mathcal{S}^\perp}\big(x^\star(x_t,s,t) - \mu\big) = 0.$$

Proof of Theorem 7. We analyze the dynamics strictly within the orthogonal noise subspace $\mathcal{S}^\perp$ by breaking the proof into three geometric steps.

1. Independence and Instantaneous Velocity. By the definition of the affine subspace, we have $\mathcal{P}_{\mathcal{S}^\perp}(x_0 - \mu) = 0$, which implies $\mathcal{P}_{\mathcal{S}^\perp} x_0 = \mathcal{P}_{\mathcal{S}^\perp}\mu$. Projecting the forward process (46) onto $\mathcal{S}^\perp$ isolates the ambient noise.

$$\mathcal{P}_{\mathcal{S}^\perp} x_t = (1-t)\mathcal{P}_{\mathcal{S}^\perp}\mu + t\mathcal{P}_{\mathcal{S}^\perp}\varepsilon \implies \mathcal{P}_{\mathcal{S}^\perp}\varepsilon = \frac{1}{t}\mathcal{P}_{\mathcal{S}^\perp}\big(x_t - (1-t)\mu\big).$$

Because ε is standard isotropic Gaussian noise, its orthogonal projection is independent of the latent coordinate z. Consequently, the conditional expectation of the instantaneous velocity field decouples in the noise subspace.

$$\mathcal{P}_{\mathcal{S}^\perp} v^\star(x_t, t) = \mathbb{E}[\mathcal{P}_{\mathcal{S}^\perp}\varepsilon - \mathcal{P}_{\mathcal{S}^\perp} x_0 \mid x_t] = \frac{1}{t}\mathcal{P}_{\mathcal{S}^\perp}\big(x_t - (1-t)\mu\big) - \mathcal{P}_{\mathcal{S}^\perp}\mu = \frac{1}{t}\mathcal{P}_{\mathcal{S}^\perp}(x_t - \mu).$$

2. Linear Noise Trajectory. The Probability Flow ODE in the noise subspace is governed by the projected instantaneous velocity.

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{P}_{\mathcal{S}^\perp}(x_\tau - \mu) = \mathcal{P}_{\mathcal{S}^\perp} v^\star(x_\tau, \tau) = \frac{1}{\tau}\mathcal{P}_{\mathcal{S}^\perp}(x_\tau - \mu).$$

Solving this fundamental linear differential equation over the integration interval $[s, t]$ yields a deterministic linear scaling trajectory for the noise components.

$$\mathcal{P}_{\mathcal{S}^\perp}(x_\tau - \mu) = \frac{\tau}{t}\mathcal{P}_{\mathcal{S}^\perp}(x_t - \mu).$$

3. Average Velocity and Target Cancellation. We now evaluate the average velocity target $u^\star$ along this established linear trajectory.

$$\mathcal{P}_{\mathcal{S}^\perp} u^\star(x_t, s, t) = \frac{1}{t-s}\int_s^t \frac{1}{\tau}\mathcal{P}_{\mathcal{S}^\perp}(x_\tau - \mu)\,\mathrm{d}\tau = \frac{1}{t-s}\int_s^t \frac{1}{\tau}\Big(\frac{\tau}{t}\mathcal{P}_{\mathcal{S}^\perp}(x_t - \mu)\Big)\,\mathrm{d}\tau.$$

Canceling τ in the integrand and integrating the constant yields the exact expression for the velocity target's noise leakage.

$$\mathcal{P}_{\mathcal{S}^\perp} u^\star(x_t, s, t) = \frac{1}{t}\mathcal{P}_{\mathcal{S}^\perp}(x_t - \mu).$$

Finally, applying the transformation defined in (48), the data-centric target naturally and algebraically eliminates this high-dimensional ambient noise component.

$$\begin{aligned}\mathcal{P}_{\mathcal{S}^\perp}\big(x^\star(x_t, s, t) - \mu\big) &= \mathcal{P}_{\mathcal{S}^\perp}(x_t - \mu) - t \cdot \mathcal{P}_{\mathcal{S}^\perp} u^\star(x_t, s, t) \\ &= \mathcal{P}_{\mathcal{S}^\perp}(x_t - \mu) - t\Big(\frac{1}{t}\mathcal{P}_{\mathcal{S}^\perp}(x_t - \mu)\Big) = 0.\end{aligned}$$

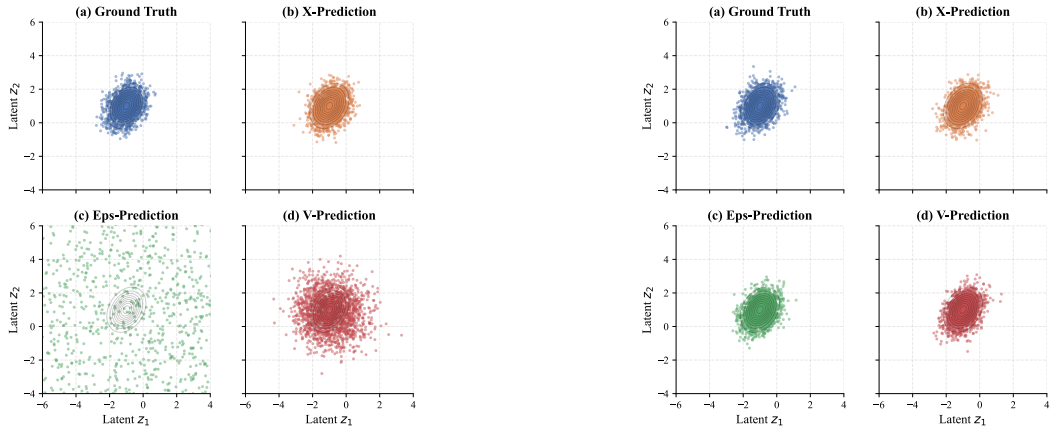
This verifies the stated difference between the projected x- and u-targets under the affine-subspace model. □

F Additional Experimental Material

This appendix supplements the main Gaussian, Swiss-roll, and ImageNet experiments with Gaussian and Gaussian-mixture sampling comparisons. Detailed configurations for the main experiments are collected in Appendix I.

F.1 Gaussian Sampling under Output Constraints

We compare samples from a two-dimensional Gaussian embedded in $\mathbb{R}^{64}$ under two output dimensions (Figure 3). The samples are consistent with the target-dependent approximation errors in the main text; this comparison is qualitative.



(a) $m = 4$ (b) $m = 32$

Figure 3: Generated samples for a 2-D Gaussian embedded into $\mathbb{R}^{64}$ under two output dimensions. At $m = 4$, the noise-based targets show larger off-subspace distortion than x-prediction; the gap narrows as the output dimension increases.

F.2 Gaussian-Mixture Shortcut Test

To examine the remedy with nonlinear Bayes targets, we embed a two-component Gaussian mixture from $\mathbb{R}^2$ into $\mathbb{R}^{16}$ and use the one-gate model (12) with output dimension $m = 8$. Its fixed output range contains the two-dimensional signal span but does not span the ambient space. Corollary 3 establishes exact representability at the ideal shortcut coefficient; this experiment examines a coefficient learned in practice. Figure 4 compares samples with and without the scalar time-dependent shortcut in (13). The shortcut visibly improves the two-component structure for both ε- and v-prediction, providing qualitative evidence for the remedy with nonlinear Bayes targets. Appendix I.3 gives the data, training, and sampling settings.

Existing assets and licenses. Our ImageNet experiments use the ImageNet dataset (Deng et al., 2009) under its official terms of access and use. We use the JiT-B/16 architecture following Li & He (2025). All existing datasets, architectures, and prior methods used in this work are properly cited, and we use them only for research purposes in accordance with their associated licenses and terms of use.

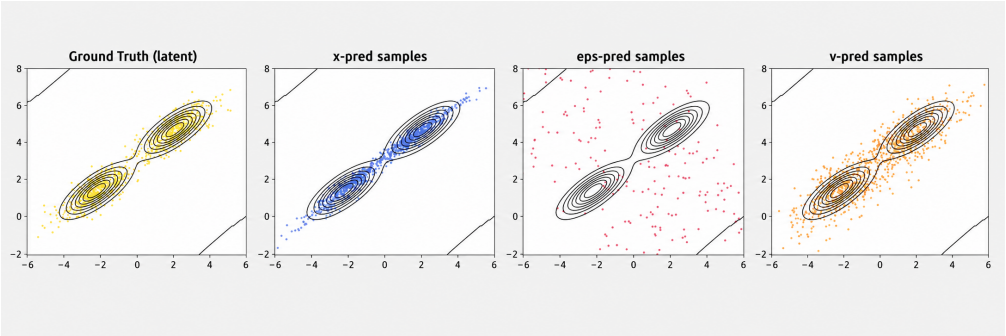
G Useful Lemmas

Lemma G.1 (MSE target $\Leftrightarrow$ Bayes-optimal target)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let X be a random variable taking values in a measurable space $\mathcal{X}$, and let $Y \in \mathbb{R}^d$ be square-integrable: $\mathbb{E}\|Y\|^2 < \infty$. For any measurable $f : \mathcal{X} \to \mathbb{R}^d$ with $\mathbb{E}\|f(X)\|^2 < \infty$, define the MSE risk*

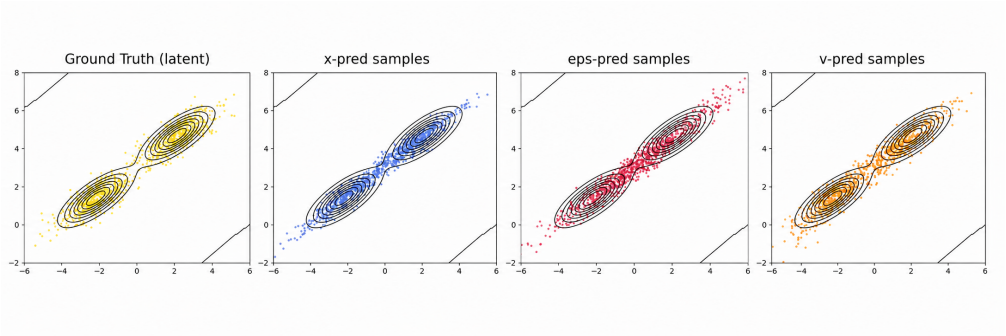
$$\mathcal{L}(f) := \mathbb{E}\big[\|f(X) - Y\|^2\big].$$

Then:

1. *The Bayes-optimal predictor is the conditional mean*
$$f^\star(X) = \mathbb{E}[Y \mid X] \quad \text{a.s.}$$
2. *(Orthogonal decomposition) For any such f,*
$$\mathcal{L}(f) = \mathbb{E}\big[\|f(X) - \mathbb{E}[Y \mid X]\|^2\big] + \mathbb{E}\big[\|Y - \mathbb{E}[Y \mid X]\|^2\big].$$
In particular, $\mathcal{L}(f)$ is minimized iff $f(X) = \mathbb{E}[Y \mid X]$ almost surely, and the excess risk equals
$$\mathcal{L}(f) - \inf_g \mathcal{L}(g) = \mathbb{E}\big[\|f(X) - \mathbb{E}[Y \mid X]\|^2\big].$$



(a) Without shortcut



(b) With shortcut

Figure 4: Gaussian-mixture samples in the two-dimensional latent coordinates. Each row shows the ground truth and samples from x-, ε-, and v-prediction, respectively. The shortcut improves the visual agreement of ε- and v-prediction with the target mixture.

Proof. Let $m(X) \coloneqq \mathbb{E}[Y \mid X]$ (well-defined in L^2 by assumption). Expand the square:

$$\|f(X) - Y\|^2 = \|f(X) - m(X)\|^2 + \|m(X) - Y\|^2 + 2\langle f(X) - m(X),\, m(X) - Y\rangle.$$

Take expectations under the joint distribution of (X, Y). The first two terms remain unchanged. For the cross term, we apply the tower property of conditional expectation:

$$\mathbb{E}\big[\langle f(X) - m(X),\, m(X) - Y\rangle\big] = \mathbb{E}_X\Big[\mathbb{E}\big[\langle f(X) - m(X),\, m(X) - Y\rangle \mid X\big]\Big].$$

Conditioning on X, the vector $f(X) - m(X)$ is X-measurable and can therefore be taken outside the conditional expectation:

$$\mathbb{E}\big[\langle f(X) - m(X),\, m(X) - Y\rangle \mid X\big] = \langle f(X) - m(X),\, \mathbb{E}[m(X) - Y \mid X]\rangle\,.$$

By the definition of the conditional mean $m(X) = \mathbb{E}[Y \mid X]$, we have

$$\mathbb{E}[m(X) - Y \mid X] = 0.$$

Hence the cross term vanishes, and we obtain the decomposition

$$\mathcal{L}(f) = \mathbb{E}\|f(X) - m(X)\|^2 + \mathbb{E}\|Y - m(X)\|^2.$$

The second term is independent of f, so the minimum of $\mathcal{L}(f)$ is achieved by $f(X) = m(X)$ almost surely, and the excess risk equals the first term.

□

Lemma G.2 (Score identity)**.** *Let $X_t = a_t X_0 + b_t \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I_D)$ independent of X_0 and $b_t > 0$. Denote p_t as the density of X_t and define the Bayes-optimal targets $x_t^\star(x) \coloneqq \mathbb{E}[X_0 \mid X_t = x]$ and $\varepsilon_t^\star(x) \coloneqq \mathbb{E}[\varepsilon \mid X_t = x]$. Then for all $x \in \mathbb{R}^D$,*

$$\nabla \log p_t(x) = \frac{1}{b_t^2}\big(a_t x_t^\star(x) - x\big) = -\frac{1}{b_t}\varepsilon_t^\star(x).$$

Moreover, define the interpolation velocity with X_0 and ε held fixed

$$v_t \coloneqq a_t' X_0 + b_t' \varepsilon, \qquad v_t^\star(x) \coloneqq \mathbb{E}[v_t \mid X_t = x].$$

Then

$$v_t^\star(x) = a_t' x_t^\star(x) + b_t' \varepsilon_t^\star(x),$$

and if

$$\Delta_t \coloneqq a_t b_t' - b_t a_t' \neq 0,$$

we have the equivalent representations

$$\varepsilon_t^\star(x) = \frac{a_t v_t^\star(x) - a_t' x}{\Delta_t}, \qquad \nabla \log p_t(x) = -\frac{1}{b_t} \cdot \frac{a_t v_t^\star(x) - a_t' x}{\Delta_t}.$$

Consequently, the score estimator induced by a v-predictor $v_\theta(\cdot, t)$ is

$$\hat{s}_t^{(v)}(x) = -\frac{1}{b_t} \cdot \frac{a_t v_\theta(x,t) - a_t' x}{\Delta_t},$$

with score error

$$\hat{s}_t^{(v)}(X_t) - \nabla \log p_t(X_t) = -\frac{a_t}{b_t\, \Delta_t} \Big(v_\theta(X_t, t) - v_t^\star(X_t) \Big).$$

Proof. Write the marginal as a convolution with the probability measure p_0; this also covers a singular clean-data distribution:

$$p_t(x) = \int \phi\!\left(\frac{x - a_t x_0}{b_t} \right) \frac{1}{b_t^D}\, p_0(dx_0),$$

where $\phi(u) = (2\pi)^{-D/2} \exp(-\|u\|^2/2)$ is the standard Gaussian density. Differentiate under the integral sign:

$$\begin{aligned}
\nabla_x p_t(x) &= \int \frac{1}{b_t^D}\, \nabla_x \phi\!\left(\frac{x - a_t x_0}{b_t} \right) p_0(dx_0) \\
&= \int \frac{1}{b_t^D}\, \frac{1}{b_t}\, \nabla_u \phi(u) \Big|_{u=(x - a_t x_0)/b_t} p_0(dx_0) \\
&= \int \frac{1}{b_t^D}\, \frac{1}{b_t}\, (-u)\, \phi(u) \Big|_{u=(x - a_t x_0)/b_t} p_0(dx_0) \\
&= -\frac{1}{b_t^2} \int (x - a_t x_0)\, \phi\!\left(\frac{x - a_t x_0}{b_t} \right) \frac{1}{b_t^D}\, p_0(dx_0).
\end{aligned}$$

Divide by $p_t(x)$ to obtain

$$\nabla \log p_t(x) = -\frac{1}{b_t^2}\, \mathbb{E}\big[x - a_t X_0 \mid X_t = x \big] = \frac{1}{b_t^2} \big(a_t\, x_t^\star(x) - x \big),$$

which proves the first identity.

For the second identity, note that $\varepsilon = (X_t - a_t X_0)/b_t$, hence conditioning on $X_t = x$ gives

$$\varepsilon_t^\star(x) = \mathbb{E}[\varepsilon \mid X_t = x] = \frac{1}{b_t} \mathbb{E}[x - a_t X_0 \mid X_t = x] = \frac{x - a_t x_t^\star(x)}{b_t},$$

and thus $\nabla \log p_t(x) = -(1/b_t) \varepsilon_t^\star(x)$.

For the v-statements, linearity of conditional expectation yields

$$v_t^\star(x) = \mathbb{E}[a_t' X_0 + b_t' \varepsilon \mid X_t = x] = a_t' \mathbb{E}[X_0 \mid X_t = x] + b_t' \mathbb{E}[\varepsilon \mid X_t = x] = a_t' x_t^\star(x) + b_t' \varepsilon_t^\star(x).$$

Next, the pair of linear relations

$$x = a_t X_0 + b_t \varepsilon, \qquad v_t = a_t' X_0 + b_t' \varepsilon$$

implies

$$\varepsilon = \frac{a_t v_t - a_t' x}{\Delta_t}, \qquad \Delta_t \coloneqq a_t b_t' - b_t a_t'.$$

Taking conditional expectation given $X_t = x$ gives

$$\varepsilon_t^\star(x) = \mathbb{E}[\varepsilon \mid X_t = x] = \frac{a_t v_t^\star(x) - a_t' x}{\Delta_t}.$$

Substituting this into $\nabla \log p_t(x) = -(1/b_t)\varepsilon_t^\star(x)$ yields

$$\nabla \log p_t(x) = -\frac{1}{b_t} \cdot \frac{a_t v_t^\star(x) - a_t' x}{\Delta_t}.$$

Finally, replacing $v_t^\star$ by v_θ in the above identity defines $\hat{s}_t^{(v)}$, and the score error follows immediately. □

Lemma G.3 (Bayes-Optimal Predictors under Gaussian Prior)**.** *Under Assumption 2, the Bayes-optimal predictors $y_\tau^\star(x_t) = \mathbb{E}[y_\tau \mid x_t]$ for the x-, ε-, and v-prediction targets are affine functions of x_t and take the following forms:*

$$x^\star(x_t) = \mu + a_t\Sigma(a_t^2\Sigma + b_t^2 I)^{-1}(x_t - a_t\mu), \tag{49}$$

$$\varepsilon^\star(x_t) = b_t(a_t^2\Sigma + b_t^2 I)^{-1}(x_t - a_t\mu), \tag{50}$$

$$v^\star(x_t) = a_t'\mu + (a_t' a_t\Sigma + b_t' b_t I)(a_t^2\Sigma + b_t^2 I)^{-1}(x_t - a_t\mu). \tag{51}$$

Proof. Given $x_0 \sim \mathcal{N}(\mu, \Sigma)$ and the forward process $x_t = a_t x_0 + b_t\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I_D)$, the joint distribution of (x_0, x_t) is Gaussian:

$$\begin{bmatrix} x_0 \\ x_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ a_t\mu \end{bmatrix}, \begin{bmatrix} \Sigma & a_t\Sigma \\ a_t\Sigma & a_t^2\Sigma + b_t^2 I \end{bmatrix}\right).$$

Using the standard conditional expectation formula for multivariate Gaussians, $\mathbb{E}[x_0 \mid x_t] = \mu_{x_0} + \Sigma_{x_0 x_t}\Sigma_{x_t x_t}^{-1}(x_t - \mu_{x_t})$, we substitute the block covariance matrices to obtain (49):

$$x^\star(x_t) = \mu + a_t\Sigma(a_t^2\Sigma + b_t^2 I)^{-1}(x_t - a_t\mu).$$

For ε-prediction, we observe from the forward process that $\varepsilon = \frac{1}{b_t}(x_t - a_t x_0)$. By the linearity of expectation:

$$\varepsilon^\star(x_t) = \mathbb{E}\left[\frac{1}{b_t}(x_t - a_t x_0) \,\middle|\, x_t\right] = \frac{1}{b_t}\left(x_t - a_t x^\star(x_t)\right).$$

Substituting (49) into the above and applying the identity $I - a_t^2\Sigma(a_t^2\Sigma + b_t^2 I)^{-1} = b_t^2(a_t^2\Sigma + b_t^2 I)^{-1}$, we arrive at (50):

$$\varepsilon^\star(x_t) = \frac{1}{b_t}\left[(x_t - a_t\mu) - a_t^2\Sigma(a_t^2\Sigma + b_t^2 I)^{-1}(x_t - a_t\mu)\right] = b_t(a_t^2\Sigma + b_t^2 I)^{-1}(x_t - a_t\mu).$$

Finally, for v-prediction defined as $v = a_t' x_0 + b_t'\varepsilon$, we have:

$$v^\star(x_t) = a_t'\mathbb{E}[x_0 \mid x_t] + b_t'\mathbb{E}[\varepsilon \mid x_t] = a_t' x^\star(x_t) + b_t'\varepsilon^\star(x_t). \tag{52}$$

Substituting (49) and (50) and grouping the terms associated with $(x_t - a_t\mu)$ yields (51):

$$v^\star(x_t) = a_t'\mu + (a_t' a_t\Sigma + b_t' b_t I)\,(a_t^2\Sigma + b_t^2 I)^{-1}(x_t - a_t\mu).$$

This completes the proof. □

Lemma G.4 (Multi-component GMM Score)**.** *Let $p_t(x) = \sum_{k=1}^M \pi_k\mathcal{N}(x; a_t\mu_k, \Sigma_{k,t})$ be a Gaussian mixture at time t. Define the global mean $\bar{\mu}_t := a_t\sum_{k=1}^M \pi_k\mu_k$ and a reference covariance $\bar{\Sigma}_t$. The score function $\nabla \log p_t(x)$ admits the following decomposition:*

$$\nabla \log p_t(x) = -\bar{\Sigma}_t^{-1}(x - \bar{\mu}_t) + \Psi_t(x), \tag{53}$$

where $\Psi_t(x) := \sum_{k=1}^M w_k(x)\left[\bar{\Sigma}_t^{-1}(x - \bar{\mu}_t) - \Sigma_{k,t}^{-1}(x - a_t\mu_k)\right]$ represents the nonlinear structural correction.

Proof. The score function of a Gaussian mixture $p_t(x) = \sum_{k=1}^{M} \pi_k \phi_k(x)$, where $\phi_k(x) = \mathcal{N}(x; a_t\mu_k, \Sigma_{k,t})$, is given by:

$$\nabla \log p_t(x) = \frac{\sum_{k=1}^{M} \pi_k \nabla \phi_k(x)}{p_t(x)} = \sum_{k=1}^{M} \frac{\pi_k \phi_k(x)}{p_t(x)} \frac{\nabla \phi_k(x)}{\phi_k(x)}. \tag{54}$$

By defining the responsibilities $w_k(x) \coloneqq \frac{\pi_k \phi_k(x)}{p_t(x)}$ and using the score of a single Gaussian $\nabla \log \phi_k(x) = -\Sigma_{k,t}^{-1}(x - a_t\mu_k)$, we have:

$$\nabla \log p_t(x) = \sum_{k=1}^{M} w_k(x) \left[-\Sigma_{k,t}^{-1}(x - a_t\mu_k) \right]. \tag{55}$$

To introduce the reference term $-\bar{\Sigma}_t^{-1}(x - \bar{\mu}_t)$, we observe that since $\sum_{k=1}^{M} w_k(x) = 1$, we can identically write:

$$-\bar{\Sigma}_t^{-1}(x - \bar{\mu}_t) = \sum_{k=1}^{M} w_k(x) \left[-\bar{\Sigma}_t^{-1}(x - \bar{\mu}_t) \right]. \tag{56}$$

Subtracting and adding this term to (55) yields:

$$\begin{aligned} \nabla \log p_t(x) &= -\bar{\Sigma}_t^{-1}(x - \bar{\mu}_t) + \sum_{k=1}^{M} w_k(x) \bar{\Sigma}_t^{-1}(x - \bar{\mu}_t) - \sum_{k=1}^{M} w_k(x) \Sigma_{k,t}^{-1}(x - a_t\mu_k) \\ &= -\bar{\Sigma}_t^{-1}(x - \bar{\mu}_t) + \sum_{k=1}^{M} w_k(x) \left[\bar{\Sigma}_t^{-1}(x - \bar{\mu}_t) - \Sigma_{k,t}^{-1}(x - a_t\mu_k) \right]. \end{aligned}$$

By defining the summation term as $\Psi_t(x)$, we arrive at the final form:

$$\nabla \log p_t(x) = -\bar{\Sigma}_t^{-1}(x - \bar{\mu}_t) + \Psi_t(x).$$

This completes the proof. □

Lemma G.5. *Let $\Sigma \succeq 0$ with $\operatorname{rank}(\Sigma) = r$, and define $\Sigma_t \coloneqq a_t^2\Sigma + b_t^2 I_D$ with $b_t > 0$. Consider the two-component mixture setting.*

Bayes-optimal x-prediction target at time t is given by

$$x_t^\star(x) = a_t \Sigma \Sigma_t^{-1}(x_t - \bar{\mu}_t) + \bar{\mu}_0 + \frac{b_t^2}{a_t} \Sigma_t^{-1} d_t \tanh\Big(\big\langle \Sigma_t^{-1} d_t,\ x - \bar{\mu}_t \big\rangle \Big). \tag{57}$$

The Bayes ε- and v-prediction targets admit the forms:

$$\begin{aligned} \varepsilon_t^\star(x) &= b_t \Sigma_t^{-1}(x_t - \bar{\mu}_t) - b_t\, \Sigma_t^{-1} d_t \tanh\Big(\big\langle \Sigma_t^{-1} d_t,\ x - \bar{\mu}_t \big\rangle \Big), \\ v_t^\star(x) &= \bar{v}_t + (a_t' a_t \Sigma + b_t' b_t I)\, \Sigma_t^{-1}(x_t - \bar{\mu}_t) \\ &\quad + \frac{a_t' b_t^2 - b_t' b_t a_t}{a_t} \Sigma_t^{-1} d_t \tanh\Big(\big\langle \Sigma_t^{-1} d_t,\ x - \bar{\mu}_t \big\rangle \Big), \end{aligned}$$

where $\bar{\mu}_0 = \frac{\mu_1 + \mu_2}{2}$ and $\bar{v}_t = a_t' \bar{\mu}_0$.

Proof. Applying Lemma G.6 and the score identity (7), we have

$$x_t^\star(x) = \frac{1}{a_t} \Big(I_D - b_t^2 \Sigma_t^{-1} \Big)(x - \bar{\mu}_t) + \bar{\mu}_0 + \frac{b_t^2}{a_t} \Sigma_t^{-1} d_t \tanh\Big(\big\langle \Sigma_t^{-1} d_t,\ x - \bar{\mu}_t \big\rangle \Big).$$

Next, we will simplify the first term. Note that

$$\begin{aligned} \Sigma_t = a_t^2 \Sigma_0 + b_t^2 I_D \implies I_d - b_t^2 \Sigma_t^{-1} &= a_t^2 \Sigma_0 \Sigma_t^{-1}, \\ \implies \frac{1}{a_t} \Big(I_d - b_t^2 \Sigma_t^{-1} \Big) &= a_t \Sigma_0 \Sigma_t^{-1}. \end{aligned}$$

The expressions for the $\varepsilon-, v-$prediction targets follow similarly by substituting the score function into (7).

Then by the definition $\psi_t(x_t) \coloneqq \Sigma_t^{-1} d_t \tanh(\langle \Sigma_t^{-1} d_t, x - \bar{\mu}_t \rangle)$ we complete the proof. □

Lemma G.6 (Score Function of Two-Component GMM)**.** *Let the data distribution be* $p_0 = \frac{1}{2}\mathcal{N}(\mu_1, \Sigma) + \frac{1}{2}\mathcal{N}(\mu_2, \Sigma)$*. Under the Gaussian diffusion process* $x_t = a_t x_0 + b_t \varepsilon$ *where* $\varepsilon \sim \mathcal{N}(0, I_D)$*, the score function* $\nabla \log p_t(x)$ *at time* t *is given by:*

$$\nabla \log p_t(x) = -\Sigma_t^{-1}(x - \bar{\mu}_t) + \Sigma_t^{-1} d_t \tanh\left(\langle \Sigma_t^{-1} d_t, x - \bar{\mu}_t \rangle\right), \tag{58}$$

where $\Sigma_t = a_t^2 \Sigma + b_t^2 I_D$ *is the time-dependent covariance,* $\bar{\mu}_t = a_t \frac{\mu_1 + \mu_2}{2}$ *is the time-dependent mean, and* $d_t = a_t \frac{\mu_1 - \mu_2}{2}$ *is the time-dependent half-difference vector.*

Proof. By the properties of Gaussian convolution, the distribution at time t is a mixture of two Gaussians:

$$p_t(x) = \frac{1}{2}\mathcal{N}(x; a_t\mu_1, \Sigma_t) + \frac{1}{2}\mathcal{N}(x; a_t\mu_2, \Sigma_t).$$

Let $\phi_1(x)$ and $\phi_2(x)$ denote the densities of the two components. Using the identity $\nabla \mathcal{N}(x; \mu, \Sigma) = \mathcal{N}(x; \mu, \Sigma)[-\Sigma^{-1}(x - \mu)]$, the gradient of the mixture density is:

$$\begin{aligned}\nabla p_t(x) &= \frac{1}{2}\nabla\phi_1(x) + \frac{1}{2}\nabla\phi_2(x) \\ &= -\frac{1}{2}\phi_1(x)\Sigma_t^{-1}(x - a_t\mu_1) - \frac{1}{2}\phi_2(x)\Sigma_t^{-1}(x - a_t\mu_2).\end{aligned}$$

Applying the score definition $\nabla \log p_t(x) = \frac{\nabla p_t(x)}{p_t(x)}$ and substituting $a_t\mu_1 = \bar{\mu}_t + d_t$ and $a_t\mu_2 = \bar{\mu}_t - d_t$, we obtain:

$$\begin{aligned}\nabla \log p_t(x) &= \frac{\phi_1(x)[-\Sigma_t^{-1}(x - \bar{\mu}_t - d_t)] + \phi_2(x)[-\Sigma_t^{-1}(x - \bar{\mu}_t + d_t)]}{\phi_1(x) + \phi_2(x)} \\ &= -\Sigma_t^{-1}(x - \bar{\mu}_t) + \Sigma_t^{-1} d_t \left(\frac{\phi_1(x) - \phi_2(x)}{\phi_1(x) + \phi_2(x)}\right).\end{aligned}$$

To simplify the term in parentheses, we calculate the log-density ratio:

$$\begin{aligned}\log \frac{\phi_1(x)}{\phi_2(x)} &= -\frac{1}{2}\|x - \bar{\mu}_t - d_t\|^2_{\Sigma_t^{-1}} + \frac{1}{2}\|x - \bar{\mu}_t + d_t\|^2_{\Sigma_t^{-1}} \\ &= \langle \Sigma_t^{-1}(x - \bar{\mu}_t), d_t \rangle + \langle \Sigma_t^{-1} d_t, x - \bar{\mu}_t \rangle \\ &= 2\langle \Sigma_t^{-1} d_t, x - \bar{\mu}_t \rangle.\end{aligned}$$

Let $z = \langle \Sigma_t^{-1} d_t, x - \bar{\mu}_t \rangle$. Then $\frac{\phi_1}{\phi_2} = e^{2z}$. The mixture weight term becomes:

$$\frac{\phi_1(x) - \phi_2(x)}{\phi_1(x) + \phi_2(x)} = \frac{e^{2z} - 1}{e^{2z} + 1} = \tanh(z).$$

Substituting this back into the expression for $\nabla \log p_t(x)$ yields the desired result. □

Lemma G.7 (Sufficient metric-entropy condition for sublinear tangent complexity)**.** *Let* $N(\mathcal{A}, \|\cdot\|_2, \eta)$ *denote the minimal number of* η*-balls in the Euclidean metric required to cover the global unit tangent bundle* $\mathcal{A}$*. If the Dudley entropy integral satisfies the sublinear scaling bound*

$$\int_0^2 \sqrt{\log N(\mathcal{A}, \|\cdot\|_2, \eta)}\, d\eta = o(\sqrt{D}),$$

then the Gaussian tangent complexity is strictly bounded by $\Gamma_{\tan} = o(D)$*.*

Proof of Lemma G.7. We analyze the supremum of the canonical Gaussian process $Z_u = \langle \varepsilon, u \rangle$ indexed by the spatial vectors $u \in \mathcal{A}$. By Dudley's chaining inequality (Dudley, 1967), the expected supremum of this process is bounded as

$$\mathbb{E}\Big[\sup_{u \in \mathcal{A}} \langle \varepsilon, u \rangle\Big] \le C \int_0^2 \sqrt{\log N(\mathcal{A}, \|\cdot\|_2, \eta)}\, d\eta = o(\sqrt{D}).$$

Here, $C > 0$ is a universal constant independent of D. Because the tangent bundle $\mathcal{A}$ is strictly contained within the unit sphere $\mathbb{S}^{D-1}$, the mapping $\varepsilon \mapsto \sup_{u \in \mathcal{A}} |\langle \varepsilon, u \rangle|$ is globally 1-Lipschitz.

By the Gaussian Poincaré inequality, the variance of any 1-Lipschitz function of a standard Gaussian vector is at most 1. Applying the standard variance decomposition $\mathbb{E}[X^2] = (\mathbb{E}[X])^2 + \mathrm{Var}(X)$ establishes the final squared bound.

$$\Gamma_{\tan} = \mathbb{E}\Big[\sup_{u\in\mathcal{A}} |\langle \varepsilon, u\rangle|^2\Big] \le \Big(o(\sqrt{D})\Big)^2 + 1 = o(D).$$

This completes the proof. □

H Forward Schedules and Score Matching

For the scalar linear drift $f(x,t) = \mathbf{f}(t)x$ in (1), the transition coefficients are

$$a_t = \exp\!\left(\int_0^t \mathbf{f}(s)\,\mathrm{d}s\right), \qquad b_t = a_t\left(\int_0^t \frac{g(s)^2}{a_s^2}\,\mathrm{d}s\right)^{1/2}.$$

These formulas follow by multiplying the SDE by a_t^{-1} and using the variance of the resulting Gaussian stochastic integral.

Variance-Preserving (VP). Here,

$$f(x_t,t) = -\frac{1}{2}\beta_t x_t, \qquad g(t) = \sqrt{\beta_t},$$

which gives

$$a_t = \exp\!\left(-\frac{1}{2}\int_0^t \beta_s\,\mathrm{d}s\right), \quad b_t = \sqrt{1-a_t^2}, \quad a_t^2 + b_t^2 = 1.$$

Variance-Exploding (VE). For a nondecreasing variance schedule σ_t^2

$$f(x_t,t) = 0, \quad g(t) = \sqrt{\frac{\mathrm{d}}{\mathrm{d}t}\sigma_t^2}, \qquad a_t = 1, \qquad b_t = \sqrt{\sigma_t^2 - \sigma_0^2}.$$

Under the convention $\sigma_0 = 0$, this reduces to $b_t = \sigma_t$. The signal remains unscaled while the perturbation variance increases.

The reverse SDE in (3), initialized from p_T, recovers the forward marginals when the true score is used and the time reversal is well posed. In practice, one estimates the score through the weighted MSE

$$\mathbb{E}_{t,x_0,\varepsilon} w_t\Big[\big\|s_\theta(x_t,t) - \nabla \log p_t(x_t)\big\|^2\Big], \tag{59}$$

where s_θ is the score model and w_t is a nonnegative time weight. Equation (7) converts each prediction parameterization into a score estimator. Its error is the corresponding prediction error relative to the Bayes target, scaled by $\lambda_t^{(\tau)}$; the irreducible Bayes risk in (6) must be separated when comparing the raw training MSEs of different targets.

I Experimental Settings

I.1 Gaussian approximation

We use a zero-mean Gaussian with signal dimension 8 and covariance $2I_8$ in the signal subspace, with $a_t = e^{-t}$ and $b_t = \sqrt{1-e^{-2t}}$. Both factors of $A_\tau \Theta_\tau \psi(x_t)$ are trained, so the output range is learned. We vary D at $m = 8$ and m at $D = 64$; cases with $m < 8$ probe ranks below the condition in Theorem 2. Excess risk is evaluated against the Bayes predictor on 20,000 samples per seed, reporting the mean and standard deviation over 10 seeds (Figure 1).

I.2 Swiss roll

Following Li & He (2025, Figure 2), we embed a planar spiral into $D \in \{32, 512\}$ dimensions using a random orthonormal projection. The core has five ReLU hidden layers of width 256, and a small

MLP parameterizes the scalar shortcut. We use the linear path $x_t = tx_0 + (1-t)\varepsilon$ and velocity MSE for all methods, converting x-predictions to velocity with denominators clamped below at 0.05. Methods share data, core initialization, and training random inputs within each seed, and sampling uses shared initial noise. Figures 2a and 2b show samples from seed 0, projected onto the signal plane. The projection is used only for visualization and is not provided to the models.

I.3 Gaussian-mixture sample comparison

We embed a balanced two-component Gaussian mixture with shared anisotropic covariance from two dimensions into $\mathbb{R}^{16}$. The fixed rank-8 output basis contains the signal span. For each target and discrete noise level, a one-gate core is trained with or without a learned scalar shortcut, using the path $x_t = tx_0 + (1-t)\varepsilon$. Methods share training batches and use sample-target MSE. Figure 4 reports seed 0.

I.4 Supplementary Gaussian checks

Appendix D.3 compares prescribed and fitted shortcut coefficients through the residual energy $\|P_{\mathcal{L}^\perp}(y_\tau^\star(X_t) - cX_t)\|^2$. We use $r = 4$, $m = 8$, and $D \in \{16, 32, 64\}$ over five seeds; Table 2 reports the results.

I.5 ImageNet

We compare x-prediction, direct v-prediction, and shortcut-assisted v-prediction using JiT-B/16 on pixel-space ImageNet 256×256 for 200 epochs. Training and sampling follow Li & He (2025), with $c_\phi(t)x_t$ added for the shortcut variant. Both the hidden width and the per-patch output dimension $16^2 \times 3$ equal 768, so the output-head dimensions do not impose a rank bottleneck. As in the Swiss-roll experiment at $D = 32$ (Figure 2b), adding the shortcut improves v-prediction over direct prediction (Table 1).