ONE STEP, PROVABLY: DISTRIBUTIONAL GUARANTEES FOR SCORE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models generate high-quality samples but require many sequential network evaluations. Score distillation enables one-step generation from a pretrained teacher diffusion model, yet theoretical guarantees for its distributional accuracy remain limited. We analyze one-step score distillation using an equivalent reverse-time diffusion whose marginals match the student's forward-noised distributions. Applying this framework to Score Identity Distillation (SiD), we establish an explicit population-level total-variation bound between the forward-noised student and data distributions at a positive noise level corresponding to early stopping in the reverse-time analysis. The bound accounts for the attained SiD objective, teacher-score and score-surrogate errors, and terminal distribution mismatches, with no sampling discretization term. For bounded generators, we further construct a single ReLU score-surrogate network with explicit size bounds for any prescribed mean-square approximation accuracy, uniformly over the analyzed noise range. Finally, a Gaussian example shows that SiD can mitigate the effect of teacher-score bias and achieve smaller distributional error than exact probability-flow sampling driven by the same teacher.

1 INTRODUCTION

Diffusion models generate high-quality samples (Song & Ermon, 2019; Ho et al., 2020; Song et al., 2020b), but their iterative sampling procedures impose substantial computational costs. Producing a single sample requires repeated evaluations of a denoising network, often over many sequential steps (Ulhaq et al., 2022; Shen et al., 2025). Advanced SDE/ODE solvers can reduce this cost (Song et al., 2020a; Karras et al., 2022; Zhang & Chen, 2022; Lu et al., 2022a;b), but generally retain multiple sequential network evaluations.

One-step distillation offers a different route: compressing a pretrained diffusion model into a generator that maps noise directly to a sample, enabling one-pass generation as in GANs (Goodfellow et al., 2020) and VAEs (Kingma & Welling, 2013). Existing approaches include trajectory-based distillation (Luhman & Luhman, 2021; Salimans & Ho, 2022; Liu et al., 2023), consistency models (Song et al., 2023), and *score distillation* (Poole et al., 2023; Wang et al., 2023; Luo et al., 2023; Yin et al., 2024a;b; Nguyen & Tran, 2024; Zhou et al., 2024; 2025b;c). Despite their empirical success, the distributional guarantees of one-step score-distillation methods remain less well understood. In particular, a fundamental question is:

When does one-step score distillation yield an accurate generator, and how is its distributional error controlled by teacher accuracy and the distillation objectives?

A substantial body of theory studies the statistical expressivity of diffusion models (Chen et al., 2023b; Oko et al., 2023; Fu et al., 2024; Cheng et al., 2025) and the convergence of their reverse-time sampling procedures (Chen et al., 2022; Lee et al., 2022; Chen et al., 2023c; Benton et al., 2023; Li et al., 2023; Li & Yan, 2025). These results concern the iterative diffusion model itself; they do not directly provide guarantees for a one-step student learned through distillation. We address this gap by studying *Score Identity Distillation* (SiD) (Zhou et al., 2024) as a representative method.

SiD trains two networks using a frozen *teacher* diffusion model. The first is a one-step *student model*, or generator, that maps Gaussian noise directly to a sample. The second is a *student score*

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.

surrogate that approximates the score of the generator's forward-noised output distribution. The surrogate learns the score induced by the current generator, while the generator is updated through the SiD objective to align the student and teacher scores along the diffusion path. The theoretical challenge is to translate these coupled training objectives into a guarantee on the generator's distribution.

We establish an explicit total-variation bound between the forward-noised student and data distributions at a positive noise level (Theorem 3.6). The bound accounts for teacher-score error, the SiD training objective, score-surrogate error, and terminal distribution mismatches. The main analytical difficulty is that these errors are controlled under different measures: teacher-score error is measured under the noised *data* distributions, whereas the distillation objective controls a score discrepancy under the noised *student* distributions.

Our analysis, as illustrated in Figure 1, addresses this mismatch by representing the data and student distributions through the exact reverse-time diffusions associated with their respective forward-noised laws. We compare each diffusion with a common teacher-driven reverse process. These two path comparisons preserve the measures under which the respective score errors are controlled and explicitly retain the terminal distribution mismatches. On the student side, the SiD score identity and the surrogate's excess denoising risk control the score discrepancy. Under the stated change-of-measure conditions, Girsanov's theorem and Pinsker's inequality then yield the distributional bound. The reverse-time processes serve only as analytical devices: generation remains one-step, and the argument introduces no sampling discretization error.

We further quantify the network complexity of the student score surrogate required by this guarantee (Theorem 3.7). Finally, we give an example in which one-step generation can implicitly neutralize a teacher-score inconsistency that accumulates during the teacher's iterative diffusion sampling (Proposition 3.9).

1.1 Related Work

Distillation of diffusion models. Methods that compress a pretrained diffusion model into a one- or few-step generator fall into three families: trajectory distillation, which trains the student model to reproduce the teacher's ODE sampler (Luhman & Luhman, 2021; Salimans & Ho, 2022; Song et al., 2023; Liu et al., 2023); adversarial distillation, which sharpens one-step samples with a discriminator (Sauer et al., 2023; Xu et al., 2024); and score distillation, which matches the student-model distribution to the teacher's through their scores. The last family originates in Score Distillation Sampling (Poole et al., 2023) and Variational Score Distillation (Wang et al., 2023) for text-to-3D generation, and was brought to one-step image generation by Diff-Instruct (Luo et al., 2023) and Distribution Matching Distillation (Yin et al., 2024b;a). Continuous semi-implicit models (CoSIM) (Yu et al., 2025) learn a continuous transition kernel for multistep distillation at the distributional level. Score Identity Distillation (SiD) (Zhou et al., 2024) is a representative and widely adopted member of this family, with state-of-the-art one-step performance and several follow-ups (Zhou et al., 2025b;c;a), so we use it as the object of our analysis.

Theory of one-step generation. Existing guarantees for one-step generators concern consistency models (Lyu et al., 2024; Dou et al., 2024; Li et al., 2025; Boffi et al., 2024). Consistency models learn the flow map of the probability-flow ODE by enforcing self-consistency along its trajectories, and their analyses bound how well the learned map tracks the ODE solution. Score distillation is different in kind: it never uses trajectories, but matches the induced score of the student model to the teacher score at every noise level through a surrogate, so the analysis must instead control how score discrepancies translate into distributional error, which is what our error decomposition does. For score distillation itself, the available theory identifies the population objective being optimized and its gradient (Luo et al., 2023; 2024; Xu et al., 2025; Zhou et al., 2024), but does not quantify how far the resulting student model is from the data distribution. To the best of our knowledge, we give the first distributional guarantee for a score-distilled one-step student model that jointly accounts for the teacher's error, the residual distillation objective, and the surrogate's approximation error, together with a bound on the network size the surrogate requires. Our guarantee also differs from the convergence theory of diffusion samplers (Chen et al., 2022; Benton et al., 2023; Li & Yan, 2025), whose bounds are dominated by the discretization error of the reverse process and scale

with the number of sampling steps, whereas a distilled student model never discretizes the reverse process, so this term is absent from our bound.

2 BACKGROUND

2.1 NOTATION

We use bold symbols to denote vectors and matrices. For $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_\infty$ denote the Euclidean and maximum norms. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|_0$ is the number of nonzero entries and $\|\mathbf{A}\|_\infty$ is the largest absolute entry. We write $[n] = \{1, \dots, n\}$ and use $O(\cdot)$ for bounds up to numerical constants. Capital P denotes a distribution and lower-case p denotes its density. $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)| \in [0, 1]$ and $\mathrm{KL}(P\|Q)$ denote the total-variation distance and the Kullback–Leibler divergence, $\mathcal{P}_2(\mathbb{R}^d)$ is the set of probability measures on $\mathbb{R}^d$ with finite second moment, and W_2 is the Euclidean 2-Wasserstein distance on $\mathcal{P}_2(\mathbb{R}^d)$.

2.2 DIFFUSION MODELS

Let $\mathbf{X}_0 \sim P^{\mathrm{data}}$ be a clean data sample, and define the diffusion forward process by

$$\mathbf{X}_t = \alpha_t \mathbf{X}_0 + \sigma_t \xi, \qquad \xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d),$$

where ξ is independent of $\mathbf{X}_0$, P_t^{data} is the distribution of $\mathbf{X}_t$, and p_t^{data} is its density. We assume that (α_t, σ_t) is the marginal schedule of the SDE $\mathrm{d}\mathbf{X}_t = b(t)\mathbf{X}_t\,\mathrm{d}t + \sqrt{a(t)}\,\mathrm{d}\mathbf{W}_t$, where $b(t) = \dot{\alpha}_t/\alpha_t$ and $a(t) = \frac{\mathrm{d}}{\mathrm{d}t}\sigma_t^2 - 2b(t)\sigma_t^2 > 0$. In the variance-exploding setting used by SiD (Zhou et al., 2024), $\alpha_t = 1$ and $a(t) = \frac{\mathrm{d}}{\mathrm{d}t}\sigma_t^2$. The error bounds below allow any such schedule with $a(t) > 0$, while the network approximation result specializes to the Brownian choice $\sigma_t^2 = t$, for which $a(t) = 1$.

As t grows, P_t^{data} approaches a tractable Gaussian. Generation reverses this process: starting from Gaussian noise at a terminal time $T > 0$, one runs the time reversal of the forward SDE (Anderson, 1982; Song et al., 2020b),

$$\mathrm{d}\mathbf{X}_t = \big[b(t)\mathbf{X}_t - a(t)\,\nabla_{\mathbf{x}} \log p_t^{\mathrm{data}}(\mathbf{X}_t)\big]\,\mathrm{d}t + \sqrt{a(t)}\;\mathrm{d}\overline{\mathbf{W}}_t, \tag{2.1}$$

backward in time from T to an early stopping time $0 < t_0 < T$, where $\overline{\mathbf{W}}_t$ is a reverse-time Brownian motion; its marginals coincide with those of the forward process, so its output at time t_0 is distributed as $P_{t_0}^{\mathrm{data}}$. Replacing the noise term by a drift correction gives the deterministic probability-flow ODE

$$\mathrm{d}\mathbf{X}_t = \big[b(t)\mathbf{X}_t - \tfrac{1}{2}a(t)\,\nabla_{\mathbf{x}} \log p_t^{\mathrm{data}}(\mathbf{X}_t)\big]\,\mathrm{d}t, \tag{2.2}$$

which has the same marginals. Either way, sampling requires the *data score*

$$\mathbf{s}_t^\star(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_t^{\mathrm{data}}(\mathbf{x}),$$

which is unknown. A diffusion model therefore trains a neural network $\mathbf{s}_t^\phi(\mathbf{x})$, the *score network*, to approximate $\mathbf{s}_t^\star$, and generates samples by running (2.1) (or a discretization of it) with $\mathbf{s}_t^\star$ replaced by $\mathbf{s}_t^\phi$.

The score network can be parameterized in several equivalent ways. A common alternative is the *denoiser* $\mathbf{f}_\phi$, which predicts the clean sample from the noisy one and is related to the score by Tweedie's formula,

$$\mathbf{f}_\phi(\mathbf{x}_t, t) := \frac{\mathbf{x}_t + \sigma_t^2 \mathbf{s}_t^\phi(\mathbf{x}_t)}{\alpha_t}. \tag{2.3}$$

In this parameterization, the network is trained by minimizing the denoising objective

$$\mathcal{L}_{\mathrm{DM}}(\phi) := \int_{t_0}^{T} w(t)\,\mathbb{E}\Big[\|\mathbf{f}_\phi(\mathbf{X}_t, t) - \mathbf{X}_0\|_2^2\Big]\,\mathrm{d}t,$$

whose population minimizer is the conditional mean $\mathbb{E}[\mathbf{X}_0 \mid \mathbf{X}_t]$ and recovers the data score through Tweedie's formula (Vincent, 2011). A typical choice for the present analysis is $w(t) = a(t)\alpha_t^2\sigma_t^{-4}$.

2.3 SCORE IDENTITY DISTILLATION (SID)

Terminology. SiD involves three networks, named here by their roles. The *teacher* is a frozen pretrained score network $\mathbf{s}_t^\phi$ used to supervise the training of a one-step student model. Throughout the analysis, we treat $\mathbf{s}_t^\phi$ as fixed and call it the teacher score. The *student model*, also called the one-step generator, is the neural network $\mathbf{g}_\theta$ distilled from the teacher and used at generation time. We write *student score* for the score $\mathbf{s}_t^\theta$ induced by its noised output distribution. The *student score surrogate*, or *surrogate* for short, is an auxiliary network $\mathbf{f}_\psi$ with associated score $\mathbf{s}_t^\psi$ that is fitted to track the student score, which is not available in closed form.

Student model. The one-step student model is a neural network $\mathbf{g}_\theta \colon \mathbb{R}^d \to \mathbb{R}^d$ that maps a standard Gaussian input directly to a sample, so that generation costs a single network evaluation. Let $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, set $\mathbf{X}_0 = \mathbf{g}_\theta(\mathbf{Z})$, and denote the distribution of $\mathbf{X}_0$ by P^θ. Distillation aims to make P^θ close to P^{data}. Since the teacher gives access to the data only through its score along the diffusion path, SiD applies the same forward process to the student model's samples:

$$\mathbf{X}_t = \alpha_t \mathbf{X}_0 + \sigma_t \xi, \qquad \xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \tag{2.4}$$

where ξ is independent of $\mathbf{Z}$. Denote the distribution and density of $\mathbf{X}_t$ by P_t^θ and p_t^θ, and let $P_{0,t}^\theta$ denote the joint distribution of $(\mathbf{X}_0, \mathbf{X}_t)$. The reverse process of (2.4) is driven by the *induced student score* $\mathbf{s}_t^\theta(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_t^\theta(\mathbf{x})$, just as (2.1) is driven by $\mathbf{s}_t^\star$; if $\mathbf{s}_t^\theta = \mathbf{s}_t^\phi$ for all t and the teacher is accurate, the two reverse processes coincide and so do P^θ and P^{data}. SiD therefore seeks to match $\mathbf{s}_t^\theta$ to $\mathbf{s}_t^\phi$.

Student score surrogate. Although $\mathbf{s}_t^\theta$ is determined by the noised student distribution, it is generally unavailable in closed form. SiD therefore introduces a *surrogate* denoiser $\mathbf{f}_\psi$, trained to track the student score through denoising score matching on the student model's own samples, whose associated score is

$$\mathbf{s}_t^\psi(\mathbf{x}_t) := \frac{\alpha_t \mathbf{f}_\psi(\mathbf{x}_t, t) - \mathbf{x}_t}{\sigma_t^2}. \tag{2.5}$$

Bilevel training. The SiD procedure is bilevel and involves two comparisons. The lower level is the *student–surrogate* comparison: for the current student model, the surrogate is fitted on pairs $(\mathbf{X}_0, \mathbf{X}_t)$ drawn from that model by minimizing the denoising loss

$$R_t^\theta(\psi) := \mathbb{E}_{(\mathbf{X}_0, \mathbf{X}_t) \sim P_{0,t}^\theta} \left[\|\mathbf{f}_\psi(\mathbf{X}_t, t) - \mathbf{X}_0\|_2^2 \right],$$

whose population minimizer has score exactly $\mathbf{s}_t^\theta$. The upper level is the *teacher–student* comparison: the student model is updated so that the surrogate's estimate of its induced score matches the teacher score. The upper-level losses studied below are

$$\ell_t^{(1)} := \mathbb{E}_{\mathbf{X}_t \sim P_t^\theta} \left[\left\| \mathbf{s}_t^\phi(\mathbf{X}_t) - \mathbf{s}_t^\psi(\mathbf{X}_t) \right\|_2^2 \right],$$

$$\ell_t^{(2)} := \mathbb{E}_{(\mathbf{X}_0, \mathbf{X}_t) \sim P_{0,t}^\theta} \left[\alpha_t \sigma_t^{-2} \left(\mathbf{s}_t^\phi(\mathbf{X}_t) - \mathbf{s}_t^\psi(\mathbf{X}_t) \right)^\top \left(\mathbf{f}_\phi(\mathbf{X}_t, t) - \mathbf{X}_0 \right) \right].$$

When the surrogate is exact, both losses equal the desired student–teacher score discrepancy $\mathbb{E}_{P_t^\theta} \left\| \mathbf{s}_t^\phi - \mathbf{s}_t^\theta \right\|_2^2$; the second form follows from the score identity $\mathbb{E}[\mathbf{X}_0 \mid \mathbf{X}_t] = \mathbf{f}_{\psi^\star}(\mathbf{X}_t, t)$ (Zhou et al., 2024). Nevertheless, $\ell_t^{(1)}$ cannot be used on its own as the training loss, since the surrogate is held fixed in the student update, θ enters $\ell_t^{(1)}$ only through the samples $\mathbf{X}_t$, and its gradient with respect to θ can vanish identically even when the student model is far from optimal (Zhou et al., 2024, Proposition 4). The projected form $\ell_t^{(2)}$ avoids this degeneracy because θ also enters through the denoising residual $\mathbf{f}_\phi(\mathbf{X}_t, t) - \mathbf{X}_0$, and SiD combines the two forms into a single fused loss.

Fused training objective. We use the following reparameterization of the SiD objective. For $\lambda \geq 0$, define

$$\ell_t^{(\lambda)} := (1 - \lambda)\ell_t^{(1)} + \lambda \ell_t^{(2)}. \tag{2.6}$$

The time-integrated objective used in the analysis is

$$\mathcal{S}_\lambda(\theta, \psi) := \int_{t_0}^{T} a(t)\, \ell_t^{(\lambda)}\, \mathrm{d}t.$$

In summary, the bilevel procedure can be written as

$$\psi^\star(\theta) \in \arg\min_{\psi} \int_{t_0}^{T} a(t)\alpha_t^2\sigma_t^{-4} R_t^\theta(\psi)\, \mathrm{d}t, \qquad \theta^\star \in \arg\min_{\theta} \mathcal{S}_\lambda\big(\theta, \psi^\star(\theta)\big).$$

In practice, the two levels are updated alternately: the surrogate is fitted to pairs $(\mathbf{X}_0, \mathbf{X}_t)$ drawn from the current student model, and the student model is then updated using the current surrogate. The analysis below keeps the errors from these two stages separate.

3 Main results

Our analysis relates the student's distributional accuracy to the attained population SiD objective, teacher-score error, and student score-surrogate error. We compare the exact reverse-time diffusions associated with the student and data laws through a common teacher-driven process. These processes serve only as analytical devices; generation remains one-step. Section 3.1 formalizes the three error quantities. Section 3.2 combines them with terminal distribution mismatches to establish a total-variation bound between the noised student and data distributions at an early-stopping time $t_0 > 0$ (Theorem 3.6), followed by a mixed Wasserstein–TV comparison for the clean student output (3.5). For bounded student models under Brownian noising, Section 3.3 constructs a single ReLU score surrogate with explicit size bounds and arbitrarily small mean-square error, uniformly over a fixed interval of positive noise levels (Theorem 3.7). Finally, Section 3.4 gives a Gaussian example in which the population SiD optimum is more accurate than exact probability-flow sampling driven by the same biased teacher.

3.1 Assumptions

The following assumptions quantify the three errors of the fixed teacher, the upper-level student update, and the lower-level surrogate update. First, we state an integrability condition that makes the denoising errors well-defined.

Assumption 3.1 (Integrability)**.** The data and student distributions have finite second moments, $P^{\mathrm{data}}, P^\theta \in \mathcal{P}_2(\mathbb{R}^d)$, with $m_2^{\mathrm{data}} := \mathbb{E}_{P^{\mathrm{data}}} \|\mathbf{X}\|_2^2$ and $m_2^\theta := \mathbb{E}\,\|\mathbf{g}_\theta(\mathbf{Z})\|_2^2$, and the surrogate denoiser is square integrable, $\mathbf{f}_\psi(\cdot, t) \in L^2(P_t^\theta; \mathbb{R}^d)$ for every $t \in [t_0, T]$.

Assumption 3.2 (Teacher-data score error)**.** There exists $\epsilon_1 \geq 0$ such that

$$\mathcal{E}_1 := \int_{t_0}^{T} a(t)\, \mathbb{E}_{\mathbf{X}_t \sim P_t^{\mathrm{data}}} \left[\left\| \mathbf{s}_t^\star(\mathbf{X}_t) - \mathbf{s}_t^\phi(\mathbf{X}_t) \right\|_2^2 \right] \mathrm{d}t \leq \epsilon_1.$$

Since the teacher is fixed, ϵ_1 is exogenous and depends only on the teacher–data discrepancy along the noised data path and is not affected by SiD training.

Assumption 3.3 (Integrated population SiD objective)**.** For some $\lambda \geq 0$, there exists $\epsilon_2 \geq 0$ such that

$$\mathcal{S}_\lambda(\theta, \psi) = \int_{t_0}^{T} a(t)\, \ell_t^{(\lambda)}\, \mathrm{d}t \leq \epsilon_2,$$

where $\ell_t^{(\lambda)}$ is defined in (2.6).

Because $\ell_t^{(\lambda)}$ is expressed through the surrogate rather than the unavailable student score, a small ϵ_2 alone does not imply that $\mathbf{s}_t^\theta$ is close to $\mathbf{s}_t^\phi$. The next assumption controls the discrepancy between the surrogate and the induced student score.

Assumption 3.4 (Surrogate–student score error)**.** There exists $\epsilon_3 \geq 0$ such that

$$\mathcal{D}_\psi^\theta := \int_{t_0}^{T} a(t)\, \alpha_t^2 \sigma_t^{-4} \left[R_t^\theta(\psi) - R_t^{\theta,\star} \right] \mathrm{d}t \leq \epsilon_3,$$

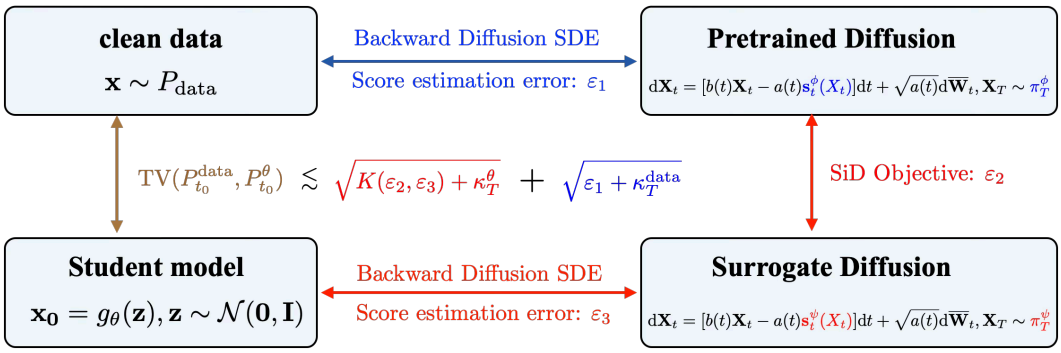

Figure 1: Our theoretical framework for analyzing score distillation. The clean data, the pretrained teacher, and the one-step student model each define a reverse-time diffusion SDE with the same diffusion coefficient and controllable terminal distribution shift. Bounding the distributional accuracy of the student therefore reduces to a score-matching problem: the teacher's score error ϵ_1 and the SiD objective ϵ_2 together with the surrogate error ϵ_3 control the two score discrepancies, and Girsanov's theorem converts them into a total-variation bound.

where

$$R_t^{\theta,\star} := \inf_{\mathbf{f} \in L^2(P_t^{\theta}; \mathbb{R}^d)} \mathbb{E}_{(\mathbf{X}_0, \mathbf{X}_t) \sim P_{0,t}^{\theta}} \left[\|\mathbf{f}(\mathbf{X}_t, t) - \mathbf{X}_0\|_2^2 \right].$$

The reference risk $R_t^{\theta,\star}$ is the Bayes denoising risk, attained by the conditional mean $\mathbb{E}[\mathbf{X}_0 \mid \mathbf{X}_t]$ over all square-integrable denoisers. By the denoising projection identity (Vincent, 2011) (Lemma A.1),

$$\alpha_t^2 \sigma_t^{-4} \left[R_t^{\theta}(\psi) - R_t^{\theta,\star} \right] = \mathbb{E}_{\mathbf{X}_t \sim P_t^{\theta}} \left\| \mathbf{s}_t^{\psi}(\mathbf{X}_t) - \mathbf{s}_t^{\theta}(\mathbf{X}_t) \right\|_2^2.$$

Consequently,

$$\mathcal{D}_{\psi}^{\theta} = \int_{t_0}^{T} a(t)\, \mathbb{E}_{\mathbf{X}_t \sim P_t^{\theta}} \left\| \mathbf{s}_t^{\psi}(\mathbf{X}_t) - \mathbf{s}_t^{\theta}(\mathbf{X}_t) \right\|_2^2 \mathrm{d}t,$$

so ϵ_3 bounds the integrated surrogate-score error under the noised student distributions. This identity connects the surrogate's denoising objective to the score discrepancy required by our analysis.

Terminal mismatches. Our distributional comparison uses a teacher-driven reverse process initialized at $\gamma_T = \mathcal{N}(\mathbf{0}, \sigma_T^2 \mathbf{I}_d)$. The exact data and student reversals instead start from P_T^{data} and P_T^{θ}, respectively. We record these terminal mismatches as

$$\kappa_T^{\text{data}} := \mathrm{KL}(P_T^{\text{data}} \| \gamma_T), \qquad \kappa_T^{\theta} := \mathrm{KL}(P_T^{\theta} \| \gamma_T).$$

Minor technical assumptions are stated in Assumption A.2 in the appendix.

3.2 Main Theoretical Framework: Backward SDE as a Bridge

For a time-dependent vector field $\mathbf{s} = (\mathbf{s}_t)_{t \in [t_0, T]}$ and a terminal law π_T, consider the reverse-time diffusion SDE

$$\mathrm{d}\mathbf{X}_t = \left[b(t)\mathbf{X}_t - a(t)\, \mathbf{s}_t(\mathbf{X}_t) \right] \mathrm{d}t + \sqrt{a(t)}\, \mathrm{d}\overline{\mathbf{W}}_t, \qquad \mathbf{X}_T \sim \pi_T, \tag{3.1}$$

run backward from T to t_0. Under the reverse-process conditions in Assumption A.2, if $\mathbf{s} = \mathbf{s}^q$ is the score of the noised laws $(P_t^q)_t$ and $\pi_T = P_T^q$, then the solution has marginals P_t^q. This applies to the data ($q = \text{data}$) and the student ($q = \theta$). The teacher-driven comparison process instead uses $\mathbf{s}_t = \mathbf{s}_t^{\phi}$ and $\pi_T = \gamma_T$; denote its marginals by $\widehat{P}_t^{\phi}$. The three processes share the diffusion coefficient but generally differ in their drift and terminal law (Figure 1). Comparing the data and student reversals with this common teacher-driven process allows us to control their distributional difference through score discrepancies and terminal mismatches. These processes serve only as analytical devices: the student generates samples directly, so no sampling discretization term enters the bound.

Assumption 3.2 controls the teacher–data score discrepancy under the noised data distributions. The student–teacher discrepancy, measured under the noised student distributions, must be controlled jointly through the population SiD objective and the surrogate's approximation error. The following lemma establishes this connection, providing the student-side score bound needed for the reverse-process comparison.

Lemma 3.5 (Student–teacher score error bound)**.** Define the student–teacher score error as

$$\mathcal{M}_{\theta,\phi} := \int_{t_0}^{T} a(t)\, \mathbb{E}_{\mathbf{X}_t \sim P_t^\theta} \left[\left\| \mathbf{s}_t^\theta(\mathbf{X}_t) - \mathbf{s}_t^\phi(\mathbf{X}_t) \right\|_2^2 \right] \mathrm{d}t.$$

Under Assumptions 3.3 and 3.4, we have

$$\mathcal{M}_{\theta,\phi} \le K_\lambda(\epsilon_2, \epsilon_3) := \left(\frac{|2-\lambda|\sqrt{\epsilon_3} + \sqrt{\lambda^2 \epsilon_3 + 4\epsilon_2}}{2} \right)^2. \tag{3.2}$$

The proof is provided in Appendix A.2. Lemma 3.5 supplies the student-side score error of the framework; combining it with the teacher-side error of Assumption 3.2 along the common backward SDE (3.1) yields our main result.

Theorem 3.6 (Total-variation guarantee)**.** Under Assumptions 3.2–3.4 and the reverse-process conditions in Assumption A.2,

$$\mathrm{TV}\big(P_{t_0}^\theta, P_{t_0}^{\mathrm{data}}\big) \le \underbrace{\sqrt{\frac{1}{2}\kappa_T^\theta + \frac{1}{4}K_\lambda(\epsilon_2,\epsilon_3)} + \sqrt{\frac{1}{2}\kappa_T^{\mathrm{data}} + \frac{1}{4}\epsilon_1}}_{\Delta_{t_0,T}}. \tag{3.3}$$

Moreover, under Assumption 3.1 the terminal mismatches satisfy

$$\kappa_T^{\mathrm{data}} \le \frac{\alpha_T^2\, m_2^{\mathrm{data}}}{2\sigma_T^2}, \qquad \kappa_T^\theta \le \frac{\alpha_T^2\, m_2^\theta}{2\sigma_T^2}, \tag{3.4}$$

For the variance-exploding schedule ($\alpha_t = 1$) this gives $\kappa_T^q \le m_2^q/(2\sigma_T^2)$, that is, $m_2^q/(2T)$ for the Brownian choice $\sigma_t^2 = t$ used in Theorem 3.7. When P^{data} is a Dirac data distribution, the variance-exploding schedule gives an exact $O(1/T)$ rate and cannot be improved in general, whereas for the variance-preserving schedule the bound is exponentially small in T.

Explanation of Theorem 3.6. Since $K_\lambda(\epsilon_2,\epsilon_3) = O(\epsilon_2+\epsilon_3)$, the bound reads $\mathrm{TV}(P_{t_0}^\theta, P_{t_0}^{\mathrm{data}}) = O(\sqrt{\epsilon_1} + \sqrt{\epsilon_2} + \sqrt{\epsilon_3})$ once the terminal mismatches (3.4) are made negligible by a long horizon: the teacher's error, the attained SiD objective, and the surrogate error enter additively and on equal footing. Unlike convergence guarantees for diffusion samplers, none of the three terms involves a discretization error, because the reverse SDE serves only as the analytical bridge as shown in Figure 1. Beyond the teacher's own error ϵ_1, the approximation error of the distilled model rests entirely on $K_\lambda(\epsilon_2,\epsilon_3)$, which adapts to the expressivity of the two trained networks. Heuristically, in the worst case, a deep one-step model can emulate a discretized sampler inside its layers, one layer per DDPM or ODE step, and thereby attain at least the accuracy of that sampler; but a network can also learn shortcuts and adapt to the data, so the attained error can be far smaller than any discretization error. The proof is provided in Appendix A.4.

Early stopping and the clean output. Theorem 3.6 compares the two *noised* laws at the early-stopping time $t_0 > 0$; the price of the positive cutoff is a Wasserstein perturbation of the clean laws. Precisely, under Assumption 3.1 there is a comparison distribution $R_{\theta,t_0} \in \mathcal{P}_2(\mathbb{R}^d)$ with

$$\mathrm{TV}\big(P^\theta, R_{\theta,t_0}\big) \le \min\{1, \Delta_{t_0,T}\}, \qquad W_2\big(R_{\theta,t_0}, P^{\mathrm{data}}\big) \le 2\sqrt{d}\,\sigma_{t_0}/\alpha_{t_0}, \tag{3.5}$$

where $\Delta_{t_0,T}$ is defined in Eq. (3.3). Thus, the clean student law $P^\theta = \mathcal{L}(\mathbf{g}_\theta(\mathbf{Z}))$ lies within total variation $\Delta_{t_0,T}$ of the W_2-ball of radius $2\sqrt{d}\,\sigma_{t_0}/\alpha_{t_0}$ around the data, which is $2\sqrt{dt_0}$ for the Brownian schedule. The early-stopping error thus enters only through this Wasserstein radius, with the TV budget of Theorem 3.6 unchanged. This mixed Wasserstein–TV form is standard in diffusion sampling theory (Lee et al., 2022; Chen et al., 2022; 2023a).

3.3 NETWORK COMPLEXITY OF THE SURROGATE

Theorem 3.6 treats the surrogate error ϵ_3 as an input. We now show that, for a bounded student model under the Brownian schedule, any such error level is realized by one ReLU network shared across all $t \in [t_0, T]$, with explicit size as a function of the target accuracy. To this end, let $d_0 = d + 1$, $d_L = d$, and define the ReLU class

$$\mathcal{F}(M, W, \kappa, L, K) := \Big\{ \mathbf{s}(\mathbf{x}, t) = \big(\mathbf{A}_L \,\mathrm{ReLU}(\,\cdot\,) + \mathbf{b}_L\big) \circ \cdots \circ \big(\mathbf{A}_1 [\mathbf{x}^\top, t]^\top + \mathbf{b}_1\big) : \\ \mathbf{A}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}, \quad \mathbf{b}_\ell \in \mathbb{R}^{d_\ell}, \quad \max_{0 \le \ell \le L} d_\ell \le W, \\ \sup_{(\mathbf{x},t) \in \mathbb{R}^d \times [t_0, T]} \|\mathbf{s}(\mathbf{x}, t)\|_\infty \le M, \quad \max_{1 \le \ell \le L} \big(\|\mathbf{A}_\ell\|_\infty \vee \|\mathbf{b}_\ell\|_\infty\big) \le \kappa, \\ \sum_{\ell=1}^{L} \big(\|\mathbf{A}_\ell\|_0 + \|\mathbf{b}_\ell\|_0\big) \le K \Big\},$$

where $\mathrm{ReLU}(u) = \max\{u, 0\}$ acts coordinatewise, and M, W, κ, L, K bound, respectively, the output magnitude (uniformly over $(\mathbf{x}, t)$, since $\mathbf{s}$ is one joint space–time network), the width, the entrywise weight magnitude, the depth, and the number of nonzero parameters.

Theorem 3.7 (Joint ReLU approximation of the student score)**.** Assume the Brownian variance-exploding schedule $\alpha_t = 1$, $\sigma_t^2 = t$, and suppose that $\|\mathbf{g}_\theta(\mathbf{z})\|_\infty \le B_g$ for every $\mathbf{z} \in \mathbb{R}^d$. There is a constant $C \ge 3$, depending only on the fixed quantities d, t_0, T, and B_g, such that the following holds. For every approximation error $\epsilon \in (0, 1)$, there is one network $\mathbf{s}_\psi \in \mathcal{F}(M_\epsilon, W_\epsilon, \kappa_\epsilon, L_\epsilon, K_\epsilon)$ with parameters satisfying

$$M_\epsilon \le C \log^{1/2}\left(\frac{C}{\epsilon}\right), \qquad W_\epsilon \le C\epsilon^{-3} \log^{d+10}\left(\frac{C}{\epsilon}\right), \\ \kappa_\epsilon \le \exp\left(C \log^4\left(\frac{C}{\epsilon}\right)\right), \qquad L_\epsilon \le C \log^4\left(\frac{C}{\epsilon}\right), \\ K_\epsilon \le C\epsilon^{-3} \log^{d+12}\left(\frac{C}{\epsilon}\right). \tag{3.6}$$

The network satisfies

$$\sup_{t \in [t_0, T]} \mathbb{E}_{\mathbf{X}_t \sim P_t^\theta} \left\|\mathbf{s}_\psi(\mathbf{X}_t, t) - \mathbf{s}_t^\theta(\mathbf{X}_t)\right\|_2^2 \le \epsilon. \tag{3.7}$$

Thus all dependence on ϵ is displayed in (3.6); the constant C is independent of ϵ. Consequently, for any $\epsilon_3 > 0$, choosing $\epsilon = \frac{1}{2}\min\{1, \epsilon_3/(T - t_0)\}$ makes the surrogate denoiser $\mathbf{f}_\psi(\mathbf{x}, t) = \mathbf{x} + t\mathbf{s}_\psi(\mathbf{x}, t)$ satisfy Assumption 3.4.

The proof is provided in Appendix B.6. Taken together, Theorems 3.6 and 3.7 show that, up to the SiD optimization error, the student model is as close to the data as the teacher score allows.

3.4 BEYOND THE TEACHER: A COMPARISON WITH TEACHER-DRIVEN SAMPLING

A further question is whether it can do better than the teacher. This section gives a positive answer: because SiD matches the teacher score in an integrated sense along the whole diffusion path rather than following it step by step, a systematic teacher error can partially cancel out. To isolate this effect, we consider a one-dimensional Gaussian setting in which everything except the teacher bias is exact: the surrogate is exact ($\epsilon_3 = 0$), the SiD minimizer and the probability-flow ODE are available in closed form, and the ODE is solved exactly. The only imperfection is a multiplicative bias in the teacher score, and we compare how it propagates to the outputs of SiD and of the teacher-driven probability-flow ODE.

Assumption 3.8 (Biased one-dimensional Gaussian setting)**.** Let $T > 0$, and consider the Brownian schedule $\alpha_t = 1$, $\sigma_t^2 = t$ on $[0, T]$ with $P^{\mathrm{data}} = \mathcal{N}(0, 1)$, so that $P_t^{\mathrm{data}} = \mathcal{N}(0, 1 + t)$ and $s_t^\star(x) = -x/(1+t)$. The teacher has a multiplicative bias, $s_t^\phi(x) = -cx/(1+t)$ with $c > 0$, $c \ne 1$. We constrain the student-model family to $P^\theta = \mathcal{N}(0, q)$, $q > 0$, and assume that the surrogate is exact, $s_t^\psi = s_t^\theta$. To isolate the propagation of the teacher-score error from terminal mismatch, the probability-flow ODE is initialized at the exact terminal law $P_T^{\mathrm{data}} = \mathcal{N}(0, 1 + T)$.

Because the denoiser formulas are singular at $t = 0$, we define the objective on $[0, T]$ by the improper limit $\mathcal{S}_\lambda^{[0,T]}(q) := \lim_{\varepsilon \downarrow 0} \int_\varepsilon^T \ell_t^{(\lambda)} \, \mathrm{d}t$, whose score-form integrand has a finite limit, so it is well defined for every $q > 0$.

Let q_{SiD} denote the *oracle* minimizer of $\mathcal{S}_\lambda^{[0,T]}$, that is, the minimizer of the objective after the exact surrogate has been substituted, and let $P^{\mathrm{SiD}} = \mathcal{N}(0, q_{\mathrm{SiD}})$. Let P^{PF} be the law of the output X_0^{PF} of the teacher-driven probability-flow ODE. We have the following result:

Proposition 3.9 (SiD versus probability-flow sampling)**.** Under Assumption 3.8, for every $\lambda \geq 0$ the objective $\mathcal{S}_\lambda^{[0,T]}(q)$ has a unique minimizer q_{SiD}, and the oracle SiD output satisfies, for every horizon $T > 0$,

$$\mathrm{KL}\big(P^{\mathrm{SiD}} \,\big\|\, P^{\mathrm{data}}\big) \leq \kappa(c) := \frac{1}{2}\big(c^{-2} - 1 - \log c^{-2}\big), \qquad \mathrm{TV}\big(P^{\mathrm{SiD}}, P^{\mathrm{data}}\big) \leq \sqrt{1 - e^{-\kappa(c)}} < 1. \tag{3.8}$$

By contrast, exact probability-flow ODE sampling with the same teacher satisfies

$$\mathrm{KL}\big(P^{\mathrm{PF}} \,\big\|\, P^{\mathrm{data}}\big) = \frac{1}{2}\Big[(1+T)^{1-c} - 1 - (1-c)\log(1+T)\Big], \tag{3.9}$$

which grows like $\frac{1}{2}(1+T)^{1-c}$ when $c < 1$ and like $\frac{1}{2}(c-1)\log(1+T)$ when $c > 1$; in both cases $\mathrm{KL}(P^{\mathrm{PF}} \| P^{\mathrm{data}}) \to \infty$ and $\mathrm{TV}(P^{\mathrm{PF}}, P^{\mathrm{data}}) \to 1$ as $T \to \infty$.

The proof is given in Appendix B.7. Intuitively, SiD matches the biased teacher score in an integrated sense over the whole diffusion path, so the teacher's errors at different noise levels are averaged and may partially cancel, leaving a distortion bounded uniformly in T, whereas the probability-flow ODE follows the biased drift step by step and the errors accumulate. For instance, with $c = 1/2$ the SiD output satisfies $\mathrm{KL}(P^{\mathrm{SiD}} \| P^{\mathrm{data}}) < 0.81$ for every T, whereas the teacher-driven sampler incurs a KL divergence growing like $\sqrt{T}/2$. The comparison concerns this Gaussian model with an exact surrogate and is not a uniform dominance statement, but it shows that one-step distillation can approximate the data more accurately than iterative sampling with the same teacher.

4 Conclusions and Discussions

We have developed a theoretical framework for one-step generators trained by score distillation, in which the data, the teacher, and the one-step student model each drive the same reverse-time diffusion SDE with different scores, so distributional accuracy reduces to score matching along the diffusion path. Applying this framework to SiD, we have derived an explicit total-variation bound between the noised student and data distributions in terms of the teacher's score error, the attained SiD objective, and the surrogate's score error, with no sampling discretization term. We have also transferred it to the clean one-step output in a mixed Wasserstein–TV form. We have further quantified the approximation error between the induced student score and a ReLU score surrogate, showing that a single network of explicit size drives the surrogate term in the bound to any prescribed level. Finally, we have given an intuitive Gaussian example in which SiD performs beyond its teacher: the teacher's bias is averaged out by SiD's integrated score matching, whereas it accumulates under teacher-driven sampling.

Two limitations point to future work. First, our guarantee is stated at the population level: the three error terms are treated as attained quantities, and the statistical error of estimating the surrogate and the student model from finitely many samples is left for future work. Second, we do not quantify how small the attained SiD objective ϵ_2 can be made, nor its dependence on the size of the student network; an approximation theory for the student model analogous to Theorem 3.7 remains open.

AI use statement

In this work, we used generative AI tools for assisting in the writing of proofs, formulating mathematical claims, proposing or refining hypotheses. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, providing critical ingredients for proving mathematical claims. Generating synthetic data sets, interpreting results, assisting with translation, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, designing or

providing feedback on research methodology or experiments, implementing methods are not applicable to this work. Additionally, we used generative AI tools for draft parts of a research paper. We have reviewed all AI-assisted work. We checked all LLM-generated proofs manually. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

Nicholas M Boffi, Michael S Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *arXiv preprint arXiv:2406.07507*, 2024.

Clément L. Canonne. A short note on an inequality between kl and tv, 2023.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023a.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data, 2023b. URL https://arxiv.org/abs/2302.07194.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. *arXiv preprint arXiv:2305.11798*, 2023c.

Ziheng Cheng, Tianyu Xie, Shiyue Zhang, and Cheng Zhang. Provable sample-efficient transfer learning conditional diffusion models via representation learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 79411–79460. Curran Associates, Inc., 2025. doi: 10.52202/085713-2655. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/72176f95680c3fb27a0966f36d5d0c53-Paper-Conference.pdf.

Zehao Dou, Minshuo Chen, Mengdi Wang, and Zhuoran Yang. Theory of consistency diffusion models: Distribution estimation meets fast sampling. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory, 2024. URL https://arxiv.org/abs/2403.11968.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. *arXiv preprint arXiv:2209.12381*, 2022.

Gen Li and Yuling Yan. $O(d/T)$ convergence theory for diffusion probabilistic models under minimal assumptions. *Journal of Machine Learning Research*, 26(292):1–55, 2025.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.

Gen Li, Zhihan Huang, and Yuting Wei. Towards a mathematical theory for consistency training in diffusion models. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*. PMLR, 2025.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. *arXiv preprint arXiv:2309.06380*, 2023.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022a.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022b.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-Instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023.

Weijian Luo, Zemin Huang, Zhengyang Geng, J Zico Kolter, and Guo-Jun Qi. One-step diffusion distillation through score implicit matching. *Advances in Neural Information Processing Systems*, 37, 2024.

Junlong Lyu, Zhitang Chen, and Shoubo Feng. Sampling is as easy as keeping the consistency: Convergence guarantee for consistency models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 33664–33685. PMLR, 2024.

Thuan Hoang Nguyen and Anh Tran. SwiftBrush: One-step text-to-image diffusion model with variational score distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. *arXiv preprint arXiv:2303.01861*, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *The Eleventh International Conference on Learning Representations*, 2023.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

Hui Shen, Jingxuan Zhang, Boning Xiong, Rui Hu, Shoufa Chen, Zhongwei Wan, Xin Wang, Yu Zhang, Zixuan Gong, Guangyin Bao, Chaofan Tao, Yongfeng Huang, Ye Yuan, and Mi Zhang. Efficient diffusion models: A survey. *Transactions on Machine Learning Research*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

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. *arXiv preprint arXiv:2011.13456*, 2020b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Anwaar Ulhaq, Naveed Akhtar, and Ganna Pogrebna. Efficient diffusion models for vision: A survey. *arXiv preprint arXiv:2210.09292*, 2022.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2023.

Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. UFOGen: You forward once large scale text-to-image generation via diffusion GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching. *arXiv preprint arXiv:2502.15681*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in Neural Information Processing Systems*, 37, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Longlin Yu, Jiajun Zha, Tong Yang, Tianyu Xie, Xiangyu Zhang, S.-H. Chan, and Cheng Zhang. Continuous semi-implicit models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73375–73400. PMLR, 2025. URL https://proceedings.mlr.press/v267/yu25v.html.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62307–62331. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhou24x.html.

Mingyuan Zhou, Yi Gu, and Zhendong Wang. Few-step diffusion via score identity distillation. *arXiv preprint arXiv:2505.12674*, 2025a.

Mingyuan Zhou, Zhendong Wang, Huangjie Zheng, and Hai Huang. Guided score identity distillation for data-free one-step text-to-image generation. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://arxiv.org/abs/2406.01561.

Mingyuan Zhou, Huangjie Zheng, Yi Gu, Zhendong Wang, and Hai Huang. Adversarial score identity distillation: Rapidly surpassing the teacher in one step. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://arxiv.org/abs/2410.14919.

A PROOFS FOR THE POPULATION ERROR DECOMPOSITION

The section proves the three steps used in the main text:

1. the equivalence between the surrogate's excess denoising error and its student-score error;
2. the population SiD error bound in Lemma 3.5;
3. the total-variation guarantee in Theorem 3.6.

Lemma A.1 (Denoising projection identity)**.** Suppose that Assumption 3.1 holds, that is, $\mathbb{E}\,\|\mathbf{g}_\theta(\mathbf{Z})\|_2^2 < \infty$ and $\mathbf{f}_\psi(\cdot, t) \in L^2(P_t^\theta; \mathbb{R}^d)$. For each $t \in [t_0, T]$, let $\mathbf{m}_t^\theta(\mathbf{x}) = \mathbb{E}[\mathbf{X}_0 \mid \mathbf{X}_t = \mathbf{x}]$. Then

$$R_t^\theta(\psi) - R_t^{\theta,\star} = \mathbb{E}_{\mathbf{X}_t \sim P_t^\theta} \left\| \mathbf{f}_\psi(\mathbf{X}_t, t) - \mathbf{m}_t^\theta(\mathbf{X}_t) \right\|_2^2 . \tag{A.1}$$

Moreover,

$$\alpha_t^2 \sigma_t^{-4} \left[R_t^\theta(\psi) - R_t^{\theta,\star} \right] = \mathbb{E}_{\mathbf{X}_t \sim P_t^\theta} \left\| \mathbf{s}_t^\psi(\mathbf{X}_t) - \mathbf{s}_t^\theta(\mathbf{X}_t) \right\|_2^2 . \tag{A.2}$$

A.1 PROOF OF LEMMA A.1

Proof. Fix t, and let $\mathbf{X}_0, \mathbf{X}_t$ be the random vectors in (2.4). Since $\mathbb{E}\,\|\mathbf{X}_0\|_2^2 = \mathbb{E}\,\|\mathbf{g}_\theta(\mathbf{Z})\|_2^2 < \infty$, the conditional mean $\mathbf{m}_t^\theta(\mathbf{X}_t) = \mathbb{E}[\mathbf{X}_0 \mid \mathbf{X}_t]$ is well defined and square integrable, and every risk below is finite. (The integrability of $\mathbf{X}_0$ cannot be dropped: if $\mathbb{E}\,\|\mathbf{X}_0\|_2^2 = \infty$, then $\mathbb{E}\,\|\mathbf{f}(\mathbf{X}_t, t) - \mathbf{X}_0\|_2^2 = \infty$ for every square-integrable $\mathbf{f}$, since otherwise $\mathbf{X}_0 = (\mathbf{X}_0 - \mathbf{f}) + \mathbf{f}$ would be square integrable, and the excess risk in (A.1) would be undefined.) For any square-integrable function $\mathbf{f}(\mathbf{X}_t, t)$, adding and subtracting $\mathbf{m}_t^\theta(\mathbf{X}_t)$ gives

$$\mathbb{E}\left[\|\mathbf{f}(\mathbf{X}_t, t) - \mathbf{X}_0\|_2^2 \right] \tag{A.3}$$

$$= \mathbb{E}\left[\left\|\mathbf{f}(\mathbf{X}_t, t) - \mathbf{m}_t^\theta(\mathbf{X}_t)\right\|_2^2 \right] + \mathbb{E}\left[\left\|\mathbf{m}_t^\theta(\mathbf{X}_t) - \mathbf{X}_0\right\|_2^2 \right] + 2\mathbb{E}\left[\left(\mathbf{f}(\mathbf{X}_t, t) - \mathbf{m}_t^\theta(\mathbf{X}_t)\right)^\top \left(\mathbf{m}_t^\theta(\mathbf{X}_t) - \mathbf{X}_0\right) \right]. \tag{A.4}$$

The cross term is zero. Indeed, the tower property gives

$$\begin{aligned} &\mathbb{E}\left[\left(\mathbf{f}(\mathbf{X}_t, t) - \mathbf{m}_t^\theta(\mathbf{X}_t)\right)^\top \left(\mathbf{m}_t^\theta(\mathbf{X}_t) - \mathbf{X}_0\right) \right] \\ &= \mathbb{E}\left[\left(\mathbf{f}(\mathbf{X}_t, t) - \mathbf{m}_t^\theta(\mathbf{X}_t)\right)^\top \mathbb{E}\left[\mathbf{m}_t^\theta(\mathbf{X}_t) - \mathbf{X}_0 \mid \mathbf{X}_t \right] \right] \\ &= 0, \end{aligned}$$

because $\mathbb{E}[\mathbf{X}_0 \mid \mathbf{X}_t] = \mathbf{m}_t^\theta(\mathbf{X}_t)$. The second term on the right of (A.4) is independent of $\mathbf{f}$, and the first term is nonnegative. The minimizer is therefore $\mathbf{f} = \mathbf{m}_t^\theta$. Subtracting the minimum risk from the risk of $\mathbf{f}_\psi$ proves (A.1).

We next connect the conditional mean to the score. The conditional density of $\mathbf{X}_t$ given $\mathbf{X}_0 = \mathbf{x}_0$ is

$$q_t(\mathbf{x} \mid \mathbf{x}_0) = (2\pi\sigma_t^2)^{-d/2} \exp\left(-\frac{\|\mathbf{x} - \alpha_t \mathbf{x}_0\|_2^2}{2\sigma_t^2} \right).$$

Therefore

$$p_t^\theta(\mathbf{x}) = \int q_t(\mathbf{x} \mid \mathbf{x}_0) P^\theta(\,\mathrm{d}\mathbf{x}_0). \tag{A.5}$$

Every Gaussian kernel is positive, so $p_t^\theta(\mathbf{x}) > 0$. For each coordinate j,

$$\frac{\partial}{\partial x_j} q_t(\mathbf{x} \mid \mathbf{x}_0) = -\frac{x_j - \alpha_t (x_0)_j}{\sigma_t^2} q_t(\mathbf{x} \mid \mathbf{x}_0).$$

The derivative is bounded uniformly in $\mathbf{x}_0$ by a finite constant proportional to $(2\pi\sigma_t^2)^{-d/2}\sigma_t^{-1}$. Thus differentiation may be moved through the integral in (A.5). Consequently,

$$\begin{aligned} \nabla_{\mathbf{x}} p_t^\theta(\mathbf{x}) &= -\frac{1}{\sigma_t^2} \int (\mathbf{x} - \alpha_t \mathbf{x}_0) q_t(\mathbf{x} \mid \mathbf{x}_0) P^\theta(\,\mathrm{d}\mathbf{x}_0) \\ &= \frac{p_t^\theta(\mathbf{x})}{\sigma_t^2} \left(\alpha_t \mathbb{E}[\mathbf{X}_0 \mid \mathbf{X}_t = \mathbf{x}] - \mathbf{x} \right). \end{aligned} \tag{A.6}$$

For the second equality, Bayes' formula says that the conditional distribution of $\mathbf{X}_0$ given $\mathbf{X}_t = \mathbf{x}$ has density $q_t(\mathbf{x} \mid \mathbf{x}_0)/p_t^\theta(\mathbf{x})$ relative to P^θ. Dividing (A.6) by $p_t^\theta(\mathbf{x})$ yields Tweedie's identity

$$\mathbf{s}_t^\theta(\mathbf{x}) = \frac{\alpha_t \mathbf{m}_t^\theta(\mathbf{x}) - \mathbf{x}}{\sigma_t^2}. \tag{A.7}$$

Equations (2.5) and (A.7) give

$$\mathbf{s}_t^\psi(\mathbf{X}_t) - \mathbf{s}_t^\theta(\mathbf{X}_t) = \frac{\alpha_t}{\sigma_t^2}\left(\mathbf{f}_\psi(\mathbf{X}_t, t) - \mathbf{m}_t^\theta(\mathbf{X}_t)\right).$$

Taking squared Euclidean norms and expectations and then using (A.1) proves (A.2). □

A.2 Proof of Lemma 3.5

Proof. For every t, define the two score differences

$$\mathbf{U}_t(\mathbf{x}) := \mathbf{s}_t^\phi(\mathbf{x}) - \mathbf{s}_t^\theta(\mathbf{x}), \qquad \boldsymbol{\Delta}_t(\mathbf{x}) := \mathbf{s}_t^\psi(\mathbf{x}) - \mathbf{s}_t^\theta(\mathbf{x}).$$

The difference used in the population SiD objective is $\boldsymbol{\delta}_t = \mathbf{U}_t - \boldsymbol{\Delta}_t$. Set

$$\begin{aligned} M_t &:= \mathbb{E}_{\mathbf{X}_t \sim P_t^\theta} \|\mathbf{U}_t(\mathbf{X}_t)\|_2^2, \\ D_t &:= \mathbb{E}_{\mathbf{X}_t \sim P_t^\theta} \|\boldsymbol{\Delta}_t(\mathbf{X}_t)\|_2^2, \\ C_t &:= \mathbb{E}_{\mathbf{X}_t \sim P_t^\theta} \left[\mathbf{U}_t(\mathbf{X}_t)^\top \boldsymbol{\Delta}_t(\mathbf{X}_t)\right]. \end{aligned}$$

Expanding $\|\mathbf{U}_t - \boldsymbol{\Delta}_t\|_2^2$ gives

$$\ell_t^{(1)} = M_t + D_t - 2C_t. \tag{A.8}$$

We now compute $\ell_t^{(2)}$. First condition on $\mathbf{X}_t$. By the definition of $\mathbf{f}_\phi$ in (2.3) and Tweedie's identity (A.7),

$$\begin{aligned} &\mathbb{E}\left[\alpha_t \sigma_t^{-2}\big(\mathbf{f}_\phi(\mathbf{X}_t, t) - \mathbf{X}_0\big) \mid \mathbf{X}_t\right] \\ &= \alpha_t \sigma_t^{-2}\big(\mathbf{f}_\phi(\mathbf{X}_t, t) - \mathbf{m}_t^\theta(\mathbf{X}_t)\big) \\ &= \sigma_t^{-2}\left(\mathbf{X}_t + \sigma_t^2 \mathbf{s}_t^\phi(\mathbf{X}_t) - \alpha_t \mathbf{m}_t^\theta(\mathbf{X}_t)\right) \\ &= \mathbf{s}_t^\phi(\mathbf{X}_t) - \mathbf{s}_t^\theta(\mathbf{X}_t) = \mathbf{U}_t(\mathbf{X}_t). \end{aligned}$$

Because $\boldsymbol{\delta}_t(\mathbf{X}_t)$ is a function of $\mathbf{X}_t$, the tower property gives

$$\begin{aligned} \ell_t^{(2)} &= \mathbb{E}_{\mathbf{X}_t \sim P_t^\theta}\left[\big(\mathbf{U}_t(\mathbf{X}_t) - \boldsymbol{\Delta}_t(\mathbf{X}_t)\big)^\top \mathbf{U}_t(\mathbf{X}_t)\right] \\ &= M_t - C_t. \end{aligned} \tag{A.9}$$

Substituting (A.8) and (A.9) into the definition of $\ell_t^{(\lambda)}$ yields the exact pointwise identity

$$\ell_t^{(\lambda)} = M_t + (1-\lambda)D_t - (2-\lambda)C_t. \tag{A.10}$$

Integrate (A.10) against $a(t)\,\mathrm{d}t$. Lemma A.1 identifies $\int a(t) D_t \,\mathrm{d}t$ with $\mathcal{D}_\psi^\theta$. Therefore

$$\mathcal{S}_\lambda = \mathcal{M}_{\theta,\phi} + (1-\lambda)\mathcal{D}_\psi^\theta - (2-\lambda)\mathcal{C}, \qquad \mathcal{C} := \int_{t_0}^T a(t) C_t \,\mathrm{d}t. \tag{A.11}$$

It is convenient to view the integrated quantities as squared norms in the Hilbert space $\mathcal{H}$ of measurable $\mathbf{h} = (\mathbf{h}_t)_t$ with

$$\|\mathbf{h}\|_{\mathcal{H}}^2 := \int_{t_0}^T a(t)\, \mathbb{E}_{\mathbf{X}_t \sim P_t^\theta} \|\mathbf{h}_t(\mathbf{X}_t)\|_2^2 \,\mathrm{d}t,$$

so that $\mathcal{M}_{\theta,\phi} = \|\mathbf{U}\|_{\mathcal{H}}^2$, $\mathcal{D}_\psi^\theta = \|\boldsymbol{\Delta}\|_{\mathcal{H}}^2$, and $\mathcal{C} = \langle \mathbf{U}, \boldsymbol{\Delta}\rangle_{\mathcal{H}}$. Expanding the square shows that (A.11) is the exact completed-square identity

$$\mathcal{S}_\lambda = \left\|\mathbf{U} - \frac{2-\lambda}{2}\boldsymbol{\Delta}\right\|_{\mathcal{H}}^2 - \frac{\lambda^2}{4}\|\boldsymbol{\Delta}\|_{\mathcal{H}}^2, \tag{A.12}$$

because $\left\|\mathbf{U}-\frac{2-\lambda}{2}\boldsymbol{\Delta}\right\|_{\mathcal{H}}^2=\mathcal{M}_{\theta,\phi}-(2-\lambda)\mathcal{C}+\frac{(2-\lambda)^2}{4}\mathcal{D}_\psi^\theta$ and $\frac{(2-\lambda)^2}{4}-\frac{\lambda^2}{4}=1-\lambda$. In particular, since $\mathcal{S}_\lambda\le\epsilon_2$ and $\mathcal{D}_\psi^\theta\le\epsilon_3$ are finite, (A.12) shows that $\mathbf{U}-\frac{2-\lambda}{2}\boldsymbol{\Delta}\in\mathcal{H}$, hence $\mathbf{U}\in\mathcal{H}$ and $\mathcal{M}_{\theta,\phi}<\infty$; no a priori finiteness of $\mathcal{M}_{\theta,\phi}$ is needed. The triangle inequality in $\mathcal{H}$ now gives

$$\begin{aligned}\sqrt{\mathcal{M}_{\theta,\phi}}=\|\mathbf{U}\|_{\mathcal{H}}&\le\left\|\mathbf{U}-\frac{2-\lambda}{2}\boldsymbol{\Delta}\right\|_{\mathcal{H}}+\frac{|2-\lambda|}{2}\|\boldsymbol{\Delta}\|_{\mathcal{H}}\\&=\sqrt{\mathcal{S}_\lambda+\frac{\lambda^2}{4}\mathcal{D}_\psi^\theta}+\frac{|2-\lambda|}{2}\sqrt{\mathcal{D}_\psi^\theta}\le\sqrt{\epsilon_2+\frac{\lambda^2}{4}\epsilon_3}+\frac{|2-\lambda|}{2}\sqrt{\epsilon_3},\end{aligned}\tag{A.13}$$

where the last step uses that both terms are nondecreasing in $\mathcal{S}_\lambda$ and $\mathcal{D}_\psi^\theta$. Squaring (A.13) and writing $\sqrt{\epsilon_2+\lambda^2\epsilon_3/4}=\frac{1}{2}\sqrt{\lambda^2\epsilon_3+4\epsilon_2}$ proves (3.2). □

A.3 Reverse-path KL and elementary total-variation tools

Use the increasing reverse-time variable $u=T-t$. For the terminal distribution π_T fixed in the main text, define the teacher reverse process by

$$\mathrm{d}\mathbf{Y}_u^\phi=\left[-b(T-u)\mathbf{Y}_u^\phi+a(T-u)\mathbf{s}_{T-u}^\phi(\mathbf{Y}_u^\phi)\right]\mathrm{d}u+\sqrt{a(T-u)}\,\mathrm{d}\overline{\mathbf{W}}_u,\qquad\mathbf{Y}_0^\phi\sim\pi_T.\tag{A.14}$$

The distribution of $\mathbf{Y}_{T-t}^\phi$ is denoted by $\widehat{P}_t^\phi$.

Assumption A.2 (Reverse-process regularity)**.** The teacher reverse SDE (A.14) has a weak solution whose distribution on path space is unique; denote the path measure by $\mathbb{Q}^\phi$. For $q\in\{\mathrm{data},\theta\}$, set $\mathbf{s}_t^q=\mathbf{s}_t^\star$ for $q=\mathrm{data}$ and $\mathbf{s}_t^q=\mathbf{s}_t^\theta$ for $q=\theta$. The reverse SDE obtained from (A.14) by replacing $\mathbf{s}_t^\phi$ with $\mathbf{s}_t^q$ and starting from P_T^q has a weak solution whose distribution is unique, with time-t marginal P_t^q; denote its path measure by $\mathbb{Q}^q$. Moreover, $P_T^q\ll\pi_T$ and $\kappa_T^q=\mathrm{KL}(P_T^q\|\pi_T)<\infty$; with the normalized drift difference $\mathbf{h}_u^q(\mathbf{y}):=\sqrt{a(T-u)}\,(\mathbf{s}_{T-u}^q-\mathbf{s}_{T-u}^\phi)(\mathbf{y})$, the likelihood process

$$L_u:=\frac{\mathrm{d}P_T^q}{\mathrm{d}\pi_T}(\mathbf{Y}_0^\phi)\,\exp\left\{\int_0^u\mathbf{h}_v^q(\mathbf{Y}_v^\phi)^\top\,\mathrm{d}\mathbf{W}_v^\phi-\frac{1}{2}\int_0^u\left\|\mathbf{h}_v^q(\mathbf{Y}_v^\phi)\right\|_2^2\,\mathrm{d}v\right\},\qquad u\in[0,T-t_0],$$

is a true $\mathbb{Q}^\phi$-martingale, so that $\mathbb{E}_{\mathbb{Q}^\phi}L_{T-t_0}=1$, and the corresponding energy is finite:

$$\int_{t_0}^T a(t)\mathbb{E}_{\mathbf{X}_t\sim P_t^q}\left\|\mathbf{s}_t^q(\mathbf{X}_t)-\mathbf{s}_t^\phi(\mathbf{X}_t)\right\|_2^2\,\mathrm{d}t<\infty.$$

Novikov's condition is one sufficient condition for the martingale requirement. We stress that the martingale requirement is a genuine change-of-measure condition: it is not implied by finite score energy or by the finite second moments of Assumption 3.1 alone, and it is the property used in the proof of Lemma A.4 below.

The finite second moments of Assumption 3.1 are not needed for the change-of-drift identity, but they make the terminal mismatches explicit for the natural Gaussian reference.

Lemma A.3 (Terminal mismatch under a Gaussian reference)**.** Let $q\in\{\mathrm{data},\theta\}$, let $P^q\in\mathcal{P}_2(\mathbb{R}^d)$ with $m_2^q:=\mathbb{E}_{P^q}\|\mathbf{X}\|_2^2$, and let $\pi_T=\gamma_T=\mathcal{N}(\mathbf{0},\sigma_T^2\mathbf{I}_d)$. Then $P_T^q\ll\gamma_T$ and

$$\kappa_T^q=\mathrm{KL}(P_T^q\|\gamma_T)\le\frac{\alpha_T^2\,m_2^q}{2\sigma_T^2}.$$

Proof. By construction $P_T^q=\int\mathcal{N}(\alpha_T\mathbf{x},\sigma_T^2\mathbf{I}_d)\,P^q(\,\mathrm{d}\mathbf{x})$ is a mixture of Gaussians with the covariance of π_T, hence absolutely continuous with respect to π_T. The Kullback–Leibler divergence is jointly convex, so

$$\mathrm{KL}(P_T^q\|\pi_T)\le\int\mathrm{KL}\big(\mathcal{N}(\alpha_T\mathbf{x},\sigma_T^2\mathbf{I}_d)\,\big\|\,\mathcal{N}(\mathbf{0},\sigma_T^2\mathbf{I}_d)\big)\,P^q(\,\mathrm{d}\mathbf{x})=\int\frac{\|\alpha_T\mathbf{x}\|_2^2}{2\sigma_T^2}\,P^q(\,\mathrm{d}\mathbf{x})=\frac{\alpha_T^2m_2^q}{2\sigma_T^2}.$$

□

For the variance-exploding schedule ($\alpha_T = 1$) the bound reads $m_2^q/(2\sigma_T^2)$, and for the Brownian schedule $\sigma_T^2 = T$ it is $m_2^q/(2T)$. This polynomial rate cannot be improved in general: for a Dirac distribution $P^q = \delta_{\mathbf{x}}$ the bound is an equality, $\mathrm{KL}(\mathcal{N}(\mathbf{x}, \sigma_T^2 \mathbf{I}_d) \| \mathcal{N}(\mathbf{0}, \sigma_T^2 \mathbf{I}_d)) = \|\mathbf{x}\|_2^2/(2\sigma_T^2)$, reflecting the fact that the variance-exploding forward process is not ergodic. For the variance-preserving schedule ($\alpha_T = e^{-T}$, $\sigma_T^2 = 1 - e^{-2T}$) the forward process is the ergodic Ornstein–Uhlenbeck process and the bound is exponentially small in T.

We now prove the change-of-drift bound used in Theorem 3.6.

Lemma A.4 (Reverse-path KL bound)**.** Under Assumption A.2, for $q \in \{\mathrm{data}, \theta\}$,

$$\mathrm{KL}\left(P_{t_0}^q \middle\| \widehat{P}_{t_0}^\phi\right) \le \kappa_T^q + \frac{1}{2}\int_{t_0}^T a(t)\mathbb{E}_{\mathbf{X}_t \sim P_t^q}\left[\left\|\mathbf{s}_t^q(\mathbf{X}_t) - \mathbf{s}_t^\phi(\mathbf{X}_t)\right\|_2^2\right] \mathrm{d}t. \tag{A.15}$$

Proof. Put $L := T - t_0$ and use the increasing reverse-time variable $u = T - t$. Let $\mathbb{Q}^q$ be the path measure of

$$\mathrm{d}\mathbf{Y}_u^q = \left[-b(T-u)\mathbf{Y}_u^q + a(T-u)\mathbf{s}_{T-u}^q(\mathbf{Y}_u^q)\right]\mathrm{d}u + \sqrt{a(T-u)}\,\mathrm{d}\mathbf{W}_u^q, \qquad \mathbf{Y}_0^q \sim P_T^q.$$

By exact time reversal, $\mathbf{Y}_u^q \sim P_{T-u}^q$. Let $\mathbb{Q}^\phi$ be the path measure of (A.14). Define the normalized drift difference by

$$\mathbf{h}_u^q(\mathbf{y}) := \sqrt{a(T-u)}\left(\mathbf{s}_{T-u}^q(\mathbf{y}) - \mathbf{s}_{T-u}^\phi(\mathbf{y})\right).$$

Multiplying $\mathbf{h}_u^q$ by the diffusion matrix $\sqrt{a(T-u)}\mathbf{I}_d$ gives the drift difference $a(T-u)(\mathbf{s}_{T-u}^q - \mathbf{s}_{T-u}^\phi)$.

Under the teacher path measure $\mathbb{Q}^\phi$, consider the likelihood process of Assumption A.2,

$$L_u := \frac{\mathrm{d}P_T^q}{\mathrm{d}\pi_T}(\mathbf{Y}_0^\phi)\exp\left\{\int_0^u \mathbf{h}_v^q(\mathbf{Y}_v^\phi)^\top \mathrm{d}\mathbf{W}_v^\phi - \frac{1}{2}\int_0^u \left\|\mathbf{h}_v^q(\mathbf{Y}_v^\phi)\right\|_2^2 \mathrm{d}v\right\}.$$

By the martingale requirement in Assumption A.2, $\mathbb{E}_{\mathbb{Q}^\phi} L_L = 1$, so L_L defines a new probability measure by $\mathrm{d}\mathbb{Q}^q = L_L\,\mathrm{d}\mathbb{Q}^\phi$. Girsanov's theorem says that under this new measure

$$\mathbf{W}_u^q := \mathbf{W}_u^\phi - \int_0^u \mathbf{h}_v^q(\mathbf{Y}_v^\phi)\,\mathrm{d}v$$

is Brownian motion. Substituting $\mathrm{d}\mathbf{W}_u^\phi = \mathrm{d}\mathbf{W}_u^q + \mathbf{h}_u^q\,\mathrm{d}u$ into the teacher SDE adds $\sqrt{a(T-u)}\mathbf{h}_u^q$ to the drift. The factor $\mathrm{d}P_T^q/\mathrm{d}\pi_T$ changes the initial distribution from π_T to P_T^q. Uniqueness in distribution from Assumption A.2 therefore identifies the changed path measure with $\mathbb{Q}^q$.

Taking the logarithm of L_L gives, in the direction $\mathbb{Q}^q \| \mathbb{Q}^\phi$,

$$\log\frac{\mathrm{d}\mathbb{Q}^q}{\mathrm{d}\mathbb{Q}^\phi} = \log\frac{\mathrm{d}P_T^q}{\mathrm{d}\pi_T}(\mathbf{Y}_0^q) + \int_0^L \mathbf{h}_u^q(\mathbf{Y}_u^q)^\top \mathrm{d}\mathbf{W}_u^\phi - \frac{1}{2}\int_0^L \|\mathbf{h}_u^q(\mathbf{Y}_u^q)\|_2^2\,\mathrm{d}u. \tag{A.16}$$

Under $\mathbb{Q}^q$, the two Brownian motions satisfy

$$\mathrm{d}\mathbf{W}_u^\phi = \mathrm{d}\mathbf{W}_u^q + \mathbf{h}_u^q(\mathbf{Y}_u^q)\,\mathrm{d}u.$$

Substituting this relation into (A.16) gives

$$\log\frac{\mathrm{d}\mathbb{Q}^q}{\mathrm{d}\mathbb{Q}^\phi} = \log\frac{\mathrm{d}P_T^q}{\mathrm{d}\pi_T}(\mathbf{Y}_0^q) + \int_0^L \mathbf{h}_u^q(\mathbf{Y}_u^q)^\top \mathrm{d}\mathbf{W}_u^q + \frac{1}{2}\int_0^L \|\mathbf{h}_u^q(\mathbf{Y}_u^q)\|_2^2\,\mathrm{d}u.$$

Take expectation under $\mathbb{Q}^q$. The stochastic integral has mean zero by Assumption A.2; hence

$$\begin{aligned}\mathrm{KL}(\mathbb{Q}^q\|\mathbb{Q}^\phi) &= \mathrm{KL}(P_T^q\|\pi_T) + \frac{1}{2}\mathbb{E}_{\mathbb{Q}^q}\int_0^L \|\mathbf{h}_u^q(\mathbf{Y}_u^q)\|_2^2\,\mathrm{d}u \\ &= \kappa_T^q + \frac{1}{2}\int_{t_0}^T a(t)\mathbb{E}_{\mathbf{X}_t\sim P_t^q}\left\|\mathbf{s}_t^q(\mathbf{X}_t) - \mathbf{s}_t^\phi(\mathbf{X}_t)\right\|_2^2\,\mathrm{d}t. \end{aligned}\tag{A.17}$$

The second equality uses $t = T - u$ and $\mathbf{Y}_u^q \sim P_{T-u}^q$.

To pass from paths to endpoints, let $\mathbf{Z} = \mathbf{Y}_L$, let $R = \mathrm{d}\mathbb{Q}^q / \mathrm{d}\mathbb{Q}^\phi$, and note that the likelihood ratio of the endpoint distributions is $r(\mathbf{Z}) = \mathbb{E}_{\mathbb{Q}^\phi}[R \mid \mathbf{Z}]$. Since $x \mapsto x \log x$ is convex, conditional Jensen gives

$$\mathbb{E}_{\mathbb{Q}^\phi}[r(\mathbf{Z}) \log r(\mathbf{Z})] \le \mathbb{E}_{\mathbb{Q}^\phi}[R \log R].$$

The left side is the endpoint KL divergence and the right side is the path KL divergence. Combining this data-processing inequality with (A.17) proves (A.15). The proof is complete. □

Lemma A.5 (Pinsker's inequality)**.** For any probability measures $P \ll Q$,

$$\mathrm{TV}(P, Q) \le \sqrt{\frac{1}{2} \mathrm{KL}(P \| Q)}. \tag{A.18}$$

Proof. Let A be a positive set for the signed measure $P - Q$. Then $P(A) - Q(A) = \mathrm{TV}(P, Q)$. Mapping a sample to the indicator of A and using the data-processing argument from the preceding proof gives

$$\mathrm{KL}(P \| Q) \ge p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}, \qquad p := P(A), \quad q := Q(A).$$

For fixed $q \in (0, 1)$, define the binary divergence on the right by $k_q(p)$. Direct differentiation gives

$$k_q(q) = k_q'(q) = 0, \qquad k_q''(p) = \frac{1}{p(1 - p)} \ge 4.$$

Integrating the second-derivative inequality twice from q to p shows that $k_q(p) \ge 2(p - q)^2$. The boundary cases follow by taking limits. Hence $\mathrm{KL}(P \| Q) \ge 2\,\mathrm{TV}(P, Q)^2$, which is equivalent to (A.18). The proof is complete. □

A.4 Proof of Theorem 3.6

Proof. Apply Lemma A.4 first with $q = \mathrm{data}$. Assumption 3.2 bounds the energy integral by ϵ_1, so Lemma A.5 gives

$$\mathrm{TV}\left(P_{t_0}^{\mathrm{data}}, \widehat{P}_{t_0}^{\phi}\right) \le \sqrt{\frac{1}{2} \kappa_T^{\mathrm{data}} + \frac{1}{4} \epsilon_1}. \tag{A.19}$$

Apply the same two lemmas with $q = \theta$. The energy integral is $\mathcal{M}_{\theta,\phi}$, which Lemma 3.5 bounds by $K_\lambda(\epsilon_2, \epsilon_3)$. Therefore

$$\mathrm{TV}\left(P_{t_0}^{\theta}, \widehat{P}_{t_0}^{\phi}\right) \le \sqrt{\frac{1}{2} \kappa_T^{\theta} + \frac{1}{4} K_\lambda(\epsilon_2, \epsilon_3)}. \tag{A.20}$$

Finally, total variation satisfies the triangle inequality. This follows immediately from its L^1 form:

$$\frac{1}{2} \int |\,\mathrm{d}P - \mathrm{d}R| \le \frac{1}{2} \int |\,\mathrm{d}P - \mathrm{d}Q| + \frac{1}{2} \int |\,\mathrm{d}Q - \mathrm{d}R|\,.$$

Using the intermediate distribution $\widehat{P}_{t_0}^{\phi}$ in this inequality, and then substituting (A.19) and (A.20), proves (3.3). □

A.5 From the noised comparison to the clean output

Theorem 3.6 compares the two *noised* laws at the early-stopping time $t_0 > 0$. The following coupling lemma transfers such a comparison to the clean laws, at the price of a Wasserstein radius proportional to the noise level. It compares two smoothed laws, which is the situation of Theorem 3.6, rather than a sampler output with a smoothed target as in the diffusion-sampling literature (Lee et al., 2022; Chen et al., 2022; 2023a).

Lemma A.6 (Gaussian smoothing to a Wasserstein–TV comparison)**.** Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, let $\tau > 0$, put $\mu_\tau = \mu * \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}_d)$, $\nu_\tau = \nu * \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}_d)$, and let $\delta = \mathrm{TV}(\mu_\tau, \nu_\tau)$. Then there exists $R \in \mathcal{P}_2(\mathbb{R}^d)$ such that

$$\mathrm{TV}(\mu, R) \le \delta, \qquad W_2(R, \nu) \le 2\sqrt{d}\,\tau.$$

Proof. Let Γ be a maximal coupling of the smoothed laws: $(\mathbf{U}, \mathbf{V}) \sim \Gamma$ with $\mathbf{U} \sim \mu_\tau$, $\mathbf{V} \sim \nu_\tau$, and $\mathbb{P}(\mathbf{U} \neq \mathbf{V}) = \delta$. Let $K_\mu(\mathbf{u}, \mathrm{d}\mathbf{x})$ be the regular conditional law of a clean μ-sample given its noised observation $\mathbf{u}$, that is, the conditional law of $\mathbf{X}$ given $\mathbf{X} + \tau\boldsymbol{\xi} = \mathbf{u}$ when $\mathbf{X} \sim \mu$ and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ is independent of $\mathbf{X}$; define $K_\nu(\mathbf{v}, \mathrm{d}\mathbf{y})$ analogously. Construct $(\mathbf{X}, \mathbf{U}, \mathbf{V}, \mathbf{Y})$ with joint law

$$\mathbb{P}(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{u}, \mathrm{d}\mathbf{v}, \mathrm{d}\mathbf{y}) = \Gamma(\mathrm{d}\mathbf{u}, \mathrm{d}\mathbf{v})\, K_\mu(\mathbf{u}, \mathrm{d}\mathbf{x})\, K_\nu(\mathbf{v}, \mathrm{d}\mathbf{y}).$$

By construction, $(\mathbf{X}, \mathbf{U})$ has the clean–noised joint law of μ and $(\mathbf{Y}, \mathbf{V})$ has that of ν. In particular $\mathbf{X} \sim \mu$, $\mathbf{Y} \sim \nu$, and $\mathbb{E}\,\|\mathbf{X} - \mathbf{U}\|_2^2 = \mathbb{E}\,\|\mathbf{Y} - \mathbf{V}\|_2^2 = d\tau^2$, because $\mathbf{U} - \mathbf{X}$ and $\mathbf{V} - \mathbf{Y}$ are $\mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}_d)$ distributed. Note that the two Gaussian noises $\mathbf{U} - \mathbf{X}$ and $\mathbf{V} - \mathbf{Y}$ are in general dependent; only the two clean–noised pair laws are preserved.

Let $G = \{\mathbf{U} = \mathbf{V}\}$, define $\mathbf{H} = \mathbf{X}$ on G and $\mathbf{H} = \mathbf{Y}$ on G^c, and let R be the law of $\mathbf{H}$. Then $\mathbb{E}\,\|\mathbf{H}\|_2^2 \le \mathbb{E}\,\|\mathbf{X}\|_2^2 + \mathbb{E}\,\|\mathbf{Y}\|_2^2 < \infty$, so $R \in \mathcal{P}_2(\mathbb{R}^d)$. The coupling $(\mathbf{X}, \mathbf{H})$ gives

$$\mathrm{TV}(\mu, R) \le \mathbb{P}(\mathbf{X} \neq \mathbf{H}) \le \mathbb{P}(G^c) = \delta.$$

For the coupling $(\mathbf{H}, \mathbf{Y})$, the event G^c carries no transport cost because $\mathbf{H} = \mathbf{Y}$ there, while on G we have $\mathbf{U} = \mathbf{V}$, so

$$W_2^2(R, \nu) \le \mathbb{E}\,\|\mathbf{H} - \mathbf{Y}\|_2^2 = \mathbb{E}\big[\mathbb{1}_G\,\|\mathbf{X} - \mathbf{Y}\|_2^2\big] \le 2\mathbb{E}\big[\mathbb{1}_G\,\|\mathbf{X} - \mathbf{U}\|_2^2\big] + 2\mathbb{E}\big[\mathbb{1}_G\,\|\mathbf{V} - \mathbf{Y}\|_2^2\big] \le 4d\tau^2.$$

The expectations are unconditional and the indicator is simply dropped in the last step; no normalization by $1 - \delta$ and no conditional distribution on G is used. Taking square roots completes the proof. □

The factor 2 arises from concatenating the two displacement estimates through the common noised point on G; it is an upper bound and is not claimed to be sharp (for $\mu = \nu$ one may take $R = \mu$).

Proof of the clean-output guarantee (3.5). Let $S(\mathbf{x}) = \mathbf{x}/\alpha_{t_0}$. Since $\mathbf{X}_{t_0} = \alpha_{t_0}\mathbf{X}_0 + \sigma_{t_0}\xi$ with ξ independent of $\mathbf{X}_0$, we have $S_{\#}P_{t_0}^q = P^q * \mathcal{N}(\mathbf{0}, \tau_0^2\mathbf{I}_d)$ for $q \in \{\theta, \mathrm{data}\}$, where $\tau_0 = \sigma_{t_0}/\alpha_{t_0}$. Because S is the same measurable bijection for both laws,

$$\mathrm{TV}\big(P^\theta * \mathcal{N}(\mathbf{0}, \tau_0^2\mathbf{I}_d),\, P^{\mathrm{data}} * \mathcal{N}(\mathbf{0}, \tau_0^2\mathbf{I}_d)\big) = \mathrm{TV}\big(P_{t_0}^\theta, P_{t_0}^{\mathrm{data}}\big) \le \min\{1, \Delta_{t_0,T}\},$$

by Theorem 3.6 and $\mathrm{TV} \le 1$. Assumption 3.1 gives $P^\theta, P^{\mathrm{data}} \in \mathcal{P}_2(\mathbb{R}^d)$, so Lemma A.6 with $\mu = P^\theta$, $\nu = P^{\mathrm{data}}$, and $\tau = \tau_0$ yields $R_{\theta,t_0} \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mathrm{TV}(P^\theta, R_{\theta,t_0}) \le \Delta_{t_0,T}$ and $W_2(R_{\theta,t_0}, P^{\mathrm{data}}) \le 2\sqrt{d}\,\tau_0$. This proves (3.5); in particular $\inf\{\mathrm{TV}(P^\theta, R) : W_2(R, P^{\mathrm{data}}) \le 2\sqrt{d}\,\tau_0\} \le \min\{1, \Delta_{t_0,T}\}$. For the Brownian schedule, $\tau_0 = \sqrt{t_0}$, so the Wasserstein radius is $2\sqrt{dt_0}$, and Lemma A.3 with $\pi_T = \mathcal{N}(\mathbf{0}, T\mathbf{I}_d)$ gives $\kappa_T^q \le m_2^q/(2T)$, so that $\Delta_{t_0,T} \le \frac{1}{2}\sqrt{K_\lambda(\epsilon_2, \epsilon_3) + m_2^\theta/T} + \frac{1}{2}\sqrt{\epsilon_1 + m_2^{\mathrm{data}}/T}$. □

B ReLU approximation of the student score

The section proves Theorem 3.7. We use the Brownian variance-exploding schedule $\alpha_t = 1$, $\sigma_t^2 = t$, and $a(t) = 1$. The proof first gives fixed-time estimates needed for the ratio argument and then uses one set of Monte-Carlo centers for every $t \in [t_0, T]$.

B.1 Fixed-time Monte-Carlo ingredients

Fix $t \in [t_0, T]$ throughout this subsection. Let

$$\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \qquad \mathbf{X}_t = \alpha_t\mathbf{g}_\theta(\mathbf{Z}) + \sigma_t\xi, \qquad \xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d),$$

where $\mathbf{Z}$ and ξ are independent. Define

$$c_t := (2\pi\sigma_t^2)^{-d/2}, \qquad \varphi_t(\mathbf{x}; \mathbf{y}) := c_t \exp\left(-\frac{\|\mathbf{x} - \alpha_t\mathbf{y}\|_2^2}{2\sigma_t^2}\right).$$

Lemma B.1 (Density and score of the noised student model)**.** The density and score numerator of P_t^θ are

$$p_t^\theta(\mathbf{x}) = \mathbb{E}_{\mathbf{Z}}\left[\varphi_t(\mathbf{x}; \mathbf{g}_\theta(\mathbf{Z}))\right], \tag{B.1}$$

$$\mathbf{u}_t^\theta(\mathbf{x}) = \mathbb{E}_{\mathbf{Z}}\left[-\frac{\mathbf{x} - \alpha_t \mathbf{g}_\theta(\mathbf{Z})}{\sigma_t^2}\varphi_t(\mathbf{x}; \mathbf{g}_\theta(\mathbf{Z}))\right]. \tag{B.2}$$

Moreover,

$$\mathbf{s}_t^\theta(\mathbf{x}) = \frac{\mathbf{u}_t^\theta(\mathbf{x})}{p_t^\theta(\mathbf{x})} = -\frac{\mathbf{x} - \alpha_t \mathbb{E}[\mathbf{g}_\theta(\mathbf{Z}) \mid \mathbf{X}_t = \mathbf{x}]}{\sigma_t^2}. \tag{B.3}$$

Proof. Conditionally on $\mathbf{Z} = \mathbf{z}$, $\mathbf{X}_t$ has Gaussian density $\varphi_t(\mathbf{x}; \mathbf{g}_\theta(\mathbf{z}))$. Averaging over $\mathbf{Z}$ proves (B.1). Differentiating the Gaussian kernel gives

$$\nabla_{\mathbf{x}} \varphi_t(\mathbf{x}; \mathbf{y}) = -\frac{\mathbf{x} - \alpha_t \mathbf{y}}{\sigma_t^2}\varphi_t(\mathbf{x}; \mathbf{y}).$$

As in the proof of Lemma A.1, the derivative is uniformly bounded in $\mathbf{y}$, so differentiation may pass through the expectation. Hence $\nabla_{\mathbf{x}} p_t^\theta = \mathbf{u}_t^\theta$. Since $p_t^\theta(\mathbf{x}) > 0$, division by the density gives the first equality in (B.3). Bayes' formula gives the second equality as in (A.6). □

For independent centers $\mathbf{Z}_1, \ldots, \mathbf{Z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, put $\mathbf{y}_i = \mathbf{g}_\theta(\mathbf{Z}_i)$, and define

$$\widehat{p}_{n,t}(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n \varphi_t(\mathbf{x}; \mathbf{y}_i),$$

$$\widehat{\mathbf{u}}_{n,t}(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n \left(-\frac{\mathbf{x} - \alpha_t \mathbf{y}_i}{\sigma_t^2}\right)\varphi_t(\mathbf{x}; \mathbf{y}_i).$$

For $R > 0$ and a density threshold $\varepsilon_{\text{low}} > 0$, define the good region

$$G_t(R, \varepsilon_{\text{low}}) := \left\{\mathbf{x} : \|\mathbf{x}\|_\infty \le R, \quad p_t^\theta(\mathbf{x}) \ge 2\varepsilon_{\text{low}}\right\}.$$

The complement consists of the tail outside the cube and the low-density part inside the cube.

B.2 Uniform concentration of the Monte-Carlo numerator and denominator

We begin with elementary bounds for one Gaussian feature.

Lemma B.2 (Gaussian feature envelopes)**.** There is a numerical constant $C_0 \ge 1$ such that, with

$$B_{0,t} := c_t, \qquad B_{1,t} := C_0 c_t \sigma_t^{-1}, \qquad \Lambda_t := C_0 c_t \sqrt{d}\left(\sigma_t^{-1} + \sigma_t^{-2}\right).$$

the following statements hold for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$0 \le \varphi_t(\mathbf{x}; \mathbf{y}) \le B_{0,t},$$

$$\left\|-\frac{\mathbf{x} - \alpha_t \mathbf{y}}{\sigma_t^2}\varphi_t(\mathbf{x}; \mathbf{y})\right\|_\infty \le B_{1,t}.$$

Furthermore, the density feature and each coordinate of the numerator feature are Λ_t-Lipschitz with respect to $\|\cdot\|_\infty$.

Proof. Put $\mathbf{r} = (\mathbf{x} - \alpha_t \mathbf{y})/\sigma_t$. The density bound follows from $e^{-\|\mathbf{r}\|_2^2/2} \le 1$. For the numerator, its j-th coordinate has absolute value

$$\frac{c_t}{\sigma_t}|r_j| e^{-\|\mathbf{r}\|_2^2/2} \le \frac{c_t}{\sigma_t}\sup_{\mathbf{v}\in\mathbb{R}^d}\|\mathbf{v}\|_2 e^{-\|\mathbf{v}\|_2^2/2} \le C\frac{c_t}{\sigma_t}.$$

For the Lipschitz bounds, recall the multivariable mean-value inequality

$$|F(\mathbf{x}) - F(\mathbf{x}')| \le \sup_{\boldsymbol{\xi}} \|\nabla F(\boldsymbol{\xi})\|_1 \|\mathbf{x} - \mathbf{x}'\|_\infty.$$

The density gradient satisfies

$$\|\nabla_{\mathbf{x}}\varphi_t(\mathbf{x};\mathbf{y})\|_2 = \frac{c_t}{\sigma_t}\|\mathbf{r}\|_2 e^{-\|\mathbf{r}\|_2^2/2} \le Cc_t\sigma_t^{-1}.$$

Since $\|\mathbf{v}\|_1 \le \sqrt{d}\,\|\mathbf{v}\|_2$, its ℓ_1-gradient norm is at most $Cc_t\sqrt{d}\sigma_t^{-1}$.

For the j-th numerator feature $H_j(\mathbf{x};\mathbf{y}) = -(x_j - \alpha_t y_j)\sigma_t^{-2}\varphi_t(\mathbf{x};\mathbf{y})$, direct differentiation gives

$$\nabla_{\mathbf{x}}H_j(\mathbf{x};\mathbf{y}) = \frac{c_t}{\sigma_t^2}(-\mathbf{e}_j + r_j\mathbf{r})\,e^{-\|\mathbf{r}\|_2^2/2}.$$

Both $e^{-\|\mathbf{r}\|_2^2/2}$ and $|r_j|\,\|\mathbf{r}\|_2 e^{-\|\mathbf{r}\|_2^2/2}$ are bounded by numerical constants. Thus $\|\nabla H_j\|_2 \le Cc_t\sigma_t^{-2}$, and hence $\|\nabla H_j\|_1 \le Cc_t\sqrt{d}\sigma_t^{-2}$. Increasing C_0 proves all claims. The proof is complete. □

Lemma B.3 (Uniform Monte-Carlo concentration on a spatial cube)**.** Fix $R > 0$, $\tau \in (0,1)$, and $\delta \in (0,1)$, and let $B_t := B_{0,t} \vee B_{1,t}$. If

$$n \ge C\frac{B_t^2}{\tau^2}\left[d\log\left(1 + \frac{CR\Lambda_t}{\tau}\right) + \log\left(\frac{2(d+1)}{\delta}\right)\right], \tag{B.4}$$

then, with probability at least $1-\delta$,

$$\sup_{\|\mathbf{x}\|_\infty \le R}\left|\widehat{p}_{n,t}(\mathbf{x}) - p_t^\theta(\mathbf{x})\right| \le \tau, \tag{B.5}$$

$$\max_{j\in[d]}\sup_{\|\mathbf{x}\|_\infty \le R}\left|[\widehat{\mathbf{u}}_{n,t}(\mathbf{x}) - \mathbf{u}_t^\theta(\mathbf{x})]_j\right| \le \tau. \tag{B.6}$$

Proof. We divide the proof into a finite-grid step and a Lipschitz-extension step.

Step 1: concentration at one fixed point. For fixed $\mathbf{x}$, the random variables $\varphi_t(\mathbf{x};\mathbf{g}_\theta(\mathbf{z}_i))$ are independent, lie in $[0, B_{0,t}]$, and have mean $p_t^\theta(\mathbf{x})$. Hoeffding's inequality therefore gives

$$\mathbb{P}\left(\left|\widehat{p}_{n,t}(\mathbf{x}) - p_t^\theta(\mathbf{x})\right| > \frac{\tau}{2}\right) \le 2\exp\left(-c\frac{n\tau^2}{B_{0,t}^2}\right).$$

For a numerator coordinate, the corresponding variables lie in $[-B_{1,t}, B_{1,t}]$. The same inequality, with a changed numerical constant, gives

$$\mathbb{P}\left(\left|[\widehat{\mathbf{u}}_{n,t}(\mathbf{x}) - \mathbf{u}_t^\theta(\mathbf{x})]_j\right| > \frac{\tau}{2}\right) \le 2\exp\left(-c\frac{n\tau^2}{B_{1,t}^2}\right).$$

Step 2: a finite net of the cube. Set $\eta = \tau/(4\Lambda_t)$, and take a maximal η-separated subset $\mathcal{N}_\eta$ of the cube in ℓ_∞. Maximality makes it an η-net: otherwise a point farther than η from the set could be added. The ℓ_∞-balls of radius $\eta/2$ around its points are disjoint and lie inside $[-R-\eta/2, R+\eta/2]^d$. Comparing their volumes gives

$$|\mathcal{N}_\eta| \le \left(1 + \frac{2R}{\eta}\right)^d \le \left(1 + \frac{8R\Lambda_t}{\tau}\right)^d. \tag{B.7}$$

The argument also covers $\eta > 2R$, when a one-point net is enough.

Step 3: union bound on the net. There are $d+1$ scalar quantities to control at each net point. The union bound and (B.7) show that the probability of any failure on the net is at most

$$2(d+1)\,|\mathcal{N}_\eta|\exp\left(-c\frac{n\tau^2}{B_t^2}\right).$$

Condition (B.4) makes this expression at most δ.

Step 4: extend from the net to every point. On the resulting event, take any $\mathbf{x}$ in the cube and a net point $\mathbf{x}^\sharp$ satisfying $\left\|\mathbf{x} - \mathbf{x}^\sharp\right\|_\infty \le \eta$. Each individual feature is Λ_t-Lipschitz by Lemma B.2; therefore its empirical average and its expectation are also Λ_t-Lipschitz. Hence

$$\left|\widehat{p}_{n,t}(\mathbf{x}) - p_t^\theta(\mathbf{x})\right| \le \Lambda_t\eta + \frac{\tau}{2} + \Lambda_t\eta = \tau.$$

The identical argument applies to every numerator coordinate. This proves (B.5)–(B.6). The proof is complete. □

B.3 From uniform concentration to an L^2 score bound

The next lemma contains the good-region ratio calculation, the small-density volume argument, and the Gaussian tail calculation.

Lemma B.4 (Ratio conversion and truncation)**.** Assume $\|\mathbf{g}_\theta(\mathbf{z})\|_\infty \le B_g$ for every $\mathbf{z} \in \mathbb{R}^d$. Fix $r \ge 3$, and set

$$R := \alpha_t B_g + \sigma_t r, \qquad M := \frac{R + \alpha_t B_g}{\sigma_t^2}, \qquad \Gamma_t := d^2(1 + \alpha_t^2 B_g^2)(1 + \sigma_t^{-4}).$$

Suppose (B.5) and (B.6) hold with tolerance τ, where $0 < \tau \le \varepsilon_{\text{low}}$. Define the clipped score by

$$\widehat{\mathbf{s}}_{n,t}(\mathbf{x}) := \text{clip}_M\left(\frac{\widehat{\mathbf{u}}_{n,t}(\mathbf{x})}{\widehat{p}_{n,t}(\mathbf{x}) \vee \varepsilon_{\text{low}}}\right).$$

where clipping is coordinatewise. Then

$$\begin{aligned}&\mathbb{E}_{\mathbf{X}_t \sim P_t^\theta} \left\|\widehat{\mathbf{s}}_{n,t}(\mathbf{X}_t) - \mathbf{s}_t^\theta(\mathbf{X}_t)\right\|_2^2 \\ &\quad \le C\left[d\left(\frac{\tau}{\varepsilon_{\text{low}}} + \frac{B_{1,t}\tau}{\varepsilon_{\text{low}}^2}\right)^2 + dM^2\varepsilon_{\text{low}}(2R)^d + \Gamma_t e^{-r^2/8}\right]. \end{aligned} \tag{B.8}$$

Proof. **Step 1: the ratio on the good region.** Let $\mathbf{x} \in G_t(R, \varepsilon_{\text{low}})$. Since $p_t^\theta(\mathbf{x}) \ge 2\varepsilon_{\text{low}}$ and $\left|\widehat{p}_{n,t}(\mathbf{x}) - p_t^\theta(\mathbf{x})\right| \le \tau \le \varepsilon_{\text{low}}$,

$$\widehat{p}_{n,t}(\mathbf{x}) \ge \varepsilon_{\text{low}}.$$

Thus the lower clipping of the empirical denominator is inactive at $\mathbf{x}$. For each coordinate j, add and subtract $[\mathbf{u}_t^\theta(\mathbf{x})]_j / \widehat{p}_{n,t}(\mathbf{x})$ to obtain

$$\begin{aligned}\left|\frac{[\widehat{\mathbf{u}}_{n,t}(\mathbf{x})]_j}{\widehat{p}_{n,t}(\mathbf{x})} - \frac{[\mathbf{u}_t^\theta(\mathbf{x})]_j}{p_t^\theta(\mathbf{x})}\right| &\le \frac{\left|[\widehat{\mathbf{u}}_{n,t}(\mathbf{x}) - \mathbf{u}_t^\theta(\mathbf{x})]_j\right|}{\widehat{p}_{n,t}(\mathbf{x})} \\ &\quad + \left|[\mathbf{u}_t^\theta(\mathbf{x})]_j\right| \frac{\left|\widehat{p}_{n,t}(\mathbf{x}) - p_t^\theta(\mathbf{x})\right|}{\widehat{p}_{n,t}(\mathbf{x}) p_t^\theta(\mathbf{x})} \\ &\le \frac{\tau}{\varepsilon_{\text{low}}} + \frac{B_{1,t}\tau}{\varepsilon_{\text{low}}^2}.\end{aligned}$$

The last denominator could be sharpened by a factor of two, which is not needed. Summing the squared coordinate bounds gives

$$\sup_{\mathbf{x} \in G_t(R, \varepsilon_{\text{low}})} \left\|\widehat{\mathbf{s}}_{n,t}(\mathbf{x}) - \mathbf{s}_t^\theta(\mathbf{x})\right\|_2^2 \le d\left(\frac{\tau}{\varepsilon_{\text{low}}} + \frac{B_{1,t}\tau}{\varepsilon_{\text{low}}^2}\right)^2. \tag{B.9}$$

Coordinatewise clipping does not increase this error. Indeed, (B.3) and the student-model bound imply that, whenever $\|\mathbf{x}\|_\infty \le R$,

$$\left\|\mathbf{s}_t^\theta(\mathbf{x})\right\|_\infty \le \frac{R + \alpha_t B_g}{\sigma_t^2} = M. \tag{B.10}$$

Projection onto $[-M, M]^d$ is nonexpansive around every point of that cube.

Step 2: the small-density part of the cube. Let

$$A := \{\|\mathbf{X}_t\|_\infty \le R,\ p_t^\theta(\mathbf{X}_t) < 2\varepsilon_{\text{low}}\}.$$

Since $\mathbf{X}_t$ has density p_t^θ,

$$\begin{aligned}\mathbb{P}(A) &= \int_{\{\|\mathbf{x}\|_\infty \le R,\, p_t^\theta(\mathbf{x}) < 2\varepsilon_{\text{low}}\}} p_t^\theta(\mathbf{x})\, \mathrm{d}\mathbf{x} \\ &\le 2\varepsilon_{\text{low}}(2R)^d. \end{aligned} \tag{B.11}$$

On this event, both the true score and the clipped empirical score have coordinate magnitude at most M, by (B.10). Their squared difference is therefore at most $4dM^2$. Multiplying this bound by (B.11) gives a contribution at most $8dM^2\varepsilon_{\text{low}}(2R)^d$.

Step 3: tail probability of $\mathbf{X}_t$. From the coupling $\mathbf{X}_t = \alpha_t \mathbf{g}_\theta(\mathbf{Z}) + \sigma_t \xi$ and the student-model bound,

$$\|\mathbf{X}_t\|_\infty \le \alpha_t B_g + \sigma_t \|\xi\|_\infty .$$

Consequently,

$$\mathbb{P}(\|\mathbf{X}_t\|_\infty > R) \le \mathbb{P}(\|\xi\|_\infty > r) \le 2de^{-r^2/2}, \tag{B.12}$$

where the last step is the union bound and the standard one-dimensional Gaussian tail inequality.

Step 4: second moment of the true score on the tail. By (B.3) and Jensen's inequality,

$$\begin{aligned}\left\|\mathbf{s}_t^\theta(\mathbf{X}_t)\right\|_2^2 &\le \frac{2\|\mathbf{X}_t\|_2^2 + 2\alpha_t^2 dB_g^2}{\sigma_t^4} \\ &\le \frac{C\alpha_t^2 dB_g^2}{\sigma_t^4} + \frac{C}{\sigma_t^2}\|\xi\|_2^2 .\end{aligned}$$

For a scalar $G \sim \mathcal{N}(0,1)$, integration by parts gives the following bound. If $\varphi(x) = (2\pi)^{-1/2}e^{-x^2/2}$, then

$$\begin{aligned}\mathbb{E}[G^2 \mathbb{1}\{|G| > r\}] &= 2\int_r^\infty x^2\varphi(x)\,\mathrm{d}x \\ &= 2r\varphi(r) + 2\int_r^\infty \varphi(x)\,\mathrm{d}x \\ &\le C(1+r^2)e^{-r^2/2}.\end{aligned}$$

In the second line we used $\varphi'(x) = -x\varphi(x)$; in the last line we used $\int_r^\infty \varphi(x)\,\mathrm{d}x \le \varphi(r)/r$ for $r > 0$.

We now expand the vector calculation without suppressing the indices. The union bound for indicators says

$$\mathbb{1}\{\|\xi\|_\infty > r\} \le \sum_{k=1}^d \mathbb{1}\{|\Xi_k| > r\}.$$

Therefore

$$\mathbb{E}\left[\|\xi\|_2^2\,\mathbb{1}\{\|\xi\|_\infty > r\}\right] \le \sum_{j=1}^d\sum_{k=1}^d \mathbb{E}\left[\Xi_j^2 \mathbb{1}\{|\Xi_k| > r\}\right].$$

When $j = k$, the scalar second-moment bound above applies. When $j \ne k$, independence gives

$$\mathbb{E}[\Xi_j^2 \mathbb{1}\{|\Xi_k| > r\}] = \mathbb{E}[\Xi_j^2]\mathbb{P}(|\Xi_k| > r) \le 2e^{-r^2/2}.$$

There are d diagonal terms and $d(d-1)$ off-diagonal terms, so

$$\mathbb{E}\left[\|\xi\|_2^2\,\mathbb{1}\{\|\xi\|_\infty > r\}\right] \le Cd(d+r^2)e^{-r^2/2} \le Cd^2e^{-r^2/4}. \tag{B.13}$$

The last inequality uses $r \ge 3$ and $r^2e^{-r^2/2} \le Ce^{-r^2/4}$. Combining (B.12)– (B.13) gives

$$\mathbb{E}\left[\left\|\mathbf{s}_t^\theta(\mathbf{X}_t)\right\|_2^2\,\mathbb{1}\{\|\mathbf{X}_t\|_\infty > R\}\right] \le C\Gamma_t e^{-r^2/8}.$$

The empirical score is clipped at level M, so its tail contribution is at most $dM^2\mathbb{P}(\|\mathbf{X}_t\|_\infty > R)$. From $M = 2\alpha_t B_g\sigma_t^{-2} + r\sigma_t^{-1}$ and (B.12), this quantity is also at most $C\Gamma_t e^{-r^2/8}$.

Finally, decompose the expectation over the good region, the event A, and the spatial tail. Apply (B.9), the small-density bound from Step 2, and the two tail estimates from Step 4. Absorbing numerical constants proves (B.8). The proof is complete. □

Fix $\epsilon_t, \delta_t \in (0,1)$ and assume $\|\mathbf{g}_\theta(\mathbf{z})\|_\infty \le B_g$. We now make all parameters explicit. Let $C_1 \ge 1$ be a sufficiently large numerical constant and $c_1 \in (0,1)$ a sufficiently small one, and define

$$\begin{aligned}
\Gamma_t &:= d^2(1+\alpha_t^2 B_g^2)(1+\sigma_t^{-4}),\\
r_t &:= \max\left\{3, \sqrt{8\log\left(\frac{C_1\Gamma_t}{\epsilon_t}\right)}\right\},\\
R_t &:= \alpha_t B_g + \sigma_t r_t, \qquad M_t := \frac{R_t + \alpha_t B_g}{\sigma_t^2},\\
\varepsilon_{\mathrm{low},t} &:= \min\left\{\frac{1}{2}, \frac{\epsilon_t}{64 d M_t^2 (2R_t)^d}\right\},\\
\tau_t &:= c_1 \min\left\{\varepsilon_{\mathrm{low},t}, \frac{\varepsilon_{\mathrm{low},t}^2\sqrt{\epsilon_t}}{\sqrt{d}(B_{1,t}+\varepsilon_{\mathrm{low},t})}\right\}.
\end{aligned} \tag{B.14}$$

The threshold is in $(0,1)$, and $0 < \tau_t < 1$. Define the required sample size by

$$n \ge C\frac{B_t^2}{\tau_t^2}\left[d\log\left(1+\frac{CR_t\Lambda_t}{\tau_t}\right) + \log\left(\frac{2(d+1)}{\delta_t}\right)\right]. \tag{B.15}$$

Proposition B.5 (Fixed-time Monte-Carlo score)**.** Choose the parameters as in (B.14). If (B.15) holds, then, with probability at least $1-\delta_t$,

$$\mathbb{E}_{\mathbf{X}_t\sim P_t^\theta}\left\|\widehat{\mathbf{s}}_{n,t}(\mathbf{X}_t) - \mathbf{s}_t^\theta(\mathbf{X}_t)\right\|_2^2 \le C\epsilon_t. \tag{B.16}$$

Proof. Lemma B.3 and (B.15) give the uniform event with probability at least $1-\delta_t$. On this event, the choice of τ_t implies

$$d\left(\frac{\tau_t}{\varepsilon_{\mathrm{low},t}} + \frac{B_{1,t}\tau_t}{\varepsilon_{\mathrm{low},t}^2}\right)^2 \le C\epsilon_t.$$

To check the small-density term, let $A = dM_t^2(2R_t)^d$. If the second argument of the minimum defining $\varepsilon_{\mathrm{low},t}$ is selected, then $A\varepsilon_{\mathrm{low},t} = \epsilon_t/64$. If $1/2$ is selected, the inequality $\epsilon_t/(64A) \ge 1/2$ implies $A\varepsilon_{\mathrm{low},t} \le \epsilon_t/64$. Finally, the choice of r_t and a sufficiently large C_1 give $\Gamma_t e^{-r_t^2/8} \le C\epsilon_t$. Substitution into (B.8) proves (B.16). The proof is complete. □

B.4 One set of Monte-Carlo centers for all times

The fixed-time proposition does not by itself provide one surrogate network for a continuum of times. We now prove the required uniform statement. Under the Brownian variance-exploding schedule, put

$$\kappa(\mathbf{x},t;\mathbf{y}) := \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2t}\right), \qquad \mathbf{v}(\mathbf{x},t;\mathbf{y}) := -\frac{\mathbf{x}-\mathbf{y}}{t}\kappa(\mathbf{x},t;\mathbf{y}).$$

The Gaussian normalizing factor cancels in a score ratio. Thus, if

$$\overline{p}(\mathbf{x},t) := \mathbb{E}_{\mathbf{Z}}\kappa(\mathbf{x},t;\mathbf{g}_\theta(\mathbf{Z})), \qquad \overline{\mathbf{u}}(\mathbf{x},t) := \mathbb{E}_{\mathbf{Z}}\mathbf{v}(\mathbf{x},t;\mathbf{g}_\theta(\mathbf{Z})),$$

then $p_t^\theta(\mathbf{x}) = (2\pi t)^{-d/2}\overline{p}(\mathbf{x},t)$ and $\mathbf{s}_t^\theta(\mathbf{x}) = \overline{\mathbf{u}}(\mathbf{x},t)/\overline{p}(\mathbf{x},t)$.

For shared centers $\mathbf{y}_i = \mathbf{g}_\theta(\mathbf{Z}_i)$, define

$$\overline{p}_n(\mathbf{x},t) := \frac{1}{n}\sum_{i=1}^n \kappa(\mathbf{x},t;\mathbf{y}_i), \qquad \overline{\mathbf{u}}_n(\mathbf{x},t) := \frac{1}{n}\sum_{i=1}^n \mathbf{v}(\mathbf{x},t;\mathbf{y}_i).$$

Proposition B.6 (Uniform-in-time Monte-Carlo score)**.** Suppose $\|\mathbf{g}_\theta(\mathbf{z})\|_\infty \le B_g$ for every $\mathbf{z}$. For any $\epsilon, \delta \in (0,1)$, there exist finite $n, R, M, \varepsilon_{\mathrm{low}}$, and $\tau > 0$ such that, with probability at least $1-\delta$ over the shared centers,

$$\sup_{t\in[t_0,T]} \mathbb{E}_{\mathbf{X}_t\sim P_t^\theta}\left\|\overline{\mathbf{s}}_n(\mathbf{X}_t,t) - \mathbf{s}_t^\theta(\mathbf{X}_t)\right\|_2^2 \le \epsilon,$$

where

$$\overline{\mathbf{s}}_n(\mathbf{x},t) := \mathrm{clip}_M\left(\frac{\overline{\mathbf{u}}_n(\mathbf{x},t)}{\overline{p}_n(\mathbf{x},t)\vee\varepsilon_{\mathrm{low}}}\right).$$

Proof. The proof has four steps.

Step 1: bounds and Lipschitz constants. Write $\mathbf{r}=\mathbf{x}-\mathbf{y}$. Since $t\ge t_0>0$,

$$0\le\kappa\le 1,\qquad \|\mathbf{v}\|_\infty\le Ct_0^{-1/2}.$$

For example, the second inequality follows from

$$\frac{|r_j|}{t}\exp\left(-\frac{\|\mathbf{r}\|_2^2}{2t}\right)\le\frac{1}{\sqrt{t}}\sup_{a\ge0}ae^{-a^2/2}\le Ct_0^{-1/2}.$$

We next differentiate the two features. For $j,k\in[d]$,

$$\frac{\partial\kappa}{\partial x_k}=-\frac{r_k}{t}\kappa,\qquad \frac{\partial\kappa}{\partial t}=\frac{\|\mathbf{r}\|_2^2}{2t^2}\kappa,$$
$$\frac{\partial v_j}{\partial x_k}=\left(-\frac{\mathbb{1}\{j=k\}}{t}+\frac{r_jr_k}{t^2}\right)\kappa,\qquad \frac{\partial v_j}{\partial t}=\left(\frac{r_j}{t^2}-\frac{r_j\|\mathbf{r}\|_2^2}{2t^3}\right)\kappa.$$

The elementary function $a^me^{-a^2/2}$ is bounded on $[0,\infty)$ for each fixed integer m. Substituting $a=\|\mathbf{r}\|_2/\sqrt{t}$ in the derivatives therefore gives, for $t\ge t_0$,

$$|\partial_{x_k}\kappa|\le Ct_0^{-1/2},\qquad |\partial_t\kappa|\le Ct_0^{-1},\qquad |\partial_{x_k}v_j|\le Ct_0^{-1},\qquad |\partial_tv_j|\le Ct_0^{-3/2},$$

so that κ and every coordinate of $\mathbf{v}$ are

$$\Lambda:=C(d+1)\left(t_0^{-1/2}+t_0^{-1}+t_0^{-3/2}\right)$$

Lipschitz in $(\mathbf{x},t)$ under the maximum norm (the factor $d+1$ bounds the ℓ_1 norm of the $(d+1)$-dimensional gradient; the $t_0^{-1/2}$ term is the contribution of the spatial derivative of the density feature and cannot be dropped for large t_0). The same Lipschitz bound holds for empirical averages and expectations.

Step 2: concentration on an $(\mathbf{x},t)$-net. Fix $R>0$ and $\tau>0$, whose values will be chosen in Step 4. Let

$$\eta:=\frac{\tau}{4\Lambda}.$$

A grid of $[-R,R]^d\times[t_0,T]$ with maximum-norm spacing at most η can be chosen with at most

$$N_{\mathrm{net}}\le\left(1+\frac{8R\Lambda}{\tau}\right)^d\left(1+\frac{4(T-t_0)\Lambda}{\tau}\right)$$

points. At each grid point, the density features lie in $[0,1]$, and every numerator coordinate lies in $[-B_1,B_1]$, where $B_1=Ct_0^{-1/2}$. Hoeffding's inequality and a union bound over the $d+1$ quantities and the N_{net} grid points show that

$$n\ge\frac{C(1+B_1^2)}{\tau^2}\log\left(\frac{2(d+1)N_{\mathrm{net}}}{\delta}\right)$$

is sufficient for the grid-point errors to be at most $\tau/2$. For any point between grid points, the empirical average and its expectation each change by at most $\Lambda\eta=\tau/4$. The triangle inequality therefore gives, simultaneously for all $(\mathbf{x},t)\in[-R,R]^d\times[t_0,T]$,

$$|\overline{p}_n(\mathbf{x},t)-\overline{p}(\mathbf{x},t)|\le\tau,$$
$$\|\overline{\mathbf{u}}_n(\mathbf{x},t)-\overline{\mathbf{u}}(\mathbf{x},t)\|_\infty\le\tau.$$

Step 3: ratio error on the retained region. Let

$$M := \frac{R + B_g}{t_0}, \qquad c_{\max} := (2\pi t_0)^{-d/2}.$$

For $\|\mathbf{x}\|_\infty \le R$, the posterior formula in (B.3) and the student-model bound imply $\|\mathbf{s}_t^\theta(\mathbf{x})\|_\infty \le M$. Consider a point satisfying $\overline{p}(\mathbf{x}, t) \ge 2\varepsilon_{\text{low}}$, and suppose that $\tau \le \varepsilon_{\text{low}}$. Then $\overline{p}_n(\mathbf{x}, t) \ge \varepsilon_{\text{low}}$. For every coordinate j, adding and subtracting $\overline{u}_j(\mathbf{x}, t)/\overline{p}_n(\mathbf{x}, t)$ yields

$$\begin{aligned}\left|\frac{\overline{u}_{n,j}}{\overline{p}_n} - \frac{\overline{u}_j}{\overline{p}}\right| &\le \frac{|\overline{u}_{n,j} - \overline{u}_j|}{\overline{p}_n} + \frac{|\overline{u}_j|\,|\overline{p}_n - \overline{p}|}{\overline{p}_n \overline{p}} \\ &\le \frac{\tau}{\varepsilon_{\text{low}}} + M\frac{\tau}{\varepsilon_{\text{low}}} = \frac{(1+M)\tau}{\varepsilon_{\text{low}}}.\end{aligned}$$

Coordinatewise clipping cannot increase the distance to a vector already in $[-M, M]^d$. Hence the squared error on this retained region is at most

$$d\left(\frac{(1+M)\tau}{\varepsilon_{\text{low}}}\right)^2.$$

Step 4: low density, tails, and parameter choices. Inside the cube, define the low-density set

$$A_t := \{\mathbf{x} : \|\mathbf{x}\|_\infty \le R,\ \overline{p}(\mathbf{x}, t) < 2\varepsilon_{\text{low}}\}.$$

Because $p_t^\theta(\mathbf{x}) = (2\pi t)^{-d/2}\overline{p}(\mathbf{x}, t)$,

$$P_t^\theta(A_t) = \int_{A_t} p_t^\theta(\mathbf{x})\,\mathrm{d}\mathbf{x} \le 2\varepsilon_{\text{low}} c_{\max}(2R)^d.$$

Both the true score on the cube and the clipped empirical score have maximum norm at most M. The contribution of A_t to the squared error is therefore at most

$$8dM^2\varepsilon_{\text{low}} c_{\max}(2R)^d.$$

For the tail, write $\mathbf{X}_t = \mathbf{g}_\theta(\mathbf{Z}) + \sqrt{t}\,\xi$. A Gaussian union bound gives, uniformly in $t \in [t_0, T]$,

$$P_t^\theta(\|\mathbf{X}_t\|_\infty > R) \le 2d\exp\left(-\frac{(R - B_g)^2}{2T}\right).$$

Tweedie's identity also gives $\mathbf{s}_t^\theta(\mathbf{X}_t) = -t^{-1/2}\mathbb{E}[\xi \mid \mathbf{X}_t]$. Conditional Jensen and Cauchy–Schwarz imply

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{s}_t^\theta(\mathbf{X}_t)\|_2^2\,\mathbb{1}\{\|\mathbf{X}_t\|_\infty > R\}\right] \\ &\quad\le \left(\frac{\mathbb{E}\,\|\xi\|_2^4}{t_0^2}\right)^{1/2} P_t^\theta(\|\mathbf{X}_t\|_\infty > R)^{1/2} \\ &\quad= \frac{\sqrt{d(d+2)}}{t_0} P_t^\theta(\|\mathbf{X}_t\|_\infty > R)^{1/2}.\end{aligned}$$

The empirical score contributes at most $dM^2 P_t^\theta(\|\mathbf{X}_t\|_\infty > R)$ on the tail.

We now choose the parameters in order. Let $C_R \ge 3$ be a sufficiently large constant depending only on d, t_0, T, and B_g, and set

$$R := B_g + C_R\sqrt{\log\left(\frac{C_R}{\epsilon}\right)}.$$

The Gaussian tail bound then makes each of the two tail terms at most $\epsilon/8$. To see this directly, substitute the chosen radius into the preceding tail estimate:

$$\sup_{t\in[t_0,T]} P_t^\theta(\|\mathbf{X}_t\|_\infty > R) \le 2d\left(\frac{\epsilon}{C_R}\right)^{C_R^2/(2T)}.$$

The true-score tail term is bounded by a constant times the square root of this probability, whereas the empirical-score tail term is bounded by dM^2 times the probability. Moreover, $M \le C(1 + \sqrt{\log(C_R/\epsilon)})$. Choosing C_R sufficiently large therefore makes both tail terms at most $\epsilon/8$. Next choose

$$\varepsilon_{\mathrm{low}} := \min\left\{\frac{1}{2},\ \frac{\epsilon}{64dM^2 c_{\max}(2R)^d}\right\}.$$

Then choose

$$\tau := \min\left\{\varepsilon_{\mathrm{low}},\ \frac{\varepsilon_{\mathrm{low}}}{1+M}\sqrt{\frac{\epsilon}{8d}}\right\}.$$

Finally choose n according to Step 2. The retained-region, low-density, and tail bounds then sum to at most ϵ, uniformly over $t \in [t_0, T]$. □

B.5 ReLU realization of the Monte-Carlo score

We first prove the scalar approximation tools used in the construction. The bounds below contain generous logarithmic slack so that the bookkeeping in the final parallel network is easy to verify.

Lemma B.7 (Elementary ReLU toolbox)**.** Let $0 < \zeta < 1/2$, $H \ge 1$, $Q \ge 0$, and $0 < \underline{a} \le \overline{a}$. Suppose

$$\mathfrak{A} \ge \max\left\{2, H, e^Q, \underline{a}^{-1}, \overline{a}, \zeta^{-1}\right\}. \tag{B.17}$$

There are scalar ReLU networks with the following properties:

1. $\mathrm{Mult}_{H,\zeta}(x,y)$ approximates xy within ζ on $[-H,H]^2$.
2. $\mathrm{Exp}_{Q,\zeta}(x)$ approximates e^{-x} within ζ on $[0,Q]$, and its output lies in $[0,1]$.
3. $\mathrm{Inv}_{\underline{a},\overline{a},\zeta}(x)$ approximates x^{-1} within ζ on $[\underline{a},\overline{a}]$.
4. Affine maps, finite sums, parallel concatenation, the operations $\min$, $\max$, and clipping to a fixed interval are represented exactly.

Each approximate scalar block may be chosen with

$$\begin{aligned} W &\le C\log^6 \mathfrak{A}, & L &\le C\log^4 \mathfrak{A}, \\ K &\le C\log^8 \mathfrak{A}, & \kappa &\le \exp\left(C\log^4 \mathfrak{A}\right). \end{aligned} \tag{B.18}$$

Proof. **Step 1: exact piecewise-linear operations.** The identities

$$\max\{x,y\} = y + \mathrm{ReLU}(x-y), \qquad \min\{x,y\} = x - \mathrm{ReLU}(x-y)$$

give exact two-input maximum and minimum networks. Therefore

$$\mathrm{clip}_{[a,b]}(x) = \min\{b, \max\{a,x\}\}$$

is exact as well. Parallelization is obtained by placing weight matrices on block diagonals. More explicitly, if branch r has one layer $\mathbf{h}^{(r)} = \mathrm{ReLU}(\mathbf{A}^{(r)}\mathbf{v}^{(r)} + \mathbf{b}^{(r)})$, then the concatenated layer is

$$\begin{pmatrix} \mathbf{h}^{(1)} \\ \vdots \\ \mathbf{h}^{(m)} \end{pmatrix} = \mathrm{ReLU}\left(\mathrm{diag}(\mathbf{A}^{(1)},\dots,\mathbf{A}^{(m)}) \begin{pmatrix} \mathbf{v}^{(1)} \\ \vdots \\ \mathbf{v}^{(m)} \end{pmatrix} + \begin{pmatrix} \mathbf{b}^{(1)} \\ \vdots \\ \mathbf{b}^{(m)} \end{pmatrix}\right).$$

A final row block $(\mathbf{I},\dots,\mathbf{I})$ sums the branch outputs. If branches have unequal depths, a signed scalar x is carried by the two nonnegative channels $(\mathrm{ReLU}(x), \mathrm{ReLU}(-x))$ and reconstructed as their difference, since $x = \mathrm{ReLU}(x) - \mathrm{ReLU}(-x)$. Repeating these two channels pads any desired number of layers. The displayed matrices show directly that widths and nonzero parameter counts add under parallelization, while depths add under composition.

Step 2: a deep square network. On $[0,1]$, define the tent map

$$g(x) = 2\,\mathrm{ReLU}(x) - 4\,\mathrm{ReLU}\left(x - \frac{1}{2}\right) + 2\,\mathrm{ReLU}(x-1).$$

Let g_s be its s-fold composition and set

$$S_m(x) := x - \sum_{s=1}^{m} 4^{-s} g_s(x).$$

An induction over the dyadic refinements shows that S_m is the linear interpolant of x^2 on the grid $\{k2^{-m} : 0 \le k \le 2^m\}$. Here is the induction explicitly: for $m = 0$, the interpolant is the line $S_0(x) = x$; when each grid interval is split in half, the tent g_{m+1} equals one at the new midpoint and zero at all old grid points, so subtracting $4^{-(m+1)} g_{m+1}$ makes the value at every new midpoint equal to its square without changing the old interpolation values. On an interval of length h, the linear interpolation error for $f(x) = x^2$ is at most $h^2 \sup |f''| /8 = h^2/4$. Therefore

$$\sup_{x \in [0,1]} \left| S_m(x) - x^2 \right| \le 4^{-(m+1)}.$$

The network can carry $g_s(x)$ and the current partial sum through its layers, so it has constant width, depth $O(m)$, $O(m)$ nonzero parameters, and weights bounded by four.

For $x \in [-H, H]$, apply this construction to $|x| /H = [\mathrm{ReLU}(x) + \mathrm{ReLU}(-x)]/H$ and multiply the output by H^2. Choosing $m \ge \log_4(2H^2/\zeta)$ produces a square approximation with error at most $\zeta/2$.

Step 3: multiplication. The polarization identity

$$xy = \frac{(x+y)^2 - (x-y)^2}{4}$$

turns two square approximators on $[-2H, 2H]$ into a multiplication approximator on $[-H, H]^2$. Assigning error 2ζ across the two squares gives total error at most ζ. This construction has depth and parameter count $O(\log(H^2/\zeta))$, which is dominated by (B.18).

Step 4: the exponential. Let

$$P_m(x) = \sum_{k=0}^{m} \frac{(-x)^k}{k!}.$$

Taylor's theorem and the fact that every derivative of e^{-x} has absolute value at most one on $[0, \infty)$ give

$$\sup_{0 \le x \le Q} \left| e^{-x} - P_m(x) \right| \le \frac{Q^{m+1}}{(m+1)!} \le \left(\frac{eQ}{m+1} \right)^{m+1}.$$

Choose $m + 1 \ge \max\{2eQ, \log_2(4/\zeta)\}$. The last display is then at most $\zeta/4$.

We compute the powers recursively and verify the input range of every multiplication block. Put

$$M_Q := \max\{1, Q\}, \qquad H_{\mathrm{int}} := 1 + M_Q^m.$$

Starting from $\widetilde{x}^0 = 1$, obtain $\widetilde{x}^k$ by approximately multiplying $\widetilde{x}^{k-1}$ by x. If each multiplication has error at most δ, and $E_k := \left| \widetilde{x}^k - x^k \right|$, the triangle inequality gives

$$E_k \le \delta + Q E_{k-1} \le \delta k M_Q^{k-1}.$$

For completeness, the second inequality is an induction. It is true at $k = 1$. If it holds at $k - 1$, then

$$E_k \le \delta + Q\delta(k-1)M_Q^{k-2} \le \delta + \delta(k-1)M_Q^{k-1} \le \delta k M_Q^{k-1}.$$

Thus the accumulated polynomial error is at most $\zeta/4$ when, for example,

$$\delta = \frac{\zeta}{4(m+1)^2 M_Q^m}.$$

This choice also gives $E_k < 1$ for all $k \le m$, and hence

$$\left| \widetilde{x}^k \right| \le |x|^k + E_k \le M_Q^m + 1 = H_{\mathrm{int}}.$$

Thus all approximate products are indeed evaluated in $[-H_{\text{int}}, H_{\text{int}}]^2$; the error recursion does not presuppose an unverified input bound. The remaining affine summation is exact. Finally clip the polynomial output to $[0, 1]$; because $e^{-x} \in [0, 1]$, clipping cannot increase the error.

Condition $e^Q \le \mathfrak{A}$ implies $Q \le \log \mathfrak{A}$, so $m = O(\log \mathfrak{A})$. The largest power handled internally has logarithm at most $m \log M_Q = O(\log^2 \mathfrak{A})$, while $\log H_{\text{int}} = O(\log^2 \mathfrak{A})$ and $\log \delta^{-1} = O(\log^2 \mathfrak{A})$. By Steps 2–3, each power multiplication therefore has constant width and depth and sparsity $O(\log^2 \mathfrak{A})$. There are $m = O(\log \mathfrak{A})$ such products in sequence, so the full exponential block has depth and sparsity $O(\log^3 \mathfrak{A})$. Its weights are bounded by $\exp(O(\log^2 \mathfrak{A}))$. Each of these bounds is smaller than the corresponding bound in (B.18).

Step 5: the reciprocal. Write $z = x/\overline{a}$ and $r = \underline{a}/\overline{a}$, so $z \in [r, 1]$. Newton's iteration

$$y_0 = 1, \qquad y_{k+1} = y_k(2 - z y_k)$$

obeys the exact residual identity

$$1 - z y_{k+1} = (1 - z y_k)^2.$$

It follows by induction that $1 - z y_k = (1 - z)^{2^k} \le e^{-r 2^k}$. Since $x^{-1} = 1/(\overline{a} z)$, the exact iterate has error at most $e^{-r 2^k}/\underline{a}$. Choose an integer $k \ge 1$ so that

$$2^k \ge r^{-1} \log\left(\frac{4}{\underline{a}\zeta}\right).$$

Then the exact-iteration error is at most $\zeta/4$.

For approximate multiplication, clip every computed iterate to $[0, 2/r]$; the exact iterates lie in $[0, 1/r]$, so this clipping does not alter them. On the larger interval the map $F_z(y) = y(2 - zy)$ has derivative bounded by $6/r$. Each Newton step uses two multiplication blocks, first zy and then $y \cdot (2 - zy)$; give each of them accuracy $\mu := \delta/(1 + 2/r)$. The error of the inner product is amplified by at most $|y| \le 2/r$ in the outer product, and the outer product adds at most μ, so the two multiplications together introduce error at most $(1 + 2/r)\mu = \delta$ per step. Induction and the mean-value theorem then give

$$|\widetilde{y}_k - y_k| \le Ck \left(\frac{6}{r}\right)^k \delta.$$

Taking

$$\delta = \frac{\zeta \overline{a}}{4Ck(6/r)^k}$$

makes the output error after division by $\overline{a}$ at most $\zeta/4$. A final clip to $[0, \underline{a}^{-1}]$ is nonexpansive around the true reciprocal.

From (B.17), $r^{-1} \le \mathfrak{A}^2$, so $k = O(\log \mathfrak{A})$, and $\log \delta^{-1} = O(\log^2 \mathfrak{A})$. The multiplication inputs are bounded by a polynomial in r^{-1}, and hence by a polynomial in $\mathfrak{A}$. The multiplication construction and the composition rules from Step 1 again give (B.18). This proves all parts of the lemma. □

Lemma B.8 (Subnetworks for the Brownian VE schedule)**.** For every $\zeta \in (0, 1)$, ReLU networks on $[t_0, T]$ can represent $\alpha_t = 1$ and $\sigma_t^2 = t$ exactly and can approximate $\sigma_t = \sqrt{t}$ and $\sigma_t^{-2} = t^{-1}$ uniformly within ζ.

Proof. The constant function 1 and the identity function t are affine, so one affine layer represents α_t and σ_t^2 exactly. Lemma B.7 with $\underline{a} = t_0$ and $\overline{a} = T$ provides the reciprocal network for t^{-1}.

It remains to approximate $\sqrt{t}$. Partition $[t_0, T]$ into m intervals of length at most $h = (T - t_0)/m$, and let S_m be the linear interpolant of $f(t) = \sqrt{t}$ on the grid. Since

$$|f''(t)| = \frac{1}{4t^{3/2}} \le \frac{1}{4t_0^{3/2}},$$

the elementary linear-interpolation remainder on an interval of length h gives

$$\sup_{t \in [t_0, T]} \left| S_m(t) - \sqrt{t} \right| \le \frac{h^2}{32 t_0^{3/2}}.$$

Choose m large enough that the right-hand side is at most ζ. Every continuous piecewise-linear function with breakpoints $a_1, \dots, a_{m-1}$ has the representation

$$S_m(t) = c_0 + c_1 t + \sum_{k=1}^{m-1} c_{k+1} \operatorname{ReLU}(t - a_k).$$

The last display is a two-layer ReLU network, which completes the construction. □

For the Monte-Carlo features, define

$$C_{r,t} := R_t + \alpha_t B_g, \qquad Q_t := \frac{d C_{r,t}^2}{2\sigma_t^2}.$$

The single internal accuracy used below is

$$\eta_{\mathrm{net},t} := c_2 \min\left\{ \frac{1}{4}, \frac{\varepsilon_{\mathrm{low},t}}{2}, \frac{\varepsilon_{\mathrm{low},t}^2 \sqrt{\epsilon_t}}{\sqrt{d}(B_{1,t} + \varepsilon_{\mathrm{low},t} + 1)} \right\}, \tag{B.19}$$

where $c_2 > 0$ is a sufficiently small numerical constant. Finally set

$$\mathfrak{N}_t := \max\left\{ 2, e^{Q_t}, n, d, c_t, C_{r,t}, \sigma_t^{-1}, M_t, \varepsilon_{\mathrm{low},t}^{-1}, \epsilon_t^{-1}, \delta_t^{-1}, \eta_{\mathrm{net},t}^{-1} \right\}.$$

All scalar ranges and reciprocal accuracies used in the construction have logarithms bounded by a fixed power of $\log \mathfrak{N}_t$.

Proposition B.9 (Fixed-time ReLU realization)**.** Under the conditions and parameter choices of Proposition B.5, there is a ReLU network $\mathbf{s}_{\psi,t} : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\mathbb{E}_{\mathbf{X}_t \sim P_t^\theta} \left\| \mathbf{s}_{\psi,t}(\mathbf{X}_t) - \mathbf{s}_t^\theta(\mathbf{X}_t) \right\|_2^2 \le C\epsilon_t. \tag{B.20}$$

The network can be chosen with output bound M_t, width $O(nd \log^7 \mathfrak{N}_t)$, depth $O(\log^4 \mathfrak{N}_t)$, sparsity $O(nd \log^9 \mathfrak{N}_t)$, and parameter bound $\exp(O(\log^4 \mathfrak{N}_t))$.

Proof. Condition on a realization of the centers for which the conclusion of Proposition B.5 holds. This event has probability at least $1 - \delta_t$. We construct a network for these fixed centers and then count its parameters.

Step 1: one Gaussian feature. For each $i \in [n]$, the network first forms the residual vector exactly:

$$\mathbf{r}_i(\mathbf{x}) := \mathbf{x} - \alpha_t \mathbf{y}_i.$$

On $\|\mathbf{x}\|_\infty \le R_t$, every coordinate satisfies $|[\mathbf{r}_i(\mathbf{x})]_j| \le C_{r,t}$. Next the network computes

$$q_i(\mathbf{x}) := \frac{1}{2\sigma_t^2} \sum_{j=1}^{d} [\mathbf{r}_i(\mathbf{x})]_j^2, \qquad 0 \le q_i(\mathbf{x}) \le Q_t. \tag{B.21}$$

using d square blocks in parallel and one exact affine summation. Because approximate squaring can move the result slightly outside its true range, we clip the computed value exactly to $[0, Q_t]$ before applying the exponential block. We then form

$$\begin{aligned} e_i(\mathbf{x}) &:= e^{-q_i(\mathbf{x})}, \\ \phi_i(\mathbf{x}) &:= c_t e_i(\mathbf{x}), \\ \mathbf{a}_i(\mathbf{x}) &:= -\sigma_t^{-2} \mathbf{r}_i(\mathbf{x}) \phi_i(\mathbf{x}). \end{aligned} \tag{B.22}$$

Here is an explicit error allocation. Set

$$\zeta_t := \frac{\eta_{\mathrm{net},t}}{C(1+d)(1+c_t)(1+C_{r,t})(1+\sigma_t^{-2})^2},$$

with a sufficiently large numerical C. Assign accuracy ζ_t to every square, exponential, and multiplication block in (B.21)–(B.22). The quadratic-form error is at most $d\zeta_t/(2\sigma_t^2)$. Clipping is one-Lipschitz, and $x \mapsto e^{-x}$ is one-Lipschitz on $[0,\infty)$, so the exponential input error is not amplified. Multiplication by c_t is exact. For a numerator coordinate, the product error is at most

$$\sigma_t^{-2}\left(\zeta_t + C_{r,t}\left|\phi_i^{\mathrm{NN}}(\mathbf{x}) - \phi_i(\mathbf{x})\right|\right).$$

Substitution of the definition of ζ_t into these three displayed bounds gives, for all $\|\mathbf{x}\|_\infty \le R_t$,

$$\left|\phi_i^{\mathrm{NN}}(\mathbf{x}) - \phi_i(\mathbf{x})\right| \le \eta_{\mathrm{net},t}, \qquad \left\|\mathbf{a}_i^{\mathrm{NN}}(\mathbf{x}) - \mathbf{a}_i(\mathbf{x})\right\|_\infty \le \eta_{\mathrm{net},t}.$$

Every range and every reciprocal accuracy involved here is bounded by a fixed power of $\mathfrak{N}_t$; in particular, $\zeta_t^{-1} \le \mathfrak{N}_t^C$. Thus Lemma B.7 applies with a complexity parameter that changes $\log \mathfrak{N}_t$ only by a numerical factor.

Step 2: averaging the features. Run the n feature networks in parallel and add one affine averaging layer:

$$\widehat{p}_{n,t}^{\mathrm{NN}}(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n \phi_i^{\mathrm{NN}}(\mathbf{x}), \qquad \widehat{\mathbf{u}}_{n,t}^{\mathrm{NN}}(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n \mathbf{a}_i^{\mathrm{NN}}(\mathbf{x}).$$

Averaging does not enlarge a uniform error. Hence

$$\left|\widehat{p}_{n,t}^{\mathrm{NN}}(\mathbf{x}) - \widehat{p}_{n,t}(\mathbf{x})\right| \le \eta_{\mathrm{net},t}, \qquad \left\|\widehat{\mathbf{u}}_{n,t}^{\mathrm{NN}}(\mathbf{x}) - \widehat{\mathbf{u}}_{n,t}(\mathbf{x})\right\|_\infty \le \eta_{\mathrm{net},t} \tag{B.23}$$

on the cube.

Step 3: the thresholded reciprocal. Define

$$D(\mathbf{x}) := \widehat{p}_{n,t}(\mathbf{x}) \vee \varepsilon_{\mathrm{low},t}, \qquad D^{\mathrm{NN}}(\mathbf{x}) := \widehat{p}_{n,t}^{\mathrm{NN}}(\mathbf{x}) \vee \varepsilon_{\mathrm{low},t}.$$

The maximum operation is exact and one-Lipschitz, so

$$\left|D^{\mathrm{NN}}(\mathbf{x}) - D(\mathbf{x})\right| \le \eta_{\mathrm{net},t}. \tag{B.24}$$

Both denominators are at least $\varepsilon_{\mathrm{low},t}$. Also $D^{\mathrm{NN}} \le c_t + 1$, because the exponential feature is clipped to $[0,1]$. Apply the reciprocal block on $[\varepsilon_{\mathrm{low},t}, c_t+1]$ with approximation error $\eta_{\mathrm{net},t}$. Denoting its output by $I^{\mathrm{NN}}(\mathbf{x})$, the identity $\left|a^{-1}-b^{-1}\right| = |a-b|/(ab)$ and (B.24) give

$$\left|I^{\mathrm{NN}}(\mathbf{x}) - D(\mathbf{x})^{-1}\right| \le \eta_{\mathrm{net},t} + \frac{\eta_{\mathrm{net},t}}{\varepsilon_{\mathrm{low},t}^2}. \tag{B.25}$$

Step 4: the ratio and final clipping. For each numerator coordinate, use one multiplication block of accuracy $\eta_{\mathrm{net},t}$ to multiply $\widehat{\mathbf{u}}_{n,t}^{\mathrm{NN}}$ by I^{NN}. Since $\|\widehat{\mathbf{u}}_{n,t}\|_\infty \le B_{1,t}$, (B.23)– (B.25) imply

$$\begin{aligned}&\left\|\mathrm{Mult}^{\mathrm{NN}}(\widehat{\mathbf{u}}_{n,t}^{\mathrm{NN}}, I^{\mathrm{NN}}) - \frac{\widehat{\mathbf{u}}_{n,t}}{D}\right\|_\infty \\ &\quad\le C\eta_{\mathrm{net},t}\left(1 + \varepsilon_{\mathrm{low},t}^{-1} + B_{1,t} + B_{1,t}\varepsilon_{\mathrm{low},t}^{-2}\right) \le \sqrt{\frac{\epsilon_t}{d}}.\end{aligned}$$

The last inequality is exactly the reason for the choice (B.19); taking c_2 small enough absorbs the numerical constant. Finally, clip every output coordinate exactly to $[-M_t, M_t]$. Clipping is nonexpansive, and the same clipping defines $\widehat{\mathbf{s}}_{n,t}$. Therefore

$$\sup_{\|\mathbf{x}\|_\infty \le R_t} \|\mathbf{s}_{\psi,t}(\mathbf{x}) - \widehat{\mathbf{s}}_{n,t}(\mathbf{x})\|_2^2 \le \epsilon_t.$$

The final clip also proves the global output bound $M_{\mathrm{out},t} = M_t$.

Step 5: pass from the cube to $L^2(P_t^\theta)$. Outside the cube, both $\mathbf{s}_{\psi,t}$ and $\widehat{\mathbf{s}}_{n,t}$ have coordinate magnitude at most M_t. Hence

$$\mathbb{E}\,\|\mathbf{s}_{\psi,t}(\mathbf{X}_t) - \widehat{\mathbf{s}}_{n,t}(\mathbf{X}_t)\|_2^2 \le \epsilon_t + 4dM_t^2\mathbb{P}(\|\mathbf{X}_t\|_\infty > R_t) \le C\epsilon_t,$$

where the last inequality follows from the tail calculation in Lemma B.4 and the choice of r_t. The vector inequality $\|\mathbf{v}+\mathbf{w}\|_2^2 \le 2\|\mathbf{v}\|_2^2 + 2\|\mathbf{w}\|_2^2$, together with Proposition B.5, now gives

$$\mathbb{E}\left\|\mathbf{s}_{\psi,t}(\mathbf{X}_t) - \mathbf{s}_t^\theta(\mathbf{X}_t)\right\|_2^2 \le 2\mathbb{E}\left\|\mathbf{s}_{\psi,t}(\mathbf{X}_t) - \widehat{\mathbf{s}}_{n,t}(\mathbf{X}_t)\right\|_2^2 + 2\mathbb{E}\left\|\widehat{\mathbf{s}}_{n,t}(\mathbf{X}_t) - \mathbf{s}_t^\theta(\mathbf{X}_t)\right\|_2^2 \le C\epsilon_t.$$

This proves (B.20).

Step 6: width, depth, sparsity, and weights. One feature branch uses d square blocks and d numerator-product blocks, plus one exponential block. Lemma B.7 and parallelization therefore give, for one branch,

$$W_i \le Cd\log^6 \mathfrak{N}_t, \qquad L_i \le C\log^4 \mathfrak{N}_t, \qquad K_i \le Cd\log^8 \mathfrak{N}_t.$$

Running n branches in parallel multiplies width and sparsity by n but does not multiply depth. We list the remaining counts explicitly. The averaging layer has $O(nd)$ nonzero weights. The exact denominator maximum and the d output clips use $O(d)$ neurons and nonzero weights. The one reciprocal block uses at most $C\log^6 \mathfrak{N}_t$ width and $C\log^8 \mathfrak{N}_t$ nonzero parameters, while the d final multiplication blocks use at most $Cd\log^6 \mathfrak{N}_t$ width and $Cd\log^8 \mathfrak{N}_t$ nonzero parameters. These post-processing blocks are used a fixed number of times in sequence, so their depths add to at most $C\log^4 \mathfrak{N}_t$.

To carry an affine input through a deeper parallel branch, use the exact two-channel representation $x = \mathrm{ReLU}(x) - \mathrm{ReLU}(-x)$. Across at most $C\log^4 \mathfrak{N}_t$ layers, these identity channels add at most $Cnd\log^4 \mathfrak{N}_t$ nonzero parameters. Adding all the displayed counts first yields the stronger orders $W_t = O(nd\log^6 \mathfrak{N}_t)$ and $K_t = O(nd\log^8 \mathfrak{N}_t)$. Allowing one logarithmic factor of slack gives the stated convenient bounds

$$W_t \le Cnd\log^7 \mathfrak{N}_t, \qquad L_t \le C\log^4 \mathfrak{N}_t, \qquad K_t \le Cnd\log^9 \mathfrak{N}_t.$$

All affine coefficients, including the fixed centers and schedule rescalings, are bounded by a fixed power of $\mathfrak{N}_t$. The scalar toolbox uses weights at most $\exp(C\log^4 \mathfrak{N}_t)$, which dominates those affine coefficients. Thus $\kappa_t \le \exp(C\log^4 \mathfrak{N}_t)$, as stated in Proposition B.9. □

B.6 Proof of Theorem 3.7

Proof. Fix $\epsilon \in (0,1)$. Apply Proposition B.6 with target error $\epsilon/4$ and failure probability $\delta = 1/2$. The proposition gives a nonempty event of center realizations on which

$$\sup_{t\in[t_0,T]} \mathbb{E}\left\|\bar{\mathbf{s}}_n(\mathbf{X}_t,t) - \mathbf{s}_t^\theta(\mathbf{X}_t)\right\|_2^2 \le \frac{\epsilon}{4}.$$

Choose and fix one realization in that event. This use of a positive probability event is the probabilistic method: at least one deterministic set of centers has the stated property. In Step 4 of the proposition, the target error is $\epsilon/4$, so the truncation radius is chosen such that

$$dM^2 \sup_{t\in[t_0,T]} P_t^\theta(\|\mathbf{X}_t\|_\infty > R) \le \frac{\epsilon}{32}, \qquad M = \frac{R+B_g}{t_0}.$$

Multiplying the last probability bound by four gives the $\epsilon/8$ tail bound used below.

We construct a network that approximates $\bar{\mathbf{s}}_n$ on $[-R,R]^d \times [t_0,T]$. Define

$$Q := \frac{d(R+B_g)^2}{2t_0}, \qquad H := 2 + R + B_g + t_0^{-1} + T + Q + M.$$

All quantities entering a feature branch lie in an interval whose absolute value is at most a fixed polynomial in H.

Step 1: schedule subnetworks. Lemma B.8 represents $\alpha_t = 1$ and $\sigma_t^2 = t$ exactly, approximates $\sigma_t = \sqrt{t}$, and supplies a reciprocal network $I(t)$ satisfying

$$\sup_{t\in[t_0,T]} \left|I(t) - t^{-1}\right| \le \eta.$$

The score feature uses t and t^{-1}; the square-root subnetwork covers the sampling coefficient σ_t, although the normalizing factor cancels from the score ratio.

Step 2: one joint $(\mathbf{x}, t)$-feature. For each fixed center $\mathbf{y}_i$, an affine layer forms $\mathbf{r}_i = \mathbf{x} - \mathbf{y}_i$. Parallel square networks and an affine summation approximate $\|\mathbf{r}_i\|_2^2$. A multiplication block then forms

$$q_i(\mathbf{x}, t) = \frac{1}{2} \|\mathbf{r}_i\|_2^2 \, t^{-1}.$$

On the cube, $0 \le q_i \le Q$. We clip the approximate value to $[0, Q]$, apply the exponential network, and obtain an approximation to

$$k_i(\mathbf{x}, t) = e^{-q_i(\mathbf{x}, t)}.$$

Finally, multiplication blocks form every coordinate of

$$\mathbf{v}_i(\mathbf{x}, t) = -\mathbf{r}_i \, t^{-1} k_i(\mathbf{x}, t).$$

We record the error propagation. Give every square, reciprocal, exponential, and multiplication block accuracy at most η. The squared-norm error is at most $d\eta$. The product with the reciprocal has error at most $CdH^2\eta$. Clipping is one-Lipschitz, and $q \mapsto e^{-q}$ is one-Lipschitz on $[0, \infty)$, so the exponential-feature error is at most $CdH^2\eta$. For the numerator we first form the scaled residual

$$w_{i,j}(\mathbf{x}, t) := \frac{x_j - y_{i,j}}{t} = r_{i,j} \, t^{-1}$$

by one multiplication of $r_{i,j}$ with the reciprocal network $I(t)$, and only then multiply by the exponential feature, $v_{i,j} = -w_{i,j} k_i$. On the cube $|w_{i,j}| \le (R + B_g)/t_0 = M \le H$, and $w_{i,j}$ carries error at most $(R + B_g)\eta + \eta \le CH\eta$, so the final product has error at most $|w_{i,j}| \cdot CdH^2\eta + |k_i| \cdot CH\eta + \eta \le CdH^3\eta$: the exponential-feature error is amplified only by the factor $|w_{i,j}| \le H$. Thus, after increasing the numerical constant,

$$\begin{aligned} \left| k_i^{\mathrm{NN}}(\mathbf{x}, t) - k_i(\mathbf{x}, t) \right| &\le CdH^3\eta, \\ \left\| \mathbf{v}_i^{\mathrm{NN}}(\mathbf{x}, t) - \mathbf{v}_i(\mathbf{x}, t) \right\|_\infty &\le CdH^3\eta \end{aligned}$$

uniformly on the space–time cube.

Step 3: averages and the ratio. Run the n feature branches in parallel and average their outputs with one affine layer. Averaging does not enlarge a uniform error. Denote the two resulting network outputs by $\overline{p}_n^{\mathrm{NN}}$ and $\overline{\mathbf{u}}_n^{\mathrm{NN}}$. The operation $\overline{p}_n^{\mathrm{NN}} \vee \varepsilon_{\mathrm{low}}$ is exact. Apply the reciprocal network on $[\varepsilon_{\mathrm{low}}, 2]$, multiply by each numerator coordinate, and clip the final output to $[-M, M]^d$.

The identity $\left| a^{-1} - b^{-1} \right| = |a - b| / (ab)$ shows that a denominator error ξ changes its reciprocal by at most $\xi / \varepsilon_{\mathrm{low}}^2$. The numerator is bounded on the cube by a fixed polynomial in H. Consequently, the ratio and final clipping produce a network $\mathbf{s}_\psi^{\mathrm{NN}}$ satisfying

$$\sup_{\substack{\|\mathbf{x}\|_\infty \le R \\ t \in [t_0, T]}} \left\| \mathbf{s}_\psi^{\mathrm{NN}}(\mathbf{x}, t) - \overline{\mathbf{s}}_n(\mathbf{x}, t) \right\|_\infty \le CdH^4 \varepsilon_{\mathrm{low}}^{-2} \eta.$$

Choose the internal accuracy to be

$$\eta := \frac{\varepsilon_{\mathrm{low}}^2}{CdH^4} \sqrt{\frac{\epsilon}{16d}},$$

where the constant C is taken sufficiently large. The squared Euclidean error on the cube is then at most $\epsilon/16$. Outside the cube, both outputs have maximum norm at most M, so the choice of R gives

$$\sup_{t \in [t_0, T]} \mathbb{E} \left\| \mathbf{s}_\psi^{\mathrm{NN}}(\mathbf{X}_t, t) - \overline{\mathbf{s}}_n(\mathbf{X}_t, t) \right\|_2^2 \le \frac{\epsilon}{4}.$$

Step 4: total error and network size. For vectors $\mathbf{a}, \mathbf{b}$, $\|\mathbf{a} + \mathbf{b}\|_2^2 \le 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$. Applying this inequality to the network-to-Monte-Carlo and Monte-Carlo-to-score differences gives

$$\sup_{t \in [t_0, T]} \mathbb{E} \left\| \mathbf{s}_\psi^{\mathrm{NN}}(\mathbf{X}_t, t) - \mathbf{s}_t^\theta(\mathbf{X}_t) \right\|_2^2 \le \epsilon.$$

For completeness, define

$$\mathfrak{A} := \max\left\{2, H, e^Q, t_0^{-1}, T, \varepsilon_{\text{low}}^{-1}, \eta^{-1}, n, d, M\right\}.$$

We now make the dependence on ϵ explicit. Put $\ell_\epsilon := \log(C/\epsilon)$, increasing C when necessary. The radius chosen in Step 4 of Proposition B.6, applied there with error $\epsilon/4$, gives

$$R \le C\ell_\epsilon^{1/2}, \qquad M = \frac{R + B_g}{t_0} \le C\ell_\epsilon^{1/2}.$$

Since $c_{\max}$ is fixed, the explicit low-density threshold in that proposition satisfies

$$\varepsilon_{\text{low}}^{-1} \le C\epsilon^{-1}\ell_\epsilon^{d/2+1}.$$

The chosen concentration tolerance then gives

$$\tau^{-2} \le C\epsilon^{-3}\ell_\epsilon^{d+3}.$$

Indeed, the factor $\varepsilon_{\text{low}}^{-2}$ contributes $\epsilon^{-2}\ell_\epsilon^{d+2}$, the factor $(1+M)^2$ contributes at most $C\ell_\epsilon$, and the square of $\sqrt{\epsilon}$ contributes ϵ in the denominator.

For the space–time grid in Step 2 of the same proposition, $\log N_{\text{net}} \le C\ell_\epsilon$. Choose n to be the smallest integer satisfying the Hoeffding sample-size bound. It follows that

$$n \le C\epsilon^{-3}\ell_\epsilon^{d+4}.$$

Furthermore,

$$Q \le C\ell_\epsilon, \qquad H \le C\ell_\epsilon, \qquad \eta^{-1} \le C\epsilon^{-5/2}\ell_\epsilon^{d+6}.$$

Every quantity in the maximum defining $\mathfrak{A}$ is therefore at most a fixed power of C/ϵ. Hence

$$\log \mathfrak{A} \le C\ell_\epsilon.$$

One feature branch uses $O(d)$ scalar blocks from Lemma B.7. Running n branches in parallel, followed by a constant number of ratio blocks, gives the score branch width $Cnd\log^6 \mathfrak{A}$ and sparsity $Cnd\log^8 \mathfrak{A}$. The piecewise-linear square-root subnetwork in Lemma B.8 uses at most $C\eta^{-1/2}$ neurons and nonzero parameters. Including that schedule branch, the complete network satisfies

$$W \le C\left(nd\log^6 \mathfrak{A} + \eta^{-1/2}\right), \qquad L \le C\log^4 \mathfrak{A},$$
$$K \le C\left(nd\log^8 \mathfrak{A} + \eta^{-1/2}\right), \qquad \kappa \le \exp(C\log^4 \mathfrak{A}).$$

Here $\eta^{-1/2} \le C\epsilon^{-5/4}\ell_\epsilon^{d/2+3}$, which is no larger than the $\epsilon^{-3}\ell_\epsilon^{d+10}$ and $\epsilon^{-3}\ell_\epsilon^{d+12}$ budgets in (3.6). The final clipping gives the output bound M. Hence $\mathbf{s}_\psi^{\text{NN}} \in \mathcal{F}(M, W, \kappa, L, K)$, and the network has the joint input $(\mathbf{x}, t)$. Substituting the bounds for n and $\log \mathfrak{A}$ proves (3.6); the error calculation proves (3.7). □

B.7 Proof of Proposition 3.9

Throughout, q_{SiD} denotes the *oracle population SiD-objective minimizer*: the minimizer of $\mathcal{S}_\lambda^{[0,T]}(q)$ after the exact surrogate $s_t^\psi = s_t^\theta$ has been substituted into the objective. The proof has two parts. We first compute the two output variances in closed form,

$$q_{\text{SiD}} = \frac{\sqrt{T^2 + 4(1+T)/c^2} - T}{2}, \tag{B.26}$$

and

$$\text{Var}(X_0^{\text{PF}}) = (1+T)^{1-c}, \tag{B.27}$$

so that $P^{\text{SiD}} = \mathcal{N}(0, q_{\text{SiD}})$ and $P^{\text{PF}} = \mathcal{N}(0, (1+T)^{1-c})$. We then convert these variances into the KL and total-variation statements of the proposition.

Proof of (B.26) *and* (B.27). Under Assumption 3.8, the noised student distribution and its score are

$$P_t^\theta = \mathcal{N}(0, q+t), \qquad s_t^\theta(x) = -\frac{x}{q+t}.$$

Because the surrogate is exact, $\delta_t = s_t^\phi - s_t^\theta$, and the calculations in (A.8)–(A.9) give $\ell_t^{(1)} = \ell_t^{(2)}$. Thus $\ell_t^{(\lambda)}$ has the same value for every λ, namely

$$\begin{aligned}\mathcal{S}_\lambda^{[0,T]}(q) &= \int_0^T \mathbb{E}_{X_t \sim P_t^\theta}\left[\left|s_t^\phi(X_t) - s_t^\theta(X_t)\right|^2\right] \mathrm{d}t \\ &= \int_0^T (q+t)\left(\frac{1}{q+t} - \frac{c}{1+t}\right)^2 \mathrm{d}t =: \mathcal{S}(q).\end{aligned}$$

To find its minimum, expand the integrand:

$$(q+t)\left(\frac{1}{q+t} - \frac{c}{1+t}\right)^2 = \frac{1}{q+t} - \frac{2c}{1+t} + \frac{c^2(q+t)}{(1+t)^2}.$$

Differentiating under the finite integral gives

$$\begin{aligned}\mathcal{S}'(q) &= \int_0^T \left[-\frac{1}{(q+t)^2} + \frac{c^2}{(1+t)^2}\right] \mathrm{d}t, \\ \mathcal{S}''(q) &= 2\int_0^T \frac{1}{(q+t)^3}\,\mathrm{d}t > 0.\end{aligned}$$

Hence $\mathcal{S}$ is strictly convex, so any critical point is its unique global minimizer. Setting $\mathcal{S}'(q) = 0$ and evaluating the two elementary integrals gives

$$\frac{T}{q(q+T)} = \int_0^T \frac{\mathrm{d}t}{(q+t)^2} = c^2 \int_0^T \frac{\mathrm{d}t}{(1+t)^2} = \frac{c^2 T}{1+T}.$$

Since $T > 0$, cancellation yields $q(q+T) = (1+T)/c^2$. The positive root of this quadratic equation is exactly (B.26).

For the probability-flow ODE, the VE diffusion has unit diffusion coefficient, so the teacher-driven ODE is

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = -\frac{1}{2}s_t^\phi(X_t) = \frac{c}{2(1+t)}X_t.$$

Separating variables and integrating backward from T to zero gives

$$\log\frac{X_0}{X_T} = \int_T^0 \frac{c}{2(1+t)}\,\mathrm{d}t = -\frac{c}{2}\log(1+T),$$

or $X_0 = (1+T)^{-c/2}X_T$. Since $\mathrm{Var}(X_T) = 1+T$,

$$\mathrm{Var}(X_0^{\mathrm{PF}}) = (1+T)^{-c}(1+T) = (1+T)^{1-c}.$$

This proves (B.27). □

Proof of (3.8) *and* (3.9). For centered one-dimensional Gaussians, a direct computation gives

$$\mathrm{KL}\big(\mathcal{N}(0,v)\,\big\|\,\mathcal{N}(0,1)\big) = h(v) := \frac{1}{2}\big(v - 1 - \log v\big), \qquad v > 0, \tag{B.28}$$

where h is strictly decreasing on $(0,1)$, strictly increasing on $(1,\infty)$, and vanishes only at $v = 1$.

SiD output. Put $r := c^{-2}$. From the proof of (B.26), q_{SiD} is the unique positive solution of $q(q+T) = r(1+T)$. The map $q \mapsto q(q+T)$ is strictly increasing on $(0,\infty)$, takes the value $1+T$ at $q = 1$ and the value $r(r+T) = r(1+T)+r(r-1)$ at $q = r$. If $r > 1$, then $1+T < r(1+T) < r(r+T)$, so $1 < q_{\mathrm{SiD}} < r$; if $r < 1$, the inequalities reverse and $r < q_{\mathrm{SiD}} < 1$. In both cases q_{SiD} lies between 1 and r, and since h is monotone on each side of 1,

$$\mathrm{KL}\big(P^{\mathrm{SiD}}\,\big\|\,P^{\mathrm{data}}\big) = h(q_{\mathrm{SiD}}) \le h(r) = \frac{1}{2}\big(c^{-2} - 1 - \log c^{-2}\big) = \kappa(c),$$

uniformly in $T > 0$. The total-variation bound follows from the Bretagnolle–Huber inequality $\mathrm{TV}(P, Q) \le \sqrt{1 - \exp(-\mathrm{KL}(P\|Q))}$ (Canonne, 2023), which gives

$$\mathrm{TV}\big(P^{\mathrm{SiD}}, P^{\mathrm{data}}\big) \le \sqrt{1 - e^{-h(q_{\mathrm{SiD}})}} \le \sqrt{1 - e^{-\kappa(c)}} < 1.$$

Probability-flow output. By (B.27) and (B.28) with $v = (1+T)^{1-c}$, so that $\log v = (1-c)\log(1+T)$,

$$\mathrm{KL}\big(P^{\mathrm{PF}} \,\big\|\, P^{\mathrm{data}}\big) = \frac{1}{2}\Big[(1+T)^{1-c} - 1 - (1-c)\log(1+T)\Big],$$

which is (3.9). If $c < 1$, the power $(1+T)^{1-c}$ dominates the logarithm and the expression grows like $\frac{1}{2}(1+T)^{1-c}$; if $c > 1$, the power tends to zero and the expression grows like $\frac{1}{2}(c-1)\log(1+T)$. In either case it diverges as $T \to \infty$.

It remains to show that $\mathrm{TV}(P^{\mathrm{PF}}, P^{\mathrm{data}}) \to 1$. Total variation is invariant under the bijection $x \mapsto x/\sqrt{v}$ applied to both measures, which maps $\mathcal{N}(0, v)$ to $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 1)$ to $\mathcal{N}(0, 1/v)$; hence $\mathrm{TV}(\mathcal{N}(0, v), \mathcal{N}(0, 1)) = g(\max\{v, 1/v\})$ with $g(u) := \mathrm{TV}(\mathcal{N}(0, u), \mathcal{N}(0, 1))$ for $u \ge 1$. For any $\tau > 0$, taking the event $\{|x| \le \tau\}$ in the definition of total variation gives

$$g(u) \ge \mathbb{P}_{Z\sim\mathcal{N}(0,1)}\big(|Z| \le \tau\big) - \mathbb{P}_{Z\sim\mathcal{N}(0,1)}\big(|Z| \le \tau/\sqrt{u}\big) \ge 1 - 2e^{-\tau^2/2} - \sqrt{\tfrac{2}{\pi}}\,\frac{\tau}{\sqrt{u}},$$

where we used the Gaussian tail bound $\mathbb{P}(|Z| > \tau) \le 2e^{-\tau^2/2}$ and the density bound $\mathbb{P}(|Z| \le \delta) \le \sqrt{2/\pi}\,\delta$. Choosing $\tau = u^{1/4}$ shows $g(u) \to 1$ as $u \to \infty$. Since $v = (1+T)^{1-c}$ satisfies $\max\{v, 1/v\} = (1+T)^{|1-c|} \to \infty$ as $T \to \infty$ for every $c \ne 1$, we conclude that $\mathrm{TV}(P^{\mathrm{PF}}, P^{\mathrm{data}}) \to 1$, which completes the proof. □