Fast Statistical Rates for Flow Matching without Curvature Assumptions

Hari Krishna Sahoo
Purdue University
West Lafayette, IN 47906
hsahoo@purdue.edu

Mudit Gaur
Purdue University
West Lafayette, IN 47906
mgaur@purdue.edu

Vaneet Aggarwal
Purdue University
West Lafayette, IN 47906
vaneet@purdue.edu

Abstract

Flow-based generative models have emerged as an efficient alternative to diffusion models. We study the sample complexity of learning a target distribution in Wasserstein distance via flow matching, which learns a velocity field transporting a base distribution to the target. Under exact empirical risk minimization over a fixed network class and standard assumptions on the architecture and data distribution, we show that any squared-loss flow-matching variant including rectified flow with bounded derivatives achieves $\widetilde{O}(\varepsilon^{-2})$ sample complexity for the statistical term with the Wasserstein guarantee taking the form $W_2 \leq K(\varepsilon + C\sqrt{\varepsilon_{\text{app}}})$, where ε_{app} is the approximation error of the network class. This improves upon existing $\widetilde{O}(\varepsilon^{-4})$ guarantees and does not require the Polyak–Łojasiewicz condition in the statistical analysis. Instead, we exploit a structural property of the squared-loss objective - it induces an approximate Bernstein condition that directly ties the variance of the excess loss to the excess risk up to additive approximation-error. This condition enables a localized Rademacher complexity analysis yielding fast $O(1/n)$ rates. The flow-matching schedule enters the leading statistical prefactor through the Lipschitz constant of the pointwise loss, which we characterize explicitly for standard schedules.

1 Introduction

Generative modeling has become a central tool in modern machine learning, driving advances across scientific discovery, vision, language, and biology [26, 29, 30]. A major milestone was the development of denoising diffusion probabilistic models (DDPMs) [16]. Score-based formulations followed, accelerating sampling via ODE representations [27]. These approaches remain computationally intensive. Flow-matching models [21] offer a continuous-time alternative: they learn a velocity field that transports a base distribution to the target. Variants differ in their interpolation path and regression target, but all share a squared-loss training objective. Prominent instances include rectified flow [23], which uses linear interpolation, and variance-preserving (VP) formulations with trigonometric or β-schedules.

We study the sample complexity of distribution learning via flow-matching models. Existing analyses establish $\widetilde{O}(\varepsilon^{-4})$ rates [14, 13]. An $\Omega(\varepsilon^{-2})$ lower bound holds for integrated velocity estimation (Appendix E). Many prior analyses that study this problem also rely on the Polyak–Łojasiewicz (PL) condition, which is a strong curvature assumption on the loss landscape, and its validity for neural networks is difficult to verify in general. This motivates our central question:

> Can flow-matching generative models achieve $\widetilde{O}(\varepsilon^{-2})$ sample complexity for the statistical term, without relying on curvature assumptions such as PL?

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Work	Algorithm	Sample Complexity in $\widetilde{O}(\cdot)$	PL	Key Assumptions
[8]	DDPM	$\varepsilon^{-O(d)}$	No	Supported on unit cube
[6]	DDPM	$\frac{R^{6d}+R^{2d}}{\varepsilon^4}\frac{(\Theta^2 P)^D}{\sqrt{D}}$	No	Ball-supported class approximates score
[15]	DDPM	$\frac{d^2 PD}{\varepsilon^5}\cdot\log\Theta$	No	2nd moment bounded; ERM
[14]	DDPM	$\frac{W^{2D}\cdot d^2}{\varepsilon^4}$	Yes	Sub-Gaussian; no ERM
[13]	Flow matching	$\frac{W^{2D-2}\,d^2}{\varepsilon^4}$	Yes	Bounded; no ERM
Our Work	**FM variant**	$\frac{L_\ell^2 P}{\varepsilon^2}$	No	Sub-Gaussian; ERM

Table 1: Comparison of sample complexity results for distribution learning. The guarantees use different metrics and optimization assumptions.

We answer in the affirmative. The key observation is structural: the bias–variance decomposition of the squared loss yields an *approximate Bernstein condition*. For any θ in the hypothesis class,

$$\mathbb{E}\|v_\theta - v_{\theta^*}\|^2 \le 2(\mathcal{L}(\theta) - \mathcal{L}(\theta^*)) + 4\varepsilon_{\mathrm{app}}, \tag{1}$$

where v_{θ^*} is the population minimizer in the parametric class $\{v_\theta : \theta \in \Theta\}$. The term $\varepsilon_{\mathrm{app}} = \mathbb{E}\|v_{\theta^*} - v^*\|^2$ is the approximation error: the gap between the best in-class velocity field and the true regression function v^*. The condition holds for any squared-loss regression, so the analysis covers all flow-matching variants uniformly.

Local Rademacher under approximate Bernstein. Equation (1) is the engine for our main technical result. The standard Bartlett–Bousquet–Mendelson framework [4] for local Rademacher complexity requires the *exact* Bernstein condition. We extend it to the approximate setting. The residual $\rho = 4L_\ell^2\varepsilon_{\mathrm{app}}$ is carried through Talagrand's inequality and the peeling argument directly, where L_ℓ is the Lipschitz constant of the pointwise loss. The resulting inequality acquires only an additive $O(\varepsilon_{\mathrm{app}})$ floor. This gives a statistical error of $\widetilde{O}(L_\ell^2 P/n) + O(\varepsilon_{\mathrm{app}})$, with P the number of network parameters. The Benton et al. reduction [5] then yields the Wasserstein bound

$$W_2(\hat{\pi}_1, \pi_1) \le K\big(O(\varepsilon) + O(\sqrt{\varepsilon_{\mathrm{app}}})\big),$$

which requires $n = \widetilde{O}(L_\ell^2 P/\varepsilon^2)$ samples. The approximation error $\varepsilon_{\mathrm{app}}$ is set by the expressiveness of the network class. The leading statistical prefactor depends on the schedule through L_ℓ, which controls the pointwise loss on the truncated domain.

In short, two structural facts drive the rate. First, the squared-loss bias–variance decomposition yields approximate Bernstein for free, with no curvature assumption. Second, localization carrying the residual yields a fast rate up to a model-misspecification floor. The schedule-dependent constant L_ℓ provides a quantitative explanation for why some schedules outperform others in practice.

Complementary end-to-end analyses of flow matching [12] and diffusion models [24, 11] control approximation and estimation under regularity assumptions. We focus on the statistical error, since this is where the rate improvement arises. The approximation error is controlled by network expressiveness. Our contributions are:

- Any squared-loss flow-matching variant achieves statistical error $\widetilde{O}(L_\ell^2 P/n) + O(\varepsilon_{\mathrm{app}})$. This yields a Wasserstein guarantee $W_2 \le K(\varepsilon + C\sqrt{\varepsilon_{\mathrm{app}}})$ with $n = \widetilde{O}(L_\ell^2 P/\varepsilon^2)$ samples. The result improves the ε-scaling from $\widetilde{O}(\varepsilon^{-4})$ [14, 13] and makes model misspecification explicit.
- We use the squared-loss *approximate Bernstein condition* (Lemma 1) in place of the PL-based exact Bernstein condition in [13]; the residual-aware localization is developed in Theorem 3.
- We develop a local Rademacher theory under approximate Bernstein (Theorem 3). The Bartlett–Bousquet–Mendelson framework [4] is extended by carrying the additive residual ρ through Talagrand's inequality and the peeling argument. Only an $O(\varepsilon_{\mathrm{app}})$ floor is incurred in the final inequality. This broadens the scope of local Rademacher results.
- We characterize the schedule-dependent constant L_ℓ explicitly (Section 6), giving a quantitative explanation for performance differences across schedules.
- We complement the upper bounds with a lower bound having the same ε exponent (Appendix E): $\Omega(\varepsilon^{-2})$ samples are necessary for expected integrated squared velocity error at most ε^2.

2 Preliminaries

In this section, we introduce the mathematical framework underlying flow-matching generative models, describe the general formulation, and present the associated learning objective.

Let π_0 and π_1 be probability distributions on $\mathbb{R}^d$. We assume that π_0 is a known and easily sampleable reference distribution (e.g., a standard Gaussian), while π_1 is observed only through i.i.d. samples. A flow model looks for a deterministic map pushing π_0 to π_1 through a continuous trajectory,

$$X : [0,1] \to \mathbb{R}^d, \qquad t \mapsto X_t, \tag{2}$$

governed by a time-dependent velocity field $v : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$. The trajectory evolves according to the ordinary differential equation

$$dX_t = v(X_t, t)\, dt, \qquad X_0 = z. \tag{3}$$

Algorithm 1 Flow Matching

Require: Samples from π_0 and π_1, interpolation I_t, target Z_t, network v_θ, iterations T

Draw n i.i.d. pairs (X_0^i, X_1^i) with $X_0^i \sim \pi_0$, $X_1^i \sim \pi_1$
for each pair (X_0^i, X_1^i) **do**
  Sample $t_i \sim \mathrm{Uniform}[0,1]$
  Set $X_{t_i}^i = I_{t_i}(X_0^i, X_1^i)$
  Set $Z_{t_i}^i = Z_{t_i}(X_0^i, X_1^i)$
end for
for iteration $j = 1$ to T **do**
  Update θ to minimize:
  $\hat{\mathcal{L}}(\theta) = \frac{1}{n}\sum_{i=1}^n \|v_\theta(X_{t_i}^i, t_i) - Z_{t_i}^i\|^2$
end for
Sample $Z_0 \sim \pi_0$ and solve $\frac{dZ_t}{dt} = v_\theta(Z_t, t)$
return Z_1

Sampling $X_0 \sim \pi_0$ and integrating this ODE yields a terminal variable X_1 whose distribution ideally matches π_1.

A *flow-matching scheme* (see Algorithm 1) specifies an interpolation map $I_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ and a target map $Z_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$. Given paired samples (X_0, X_1) from π_0 and π_1 For the distributional guarantees we consider the function $t \mapsto I_t(x_0, x_1)$ is continuously differentiable and satisfies $I_0(x_0, x_1) = x_0$ and $I_1(x_0, x_1) = x_1$. We denote interpolated point as $X_t = I_t(X_0, X_1)$ and regression target as $Z_t = \partial_t I_t(X_0, X_1)$. The population-optimal velocity field is

$$v^*(x,t) = \mathbb{E}[\, Z_t \mid X_t = x,\, t\,]. \tag{4}$$

In practice, v^* is approximated by a neural network v_θ, trained by minimizing the population loss

$$\mathcal{L}(\theta) = \mathbb{E}_{t, X_0, X_1}\left[\|v_\theta(X_t, t) - Z_t\|^2\right], \tag{5}$$

where $t \sim \mathrm{Uniform}[0,1]$. Its empirical counterpart is

$$\hat{\mathcal{L}}(\theta) = \frac{1}{n}\sum_{i=1}^n \|v_\theta(X_t^i, t_i) - Z_t^i\|^2, \tag{6}$$

with (X_0^i, X_1^i, t_i) drawn i.i.d. from $\pi_0 \otimes \pi_1 \otimes \mathrm{Uniform}[0,1]$.

Example 1. Standard flow-matching schemes

(a) **Rectified flow / OT path** [23]. $X_t = (1-t)X_0 + tX_1, \quad Z_t = X_1 - X_0$.

(b) **VP (sin/cos).** $X_t = \cos(\frac{\pi}{2}t)X_0 + \sin(\frac{\pi}{2}t)X_1, \quad Z_t = \frac{\pi}{2}(-\sin(\frac{\pi}{2}t)X_0 + \cos(\frac{\pi}{2}t)X_1)$.

(c) **VP (β-schedule).** $X_t = \alpha_t X_0 + \sigma_t X_1$ with $\alpha_t = e^{-\frac{1}{2}\int_0^t \beta(s)ds}$, $\sigma_t = \sqrt{1-\alpha_t^2}$, so that $X_0 \sim \pi_0$ at $t = 0$ and $X_t \to X_1 \sim \pi_1$ as $t \to \infty$ (or $t = 1$ after rescaling). If β is continuous with $\beta(0) > 0$, then $|\dot{\sigma}_t/\alpha_t| \to \infty$ as $t \to 0$.

Remark 1. In rectified flow, the regression target $Z_t = X_1 - X_0$ is independent of t. In general flow matching, Z_t depends on t through the schedule derivatives. This distinction does not affect our statistical analysis, the squared-loss bias-variance decomposition holds regardless, but it does affect the Lipschitz constant of the pointwise loss (see Section 6).

3 Assumptions

To establish our sample-complexity results, we introduce assumptions on the data distribution, the flow-matching schedule, the neural-network architecture, and the approximation quality. These conditions are standard in theoretical analyses of generative models. Notably, we do **not** assume the Polyak–Łojasiewicz condition.

Assumption 1 (Data distributions). Samples (X_0^i, X_1^i, t_i) are i.i.d. draws from $\pi_0 \otimes \pi_1 \otimes \mathrm{Uniform}[0,1]$. The random variables $X_0 \sim \pi_0$ and $X_1 \sim \pi_1$ satisfy sub-Gaussian tail bounds: there exist constants $c, \sigma > 0$ such that for all $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$,

$$\Pr\left(|\langle u, X_j\rangle| \geq t\right) \leq 2\exp\left(-\frac{t^2}{c\sigma^2}\right), \qquad j \in \{0,1\}. \tag{7}$$

Remark 2. The i.i.d. assumption enables concentration of empirical averages, which is required for uniform convergence. Sub-Gaussianity rules out heavy tails and provides the concentration properties needed for our analysis. These conditions are standard in recent theoretical work on diffusion models [14, 15].

Assumption 2 (Schedule regularity). There exists a constant $c_Z > 0$ such that for all $t \in [0,1]$,

$$\|Z_t(x_0, x_1)\| \leq c_Z(\|x_0\| + \|x_1\| + 1). \tag{8}$$

The interpolation also satisfies $\|I_t(x_0, x_1)\| \leq c_I(\|x_0\| + \|x_1\| + 1)$ for a constant $c_I > 0$.

Remark 3. This assumption controls the growth of the regression target and interpolation. It holds with $c_Z = 1$ for rectified flow, $c_Z = \pi/2$ for VP sin/cos, and fails for VP β-schedule with $\beta(0) > 0$ without endpoint clipping. All listed interpolations have $c_I = 1$. The constant c_Z enters the bound through L_ℓ; see Section 6 for a detailed discussion.

Assumption 3 (Neural-network architecture). Let $\mathcal{F}$ denote the class of $\mathbb{R}^d$-valued functions represented by feed-forward neural networks with the following properties:

- It has $D \geq 2$ affine layers, a linear output layer, and connections only between successive layers.
- It contains a total of P trainable parameters.
- Each hidden layer output is bounded in $[-b, b]$.
- The activation functions in the first hidden layer are coordinate-wise non-decreasing.
- For each layer, the corresponding weight matrix W satisfies $\|W\|_{\mathrm{op}} \leq V$ (spectral/operator norm); biases, when present, are included in this constrained matrix by augmenting the input with a constant coordinate. Hidden-layer activations are differentiable and L_ϕ-Lipschitz.

Remark 4. These conditions are standard in neural network theory [2]. Common activations satisfying these conditions include $\tanh$ and $\mathrm{sigmoid}$ whose ranges are bounded. The Bernstein argument is activation-independent; extending the result to other activations requires suitable output and covering bounds.

Assumption 4 (Existence and Uniqueness). For each $x \in \mathbb{R}^d$ and $s \in [0,1]$, the flow ODEs with velocity fields $v_\theta(x,t)$ and $v^*(x,t)$ have unique solutions that are continuously differentiable in x, s, and t.

Remark 5. This ensures that the flow ODEs are well-posed. For v_θ, existence and uniqueness follow from the Lipschitz continuity in x (Corollary 3) via Picard–Lindelöf; differentiability of the flow maps, and well-posedness for v^*, are assumed directly. This assumption also appears in [5] and [13].

Assumption 5 (Approximation error). Let $\theta^* := \arg\min_{\theta\in\Theta} \mathcal{L}(\theta)$. The approximation error is

$$\varepsilon_{\mathrm{app}} := \mathbb{E}_{t,X_0,X_1}\|v_{\theta^*}(X_t,t) - v^*(X_t,t)\|^2.$$

Remark 6. We do *not* assume $\varepsilon_{\mathrm{app}}$ is small. It appears as an explicit additive floor in all our bounds including the final Wasserstein guarantee making the role of model misspecification explicit. The bound separates this approximation floor from the statistical error.

Remark 7 (No PL assumption). Unlike prior work [14, 13], we do not assume the Polyak–Łojasiewicz condition. The Bernstein-type variance control needed for fast statistical rates is derived directly from the squared-loss structure (Lemma 1), not from curvature of the loss landscape.

Assumption 6 (Access to ERM). We assume access to an empirical risk minimizer

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \|v_\theta(X_t^i, t_i) - Z_t^i\|^2.$$

That is, the parameter θ produced by the training procedure satisfies $\theta = \hat{\theta}$, so the optimization error $\mathcal{E}^{\mathrm{opt}} = \mathbb{E}\|v_{\hat{\theta}} - v_\theta\|^2 = 0$.

Remark 8 (Role of Assumption 6). The optimization error $\mathcal{E}^{\mathrm{opt}}$ measures the gap between the empirical risk minimizer (ERM) $\hat{\theta}$ and the parameter θ produced by the training procedure. Under Assumption 6, $\theta = \hat{\theta}$, so $\mathcal{E}^{\mathrm{opt}} = 0$ and the statistical rate in Theorem 4 fully characterizes the estimation error. This assumption isolates the statistical contribution of distribution learning from optimization effects and is standard in prior sample-complexity analyses of diffusion and flow-matching models [6, 15]. Without Assumption 6, an optimization guarantee must be established separately under conditions applicable to the chosen architecture, data, and training procedure. Appendix D discusses how a verified approximate-ERM guarantee can be incorporated without changing the statistical order.

The following properties are consequences of Assumption 3 on the truncated event $\mathcal{E}_M := \{\|X_0\| \leq M,\ \|X_1\| \leq M\}$ defined in Section 5.2. We mention the results here and give detailed proofs in Appendix B.

- The parameter space Θ is compact.
- The network output $v_\theta(x,t)$ is uniformly bounded by some $B_v > 0$ over $\theta \in \Theta$ and (x,t).
- For every $\theta \in \Theta$, the function $v_\theta(x,t)$ is differentiable in (x,t) and L_t-Lipschitz in the state variable x.
- Let ℓ denote the pointwise loss $\ell(v_\theta, Z_t) = \|v_\theta - Z_t\|_2^2$. Then, on $\mathcal{E}_M$, ℓ is L_ℓ-Lipschitz in its first argument, with $L_\ell = 2(B_v + c_Z(2M+1))$.

4 Main Result

In this section, we present our main theoretical results on the sample complexity of flow matching. We begin by recalling a prior result in [5] that links the estimation error of the learned velocity field to the Wasserstein distance between the generated and target distributions. Building on this connection, we introduce an error decomposition that isolates approximation, statistical, and optimization effects.

4.1 Prior Result

Theorem 1 ([5]). *Let π_0, π_1 be probability distributions on $\mathbb{R}^d$. Let v^* be the true velocity field and v_θ the learned field. Define Y_t and Z_t as the flows generated by v_θ and v^* respectively, both starting from $Y_0 = Z_0 \sim \pi_0$. Let $\hat{\pi}_1$ be the law of Y_1 and π_1 the law of Z_1.*

Under the following assumptions:

1. ***Bounded Velocity Error:*** $\int_0^1 \mathbb{E}[\|v_\theta(X_t,t) - v^*(X_t,t)\|^2]\,dt \leq \varepsilon^2$.
2. ***Regularity:*** *$v_\theta(x,t)$ is differentiable in both inputs and L_t-Lipschitz in x.*
3. ***Existence and Uniqueness:*** *Same as Assumption 4.*

The Wasserstein-2 distance satisfies:

$$W_2(\hat{\pi}_1, \pi_1) \leq \varepsilon \cdot \exp\left(\int_0^1 L_t\,dt\right). \tag{9}$$

Assumptions 2 and 3 in Theorem 1 are consequences of Assumption 3 and Assumption 4, so a small L^2 velocity error directly yields small Wasserstein discrepancy.

4.2 Error Decomposition

Thanks to Theorem 1, controlling the velocity estimation error becomes our main objective. To analyze this quantity, we introduce the population and empirical risk minimizers:

$$\theta^* = \arg\min_{\theta\in\Theta} \mathbb{E}_{t,X_0,X_1}\left[\|v_\theta(X_t,t) - Z_t\|^2\right], \tag{10}$$

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \|v_\theta(X_t^i, t_i) - Z_t^i\|^2. \tag{11}$$

Let θ denote the parameter actually produced by the optimizer.

Proposition 1 (Error Decomposition)**.** *The velocity field estimation error decomposes as:*

$$\mathbb{E}\|v_\theta(X_t,t)-v^*(X_t,t)\|^2 \le 3\underbrace{\mathbb{E}\|v^*-v_{\theta^*}\|^2}_{=\varepsilon_{\rm app}\ \textit{(approximation)}} +3\underbrace{\mathbb{E}\|v_{\theta^*}-v_{\hat\theta}\|^2}_{\mathcal{E}^{\rm stat}\ \textit{(statistical)}}+3\underbrace{\mathbb{E}\|v_{\hat\theta}-v_\theta\|^2}_{\mathcal{E}^{\rm opt}\ \textit{(optimization)}}.$$

Apply $\|a+b+c\|^2 \le 3(\|a\|^2+\|b\|^2+\|c\|^2)$ *to* $v^*-v_\theta=(v^*-v_{\theta^*})+(v_{\theta^*}-v_{\hat\theta})+(v_{\hat\theta}-v_\theta)$.

4.3 Our Main Result

Theorem 2 (Sample Complexity of Flow Matching)**.** *Let the assumptions stated in Section 3 hold. Under Assumption 6, the optimization error vanishes ($\mathcal{E}^{\rm opt}=0$); Appendix D discusses approximate ERM under separately verified optimization assumptions. A sample size $n=\widetilde{O}(L_\ell^2 P/\varepsilon^2)$ suffices to ensure, with probability at least $1-\delta$, that the velocity estimation error satisfies*

$$\int_0^1 \mathbb{E}\|v_\theta(X_t,t)-v^*(X_t,t)\|^2\,dt \le C_1\varepsilon^2+C_2\varepsilon_{\rm app}, \tag{12}$$

where $C_1, C_2 > 0$ are universal constants. Consequently, by Theorem 1,

$$W_2(\hat\pi_1,\pi_1) \le K\sqrt{C_1\varepsilon^2+C_2\varepsilon_{\rm app}} \le K\big(\sqrt{C_1}\,\varepsilon+\sqrt{C_2\varepsilon_{\rm app}}\big), \tag{13}$$

where $K=\exp(\int_0^1 L_t\,dt)$.

Proof Sketch. By Proposition 1:

$$\mathbb{E}\|v_\theta-v^*\|^2 \le 3\varepsilon_{\rm app}+3\mathcal{E}^{\rm stat}+3\mathcal{E}^{\rm opt}.$$

By Theorem 4 (proved in Section 5), with probability at least $1-\delta$ (taking $x=\log(2/\delta)$),

$$\mathcal{E}^{\rm stat} \le \widetilde{O}\left(\frac{L_\ell^2 P}{n}\right)+8\varepsilon_{\rm app}.$$

Choosing a sufficiently large $n=\widetilde{O}(L_\ell^2P/\varepsilon^2)$ gives $\mathcal{E}^{\rm stat}\le\varepsilon^2+8\varepsilon_{\rm app}$.

Substituting $\mathcal{E}^{\rm opt}=0$ (Assumption 6), we have

$$\mathbb{E}\|v_\theta-v^*\|^2 \le 3\varepsilon_{\rm app}+3(\varepsilon^2+8\varepsilon_{\rm app})=3\varepsilon^2+27\varepsilon_{\rm app},$$

which gives (12) with $C_1=3$ and $C_2=27$. The Wasserstein bound (13) follows from Theorem 1 and $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$. □

Remark 9. Theorem 2 cleanly separates two independent requirements for accurate distribution learning. The *sample complexity* $n=\widetilde{O}(L_\ell^2P/\varepsilon^2)$ controls the statistical error for the fixed class. The *approximation error* $\varepsilon_{\rm app}$ is an architecture-dependent floor controlled by the expressiveness of the network class: a richer architecture decreases $\varepsilon_{\rm app}$ at the cost of increasing P.

5 Statistical Error

This section presents the core technical contribution of the paper: a localized Rademacher complexity analysis for squared-loss flow matching that achieves fast $\widetilde{O}(1/n)$ rates without the PL condition. The argument proceeds in three steps. First, we derive an approximate Bernstein condition from the squared-loss structure. Second, we feed it into a localization argument that directly extends the standard local Rademacher framework, carrying the additive approximation residual through the proof. Third, we combine the resulting inequality with a covering bound for the network class. The most non-trivial steps are in proving the approximate Bernstein version of local Rademacher results are inspired from [4].

Since $v^*(x,t)=\mathbb{E}[Z_t \mid X_t=x,t]$ is the regression function, the bias–variance decomposition gives, for any measurable f,

$$\mathbb{E}\|f(X_t,t)-Z_t\|^2=\mathbb{E}\|f(X_t,t)-v^*(X_t,t)\|^2+\mathbb{E}\|Z_t-v^*(X_t,t)\|^2. \tag{14}$$

The second term does not depend on f. The excess risk therefore equals the difference in squared population errors:

$$\mathcal{L}(\theta)-\mathcal{L}(\theta^*)=\mathbb{E}\|v_\theta-v^*\|^2-\mathbb{E}\|v_{\theta^*}-v^*\|^2. \tag{15}$$

5.1 Approximate Bernstein condition

The first step is a Bernstein-type bound that relates $\mathbb{E}\|v_\theta - v_{\theta^*}\|^2$ to the excess risk with additional approximation error term.

Lemma 1. *For every $\theta \in \Theta$, writing $\Delta_\theta = \mathcal{L}(\theta) - \mathcal{L}(\theta^*)$,*

$$\mathbb{E}\|v_\theta(X_t,t) - v_{\theta^*}(X_t,t)\|^2 \le \left(\sqrt{\varepsilon_{\mathrm{app}}} + \sqrt{\varepsilon_{\mathrm{app}} + \Delta_\theta}\right)^2 \le 2\Delta_\theta + 4\varepsilon_{\mathrm{app}}. \tag{16}$$

Proof. Expand $v_\theta - v^* = (v_\theta - v_{\theta^*}) + (v_{\theta^*} - v^*)$ in (15) to get

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) = \mathbb{E}\|v_\theta - v_{\theta^*}\|^2 + 2\mathbb{E}\langle v_\theta - v_{\theta^*},\ v_{\theta^*} - v^*\rangle, \tag{17}$$

and rearrange:

$$\mathbb{E}\|v_\theta - v_{\theta^*}\|^2 = \mathcal{L}(\theta) - \mathcal{L}(\theta^*) - 2\mathbb{E}\langle v_\theta - v_{\theta^*},\ v_{\theta^*} - v^*\rangle. \tag{18}$$

By Cauchy–Schwarz,

$$\left|2\mathbb{E}\langle v_\theta - v_{\theta^*},\ v_{\theta^*} - v^*\rangle\right| \le 2\big(\mathbb{E}\|v_\theta - v_{\theta^*}\|^2\big)^{1/2}\sqrt{\varepsilon_{\mathrm{app}}}, \tag{19}$$

and substitution into (18) yields

$$\mathbb{E}\|v_\theta - v_{\theta^*}\|^2 \le \mathcal{L}(\theta) - \mathcal{L}(\theta^*) + 2\big(\mathbb{E}\|v_\theta - v_{\theta^*}\|^2\big)^{1/2}\sqrt{\varepsilon_{\mathrm{app}}}. \tag{20}$$

Writing $\Delta_\theta = \mathcal{L}(\theta) - \mathcal{L}(\theta^*)$, the sharper form is

$$\mathbb{E}\|v_\theta - v_{\theta^*}\|^2 \le \left(\sqrt{\varepsilon_{\mathrm{app}}} + \sqrt{\varepsilon_{\mathrm{app}} + \Delta_\theta}\right)^2.$$

This follows by solving the quadratic inequality in (20); the relaxation below is convenient for the subsequent analysis.

Young's inequality $2ab \le \frac{1}{2}a^2 + 2b^2$ with $a = (\mathbb{E}\|v_\theta - v_{\theta^*}\|^2)^{1/2}$ and $b = \sqrt{\varepsilon_{\mathrm{app}}}$ then gives

$$\mathbb{E}\|v_\theta - v_{\theta^*}\|^2 \le \mathcal{L}(\theta) - \mathcal{L}(\theta^*) + \tfrac{1}{2}\mathbb{E}\|v_\theta - v_{\theta^*}\|^2 + 2\varepsilon_{\mathrm{app}}.$$

Subtracting $\frac{1}{2}\mathbb{E}\|v_\theta - v_{\theta^*}\|^2$ from both sides and multiplying by 2 yields (16). □

To pass from a function-level bound to a variance bound at the loss level, define the pointwise excess loss

$$g_\theta := \|v_\theta(X_t,t) - Z_t\|^2 - \|v_{\theta^*}(X_t,t) - Z_t\|^2, \tag{21}$$

so that $\Delta_\theta := \mathbb{E}[g_\theta] = \mathcal{L}(\theta) - \mathcal{L}(\theta^*) \ge 0$. For bounded targets, let $h_\theta = g_\theta$; for sub-Gaussian data, let $h_\theta = g_\theta \mathbf{1}_{\mathcal{E}_M}$, where $\mathcal{E}_M$ is defined in Section 5.2. If $\|Z_t\| \le M'$ on the domain where h_θ is nonzero, set $L_\ell = 2(B_v + M')$ and $B_\ell = (B_v + M')^2$. Then $|h_\theta| \le B_\ell$, and the identity $\|a\|^2 - \|b\|^2 = \langle a - b, a + b\rangle$ gives

$$|g_\theta| \le \|v_\theta - v_{\theta^*}\|\,(2B_v + 2\|Z_t\|) \le L_\ell\|v_\theta - v_{\theta^*}\| \tag{22}$$

on that domain. Under the original sampling distribution,

$$\mathrm{Var}(h_\theta) \le Ph_\theta^2 \le L_\ell^2\mathbb{E}\|v_\theta - v_{\theta^*}\|^2. \tag{23}$$

Inserting Lemma 1 produces the loss-level Bernstein bound

$$Ph_\theta^2 \le \underbrace{2L_\ell^2}_{=:B_0}\Delta_\theta + \underbrace{4L_\ell^2\varepsilon_{\mathrm{app}}}_{=:\rho}. \tag{24}$$

5.2 Localized Rademacher complexity with approximate Bernstein

The standard localized Rademacher framework [4, 18] assumes an exact Bernstein condition. We extend it to handle (24), carrying the residual ρ through concentration and peeling. We localize the loss class itself by Ph_θ^2. The following form also retains the small population discrepancy introduced by truncation. Write $R_n(\mathcal{H}) = \mathbb{E}_\epsilon \sup_{h\in\mathcal{H}} |n^{-1}\sum_i \epsilon_i h(\xi_i)|$.

Theorem 3. *Let $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$ contain zero, with $|h_\theta| \le B_\ell$, $B_0, B_\ell > 0$, and $\rho, \tau \ge 0$. Suppose $\Delta_\theta \ge 0$, $|Ph_\theta - \Delta_\theta| \le \tau$, and $Ph_\theta^2 \le B_0\Delta_\theta + \rho$. Let ψ be a sub-root function with positive fixed point r^* satisfying*

$$\psi(r) \ge B_0 \mathbb{E}R_n\{h \in \mathcal{H} : Ph^2 \le r\}.$$

Set $J = \lceil \log_2 \max\{1, B_\ell^2/r^*\} \rceil$ and $y = x + \log(J+1)$. For $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$, every $\hat{\theta}$ satisfying $P_n h_{\hat{\theta}} \le 0$ obeys

$$\Delta_{\hat{\theta}} \le \frac{C_1 K^2 (B_0 + B_\ell) r^*}{(K-1)B_0^2} + \frac{C_2 K^2 (B_0 + B_\ell) y}{(K-1)n} + \frac{K}{K-1}\left(\frac{\rho}{2B_0} + \tau\right), \tag{25}$$

with $C_1 = 48$ and $C_2 = 9$.

The proof uses finitely many second-moment shells and retains their confidence cost $\log(J+1)$; see Appendix C.4. For bounded targets, $h = g$ and $\tau = 0$. With $K = 2$, $B_0 = 2L_\ell^2$, and $\rho = 4L_\ell^2 \varepsilon_{\text{app}}$, the approximation term is $\rho/B_0 = 2\varepsilon_{\text{app}}$.

It remains to control r^*, which requires a covering bound for the network class.

Lemma 2 (Covering-number bound)**.** *Under Assumption 3, restrict the vector-valued network class to $\|x\| \le R$, $t \in [0,1]$. If L_R is its parameter Lipschitz constant from Corollary 4, then*

$$\log \mathcal{N}_\infty(u, \mathcal{F}) \le P \log\left(1 + \frac{2V\sqrt{P}L_R}{u}\right), \qquad u > 0,$$

where the function distance uses the Euclidean norm of the vector output.

For the corresponding loss class, define

$$A_R = B_\ell + 4V\sqrt{P}\, L_\ell L_R, \qquad H_R = 1 + P \log\left(1 + \frac{A_R n}{B_\ell}\right), \qquad q_R = \frac{H_R}{n}.$$

An empirical Dudley bound with deterministic cutoff B_ℓ/n, combined with symmetrization and contraction for the squared loss class, gives

$$\mathbb{E}R_n\{h \in \mathcal{H} : Ph^2 \le r\} \le C\big(\sqrt{rq_R} + B_\ell q_R\big).$$

This step explicitly transfers the population radius to the empirical radius (Appendix C.5). A valid sub-root function is therefore

$$\psi(r) = CB_0\big(\sqrt{rq_R} + B_\ell q_R\big) + \frac{B_0^2}{n}. \tag{26}$$

Its fixed point satisfies

$$\frac{B_0^2}{n} \le r^* \le C'(B_0^2 + B_0 B_\ell)\frac{H_R}{n}. \tag{27}$$

The lower bound follows directly from the last term of (26). Since $B_\ell = B_0/8$, the shell contribution $\log(J+1)$ is absorbed by $H_R \ge 1 + \log(1+n)$.

Combining Theorem 3 with (27) and choosing $K = 2$ gives the bounded-loss bound.

Lemma 3. *Suppose $\|X_t\| \le R$ and $\|Z_t\| \le M'$ almost surely. Under Assumptions 3, 5 and 6, for every $x > 0$, with probability at least $1 - e^{-x}$,*

$$\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta^*) \le \frac{CL_\ell^2(H_R + x)}{n} + 2\varepsilon_{\text{app}}, \tag{28}$$

where C is universal. Converting to velocity error via Lemma 1 yields the dominant term $\widetilde{O}(L_\ell^2 P/n)$ and the floor $8\varepsilon_{\text{app}}$.

Extension to sub-Gaussian data. Define $\mathcal{E}_M = \{\|X_0\| \le M, \|X_1\| \le M\}$ and choose

$$M^2 = 4c\sigma^2\big(d\log 5 + x + \log(4n^2)\big),$$

where c, σ are from Assumption 1. On $\mathcal{E}_M$, set $R_M = c_I(2M+1)$ and $M' = c_Z(2M+1)$. We apply the theorem to $h_\theta = g_\theta \mathbf{1}_{\mathcal{E}_M}$ under the original sampling law. The event that all n observations lie in $\mathcal{E}_M$ has failure probability at most e^{-x}/n, and the population truncation discrepancy satisfies $\tau \le L_\ell^2 e^{-x}/n^2$ (Appendix C.8). On this event, the original ERM satisfies $P_n h_{\hat{\theta}} \le 0$.

Theorem 4 (Statistical error)**.** *Under Assumptions 1 to 3, 5 and 6, for $n \geq 2$ and every $x > 0$, with probability at least $1 - 2e^{-x}$,*

$$\mathcal{E}^{\mathrm{stat}} := \mathbb{E}\|v_{\hat{\theta}} - v_{\theta^*}\|^2 \leq \frac{CL_\ell^2(H_{R_M} + x)}{n} + 8\varepsilon_{\mathrm{app}}, \tag{29}$$

where $L_\ell = 2(B_v + c_Z(2M+1))$, H_{R_M} is defined above, and C is universal. Here $M = O(\sigma\sqrt{c(d + \log n + x)})$ and L_{R_M}, hence H_{R_M}, displays the architecture and input-radius dependence explicitly.

6 Schedule-Dependent Constants

The statistical rate in Theorem 4 is $\widetilde{O}(L_\ell^2 P/n) + O(\varepsilon_{\mathrm{app}})$. The schedule enters through $L_\ell = 2(B_v + c_Z(2M+1))$, where c_Z is the schedule constant from Assumption 2 and $M = O(\sigma\sqrt{d + \log(n/\delta)})$ is the truncation level at confidence $1 - \delta$.

Proposition 2 (Schedule-dependent constants)**.** *Under Assumption 2, $L_\ell = O(B_v + c_Z(1 + \sigma\sqrt{d + \log(n/\delta)}))$. For fixed architecture and confidence, the following convenient upper bounds apply to standard schemes:*

Schedule	c_Z	L_ℓ *scaling*
Rectified flow / OT	1	$O(\sigma\sqrt{d \log n})$
VP sin/cos	$\pi/2$	$O(\sigma\sqrt{d \log n})$
VP β-schedule (coefficient floor τ)	$O(1/\tau)$	$O(\sigma\sqrt{d \log n}/\tau)$

Remark 10. For a clipped VP β-schedule, let $\tau \in (0, 1]$ denote the endpoint coefficient floor, rather than time. Suppose its endpoint bias is $O(\tau)$ and $c_Z \leq C_Z/\tau$. Define

$$A := C\big[B_v^2 + C_Z^2(2M+1)^2\big]\big[H_{R_M} + \log(2/\delta)\big],$$

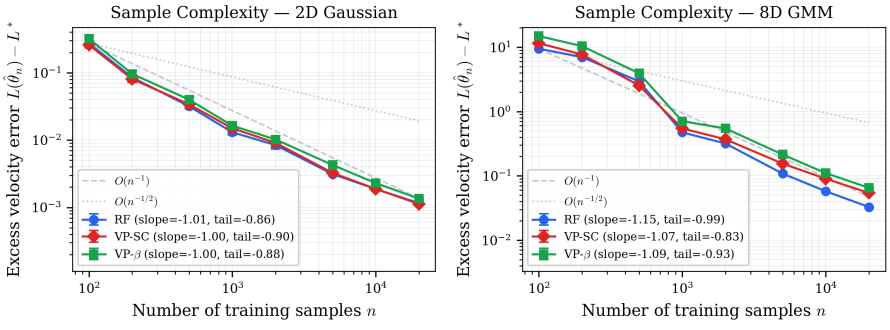
where H_{R_M} is the entropy factor used in (27), containing the parameter count P, depth, weight and activation bounds, and truncated input radius. Then Theorem 4 bounds the statistical velocity error by $A/(n\tau^2)$, up to the approximation floor. The statistical and clipping contributions to the Wasserstein bound take the form

$$K_\tau\left(\tau + \frac{\sqrt{A/n}}{\tau}\right),$$

where $K_\tau = \exp(\int L_{t,\tau}\, dt)$ is the flow-stability factor. If the endpoint-bias constant, A, and K_τ are uniformly controlled up to logarithms as $\tau \downarrow 0$, balancing gives $\tau \asymp (A/n)^{1/4}$ and a $\widetilde{O}(n^{-1/4})$ contribution. Under these conditions, $\widetilde{O}(\varepsilon^{-4})$ samples suffice for this contribution to be $O(\varepsilon)$; the approximation floor is added separately. The bias here is an endpoint error, not numerical discretization error.

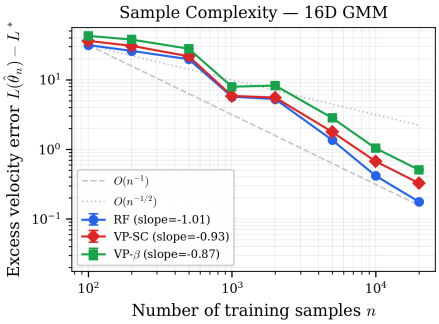
7 Numerical Experiments

To examine the scaling predicted by Theorem 2, we train the same 4-layer width-512 MLP across four data distributions of increasing dimension $d \in \{2, 8, 16, 32\}$: a 2D Gaussian and three axis-aligned GMMs. The SiLU experiments illustrate behavior beyond the bounded-activation class used in our analysis. Figure 1 shows the excess velocity error $L(\hat{\theta}_n) - L^*$ vs. training-set size n on log-log axes. Rectified flow attains slope close to the predicted -1 in every panel, consistent with the predicted $\widetilde{O}(L_\ell^2 P/n)$ scaling. The ordering RF $<$ VP-SC $<$ VP-β widens with d, consistent with the schedule-dependent constant $L_\ell = O(c_Z\sigma\sqrt{d \log n})$ from Section 6 entering quadratically into the statistical-error prefactor. VP-SC and VP-β further exhibit mild slope-flattening at higher d. A nonzero approximation floor may contribute to this behavior; a larger statistical prefactor alone does not change the predicted slope. Full setup and evaluation methodology are deferred to Appendix F.

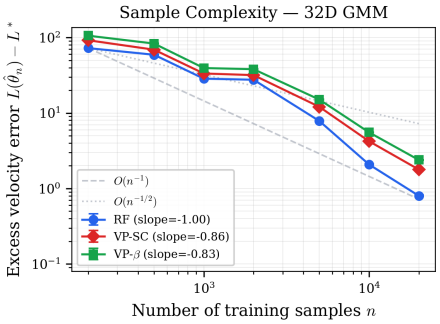


(a) 2D Gaussian

(b) 8D GMM



(c) 16D GMM



(d) 32D GMM

Figure 1: Excess velocity error vs. training-set size n for four targets of increasing dimension, same architecture (4-layer MLP, width 512). Error bars are ± 1 SEM over 7 seeds.

8 Conclusion

We show that flow matching achieves $\widetilde{O}(\varepsilon^{-2})$ sample complexity for the statistical term under ERM access, with the Wasserstein guarantee $W_2 \leq K(\varepsilon + C\sqrt{\varepsilon_{\text{app}}})$ making the approximation error explicit. This improves the ε-scaling from $\widetilde{O}(\varepsilon^{-4})$ to $\widetilde{O}(\varepsilon^{-2})$ and removes the PL condition.

We emphasize that our result does not establish a statistical separation between different flow-matching variants. The fast statistical rate $\widetilde{O}(L_\ell^2 P/n)$ holds uniformly across schedules under the stated assumptions. What differs across schedules is not the rate itself, but the constants and the approximation floor ε_{app}. In particular, rectified-flow and optimal transport (OT) paths typically have smaller target-size constants c_Z, which can reduce the estimation constants. This provides a possible explanation for their empirical advantage.

An important direction for future work is to extend these guarantees beyond the current squared-loss setting and better understand when fast rates arise for general flow-matching formulations. In particular, it would be interesting to study whether iterative linearization, trajectory straightening, or related structural reductions can induce approximate Bernstein conditions and localized complexity behavior analogous to those established here.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Science Foundation (NSF) Future Manufacturing (FM) program under Award No. 2229250. The authors also thank Dhruv Sarkar and Aprameyo Chakrabarty for their valuable input and feedback.

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[2] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, United Kingdom, 1999.

[3] Vansh Bansal, Saptarshi Roy, Purnamrita Sarkar, and Alessandro Rinaldo. On the Wasserstein convergence and straightness of rectified flow. *arXiv preprint arXiv:2410.14949*, 2024.

[4] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[5] Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024.

[6] Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and Langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

[7] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. In *Comptes Rendus Mathematique*, volume 334, pages 495–500. Elsevier, 2002.

[8] Minshuo Chen, Song Mei, Jianqing Fan, and Mengdi Wang. An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771*, 2024.

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

[10] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

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

[12] Kenji Fukumizu, Taiji Suzuki, Noboru Isobe, Kazusato Oko, and Masanori Koyama. Flow matching achieves almost minimax optimal convergence. In *International Conference on Learning Representations*, 2025.

[13] Mudit Gaur, Prashant Trivedi, Shuchin Aeron, Amrit Singh Bedi, George K Atia, and Vaneet Aggarwal. Generative modeling with continuous flows: Sample complexity of flow matching. *arXiv preprint arXiv:2512.01286*, 2025.

[14] Mudit Gaur, Prashant Trivedi, Sasidhar Kunapuli, Amrit Singh Bedi, and Vaneet Aggarwal. Sample complexity of diffusion model training without empirical risk minimizer access. *arXiv preprint arXiv:2505.18344*, 2025.

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

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

[17] Varun Kanade, Patrick Rebeschini, and Tomas Vaškevičius. Exponential tail local rademacher complexity risk bounds without the bernstein condition. *Journal of Machine Learning Research*, 25(388):1–43, 2024.

[18] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[19] Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. *Advances in Neural Information Processing Systems*, 37:63082–63109, 2024.

[20] Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1260–1285, 2015.

[21] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[22] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

[23] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

[24] Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26517–26582, 2023.

[25] Alexander Rakhlin, Karthik Sridharan, and Alexandre B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

[26] 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.

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

[28] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

[29] Mengbo Wang, Shourya Verma, Aditya Malusare, Luopin Wang, Yiyang Lu, Vaneet Aggarwal, Mario Sola, Ananth Grama, and Nadia Atallah Lanman. GeneFlow: Translation of single-cell gene expression to histopathological images via rectified flow. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[30] Shamim Yazdani, Akansha Singh, Nripsuta Saxena, Zichong Wang, Avash Palikhe, Deng Pan, Umapada Pal, Jie Yang, and Wenbin Zhang. Generative AI in depth: A survey of recent advances, model variants, and real-world applications. *Journal of Big Data*, 12(1):1–43, 2025.

Appendix

A Related Work

Some of the key related works are summarized in Table 1. The table compares the sample complexity of various generative models, including DDPMs (Denoising Diffusion Probabilistic Models), flow-matching models, and our result for general squared-loss flow matching. In these formulas, ε denotes the desired approximation error, typically measured in Wasserstein distance. The sample complexity depends on d (data dimensionality), Θ (bound on parameters in the neural network), R (hypothesis class support radius), D (depth of the neural network), W (width of the neural network), and P (number of parameters). In our setting, L_ℓ denotes the Lipschitz constant of the pointwise loss (schedule-dependent; see Section 6), and $\varepsilon_{\mathrm{app}}$ is the approximation error defined in Assumption 5. Our result achieves more efficient scaling with respect to ε compared to previous methods and removes the PL assumption. The dependence on the ambient dimension d enters through the loss bound, approximation error, and network complexity P, which for a fully connected network of depth D and width W scales as $\mathcal{O}(dW + DW^2)$.

Theoretical analysis of sample complexity for generative models has primarily focused on diffusion-based methods. [6] assume a hypothesis class $\mathcal{F}$ of $\mathbb{R}^d$-valued functions that are L^2-Lipschitz, supported on a Euclidean ball of radius R, and sufficiently expressive to uniformly approximate the true score function s_t on this ball. [9] simplified the original problem by proving that learning the score function accurately enough is sufficient for accurately sampling from the target distribution. Building on this result, [15] showed a sample complexity of $O(\varepsilon^{-5})$ under the assumption of access to the empirical risk minimizer (ERM), while [14] further improved this to $O(\varepsilon^{-4})$ for DDPMs. The latter (without ERM access) requires the PL condition.

Analogous to results for diffusion models [9], prior work [5] reduces generative modeling to the problem of learning an unknown velocity field. Building on this perspective, [13] established $\widetilde{O}(\varepsilon^{-4})$ rates using global uniform convergence arguments and the PL condition. However PL does most of the heavy lifting in their proofs involving both statistical and optimization terms.

Rectified Flow was introduced by [23] with linear interpolation paths. [19] show that few rectification steps suffice for nearly straight trajectories. [3] study asymptotic Wasserstein convergence under accurate velocity estimation. Variance-preserving formulations [27] and optimal-transport paths [21] are widely used alternatives. However, existing analyses do not address finite-sample complexity across these variants, nor do they exploit the common squared-loss structure, which is the focus of this work.

The connection between Bernstein-type variance conditions and fast rates in empirical risk minimization is well established [4, 18, 7]. We extend this machinery to the *approximate* Bernstein setting arising from squared-loss regression with model misspecification by carrying the additive residual ρ through the standard peeling and Talagrand-inequality steps (Theorem 3); the resulting oracle inequality may be of independent interest.

Approximation-based rates. [12, 24, 11] control network approximation and estimation under regularity assumptions. Our fixed-class analysis complements these end-to-end results by isolating the statistical term and retaining $\varepsilon_{\mathrm{app}}$ explicitly. The comparison of available upper bounds does not establish a statistical separation between flow matching and DDPMs.

Misspecified least squares. The inequality in Lemma 1 follows from standard squared-loss identities. Offset-complexity analyses [20, 17] and aggregation of empirical minimizers [25] provide alternative fast-rate approaches. We retain plain ERM and carry the misspecification residual through classical localization, yielding an additive approximation floor.

B Consequences of the Architectural Assumptions

In this section, we verify that several regularity properties follow directly from Assumption 3 and Assumption 2. All results hold on the truncated event $\mathcal{E}_M := \{\|X_0\| \le M,\ \|X_1\| \le M\}$.

Corollary 1 (Compactness of the parameter space)**.** *The parameter space Θ is compact.*

Proof. By Assumption 3, the total number of trainable parameters is finite, say $P < \infty$. Moreover, for each layer the corresponding weight matrix W satisfies $\|W\|_{\mathrm{op}} \le V$. Since the operator norm dominates every entry, $|W_{ij}| \le \|W\|_{\mathrm{op}} \le V$, so each individual weight parameter is bounded in absolute value by V.

Hence, the parameter space $\Theta \subset \mathbb{R}^P$ is contained in a bounded set. Since the spectral-norm constraint $\|W\|_{\mathrm{op}} \le V$ defines a closed set, Θ is closed and bounded in finite-dimensional Euclidean space. By the Heine–Borel theorem, Θ is compact. □

Corollary 2 (Uniform boundedness of v_θ)**.** *The network output v_θ is uniformly bounded over Θ and (x, t).*

Proof. By Assumption 3, the output of every hidden layer is bounded in $[-b, b]$ component-wise, so a hidden-layer output $z \in \mathbb{R}^m$ (where m denotes the relevant layer width) satisfies $\|z\|_2 \le \sqrt{m}\, b$. Let $\widetilde{W}$ denote the final affine map including its bias column, and augment z by a constant coordinate. Since $\|\widetilde{W}\|_{\mathrm{op}} \le V$,

$$\|v_\theta(x,t)\|_2 = \|\widetilde{W}(z,1)\|_2 \le V\sqrt{1+mb^2}. \tag{30}$$

Hence there exists a constant $B_v < \infty$, depending only on (b, V, m, L_ϕ, D), such that

$$\sup_{\theta\in\Theta}\sup_{(x,t)} \|v_\theta(x,t)\| \le B_v. \tag{31}$$

□

Corollary 3 (Differentiability and Lipschitz continuity in the input)**.** *For every $\theta \in \Theta$, the map $v_\theta(x, t)$ is differentiable in (x, t) and Lipschitz continuous in its first argument.*

Proof. Each layer of the network consists of an affine map followed by a differentiable activation function. Therefore, $v_\theta(x, t)$ is differentiable in (x, t) by repeated application of the chain rule.

Furthermore, each activation function is L_ϕ-Lipschitz and, by Assumption 3, each weight matrix W satisfies $\|W\|_{\mathrm{op}} \le V$. Consequently, each hidden layer is $L_\phi V$-Lipschitz and the linear output layer is V-Lipschitz with respect to the ℓ_2 norm. By composition, for any x_1, x_2,

$$\|v_\theta(x_1,t) - v_\theta(x_2,t)\| \le V(L_\phi V)^{D-1}\|x_1 - x_2\|,$$

uniformly over $\theta \in \Theta$. Hence, v_θ is L_t-Lipschitz in its first argument with $L_t \le V(L_\phi V)^{D-1}$. □

Corollary 4 (Lipschitz continuity in the parameters)**.** *Restrict the input to $\|x\| \le R$ and $t \in [0, 1]$. Let W be the maximum hidden-layer width, and define*

$$U_R = \max\{\sqrt{R^2+2}, \sqrt{1+Wb^2}\}, \qquad L_R = DU_R \max\{1, L_\phi V\}^{D-1}.$$

Then

$$\sup_{\|x\|\le R,\ t\in[0,1]} \|v_\theta(x,t) - v_{\theta'}(x,t)\| \le L_R\|\theta - \theta'\|_2.$$

Proof. Include biases through a constant input coordinate. The augmented input has norm at most $\sqrt{R^2+2}$, and every augmented hidden activation has norm at most $\sqrt{1+Wb^2}$. Replace the weight matrices one layer at a time. A perturbation of matrix j contributes at most

$$U_R(L_\phi V)^{D-j}\|\widetilde{W}_j - \widetilde{W}'_j\|_{\mathrm{F}}$$

to the output difference. Summing these contributions, bounding each propagation factor by $\max\{1, L_\phi V\}^{D-1}$, and using $\sum_j \|\widetilde{W}_j - \widetilde{W}'_j\|_{\mathrm{F}} \le D\|\theta - \theta'\|_2$ proves the claim. □

Corollary 5 (Lipschitz constant of the pointwise loss)**.** *On the truncated event $\mathcal{E}_M$, the pointwise squared loss $\ell(v_\theta, Z_t) = \|v_\theta - Z_t\|^2$ is L_ℓ-Lipschitz in v_θ, with*

$$L_\ell = 2(B_v + c_Z(2M+1)).$$

Proof. Using the identity $\|a\|^2 - \|b\|^2 = \langle a - b,\ a + b\rangle$ with $a = v_\theta - Z_t$ and $b = v_{\theta'} - Z_t$:

$$\begin{aligned} |\|v_\theta - Z_t\|^2 - \|v_{\theta'} - Z_t\|^2| &= |\langle v_\theta - v_{\theta'},\ v_\theta + v_{\theta'} - 2Z_t\rangle| \\ &\le \|v_\theta - v_{\theta'}\| \cdot (\|v_\theta\| + \|v_{\theta'}\| + 2\|Z_t\|) \\ &\le \|v_\theta - v_{\theta'}\| \cdot (2B_v + 2\|Z_t\|). \end{aligned} \tag{32}$$

On $\mathcal{E}_M$, by Assumption 2,

$$\|Z_t\| \le c_Z(\|X_0\| + \|X_1\| + 1) \le c_Z(2M + 1).$$

Therefore $L_\ell = 2B_v + 2c_Z(2M + 1) = 2(B_v + c_Z(2M + 1))$. □

C Proof of Statistical Error Bound

In this section, we present the complete proof of Theorem 4. The argument proceeds through: covering numbers (Section C.1), the local Rademacher framework (Section C.2), the approximate Bernstein condition and its consequences (Section C.3), the Dudley entropy integral (Section C.5), the fixed-point computation (Section C.6), the bounded-loss bound (Section C.7), and the sub-Gaussian extension (Section C.8).

C.1 Covering Numbers for Neural Network Classes

We use covers of the vector-valued network output, with its Euclidean norm. On $\|x\| \le R$, $t \in [0, 1]$, Corollary 4 gives

$$\|v_\theta - v_{\theta'}\|_\infty \le L_R\|\theta - \theta'\|_2.$$

The parameter set lies in the Euclidean ball of radius $V\sqrt{P}$. A maximal δ-separated subset of this set is an internal δ-cover. Comparing the disjoint balls of radius $\delta/2$ around its points with the ball of radius $V\sqrt{P} + \delta/2$ gives

$$\mathcal{N}(\delta, \Theta, \|\cdot\|_2) \le (1 + 2V\sqrt{P}/\delta)^P.$$

Consequently,

$$\log \mathcal{N}_\infty(u, \mathcal{F}) \le P \log(1 + 2V\sqrt{P}L_R/u). \tag{33}$$

There is no additional covering factor for the output coordinates: the parameter Lipschitz bound already uses the vector norm.

For the loss class $\mathcal{H} = \{h_\theta\}$,

$$\|h_\theta - h_{\theta'}\|_\infty \le L_\ell L_R\|\theta - \theta'\|_2.$$

This also holds for $h_\theta = g_\theta \mathbf{1}_{\mathcal{E}_M}$, because the common indicator decreases distances. An internal cover of a subclass can be obtained from a cover of the full class at half the radius, by replacing each occupied ball's center by a point of that subclass. Therefore, with

$$A_R = B_\ell + 4V\sqrt{P}\, L_\ell L_R, \tag{34}$$

every subclass $\mathcal{H}' \subseteq \mathcal{H}$ satisfies

$$\log \mathcal{N}(u, \mathcal{H}', L_2(P_n)) \le P \log(1 + A_R/u). \tag{35}$$

C.2 Local Rademacher Complexity Framework

We follow the localization framework of [4]. For a class $\mathcal{H}$, use the absolute Rademacher average

$$R_n(\mathcal{H}) = \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i h(\xi_i) \right|.$$

Definition 1 (Sub-root function)**.** A function $\psi : [0, \infty) \to [0, \infty)$ is sub-root if it is nonnegative, nondecreasing, and $r \mapsto \psi(r)/\sqrt{r}$ is nonincreasing for $r > 0$. A nonzero sub-root function has a unique positive fixed point.

We apply this framework to the excess-loss second-moment classes $\{h : Ph^2 \le r\}$. The approximate Bernstein residual and the population discrepancy introduced by truncation are retained explicitly in Section C.4.

C.3 Approximate Bernstein Condition

Applying Young's inequality with parameter $\eta \in (0,1)$ to (20) gives

$$\mathbb{E}\|v_\theta - v_{\theta^*}\|^2 \le \frac{1}{1-\eta}\big(\mathcal{L}(\theta) - \mathcal{L}(\theta^*)\big) + \frac{1}{\eta(1-\eta)}\varepsilon_{\text{app}}. \tag{36}$$

Choosing $\eta = 1/2$:

$$\boxed{\mathbb{E}\|v_\theta(X_t,t) - v_{\theta^*}(X_t,t)\|^2 \le 2\big(\mathcal{L}(\theta) - \mathcal{L}(\theta^*)\big) + 4\varepsilon_{\text{app}}.} \tag{37}$$

This is the approximate Bernstein condition (Lemma 1).

To obtain variance control, define the pointwise excess loss

$$g_\theta(\xi) := \|v_\theta(X_t,t) - Z_t\|^2 - \|v_{\theta^*}(X_t,t) - Z_t\|^2,$$

so that $\Delta_\theta := Pg_\theta = \mathcal{L}(\theta) - \mathcal{L}(\theta^*) \ge 0$. For bounded targets, use $h_\theta = g_\theta$. For sub-Gaussian data, use $h_\theta = g_\theta \mathbf{1}_{\mathcal{E}_M}$; all expectations remain under the original sampling law. If $\|Z_t\| \le M'$ on the domain where h_θ is nonzero, then $L_\ell = 2(B_v + M')$ gives

$$|g_\theta| \le L_\ell \|v_\theta(X_t,t) - v_{\theta^*}(X_t,t)\|$$

on that domain. Therefore

$$\text{Var}(h_\theta) \le Ph_\theta^2 \le L_\ell^2 \mathbb{E}\|v_\theta - v_{\theta^*}\|^2. \tag{38}$$

Using (37):

$$\boxed{Ph_\theta^2 \le 2L_\ell^2 \Delta_\theta + 4L_\ell^2 \varepsilon_{\text{app}} =: B_0 \Delta_\theta + \rho,} \tag{39}$$

where $B_0 := 2L_\ell^2$ and $\rho := 4L_\ell^2 \varepsilon_{\text{app}}$.

C.4 Localized Rademacher Theorem under Approximate Bernstein

Theorem 5. *Let $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$ contain zero, with $|h_\theta| \le B_\ell$, $B_0, B_\ell > 0$, and $\rho, \tau \ge 0$. Suppose*

$$\Delta_\theta \ge 0, \qquad |Ph_\theta - \Delta_\theta| \le \tau, \qquad Ph_\theta^2 \le B_0 \Delta_\theta + \rho.$$

Let ψ be a sub-root function with positive fixed point r^* such that

$$\psi(r) \ge B_0 \mathbb{E}R_n\{h \in \mathcal{H} : Ph^2 \le r\}.$$

Set

$$J = \left\lceil \log_2 \max\{1, B_\ell^2/r^*\} \right\rceil, \qquad y = x + \log(J+1).$$

For $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$, every θ satisfies

$$\begin{aligned}\Delta_\theta \le \frac{K}{K-1}(P_n h_\theta + \tau) + \frac{48K^2(B_0 + B_\ell)r^*}{(K-1)B_0^2} \\ + \frac{9K^2(B_0 + B_\ell)y}{(K-1)n} + \frac{K\rho}{2(K-1)B_0}.\end{aligned} \tag{40}$$

Proof. Write $\mathcal{H}(r) = \{h \in \mathcal{H} : Ph^2 \le r\}$ and $Q = B_0 + B_\ell$. The following $J+1$ shells cover the class:

$$\mathcal{H}_- = \mathcal{H}(r^*), \qquad \mathcal{H}_k = \{h : 2^{k-1}r^* < Ph^2 \le 2^k r^*\}, \quad k = 1, \dots, J.$$

For $R \ge r^*$, the sub-root property gives $\psi(R) \le \sqrt{Rr^*}$. By [4, Theorem 2.1], with its parameter $\alpha = 1$, with failure probability at most e^{-y},

$$\begin{aligned}\sup_{h \in \mathcal{H}(R)} (Ph - P_n h) &\le 4\mathbb{E}R_n(\mathcal{H}(R)) + \sqrt{\frac{2Ry}{n}} + \frac{8B_\ell y}{3n} \\ &\le \frac{4\sqrt{Rr^*}}{B_0} + \sqrt{\frac{2Ry}{n}} + \frac{8B_\ell y}{3n}.\end{aligned} \tag{41}$$

Apply this to $R = r^*, 2r^*, \ldots, 2^J r^*$. A union bound gives simultaneous validity with probability at least $1 - (J+1)e^{-y} = 1 - e^{-x}$.

For $h_\theta \in \mathcal{H}_k$, the lower shell endpoint implies

$$2^k r^* \le 2B_0\Delta_\theta + 2\rho \le 2Q\Delta_\theta + 2\rho. \tag{42}$$

Using $\Delta_\theta \le Ph_\theta + \tau$ and splitting square roots,

$$\begin{aligned}\Delta_\theta - P_n h_\theta \le \tau + \frac{4\sqrt{2Q\Delta_\theta r^*}}{B_0} + 2\sqrt{\frac{Q\Delta_\theta y}{n}} \\ + \frac{4\sqrt{2\rho r^*}}{B_0} + 2\sqrt{\frac{\rho y}{n}} + \frac{8B_\ell y}{3n}.\end{aligned} \tag{43}$$

Young's inequality gives

$$\frac{4\sqrt{2Q\Delta_\theta r^*}}{B_0} \le \frac{\Delta_\theta}{2K} + \frac{16KQr^*}{B_0^2}, \qquad 2\sqrt{\frac{Q\Delta_\theta y}{n}} \le \frac{\Delta_\theta}{2K} + \frac{2KQy}{n}.$$

For the residual terms, use

$$\frac{4\sqrt{2\rho r^*}}{B_0} \le \frac{32r^*}{B_0} + \frac{\rho}{4B_0}, \qquad 2\sqrt{\frac{\rho y}{n}} \le \frac{4B_0 y}{n} + \frac{\rho}{4B_0}.$$

Since $Q \ge B_0, B_\ell$ and $K > 1$, collecting terms gives

$$\Delta_\theta - P_n h_\theta \le \tau + \frac{\Delta_\theta}{K} + \frac{48KQr^*}{B_0^2} + \frac{9KQy}{n} + \frac{\rho}{2B_0}. \tag{44}$$

For the small shell, use $\sqrt{2r^* y/n} \le r^*/B_0 + B_0 y/(2n)$ to obtain

$$\Delta_\theta - P_n h_\theta \le \tau + \frac{5r^*}{B_0} + \frac{(B_0/2 + 8B_\ell/3)y}{n}.$$

This is also dominated by the right side of (44), since $\Delta_\theta \ge 0$. Dividing by $1 - 1/K = (K-1)/K$ proves the theorem. □

For ordinary excess losses, $h_\theta = g_\theta$, $\Delta_\theta = Pg_\theta$, and $\tau = 0$. For masked losses, Ph_θ need not be nonnegative; the proof only requires $\Delta_\theta \ge 0$. The same proof also covers the ordinary variance condition $\mathrm{Var}(g_\theta) \le B_0 Pg_\theta + \rho$, because

$$Pg_\theta^2 \le (B_0 + B_\ell)Pg_\theta + \rho,$$

which is the relaxed second-moment bound used in (42).

C.5 Dudley Entropy Integral

We bound the precise localized class required by the theorem:

$$\mathcal{H}(r) = \{h_\theta : Ph_\theta^2 \le r\}. \tag{45}$$

Here $h_\theta = g_\theta$ in the bounded case and $h_\theta = g_\theta \mathbf{1}_{\mathcal{E}_M}$ in the sub-Gaussian case. Let R bound the network input on the domain where h_θ is nonzero, and define

$$A_R = B_\ell + 4V\sqrt{P}\, L_\ell L_R, \qquad H_R = 1 + P\log(1 + A_R n/B_\ell), \qquad q_R = H_R/n.$$

By (35), every localized subclass has

$$\log \mathcal{N}(u, \mathcal{H}(r), L_2(P_n)) \le P\log(1 + A_R/u).$$

Set

$$\mathfrak{R}(r) = \mathbb{E}R_n(\mathcal{H}(r)), \qquad D_n(r)^2 = \sup_{h \in \mathcal{H}(r)} P_n h^2, \qquad \alpha = B_\ell/n.$$

Conditionally on the observations, apply the empirical Dudley inequality [28] to $\mathcal{H}(r) \cup (-\mathcal{H}(r))$. When $2D_n(r) > \alpha$,

$$\begin{aligned} R_n(\mathcal{H}(r)) &\le C\alpha + \frac{C}{\sqrt{n}} \int_{\alpha}^{2D_n(r)} \sqrt{\log 2 + P \log(1 + A_R/u)}\, du \\ &\le C\alpha + C\sqrt{q_R}\, D_n(r). \end{aligned} \tag{46}$$

If $2D_n(r) \le \alpha$, the direct bound $R_n(\mathcal{H}(r)) \le D_n(r) \le \alpha$ gives the same conclusion. The cutoff is deterministic and the endpoint is the empirical radius.

Symmetrization and contraction for $u \mapsto u^2$, which is $2B_\ell$-Lipschitz on $[-B_\ell, B_\ell]$, give

$$\begin{aligned} \mathbb{E}D_n(r)^2 &\le r + \mathbb{E} \sup_{h \in \mathcal{H}(r)} |(P_n - P)h^2| \\ &\le r + 2\mathbb{E}\mathbb{E}_\epsilon \sup_{h \in \mathcal{H}(r)} \left| \frac{1}{n} \sum_i \epsilon_i h(\xi_i)^2 \right| \\ &\le r + 8B_\ell \mathfrak{R}(r). \end{aligned} \tag{47}$$

Taking expectations in (46) and using Jensen,

$$\mathfrak{R}(r) \le CB_\ell/n + C\sqrt{q_R}\sqrt{r + 8B_\ell \mathfrak{R}(r)}.$$

Split the square root and use Young's inequality on the term proportional to $\sqrt{B_\ell q_R \mathfrak{R}(r)}$. Since $q_R \ge 1/n$, this yields

$$\mathfrak{R}(r) \le C\big(\sqrt{r q_R} + B_\ell q_R\big). \tag{48}$$

C.6 Fixed-Point Computation

A valid majorant from (48) is

$$\psi(r) = CB_0\big(\sqrt{r q_R} + B_\ell q_R\big) + B_0^2/n. \tag{49}$$

It is nonnegative and nondecreasing, while

$$\frac{\psi(r)}{\sqrt{r}} = CB_0\sqrt{q_R} + \frac{CB_0 B_\ell q_R + B_0^2/n}{\sqrt{r}}$$

is nonincreasing. Its positive fixed point obeys $r^* \ge B_0^2/n$ directly by construction. Writing $a = CB_0\sqrt{q_R}$ and $b_* = CB_0 B_\ell q_R + B_0^2/n$, we have

$$r^* = a\sqrt{r^*} + b_* \le r^*/2 + a^2/2 + b_*.$$

Therefore

$$\frac{B_0^2}{n} \le r^* \le C'(B_0^2 + B_0 B_\ell)\frac{H_R}{n}. \tag{50}$$

For squared loss, $B_\ell = L_\ell^2/4 = B_0/8$, so

$$r^*/B_0 \le CL_\ell^2 H_R/n. \tag{51}$$

Also $B_\ell^2/r^* \le n/64$, hence

$$J \le \lceil \log_2 \max\{1, n\} \rceil, \qquad \log(J+1) \le C\log(1+n) \le CH_R.$$

The finite-shell confidence cost is thus absorbed by H_R; the separate confidence parameter x remains explicit.

C.7 Bounded-Loss Bound

Lemma 4 (Bounded-data excess risk)**.** *Suppose $\|X_t\| \le R$ and $\|Z_t\| \le M'$ almost surely. Under Assumptions 3, 5 and 6, for every $x > 0$, with probability at least $1 - e^{-x}$,*

$$\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta^*) \le CL_\ell^2(H_R + x)/n + 2\varepsilon_{\mathrm{app}}. \tag{52}$$

Proof. Use $h = g$, $\tau = 0$, and $K = 2$ in Theorem 5. The approximation term is exactly

$$\frac{K\rho}{2(K-1)B_0} = \frac{\rho}{B_0} = 2\varepsilon_{\text{app}}.$$

The other terms follow from (51) and the preceding finite-shell estimate. □

Applying Lemma 1 gives

$$\mathbb{E}\|v_{\hat{\theta}} - v_{\theta^*}\|^2 \le CL_\ell^2(H_R + x)/n + 8\varepsilon_{\text{app}}. \tag{53}$$

Remark 11. Dimension enters through P, the input and output bounds, and the approximation error ε_{app}. The covering contribution H_R is logarithmic in the displayed parameter Lipschitz and weight bounds. Architecture and schedule quantities can also enter B_v, L_ℓ, and the ODE stability factor outside logarithms. No dimension independence of ε_{app} is assumed.

C.8 Extension to Sub-Gaussian Data

The following argument applies to the original empirical risk minimizer $\hat{\theta}$ in Assumption 6. All population expectations below are under the original distribution of $\xi = (X_0, X_1, t)$.

Write $R = \max\{\|X_0\|, \|X_1\|\}$ and $a = 4c\sigma^2$, where c is the constant in Assumption 1. A $1/2$-net of the unit sphere with at most 5^d elements satisfies $\|z\| \le 2\max_u |\langle u, z\rangle|$. Applying the directional tail bound on this net and taking a union bound over X_0, X_1 gives

$$\Pr(R > M) \le 4 \cdot 5^d \exp(-M^2/a). \tag{54}$$

For $n \ge 2$ and $x > 0$, choose

$$M^2 = a\big[d\log 5 + x + \log(4n^2)\big], \qquad p_M := e^{-x}/n^2. \tag{55}$$

Let $\mathcal{E}_M = \{R \le M\}$ denote the single-observation event, and define the sample event

$$\mathcal{A}_M := \bigcap_{i=1}^{n} \{\|X_0^i\| \le M,\ \|X_1^i\| \le M\}.$$

Then $\Pr(\mathcal{E}_M^c) \le p_M$ and

$$\Pr(\mathcal{A}_M^c) \le np_M = e^{-x}/n \le e^{-x}. \tag{56}$$

We next bound the population tail uniformly over the class. Put $H = (B_v + \|Z_t\|)^2$ and $\tau_M = \mathbb{E}[H\mathbf{1}_{\mathcal{E}_M^c}]$. Both squared losses ℓ_θ and ℓ_{θ^*} lie in $[0, H]$, so $|\ell_\theta - \ell_{\theta^*}| \le H$. Tail integration in (54) gives

$$\mathbb{E}[R^2\mathbf{1}_{\{R>M\}}] = M^2\Pr(R > M) + \int_{M^2}^{\infty} \Pr(R^2 > s)\, ds \le (M^2 + a)p_M.$$

By Assumption 2, $H \le 2(B_v + c_Z)^2 + 8c_Z^2R^2$. Consequently,

$$\begin{aligned} \tau_M &\le \big[2(B_v + c_Z)^2 + 8c_Z^2(M^2 + a)\big]p_M \\ &\le L_\ell^2 p_M, \qquad L_\ell = 2\big(B_v + c_Z(2M+1)\big), \end{aligned} \tag{57}$$

where the last inequality uses $M^2 \ge a$, valid for $d \ge 1$. In particular, $\tau_M \le L_\ell^2 e^{-x}/n^2$.

Retain the original comparator θ^* and target $\Delta_\theta = \mathbb{E}g_\theta = \mathcal{L}(\theta) - \mathcal{L}(\theta^*) \ge 0$, and use the bounded functions

$$h_\theta = g_\theta \mathbf{1}_{\mathcal{E}_M}.$$

Their means satisfy

$$|\mathbb{E}h_\theta - \Delta_\theta| \le \tau_M. \tag{58}$$

Set $B_\ell = (B_v + c_Z(2M+1))^2 = L_\ell^2/4$. The pointwise loss Lipschitz bound on $\mathcal{E}_M$, followed by the original-distribution approximate Bernstein inequality, gives

$$\begin{aligned} &|h_\theta| \le B_\ell, \\ &\mathbb{E}h_\theta^2 \le L_\ell^2\mathbb{E}\|v_\theta(X_t, t) - v_{\theta^*}(X_t, t)\|^2 \\ &\qquad \le 2L_\ell^2\Delta_\theta + 4L_\ell^2\varepsilon_{\text{app}} = B_0\Delta_\theta + \rho, \end{aligned} \tag{59}$$

where $B_0 = 2L_\ell^2$ and $\rho = 4L_\ell^2\varepsilon_{\text{app}}$. On $\mathcal{E}_M$, the network input has norm at most $R_M = c_I(2M+1)$. Multiplication by the common indicator preserves the covering bounds; hence Sections C.5 and C.6 apply to $\{h_\theta\}$ with $R = R_M$, localizing by its own population second moment.

On $\mathcal{A}_M$, the original ERM inequality implies

$$P_n h_{\hat{\theta}} = P_n(\ell_{\hat{\theta}} - \ell_{\theta^*}) \le 0.$$

Apply Theorem 5 with $K = 2$, $\tau = \tau_M$, and confidence parameter x. The concentration event fails with probability at most e^{-x}. Its intersection with $\mathcal{A}_M$ therefore has probability at least $1 - 2e^{-x}$, with no independence requirement. On this intersection,

$$\Delta_{\hat{\theta}} \le CL_\ell^2(H_{R_M} + x)/n + 2\varepsilon_{\text{app}} + 2\tau_M.$$

Using $\mathcal{E}^{\text{stat}} \le 2\Delta_{\hat{\theta}} + 4\varepsilon_{\text{app}}$ and absorbing $4\tau_M$ in the statistical term gives

$$\mathcal{E}^{\text{stat}} \le \frac{CL_\ell^2(H_{R_M} + x)}{n} + 8\varepsilon_{\text{app}}. \tag{60}$$

This proves Theorem 4.

The radius in (55) has the dependence $M = O(\sigma\sqrt{c(d + \log n + x)})$. Equivalently, for final confidence $1 - \delta$, take $x = \log(2/\delta)$ and

$$M^2 = 4c\sigma^2\big[d\log 5 + \log(8n^2/\delta)\big].$$

Then $\Pr(\mathcal{A}_M^c) \le \delta/(2n)$ and $\tau_M \le L_\ell^2\delta/(2n^2)$. In particular,

$$L_\ell^2 = O\big(B_v^2 + c_Z^2\left[1 + c\sigma^2(d + \log n + \log(1/\delta))\right]\big).$$

This displays the dimension dependence of truncation separately from the parameter count and other architecture constants.

D Optimization Error

Throughout the paper, Assumption 6 gives $\mathcal{E}^{\text{opt}} = 0$. For an approximate ERM $\theta \in \Theta$ with empirical loss gap at most η_{opt}, the localization argument can instead retain this empirical tolerance, contributing $O(\eta_{\text{opt}})$ to the excess-risk bound. Thus, an independently established η_{opt} no larger than the statistical term preserves its order. This route uses the empirical loss gap directly and does not require a parameter-distance estimate.

Conditional optimization guarantees. A PL condition, together with the applicable smoothness and step-size conditions, can yield convergence of empirical loss. Overparameterized network results provide another possible route, under their respective architecture, data, initialization, and training assumptions [1, 10, 22]. In particular, [1] analyzes ReLU networks, so its guarantee does not directly apply to the bounded, differentiable activations in Assumption 3. Such conditions must be verified separately before using these results to replace Assumption 6. If the decomposition through $\mathcal{E}^{\text{opt}}$ is used instead, a separate population prediction-error bound of the required statistical order is needed.

E Information-Theoretic Lower Bound

We establish an $\Omega(\varepsilon^{-2})$ sample-complexity lower bound for integrated velocity estimation using a Gaussian location family. Related lower-bound examples for score estimation appear in [15, Section 4].

Lemma 5 (Lower bound for velocity estimation)**.** *Fix $\sigma > 0$. Let $X_0 \sim \mathcal{N}(0, \sigma^2)$ and $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ independently, where $|\mu| \le \sigma/2$, and let $p_{\mu,t}$ be the density of $X_t = (1-t)X_0 + tX_1$. Write*

$$v_\mu(x,t) := \mathbb{E}_\mu[X_1 - X_0 \mid X_t = x].$$

For any $n \ge 1$, estimators based on n independent training triples (X_0^i, X_1^i, t_i), with $t_i \sim \text{Uniform}[0,1]$, satisfy

$$\inf_{\hat{v}} \sup_{|\mu| \le \sigma/2} \mathbb{E}_\mu\left[\int_0^1 \int_{\mathbb{R}} |\hat{v}(x,t) - v_\mu(x,t)|^2 p_{\mu,t}(x)\,dx\,dt\right] \ge \frac{\sigma^2}{32n}. \tag{61}$$

This family satisfies Assumption 1 with a uniform numerical tail constant.

Proof. Set $q_t = (1-t)^2 + t^2$. Gaussian conditioning gives

$$X_t \sim \mathcal{N}(t\mu, \sigma^2 q_t), \qquad v_\mu(x,t) = \frac{2t-1}{q_t}x + \frac{1-t}{q_t}\mu. \tag{62}$$

Indeed, $\mathbb{E}[X_1 - X_0] = \mu$ and $\mathrm{Cov}(X_1 - X_0, X_t) = \sigma^2(2t-1)$. Moreover, for $|\mu| \le \sigma/2$, Gaussian tails imply $\Pr(|X_j| \ge r) \le 2\exp(-r^2/(8\sigma^2))$ for $j \in \{0,1\}$: for $r \ge \sigma$ use $r - |\mu| \ge r/2$, and for $r < \sigma$ the bound is trivial.

Choose $a = \sigma/(2\sqrt{n})$ and the two hypotheses $\mu = \pm a$. Let Q_+ and Q_- denote their full training-data laws. The source samples and times have the same law under both hypotheses, so

$$\mathrm{KL}(Q_+ \| Q_-) = \frac{2na^2}{\sigma^2} = \frac{1}{2}, \qquad \mathrm{TV}(Q_+, Q_-) \le \frac{1}{2}$$

by Pinsker's inequality. For the test marginals, the same inequality gives

$$\mathrm{TV}(p_{a,t}, p_{-a,t}) \le \frac{ta}{\sigma\sqrt{q_t}} \le \frac{a}{\sigma} \le \frac{1}{2}.$$

Define the common density $h_t(x) := \min\{p_{a,t}(x), p_{-a,t}(x)\}$, whose mass is at least $1/2$. Since $(1-t)/q_t \ge 1$ on $[0, 1/2]$, (62) yields

$$D^2 := \int_0^1 \int_{\mathbb{R}} h_t(x) |v_a(x,t) - v_{-a}(x,t)|^2 \, dx \, dt \ge \int_0^{1/2} 4a^2 \cdot \frac{1}{2} \, dt = a^2. \tag{63}$$

To account for the hypothesis-dependent loss measures, let f_+ and f_- be densities of Q_+ and Q_- on the training sample space, and let R_+ and R_- denote the two expected losses in (61) for a fixed estimator. With s denoting the training data, nonnegativity and the squared-distance inequality give

$$\begin{aligned} R_+ + R_- &\ge \int \min\{f_+(s), f_-(s)\} \int_0^1 \int_{\mathbb{R}} h_t(x) \\ &\qquad \times \Big(|\hat{v}_s(x,t) - v_a(x,t)|^2 + |\hat{v}_s(x,t) - v_{-a}(x,t)|^2 \Big) \, dx \, dt \, ds \\ &\ge \frac{D^2}{2}\big(1 - \mathrm{TV}(Q_+, Q_-)\big) \ge \frac{a^2}{4}. \end{aligned}$$

Thus $\max\{R_+, R_-\} \ge a^2/8 = \sigma^2/(32n)$. Taking the infimum over estimators proves the result. Independent estimator randomization can be included in s without changing the argument. □

Remark 12. For fixed σ, achieving expected integrated squared velocity error at most ε^2 uniformly over this family requires $n = \Omega(\varepsilon^{-2})$. This matches, up to logarithms, the ε exponent of the statistical velocity-estimation bound. The lower bound concerns estimation of the true velocity field, while the approximation error $\varepsilon_{\mathrm{app}}$ depends on the chosen network class.

F Numerical Experiments

To examine the scaling predicted by Theorem 2, we trained the same 4-layer MLP on four data distributions of increasing dimension $d \in \{2, 8, 16, 32\}$. The rectified-flow slopes are near -1, consistent with the predicted $\widetilde{O}(L_\ell^2 P/n)$ scaling. The SiLU experiments illustrate behavior beyond the bounded-activation class used in the analysis.

For each task, schedule, and training-set size n, we train an MLP of width 512, depth 4, with SiLU activations, using Adam with cosine annealing and gradient clipping at 1.0. Results are averaged over 7 random seeds, and error bars indicate ± 1 standard error of the mean.

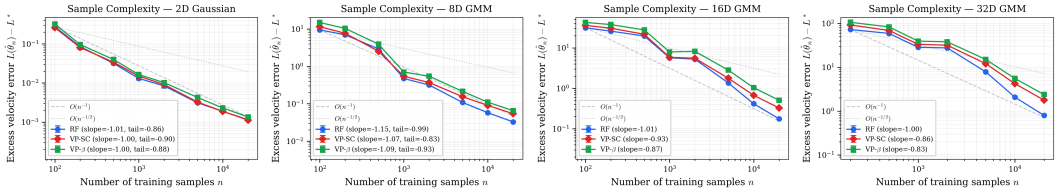
To avoid bias from estimating L^*, we exploit the closed-form Bayes-optimal velocity $v^*(x,t) = \mathbb{E}[Z_t \mid X_t = x]$ that exists when X_1 is a Gaussian mixture, since $X_t \mid$ component is Gaussian under each schedule. We compute L^* to four significant figures by Monte Carlo with $n_{\mathrm{mc}} \ge 10^6$ samples, and evaluate the excess via the paired identity $L(\hat{\theta}) - L^* = \mathbb{E}\|v_{\hat{\theta}} - v^*\|^2$ (orthogonality of the regression residual). This lets us measure excess down to 10^{-3} on the same draws.

Figure 2 plots excess velocity error against training-set size n on log-log axes for four targets of increasing dimension, all trained with the same MLP architecture:

- **2D Gaussian.** A single full-rank Gaussian, for which v^* is affine. All three schedules attain slope exactly -1.00.
- **8D GMM.** Eight axis-aligned modes at $\pm 5\, e_i$, $i = 1, \ldots, 4$. Slopes: -1.15 (RF), -1.07 (VP-SC), -1.09 (VP-β).
- **16D GMM.** Sixteen axis-aligned modes. Slopes: -1.01 (RF), -0.93 (VP-SC), -0.87 (VP-β).
- **32D GMM.** Thirty-two axis-aligned modes. Slopes: -1.00 (RF), -0.86 (VP-SC), -0.83 (VP-β).

The rectified-flow slope is at or near -1 from $d = 2$ to 32, consistent with the scaling predicted by Theorem 2. A nonzero approximation floor can contribute to the observed slope-flattening of VP-SC and VP-β at higher d; a larger statistical prefactor alone does not change the predicted slope.

Schedule ordering and the L_ℓ analysis. In every panel of Figure 2, the curves are vertically ordered RF $<$ VP-SC $<$ VP-β, and the gap between schedules widens with d. This ordering is qualitatively consistent with the schedule-dependent constant L_ℓ from Section 6, where $c_Z = 1$ for rectified flow and $c_Z = \pi/2$ for VP-SC. For VP-β, $\epsilon = 10^{-3}$ is the experimental time cutoff; the coefficient floor τ in Section 6 is instead determined by the schedule coefficients at the cutoff. Since L_ℓ enters quadratically in the statistical-error prefactor, a larger c_Z contributes a larger prefactor. Empirically, at $d = 2$ the three schedules sit within a factor of $1.3\times$ of one another at every n, whereas at $d = 32$ the gap between RF and VP-β exceeds $2\times$.



(a) 2D Gaussian (b) 8D GMM (c) 16D GMM (d) 32D GMM

Figure 2: Excess velocity error $L(\hat{\theta}_n) - L^*$ versus training-set size n, on log-log axes, for four targets of increasing dimension trained with the same architecture (4-layer MLP, width 512). Rectified flow attains a slope at or near -1 in every panel. The vertical ordering RF $<$ VP-SC $<$ VP-β is qualitatively consistent with the L_ℓ-dependent prefactor in Section 6. Reference $O(n^{-1})$ (dashed) and $O(n^{-1/2})$ (dotted) lines are included for visual reference.

Remark 13 (Compute resources). All experiments were performed on a single NVIDIA RTX 2080 Ti (11 GB) with 4 CPU threads, using PyTorch. Each rate experiment over a single dataset (7 seeds $\times$ 3 schedules $\times$ 7-8 values of n up to $n_{\max} = 20{,}000$) took approximately 3-4 hours of GPU time.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract/Introduction match Theorem 2

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: All discussed in remarks and conclusion

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Assumptions clearly stated and detailed proofs in appendix

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Numerical Experiments section and supplemental material

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code provided with readme file

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Numerical Experiments section

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Numerical Experiments section

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Single RTX 2080 and 4 CPU threads took 3-4 hours per dataset

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Theory paper with no human subjects

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Theory work, no direct impact

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No such risks

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: No assets used

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: No LLM in core method development

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.