

On Flow Matching KL Divergence

Maojiang Su^{†*1} Jerry Yao-Chieh Hu^{†*2} Sophia Pi^{†3} Han Liu^{†‡4}

[†] Center for Foundation Models and Generative AI, Northwestern University, Evanston, IL 60208, USA
Department of Computer Science, Northwestern University, Evanston, IL 60208, USA

[‡] Department of Statistics and Data Science, Northwestern University, Evanston, IL 60208, USA

We derive a deterministic, non-asymptotic upper bound on the Kullback-Leibler (KL) divergence of the flow-matching distribution approximation. In particular, if the L_2 flow-matching loss is bounded by $\epsilon^2 > 0$, then the KL divergence between the true data distribution and the estimated distribution is bounded by $A_1\epsilon + A_2\epsilon^2$. Here, the constants A_1 and A_2 depend only on the regularities of the data and velocity fields. Consequently, this bound implies statistical convergence rates of Flow Matching Transformers under the Total Variation (TV) distance. We show that, flow matching achieves nearly minimax-optimal efficiency in estimating smooth distributions. Our results make the statistical efficiency of flow matching comparable to that of diffusion models under the TV distance. Numerical studies on synthetic and learned velocities corroborate our theory.

Keywords: Flow Matching, Probability Flow ODE, Non-Asymptotic Convergence, Kullback-Leibler (KL) Divergence Error Bounds, Flow Matching Transformer

¹smj@u.northwestern.edu

²jhu@u.northwestern.edu

³sophiapi2026.1@u.northwestern.edu

⁴hanliu@northwestern.edu

*These authors contributed equally to this work. Code is available at https://github.com/MAGICS-LAB/flow_kl.

Contents

1	Introduction	1
2	Preliminaries	3
2.1	Flow Matching	3
3	Kullback-Leibler (KL) Error Bound for Flow Matching	5
4	Convergence Rates of Flow Matching in Total Variation Distance	8
5	Numerical Studies	10
5.1	Validating KL Evolution Identity (Lemma 3.1)	10
5.2	Validating KL Error Bounds (Theorem 3.1)	12
6	Discussion and Conclusion	13
A	Related Work	15
B	Proofs in the Main Text	16
B.1	Proof of Lemma 3.1	16
B.2	Proof of Theorem 3.1	18
B.3	Proof of Proposition 6.1	22
B.4	Proof of Theorem 4.1	25
B.5	Proof of Theorem 4.2	25

1 Introduction

We establish a rigorous upper bound on the Kullback-Leibler (KL) divergence between the true data distribution and the distribution estimated by flow matching, expressed in terms of the L_2 flow matching training loss. Furthermore, we prove Flow Matching Transformers achieves almost minimax optimal convergence rate under the Total Variation (TV) metric. A theoretical understanding is crucial in the current era of rapidly advancing generative AI. In particular, Flow Matching (FM) [[Lipman et al., 2022](#), [Liu et al., 2022](#), [Albergo and Vanden-Eijnden, 2022](#)] is a central paradigm for training continuous-time generative models. Instead of maximizing likelihoods, FM directly learns a deterministic velocity field that transports a simple base distribution (e.g., Gaussian) to a complex data distribution through the continuity equation. Its simple training objective and efficient inference enable state-of-the-art performance across diverse domains, including image generation [[Esser et al., 2024](#)], speech synthesis [[Le et al., 2023](#)], video generation [[Polyak et al., 2024](#)], and robotics [[Black et al., 2024](#)].

Despite its empirical success, the theoretical understanding of flow matching remains limited. Recent theoretical studies [[Chen et al., 2022](#), [Gentiloni Silveri et al., 2024](#), [Block et al., 2020](#), [De Bortoli, 2022](#)] investigate the distributional approximation error of diffusion-based generative models under stochastic differential equation (SDE) sampling. For flow-based generative models, [Albergo and Vanden-Eijnden \[2022\]](#), [Benton et al. \[2023\]](#) analyze the distributional error of flow matching under the 2-Wasserstein distance. [Albergo et al. \[2023\]](#) derive KL-based bounds for stochastic flows by injecting a small diffusion term into the sampling process. Although these results provide valuable insights, they primarily focus on stochastic sampling or Wasserstein-type

metrics, which capture geometric transport cost rather than information-theoretic discrepancy. More recently, Fukumizu et al. [2024], Jiao et al. [2024] investigate the convergence rates of flow matching under the 2-Wasserstein distance. However, the statistical convergence rate under more interpretable and task-relevant metrics, such as the Total Variation (TV) distance, is unexplored.

In contrast, our work develops a pure ODE-based theory for deterministic flow matching and establishes an explicit Kullback-Leibler (KL) divergence bound between the true and learned distributions. The KL metric is more relevant for likelihood-based generative modeling [Schulman et al., 2015, Abdolmaleki et al., 2018, Haarnoja et al., 2018, Shao et al., 2024], as it quantifies information loss and statistical discrepancy rather than geometric displacement. Building on this result, we further derive statistical convergence rates for Flow Matching Transformers under the Total Variation (TV) metric. The KL bound provides a direct and information-theoretic control of distributional error, allowing us to obtain TV convergence in a more explicit and interpretable way than studies based on the 2-Wasserstein distance. Together, our KL error bounds and convergence rates under TV distance complement existing Wasserstein-based studies, forming a unified foundation for understanding flow matching.

Contributions. Our contributions are two-fold:

- **Kullback-Leibler Error Bounds for Flow Matching.** We establish a *deterministic, non-asymptotic upper bound* on the Kullback-Leibler (KL) divergence between the true data distribution p_1 and the flow-matching-estimated distribution q_1 , expressed in terms of the L_2 flow matching training loss. Following Albergo et al. [2023], we introduce the KL Evolution Identity for Continuity Flows (Lemma 3.1) and apply Grönwall’s inequality (Lemma 3.3) to control the temporal evolution of the divergence. Under mild smoothness and regularity assumptions Assumptions 3.1 to 3.3, Theorem 3.1 shows that $\text{KL}(p_1||q_1) \leq A_1\epsilon + A_2\epsilon^2$, where A_1, A_2 are regularity and smoothness constants and ϵ^2 denotes the L_2 flow matching training loss. This result provides the first information-theoretic guarantee that links training loss to distributional approximation accuracy.
- **Convergence Rates under Total Variation Distance.** Building on the KL bound, we derive statistical convergence rates for Flow Matching Transformers under the *Total Variation (TV)* metric. We begin by introducing the *Hölder space* (Definition 4.1). Under the Hölder-smoothness assumption (Assumption 4.1), we establish the convergence rate stated in Theorem 4.1, and further show that flow matching achieves *nearly minimax-optimal convergence* under specific regularity conditions (Theorem 4.2). This result complements prior analyses based on the 2-Wasserstein distance by providing a more direct and interpretable characterization of the probabilistic approximation ability of flow matching.

Organization. Section 2 reviews the preliminary concepts of standard flow matching. Section 3 presents the KL error bound for the distributional approximation error of flow matching. Section 4 establishes the statistical convergence rates of Flow Matching Transformers and proves that flow matching achieves nearly minimax-optimal convergence. Section 5 reports the empirical results that validate our theoretical analysis. Section 6 concludes the paper and discusses the implications of our findings. The appendix provides related works (Section A), and detailed proofs of the main results (Section B).

Notation. We denote the index set $\{1, \dots, I\}$ by $[I]$. Let $x[i]$ denote the i -th component of a vector x . Let \mathbb{Z} denote integers and \mathbb{Z}_+ denote positive integers. Given random variables X and Y with marginal densities μ_x and μ_y respectively, we denote the 2-Wasserstein distance between μ_x and μ_y by $W_2(\mu_x, \mu_y)$. Given a matrix $Z \in \mathbb{R}^{d \times L}$, $\|Z\|_2$ and $\|Z\|_F$ denote the 2-norm and the Frobenius norm. Let $u^k \in \mathbb{R}^d$ be column vectors for $k \in [K]$, we denote $\text{col}(u^1, \dots, u^K) \in \mathbb{R}^{kd}$ as the vertical concatenation of u^1, \dots, u^K . Let $\text{Div} \cdot$ be the divergence operator.

2 Preliminaries

In this section, we provide an overview of the flow matching generative modeling following [Lipman et al., 2024] and [Su et al., 2025a, Section 2].

2.1 Flow Matching

Flow Model. The flow model transforms $X_0 = x_0$ drawn from a source distribution P (e.g., a Gaussian) into samples $X_1 = x_1$ from a target distribution Q . A flow $\psi : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a time-dependent mapping $\psi : (t, x) \mapsto \psi_t(x)$ that evolves the input x over time. The flow model is a continuous-time Markov process $(X_t)_{0 \leq t \leq 1}$ defined by applying a flow ψ_t to the random variable $X_0 \sim P$:

$$X_t = \psi_t(X_0), \quad t \in [0, 1].$$

On the other hand, a time-dependent velocity field $u : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$ implementing $u : (x, t) \mapsto u(\cdot, t)$ defines a unique flow ψ via the following ordinary differential equation (ODE):

$$\frac{d\psi_t}{dt} = u(\psi_t(x), t) \quad \text{with initial condition} \quad \psi_0(x) = x. \quad (2.1)$$

Given a flow ψ_t , the marginal probability density function (PDF) of flow model $X_t = \psi_t(X_0) \sim p_t$ is a continuous-time probability path $(p_t)_{0 \leq t \leq 1}$. The evolution of marginal probability densities p_t follows the push-forward equation:

$$p_t(x) = [\psi_t]_* p_0(x) := p_0(\psi_t^{-1}(x)) \cdot \left| \det \left[\frac{\partial \psi_t^{-1}}{\partial x} \right] \right|. \quad (2.2)$$

By the equivalence between flows and velocity fields [Lipman et al., 2024], any invertible C^1 diffeomorphism ψ_t induces a unique smooth conditional velocity field $u(x, t)$ given by

$$u(x, t) = \dot{\psi}_t(\psi_t^{-1}(x)), \quad \text{with} \quad \dot{\psi}_t = \frac{d}{dt} \psi_t. \quad (2.3)$$

Equations (2.2) and (2.3) characterize the relationship among the probability path p_t , the flow ψ_t , and the velocity field u . The continuity equation provides a direct link between u and p_t :

$$\frac{\partial}{\partial t} p_t(x) + \nabla \cdot (p_t(x) u(x, t)) = 0. \quad (2.4)$$

For an arbitrary probability path p_t , we define a velocity field $u(x, t)$ that *generates* p_t if its flow ψ_t satisfies (2.1). The mass conservation theorem [Lipman et al., 2024, Villani et al., 2008] ensures consistency between the continuity equation and flow ODE (2.1): a pair (u, p_t) satisfying (2.4) for $t \in [0, 1]$ corresponds to a field velocity $u(x, t)$ that generates p_t .

Continuous Normalizing Flow [Chen et al., 2018] models the velocity field $u(x, t)$ with a neural network u_θ . Once we obtain a well-trained u_θ , we generate samples from solving ODE (2.1).

Flow Matching. Instead of training flow model by maximizing the log-likelihood of training data [Chen et al., 2018], flow matching [Lipman et al., 2022] is a simulation-free framework to train flow generative models without the need of solving ODEs during training. The Flow Matching objective is designed to match the probability path $(p_t)_{0 \leq t \leq 1}$, which allows us to flow from source $p_0 = P$ to target $p_1 = Q$. Suppose u generates such probability path p_t , the flow matching loss is

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t, X_t \sim p_t} [\|u_\theta(X_t, t) - u(X_t, t)\|_2^2], \quad (2.5)$$

where $t \sim U[0, 1]$, $u_\theta(x, t)$ is a neural network with parameter θ . Flow Matching simplifies the problem of designing a probability path p_t and its corresponding velocity field $u(x, t)$ by adopting a conditional strategy. Formally, conditioning on any arbitrary random vector $Z \in \mathbb{R}^m$ with PDF p_Z , the marginal probability path p_t satisfies

$$p_t(x) = \int p_t(x|z)p_Z(z)dz. \quad (2.6)$$

Suppose conditional velocity field $u(x|z, t)$ generates $p_t(x|z)$, Lipman et al. [2022] show that following marginal velocity field $u(x, t)$ generates marginal probability path p_t under mild assumptions:

$$u(x, t) := \int u(x|z, t)p_{Z|t}(z|x)dz \quad \text{with} \quad p_{Z|t}(z|x) = \frac{p_t(x|z)p_Z(z)}{p_t(x)}, \quad (2.7)$$

where the second equation follows from the Bayes' rule. Combining above, the tractable conditional flow matching loss \mathcal{L}_{CFM} , which satisfies $\nabla_\theta \mathcal{L}_{\text{CFM}}(\theta) = \nabla_\theta \mathcal{L}_{\text{FM}}(\theta)$, is defined as:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, Z \sim p_Z, X_t \sim p_t(\cdot|Z)} [\|u_\theta(X_t, t) - u(X_t|Z, t)\|_2^2]. \quad (2.8)$$

Affine Conditional Flows. The conditional flow matching loss works with any choice of conditional probability path and conditional velocity fields. In this paper, we consider the affine conditional flow with independent data coupling following [Lipman et al., 2022, 2024]:

$$\psi_t(x|x_1) = \mu_t x_1 + \sigma_t x, \quad (2.9)$$

where $\mu_t, \sigma_t : [0, 1] \rightarrow [0, 1]$ are monotone smooth functions satisfying

$$\mu_0 = \sigma_1 = 0, \quad \mu_1 = \sigma_0 = 1, \quad \text{and} \quad \frac{d\mu_t}{dt}, -\frac{d\sigma_t}{dt} > 0 \quad \text{for} \quad t \in (0, 1). \quad (2.10)$$

Setting $Z = X_1 \sim Q$, $X_0 \sim N(0, I)$, the flow ψ_t induces the probability flow $p_t(X_t|X_1) = N(\mu_t X_1, \sigma_t^2 I)$ and velocity field

$$u(x|x_1, t) = \dot{\psi}_t(\psi_t^{-1}(x|x_1)|x_1) = \frac{\dot{\sigma}_t(x - \mu_t x_1)}{\sigma_t} + \dot{\mu}_t x_1. \quad (2.11)$$

Further, using the law of unconscious statistician with $X_t = \psi_t(X_0|X_1)$, the conditional flow matching loss takes the form

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, X_1 \sim q, X_0 \sim N(0, I)} [\|u_\theta(\mu_t X_1 + \sigma_t X_0, t) - (\dot{\mu}_t X_1 + \dot{\sigma}_t X_0)\|_2^2]. \quad (2.12)$$

In practice, for collected i.i.d. data points $\{x_i\}_{i=1}^n$, (2.12) is implemented with Monte-Carlo simulation. To avoid instability, we often clip the interval $[0, 1]$ with t_0 and T . Namely, for any velocity estimator u^θ , we consider the empirical loss function $\hat{\mathcal{L}}_{\text{CFM}}(\theta)$:

$$\hat{\mathcal{L}}_{\text{CFM}}(\theta) := \frac{1}{n} \sum_{i=1}^n \int_{t_0}^T \frac{1}{T - t_0} \mathbb{E}_{X_0 \sim N(0, I)} [\|u_\theta(\mu_t x_i + \sigma_t X_0, t) - (\dot{\mu}_t x_i + \dot{\sigma}_t X_0)\|_2^2] dt. \quad (2.13)$$

3 Kullback-Leibler (KL) Error Bound for Flow Matching

In this section, we provide a novel upper bound on the Kullback-Leibler (KL) divergence between the true data distribution and the Flow Matching (FM) estimated distribution. Our analysis starts from a fundamental KL evolution identity (Lemma 3.1) that characterizes how the divergence evolves along the probability flow [Albergo et al., 2023]. We then show that, under mild regularity assumptions on the velocity fields and density paths, this identity yields an explicit upper bound on the terminal KL divergence in terms of the L_2 Flow Matching training loss (Theorem 3.1).

Albergo et al. [2023] presents a *evolution identity* that characterizes the evolution of the KL divergence between the true distribution and the flow-matching estimated distribution.

Lemma 3.1 (KL Evolution Identity for Continuity Flows; Lemma 21 of [Albergo et al., 2023]). Let two velocities field $u(x, t), v(x, t) \in C([0, 1]; (C^1(\mathbb{R}^d))^d)$. Let p_t and q_t be two paths of differentiable probability densities on \mathbb{R}^d evolving under the continuity equations

$$\frac{\partial p_t(x)}{\partial t} + \nabla \cdot (p_t(x)u(x, t)) = 0, \quad \frac{\partial q_t(x)}{\partial t} + \nabla \cdot (q_t(x)v(x, t)) = 0, \quad (3.1)$$

with same initial distribution $p_0 = q_0$. Then for all $t \in [0, 1]$,

$$\frac{d}{dt} \text{KL}(p_t || q_t) = \mathbb{E}_{x \sim p_t} [(u(x, t) - v(x, t))^\top (\nabla \log p_t(x) - \nabla \log q_t(x))]. \quad (3.2)$$

Proof. For completeness, we restate the proof in Section B.1. □

Lemma 3.1 shows that the time derivative of the KL divergence between the true path p_t and

the estimated path q_t equals the expected inner product between the velocity error and the Stein score error of the two flows. Consequently, the terminal KL divergence is the path integral of the multiplication of velocity discrepancy and stein score discrepancy along the trajectory.

Lemma 3.2 (KL Difference for Continuity Flows). Let two velocities field $u(x, t), v(x, t) \in C([0, 1]; (C^1(\mathbb{R}^d))^d)$. Let p_t and q_t be two paths of differentiable probability densities on \mathbb{R}^d evolving under the continuity equations

$$\frac{\partial p_t(x)}{\partial t} + \nabla \cdot (p_t(x)u(x, t)) = 0, \quad \frac{\partial q_t(x)}{\partial t} + \nabla \cdot (q_t(x)v(x, t)) = 0,$$

with same initial distribution $p_0 = q_0$. Then for all $t \in [0, 1]$,

$$\text{KL}(p_1||q_1) = \int_0^1 \mathbb{E}_{x \sim p_t} [(u(x, t) - v(x, t))^\top (\nabla \log p_t(x) - \nabla \log q_t(x))] dt. \quad (3.3)$$

Proof. Taking the integral of both sides of (3.2) with respect to time t . □

By applying the KL Difference for Continuity Flows (Lemma 3.2), we bound terminal the KL error for flow matching $\text{KL}(p_1||q_1)$ in terms of the flow-matching loss and structural regularity of the underlying flows. To control the stein score error, we introduce additional conditions on training error and Lipschitz smoothness of velocities and data score.

Assumption 3.1 (Bound on L_2 Flow Matching Loss). Let p_t denote the probability path induced by u_t . Assume the true velocity $u(x, t)$ and its estimators $v(x, t)$ satisfy

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, x \sim p_t} [\|u(x, t) - v(x, t)\|_2^2] \leq \epsilon^2.$$

For each $t \in [0, 1]$, define the point-wise error $\epsilon(t)$ as

$$\epsilon(t) := \sqrt{\mathbb{E}_{x \sim p_t} [\|u(x, t) - v(x, t)\|_2^2]}.$$

Assumption 3.2 (Regularity of Data Path). Assume the true probability path p_t satisfy that for any $t \in [0, 1]$ and $x \in \mathbb{R}^d$,

$$\|\nabla \log p_t(x)\|_2 \leq B_p(t), \quad \|\nabla(\nabla \log p_t(x))\|_2 \leq U_p(t).$$

Assumption 3.3 (Regularity of Velocity Fields). Assume velocities $u(x, t), v(x, t) \in C^2$ satisfy for any $t \in [0, 1]$ and $x \in \mathbb{R}^d$, $|\nabla \cdot u - \nabla \cdot v| \leq K(t)$ and following hold

$$\begin{aligned} \|u(x, t)\|_2 &\leq M(t), & \|v(x, t)\|_2 &\leq M(t), \\ \|\nabla u(x, t)\|_2 &\leq L(t), & \|\nabla v(x, t)\|_2 &\leq L(t), \\ \|\nabla(\nabla \cdot u(x, t))\|_2 &\leq H(t), & \|\nabla(\nabla \cdot v(x, t))\|_2 &\leq H(t), \end{aligned}$$

Assumption 3.1 is a natural assumption on the training error. Assumption 3.2 and Assumption 3.3

is essential for applying the Cauchy-Schwarz inequality to expectations and ensuring integrability.

Before we demonstrate our main result, we introduce the Grönwall's Inequality as a key helping lemma. It provides a way to control the score error through its derivative upper bound.

Lemma 3.3 (Grönwall's Inequality; [Gronwall, 1919]). Let $a, b \in \mathbb{R}$ with $a < b$. Suppose functions $g(t), y(t), h(t) \in C^1[a, b]$. Then, if $y(t)$ is differentiable on $[a, b]$ and satisfies:

$$\frac{d}{dt}y(t) \leq y(t)g(t) + h(t), \quad t \in [a, b].$$

Define $\Gamma(t) = \int_a^t g(s)ds$, then it holds

$$y(t) \leq y(a)e^{\Gamma(t)} + \int_a^t e^{\Gamma(t)-\Gamma(\tau)}h(\tau)d\tau.$$

We now bound the terminal KL divergence using the flow-matching loss and the regularity constants defined in [Assumption 3.1](#), [Assumption 3.2](#) and [Assumption 3.3](#).

Theorem 3.1 (Flow Matching KL Error Bounds). Assume [Assumptions 3.1](#) to [3.3](#) hold. If the initial distributions $p_0 = q_0$, then

$$\text{KL}(p_1||q_1) \leq \epsilon \sqrt{\int_0^1 \mathbb{E}_{x \sim p_t} [\|\nabla \log p_t(x) - \nabla \log q_t(x)\|_2^2] dt}.$$

Let the lipschitz constants $U_p(t), B_p(t), K(t), L(t), M(t), H(t)$ be as defined in [Assumption 3.2](#) and [Assumption 3.3](#), then we bound the terminal KL divergence by

$$\text{KL}(p_1||q_1) \leq A_1\epsilon + A_2\epsilon^2,$$

where

$$A_1 := \exp\left\{\int_0^1 L(t) + K(t) + B_p(t)M(t)dt\right\} \cdot \int_0^1 2L(t)B_p(t) + 2H(t)dt,$$

$$A_2 := \exp\left\{\int_0^1 L(t) + K(t) + B_p(t)M(t)dt\right\} \cdot \sqrt{\int_0^1 U_p(t)^2 dt}.$$

Proof Sketch. Please see [Section B.2](#) for a detailed proof. □

[Theorem 3.1](#) implies that when the velocity estimator achieves a small flow-matching loss ϵ , the terminal KL divergence is guaranteed to remain small, with constants determined by the Lipschitz and regularity properties of the underlying flows. As $\epsilon \rightarrow 0$, the KL divergence exhibits an asymptotically linear convergence rate with respect to ϵ . This result establishes the first explicit asymptotic relation between flow matching training error and KL divergence. This provides a theoretical justification for the empirical stability of flow matching generative models.

4 Convergence Rates of Flow Matching in Total Variation Distance

In this section, we establish the statistical convergence rates of Flow Matching Transformers for large sample size under the Total Variation (TV) distance. Previous studies [Fukumizu et al., 2024, Jiao et al., 2024, Su et al., 2025a] analyze the convergence properties of Flow Matching models under the 2-Wasserstein distance. By exploiting the KL error bounds (Theorem 3.1), we extend these results to the TV distance, which provides a more direct characterization of the discrepancy between probability distributions. In Theorem 4.1, we present the convergence rate of Flow Matching Transformers under a general Hölder-smoothness assumption on the data distribution (Assumption 4.1). Furthermore, in Theorem 4.2, we show that Flow Matching Transformers achieve an almost minimax-optimal convergence rate under specific regularity conditions.

Estimating probability distributions whose densities lie within a Hölder ball is a common setting in nonparametric statistics and statistical rates studies [Györfi et al., 2002, Takezawa, 2005, Fu et al., 2024, Su et al., 2025b], as Hölder functions possess well-understood smoothness and approximation properties. We begin by introducing the formal definition of the Hölder space.

Definition 4.1 (Hölder Space). Let $\alpha \in \mathbb{Z}_+^d$, and let $\beta = k_1 + \gamma$ denote the smoothness parameter, where $k_1 = \lfloor \beta \rfloor$ and $\gamma \in [0, 1)$. Given a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, the Hölder space $\mathcal{H}^\beta(\mathbb{R}^d)$ is defined as the set of α -differentiable functions satisfying: $\mathcal{H}^\beta(\mathbb{R}^d) := \{f : \mathbb{R}^d \rightarrow \mathbb{R} \mid \|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)} < \infty\}$, where the Hölder norm $\|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)}$ satisfies:

$$\|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)} := \sum_{\|\alpha\|_1 < k_1} \sup_x |\partial^\alpha f(x)| + \max_{\alpha: \|\alpha\|_1 = k_1} \sup_{x \neq x'} \frac{|\partial^\alpha f(x) - \partial^\alpha f(x')|}{\|x - x'\|_\infty^\gamma}.$$

Also, we define the Hölder ball of radius B by $\mathcal{H}^\beta(\mathbb{R}^d, B) := \{f : \mathbb{R}^d \rightarrow \mathbb{R} \mid \|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)} < B\}$.

We assume the data distribution $p_1(x)$ is Hölder-smooth and light-tail following.

Assumption 4.1 (Generic Hölder Smooth Data). The density function $p_1(x)$ belongs to Hölder ball of radius $B > 0$ with Hölder index $\beta > 0$ (Definition 4.1), denoted by $p_1(x) \in \mathcal{H}^\beta(\mathbb{R}^{d_x}, B)$. Also, there exist constant $C_1, C_2 > 0$ such that $p_1(x) \leq C_1 \exp(-C_2 \|x_1\|_2^2/2)$.

For $t_0, T \in [0, 1]$, we evaluate the performance of estimator u_θ through risk function $\mathcal{R}(u_\theta)$:

$$\mathcal{R}(u_\theta) := \frac{1}{T - t_0} \int_{t_0}^T \mathbb{E}_{x_t \sim p_t} [\|u(x, t) - u_\theta(x, t)\|_2^2] dt,$$

Let \hat{u}_θ denote the trained velocity estimator, obtained from i.i.d. samples $\{x_i\}_{i=1}^n$ drawn from the target data distribution p_1 . The following lemma establishes the convergence rate of the expected empirical risk $\mathcal{R}(\hat{u}_\theta)$ with respect to the training samples $\{x_i\}_{i=1}^n$.

Lemma 4.1 (Velocity Estimation with Transformer; Theorem 4.2 of [Su et al., 2025a]). Let d be the transformer’s feature dimension. Following standard pachify procedures [Peebles and Xie, 2023], we partition the input (data) of dimension $d_x = d \cdot L$ into a sequence of length L and (transformer) feature dimension d . Assume [Assumption 4.1](#) hold, we have

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\hat{u}_\theta)] = O(n^{-\frac{1}{10d}} (\log n)^{10d_x}).$$

[Lemma 4.1](#) characterize how the evaluation error decrease when data sample size n increases. Next, we analyze the distribution error convergence rate of the flow matching transformers \hat{u}_θ under the TV distance. The following theorem quantifies how the TV distance between the true distribution and the estimated distribution decreases with increasing sample size n .

Theorem 4.1 (Convergence Rate under Total Variation Distance). Let p_1 be the data distribution and q_1 be the flow matching estimated distribution. Let d be the feature dimension. Assume [Assumption 4.1](#) holds. Then we have

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\text{TV}(p_1, q_1)] = O(n^{-\frac{1}{20d}} (\log n)^{5d_x}).$$

Proof. Please see [Section B.4](#) for a detailed proof. \square

We now relate our rate to the classical minimax lower bound for density estimation under Hölder smoothness. Recall the minimax lower bound under the strong Hölder smoothness assumption:

Lemma 4.2 (Theorem 4 of [Yang and Barron, 1999]). Consider the task of estimating a probability distribution $P(x)$ with density function belonging to the space

$$\mathcal{P} := \{p(x) \mid p(x) = f(x) \exp(-C_2 \|x\|_2^2 / 2) : f(x) \in \mathcal{H}^\beta(\mathbb{R}^{d_x}, B), C_1 \geq f(x) \geq C\},$$

Then, given i.i.d training data set $\{x_i\}_{i=1}^n$ drawn from distribution P , we have

$$\inf_{\hat{P}} \sup_{p(x) \in \mathcal{P}} \mathbb{E}_{\{x_i\}_{i=1}^n} [\text{TV}(\hat{P}, P)] \gtrsim \Omega(n^{-\frac{\beta}{d_x + 2\beta}}),$$

where \hat{P} runs over all possible estimators constructed from the data.

We show flow matching transformers match the minimax optimal rate under specific conditions.

Theorem 4.2 (Nearly Minimax Optimality of Flow Matching Transformers). Let C , C_1 and C_2 be positive constants. Assume the data distribution satisfies $p_1(x) = \exp(-C_2 \|x\|_2^2 / 2) \cdot f(x)$, where f belongs to Hölder space $f(x) \in \mathcal{H}^\beta(\mathbb{R}^{d_x}, B)$ ([Definition 4.1](#)) and satisfies $C_1 \geq f(x) \geq C$ for all x . Then, within the Hölder distribution class under the Total Variation metric, the Flow Matching Transformer achieves the minimax-optimal convergence rate when $20d\beta = d_x + 2\beta$.

Proof. Please see [Section B.5](#) for a detailed proof. \square

[Lemma 4.2](#) characterizes the fundamental statistical limit of estimating smooth densities within the Hölder class. Building on this, [Theorem 4.2](#) demonstrates that the distributional convergence

rate of Flow Matching Transformers scales only logarithmically worse than the classical minimax lower bound [Yang and Barron, 1999]. Moreover, this convergence rate matches that of diffusion models shown by Fu et al. [2024] using MLP networks and by Hu et al. [2024] using Transformer networks. These results indicate that Flow Matching is as efficient as diffusion models in terms of minimax convergence in Hölder class under the Total Variation (TV) distance. Consequently, our analysis provides a theoretical justification for the empirical observation that deterministic flow-based generative models achieve comparable performance to diffusion-based methods.

5 Numerical Studies

Our goal is to validate Lemma 3.1 and Theorem 3.1 with toy examples.

5.1 Validating KL Evolution Identity (Lemma 3.1)

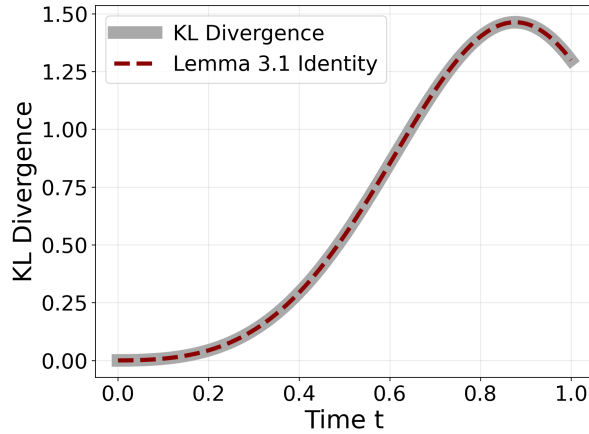


Figure 1: **Closed-Form KL Identity (Lemma 3.1) Verification without Learning.** Here p_t evolves under $a_1(t) = \sin(\pi t)$ while q_t evolves under $a_3(t) = t - \frac{1}{2}$.

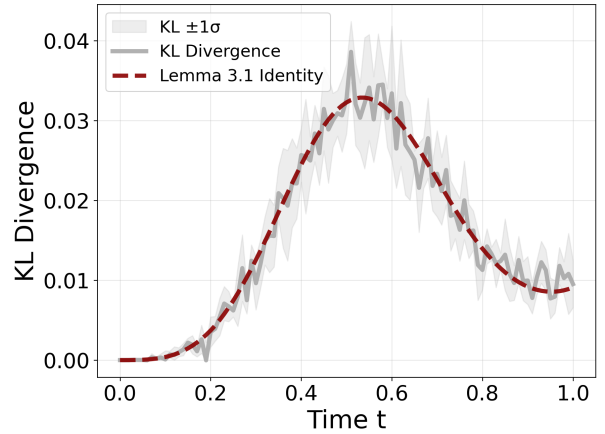


Figure 2: **KL Identity (Lemma 3.1) Verification with Learned Velocity Field c_θ .** The model is trained on $a_2(t) = 0.3 \sin(2\pi t) + 0.2$ until validation MSE ≤ 0.05 . Sampling from p_t (also under a_2), we compare the empirical KL divergence (dark grey) with the integrated RHS (dark red, dashed).

Setup. We validate the KL evolution identity that relates the time derivative of the divergence between two probability paths to an inner product between velocity and score mismatches:

$$\text{KL}(p_t||q_t) = \int_0^t \mathbb{E}_{x \sim p_s} [(u(x, s) - v(x, s))^\top (\nabla \log p_s(x) - \nabla \log q_s(x))] ds. \quad (5.1)$$

To show the generality of Lemma 3.1, we examine two settings:

- **Learn-less Flow.** Lemma 3.1 with velocity fields $v(x, s)$ given.
- **Learned Flow.** Lemma 3.1 with velocity fields $v(x, s)$ learned from synthetic data.

Data. We work in \mathbb{R}^2 with $p_0 = N(0, I_2)$ and a linear target field $u(x, t) = a(t)x$. This induces $p_t = N(0, \sigma_p(t)^2 I_2)$ with $\sigma_p(t) = \exp(\int_0^t a(s) ds)$. We use three schedules

$$\begin{aligned} a_1(t) &= \sin(\pi t), \\ a_2(t) &= 0.3 \sin(2\pi t) + 0.2, \\ a_3(t) &= t - \frac{1}{2}. \end{aligned}$$

We evaluate on a uniform grid $t_k \in [0, 1]$.

Model. We train a small MLP $v_\theta(x, t)$ by flow matching on pairs (t, x) with $t \sim U[0, 1]$ and $x \sim p_t$. For no learning baselines we also set q_t by an analytic linear field $v(x, t) = \tilde{a}(t)x$.

Baseline (Ground Truth). For the left side of (5.1), at each t , we draw $x^{(i)} \sim p_t$ via $x^{(i)} = \sigma_p(t)z^{(i)}$ with $z^{(i)} \sim N(0, I_2)$, compute $\log p_t(x^{(i)})$ in closed form, and obtain $\log q_t(x^{(i)})$ by integrating the backward initial-value problem $\dot{x}_s = -v_\theta(x_s, s)$ from (x, t) to $s = 0$ while accumulating $\ell(t) = \int_0^t \nabla \cdot v_\theta(x_s, s) ds$, then set $\log q_t(x) = \log p_0(x_0) + \ell(t)$. Then, the Monte Carlo estimate is

$$\widehat{\text{KL}}(p_t || q_t) = \frac{1}{N} \sum_{i=1}^N \log p_t(x^{(i)}) - \log q_t(x^{(i)}).$$

For the right side integrand of (5.1), we reuse the same samples and compute $u(x, t) = a(t)x$, $v_\theta(x, t)$, $s_p(x, t) = \nabla \log p_t(x) = -x/\sigma_p(t)^2$, and $s_q(x, t) = \nabla_x \log q_t(x)$ by a single autograd gradient of the scalar $\log q_t(x)$ with respect to the terminal x . Then, we have the integrand estimator

$$\widehat{g}(t) = \frac{1}{N} \sum_{i=1}^N (u(x^{(i)}, t) - v_\theta(x^{(i)}, t))^\top (s_p(x^{(i)}, t) - s_q(x^{(i)}, t)).$$

Finally, we approximate $\int_0^t \widehat{g}(s) ds$ on the time grid by the trapezoidal rule.

Results. We present our results in **Figures 1 and 2**.

- **Learnless Flow (Figure 1).** With p_t under a_1 and q_t under a_3 , both sides admit closed forms and the curves coincide within numerical precision, confirming the identity.
- **Learned Flow (Figure 2 for a_2).** For each schedule a_1, a_2, a_3 , we train v_θ to several validation error levels and compare the empirical left side $t \mapsto \widehat{\text{KL}}(p_t || q_t)$ against the numerical right side $t \mapsto \int_0^t \widehat{g}(s) ds$. The two curves track closely across t .

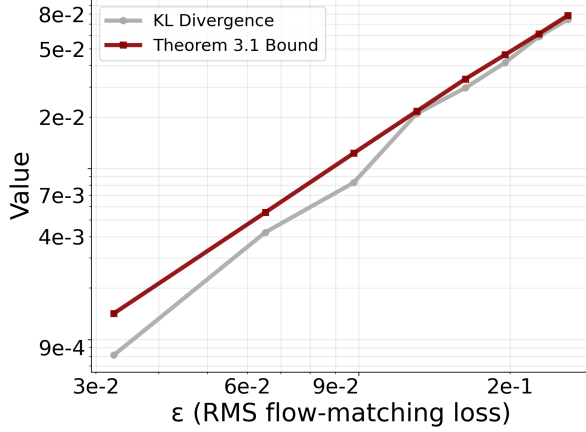


Figure 3: **Closed-Form KL Error Bound (Theorem 3.1) Verification.** Uses schedule a_3 with constant perturbations. Line plot showing $\text{KL}(p_1||q_1)$ versus $\epsilon\sqrt{S}$ for synthetic velocity fields $v(x, t) = (a_3(t) + \delta(t))x$ with $\delta(t) = \beta$, $\beta \in \{0, 0.025, \dots, 0.2\}$. Each point represents one perturbation configuration.

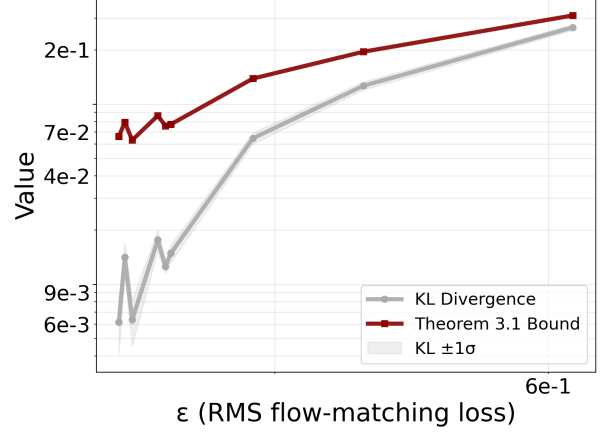


Figure 4: **KL Error Bound (Theorem 3.1) Verification with Learned Velocity Field.** Uses schedule a_1 . Line plot showing $\text{KL}(p_1||q_1^\theta)$ (dark grey) and $\epsilon_\theta\sqrt{S_\theta}$ (dark red) versus ϵ_θ (RMS flow-matching loss) on log-log axes for multiple checkpoints during training. Each point represents a checkpoint at different training stages.

5.2 Validating KL Error Bounds (Theorem 3.1)

Setup. We aim to validate the learned model bound from Theorem 3.1

$$\text{KL}(p_1||q_1^\theta) \leq \epsilon_\theta \sqrt{\int_0^1 \mathbb{E}_{x \sim p_t} \|\nabla \log p_t(x) - \nabla \log q_t^\theta(x)\|_2^2 dt}, \quad (5.2)$$

where $\epsilon_\theta^2 = \mathbb{E}_{t \sim \mathcal{U}[0,1], x \sim p_t} \|v_\theta(x, t) - u(x, t)\|_2^2$ and the time integral equals the score gap S_θ .

Data. We use same Gaussian setting as Section 5.1 with $u(x, t) = a(t)x$ and schedules

$$\begin{aligned} a_1(t) &= \sin(\pi t), \\ a_2(t) &= 0.3 \sin(2\pi t) + 0.2, \\ a_3(t) &= t - \frac{1}{2}. \end{aligned}$$

Model. We train an MLP $v_\theta(x, t)$ by flow matching using Adam with cosine decay. We save a ladder of checkpoints across training to obtain a range of ϵ_θ values.

Baseline (Ground Truth). For synthetic checks we also use $v(x, t) = (a(t) + \delta(t))x$ with constant $\delta(t) = \beta$ to produce controlled perturbations.

Estimators. For any checkpoint θ :

1. **LHS of (5.2).** We estimate $\text{KL}(p_1||q_1^\theta)$ by sampling $x \sim p_1$, computing $\log p_1(x)$ in closed form and $\log q_1^\theta(x)$ via a backward initial-value problem with divergence accumulation, then averaging $\log p_1(x) - \log q_1^\theta(x)$.
2. **Flow Error ϵ_θ .** We evaluate $\epsilon_\theta = \sqrt{\mathbb{E}\|v_\theta - u\|^2}$ on a uniform time grid with fresh Monte Carlo over $x \sim p_t$.
3. **Score Gap S_θ .** We compute $S_\theta = \int_0^1 \mathbb{E}\|s_p(x, t) - s_q^\theta(x, t)\|^2, dt$ by drawing common random numbers once, setting $x(t) = \sigma_p(t)z$, using $s_p(x, t) = -x/\sigma_p(t)^2$, and obtaining $s_q^\theta(x, t)$ by one autograd gradient of $\log q_t^\theta(x)$ with respect to terminal x . Then we do trapezoidal integration over t .

Results. We summarize our results in **Figures 3** and **4**.

- **Verification with Synthetic Perturbations (Figure 3).** With $v(x, t) = (a(t) + \delta(t))x$ under constant δ , the same picture appears without learning. The KL divergence curve lies under the curve given by $\epsilon\sqrt{S}$, validating the bound.
- **Verification with Learned Velocity Fields ϵ (Figure 4).** We plot both sides as functions of ϵ_θ on the same axes. For each checkpoint we form the triplet $(\epsilon_\theta, \text{KL}(p_1||q_1^\theta), \epsilon_\theta\sqrt{S_\theta})$ under fixed evaluation settings and a fixed common random seed, sort by ϵ_θ , and connect the points to obtain two curves. Empirically, the left side remains below the right side across checkpoints, validating the bound.

6 Discussion and Conclusion

We prove a deterministic, non-asymptotic KL upper bound for flow matching: if the L_2 flow matching loss \mathcal{L}_{FM} is bounded by $\epsilon^2 > 0$, then

$$\text{KL}(p_1||q_1) \leq A_1\epsilon + A_2\epsilon^2,$$

where A_1, A_2 depending only on regularity of the velocity fields and data scores (**Theorem 3.1**). The proof starts from the KL evolution identity for continuity flows (**Lemma 3.1**). Our contribution is to turn this pathwise identity into an deterministic, non-asymptotic bound with explicit constants via a score-PDE analysis and Grönwall’s inequality (**Lemma 3.3**). Leveraging the KL control, we derive Total Variation Distance convergence rates for Flow Matching Transformers (**Theorem 4.1**) and prove near-minimax optimality (**Theorem 4.2**). Experiments validate both the KL identity and the KL bound (**Section 5**). Overall, our results give statistical efficiency guarantees for flow matching and clarify when and how training loss controls distributional KL error.

In addition, we now prove **Theorem 3.1**’s critical dependence on the assumption that the velocity fields $u(x, t), v(x, t)$ are differentiable and strong solutions to the continuity equations (3.1). If we relax the q_t as a non differentiable function, we construct a counterexample that for any arbitrarily small flow matching loss, the resulting terminal divergence grow without bound.

Proposition 6.1 (Flow Matching Loss Fails to Control KL (Weak Version)). Fix any constants $M > \epsilon > 0$. There exist probability paths $p_t(x)$ and $q_t(x)$ with $p_0(x) = q_0(x)$, and corresponding velocity fields $u(x, t)$ and $v(x, t)$ such that for almost every $t \in [0, 1]$, [Assumptions 3.1](#) and [3.3](#) hold, and paths satisfy the weak form of continuity equations

$$\frac{\partial p_t(x)}{\partial t} + \nabla \cdot (p_t(x)u(x, t)) = 0, \quad \frac{\partial q_t(x)}{\partial t} + \nabla \cdot (q_t(x)v(x, t)) = 0.$$

Moreover,

$$\int_0^1 \mathbb{E}_{x \sim p_t} [\|u(x, t) - v(x, t)\|_2^2] dt \leq \epsilon, \quad \text{but} \quad \text{KL}(p_1 \| q_1) \geq M.$$

Proof. Please see [Section B.3](#) for a detailed proof. □

Impact Statement

By the theoretical nature of this work, we do not anticipate any negative social impact.

Acknowledgments

The authors would like to thank Minshuo Chen, Mimi Gallagher, Sara Sanchez, T.Y. Ball, Dino Feng and Andrew Chen for valuable conversations; and Mingcheng Lu, Yi-Chen Lee, Shang Wu and Zhao Song for collaborations on related topics.

This work is in remembrance of Prof. Bi-Ming Tsai (February 9, 1965 - October 21, 2025). Her wisdom in health, cooking, and the art of thoughtful living continues to inspire JH.

JH is partially supported by the Walter P. Murphy Fellowship and the Terminal Year Fellowship (Paul K. Richter Memorial Award) of Northwestern University. Han Liu is partially supported by NIH R01LM1372201, NSF AST-2421845, Simons Foundation MPS-AI-00010513, AbbVie, Dolby and Chan Zuckerberg Biohub Chicago Spoke Award. This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Appendix

A	Related Work	15
B	Proofs in the Main Text	16
B.1	Proof of Lemma 3.1	16
B.2	Proof of Theorem 3.1	18
B.3	Proof of Proposition 6.1	22
B.4	Proof of Theorem 4.1	25
B.5	Proof of Theorem 4.2	25

A Related Work

Error Bounds for Flow Matching. Several prior works have analyzed error bounds for fully-deterministic flow matching. [Albergo and Vanden-Eijnden \[2022\]](#) establish a 2-Wasserstein bound on the discrepancy between the flow-matching estimated distribution and the true data distribution with uniform Lipschitz constants. [Li et al. \[2023\]](#) derive a Total Variation (TV) bound but require the estimated Stein score error to be small. [Benton et al. \[2023\]](#) provide a 2-Wasserstein error bound while relaxing the uniform Lipschitz condition to time-dependent regularity. However, these studies focus primarily on Wasserstein Distance or TV Distance, or rely on strong smoothness assumptions on the estimated score function. In contrast, our work derives a non-asymptotic KL bound for flow matching that depends only on the regularity and smoothness of the data path and velocity fields, without requiring regularity of the estimated score path.

Statistical Rates and Minimax Optimality of Flow Models. While error bounds study deterministic approximation for a fixed model without learning effects, statistical rate studies quantifies how error decreases with sample size. [Benton et al. \[2023\]](#), [Albergo and Vanden-Eijnden \[2022\]](#) measure the convergence of flow models through the L_2 risk of the velocity field but do not provide explicit convergence rates. [Jiao et al. \[2024\]](#) derive explicit rates in the latent space of an autoencoder. However, their analysis does not account for the smoothness of the target density class. [Su et al. \[2025b\]](#) establish statistical rates for discrete flow matching by deriving a model-agnostic intrinsic error bound. [Fukumizu et al. \[2024\]](#) show that flow matching achieves nearly minimax-optimal distribution estimation rates in Besov function spaces under the 2-Wasserstein distance using ReLU network architectures. In the same vein, [Kunkel and Trabs \[2025\]](#) obtain similar results under the 1-Wasserstein distance using kernel density estimators. [Su et al. \[2025a\]](#) further establish convergence rates for high-order flow matching under the 2-Wasserstein distance. However, all of these analyses are established under Wasserstein-type metrics. In contrast, we provide convergence rates for flow matching under the Total Variation (TV) distance and prove that flow matching achieves nearly minimax-optimal efficiency in TV distance, which is comparable to diffusion models.

B Proofs in the Main Text

B.1 Proof of Lemma 3.1

Lemma B.1 (Lemma 3.1 Restated: KL Evolution Identity for Continuity Flows). Let two velocities field $u(x, t), v(x, t) \in C([0, 1]; (C^1(\mathbb{R}^d))^d)$. Let p_t and q_t be two paths of differentiable probability densities on \mathbb{R}^d evolving under the continuity equations

$$\frac{\partial p_t(x)}{\partial t} + \nabla \cdot (p_t(x)u(x, t)) = 0, \quad \frac{\partial q_t(x)}{\partial t} + \nabla \cdot (q_t(x)v(x, t)) = 0, \quad (\text{B.1})$$

with same initial distribution $p_0 = q_0$. Then for all $t \in [0, 1]$,

$$\frac{d}{dt} \text{KL}(p_t || q_t) = \mathbb{E}_{x \sim p_t} [(u(x, t) - v(x, t))^\top (\nabla \log p_t(x) - \nabla \log q_t(x))]. \quad (\text{B.2})$$

Proof. Let all probability density functions and velocity fields defined on $\Omega = \mathbb{R}^d$. Then for any differentiable function $f_t(x) : \mathbb{R}^d \rightarrow \mathbb{R}$, we have

$$\begin{aligned} & \frac{d}{dt} \int_{\Omega} p_t(x) f_t(x) dx \\ &= \int_{\Omega} \frac{\partial p_t(x)}{\partial t} f_t(x) dx + \int_{\Omega} \frac{\partial f_t(x)}{\partial t} p_t(x) dx && (\text{By the product rule}) \\ &= - \int_{\Omega} \nabla \cdot (p_t(x)u(x, t)) f_t(x) dx + \int_{\Omega} \frac{\partial f_t(x)}{\partial t} p_t(x) dx \\ & && (\text{By replacing the first term by continuity equation (B.1)}) \\ &= \int_{\Omega} (p_t(x)u(x, t))^\top \nabla f_t(x) dx - \int_{\partial \Omega} p_t(x)u(x, t) f_t(x) \cdot (\widehat{n} dS) + \int_{\Omega} \frac{\partial f_t(x)}{\partial t} p_t(x) dx \\ & && (\text{By divergence theorem}) \\ &= \int_{\Omega} (p_t(x)u(x, t))^\top \nabla f_t(x) dx + \int_{\Omega} \frac{\partial f_t(x)}{\partial t} p_t(x) dx \\ &= \int_{\Omega} p_t(x) \left[\frac{\partial}{\partial t} f_t(x) + u(x, t)^\top \nabla f_t(x) \right] dx. && (\text{B.3}) \end{aligned}$$

Now we let $f_t(x) = \log \frac{p_t(x)}{q_t(x)}$, then the derivative of KL divergence between p_t and q_t becomes

$$\begin{aligned} \frac{d}{dt} \text{KL}(p_t || q_t) &= \frac{d}{dt} \int_{\Omega} p_t(x) f_t(x) dx && (\text{By the definition of KL divergence}) \\ &= \int_{\Omega} p_t(x) \underbrace{\left[\frac{\partial}{\partial t} \log \frac{p_t(x)}{q_t(x)} \right]}_{:= (I)} + \underbrace{u(x, t)^\top \nabla \log \frac{p_t(x)}{q_t(x)}}_{:= (II)} dx. && (\text{B.4}) \end{aligned}$$

Then we calculate (I) and (II) terms in (B.4). From the continuity equations (B.1), we have

$$\begin{aligned}\frac{\partial \log p_t(x)}{\partial t} &= -u(x, t)^\top \nabla \log p_t(x) - \nabla \cdot u(x, t), \\ \frac{\partial \log q_t(x)}{\partial t} &= -v(x, t)^\top \nabla \log q_t(x) - \nabla \cdot v(x, t).\end{aligned}$$

Consequently, (I) becomes

$$\frac{\partial}{\partial t} \log \frac{p_t(x)}{q_t(x)} = -\nabla \cdot u(x, t) + \nabla \cdot v(x, t) - u(x, t)^\top \nabla \log p_t(x) + v(x, t)^\top \nabla \log q_t(x). \quad (\text{B.5})$$

For (II), by the property of logarithmic function, we have

$$u(x, t)^\top \nabla \log \frac{p_t(x)}{q_t(x)} = u(x, t)^\top [\nabla \log p_t(x) - \nabla \log q_t(x)]. \quad (\text{B.6})$$

Substituting (B.5) and (B.6) into (B.4), the derivative of KL divergence becomes

$$\begin{aligned}\frac{d}{dt} \text{KL}(p_t || q_t) &= \int_{\Omega} p_t(x) \left[\frac{\partial}{\partial t} \log \frac{p_t(x)}{q_t(x)} + u(x, t)^\top \nabla \log \frac{p_t(x)}{q_t(x)} \right] dx \\ &= \int_{\Omega} p_t(x) [-\nabla \cdot u(x, t) + \nabla \cdot v(x, t) + (-u(x, t) + v(x, t))^\top \nabla \log q_t(x)] dx.\end{aligned}$$

Notice that, for any C^1 vector field $a(x) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and the probability density function $p_t(x)$,

$$\begin{aligned}\mathbb{E}_{x \sim p_t} [\nabla \cdot a(x)] &= \int_{\Omega} (\nabla \cdot a(x)) p_t(x) dx && (\text{By the definition of expectation}) \\ &= \int_{\partial \Omega} a(x) p_t(x) \cdot (\hat{n} dS) - \int_{\Omega} a(x)^\top \nabla p_t(x) dx && (\text{By divergence theorem}) \\ &= - \int_{\Omega} a(x)^\top \nabla p_t(x) dx \\ &= - \int_{\Omega} a(x)^\top \nabla \log p_t(x) p_t(x) dx && (\text{By } \nabla \log p_t(x) = \nabla p_t(x) / p_t(x)) \\ &= - \mathbb{E}_{x \sim p_t} [a(x)^\top \nabla \log p_t(x)].\end{aligned} \quad (\text{B.7})$$

Let $a(x) = -u(x, t) + v(x, t)$. Then the derivative of KL divergence becomes

$$\begin{aligned}\frac{d}{dt} \text{KL}(p_t || q_t) &= \int_{\Omega} p_t(x) [-\nabla \cdot u(x, t) + \nabla \cdot v(x, t) + (-u(x, t) + v(x, t))^\top \nabla \log q_t(x)] dx \\ &= \int_{\Omega} p_t(x) [(u(x, t) - v(x, t))^\top (\nabla \log p_t(x) - \nabla \log q_t(x))] dx && (\text{By (B.7)})\end{aligned}$$

$$= \mathbb{E}_{x \sim p_t} [(u(x, t) - v(x, t))^\top (\nabla \log p_t(x) - \nabla \log q_t(x))].$$

This completes the proof. \square

B.2 Proof of Theorem 3.1

Theorem B.1 (Theorem 3.1 Restated: KL Error Bounds for Flow Matching). Assume Assumptions 3.1 to 3.3 hold. If the initial distributions $p_0 = q_0$, then

$$\text{KL}(p_1 || q_1) \leq \epsilon \sqrt{\int_0^1 \mathbb{E}_{x \sim p_t} [\|\nabla \log p_t(x) - \nabla \log q_t(x)\|_2^2] dt}.$$

Let the lipschitz constants $U_p(t), B_p(t), K(t), L(t), M(t), H(t)$ be as defined in Assumption 3.2 and Assumption 3.3. Then, we bound the terminal KL divergence by

$$\text{KL}(p_1 || q_1) \leq A_1 \epsilon + A_2 \epsilon^2,$$

where

$$A_1 := \exp \left\{ \int_0^1 L(t) + K(t) + B_p(t) M(t) dt \right\} \cdot \int_0^1 2L(t) B_p(t) + 2H(t) dt,$$

$$A_2 := \exp \left\{ \int_0^1 L(t) + K(t) + B_p(t) M(t) dt \right\} \cdot \sqrt{\int_0^1 U_p(t)^2 dt}.$$

Proof. Assuming the second moment of stein score error is bounded (Assumption 3.2), we bound the KL divergence $\text{KL}(p_t || q_t)$ with the Cauchy-Schwarz Inequality for expectations:

$$\begin{aligned} \text{KL}(p_1 || q_1) &= \int_0^1 \mathbb{E}_{x \sim p_t} [(u(x, t) - v(x, t))^\top (\nabla \log p_t(x) - \nabla \log q_t(x))] dt \\ &\leq \int_0^1 \sqrt{\mathbb{E}_{x \sim p_t} [\|u(x, t) - v(x, t)\|_2^2]} \cdot \sqrt{\mathbb{E}_{x \sim p_t} [\|\nabla \log p_t(x) - \nabla \log q_t(x)\|_2^2]} dt, \end{aligned}$$

where the first line is by the KL difference for continuity flows (3.3), and the second line is by Cauchy-Schwarz Inequality. Then, by Cauchy-Schwarz Inequality on L_2 space, we have

$$\begin{aligned} \text{KL}(p_1 || q_1) &\leq \sqrt{\int_0^1 \mathbb{E}_{x \sim p_t} [\|u(x, t) - v(x, t)\|_2^2] dt} \cdot \sqrt{\int_0^1 \mathbb{E}_{x \sim p_t} [\|\nabla \log p_t(x) - \nabla \log q_t(x)\|_2^2] dt} \\ &\leq \epsilon \sqrt{\int_0^1 \mathbb{E}_{x \sim p_t} [\|\nabla \log p_t(x) - \nabla \log q_t(x)\|_2^2] dt}, \end{aligned} \tag{B.8}$$

where the last step follows from Assumption 3.1.

To bound the mean-squared expected score discrepancy term $\mathbb{E}_{x \sim p_t} [\|\nabla \log p_t(x) - \nabla \log q_t(x)\|_2^2]$,

we exploit the score PDE induced by (B.3) and the continuity equations (B.1):

$$\begin{aligned}\frac{\partial \nabla \log p_t(x)}{\partial t} &= \nabla [-u(x, t)^\top \nabla \log p_t(x) - \nabla \cdot u(x, t)] \\ &= -(\nabla u(x, t))^\top \nabla \log p_t(x) - \nabla(\nabla \log p_t(x))u(x, t) - \nabla(\nabla \cdot u(x, t)),\end{aligned}$$

where the first line is by the continuity equations (B.1).

For simplicity, we define the scores for true probability paths p_t as and estimated probability paths q_t as $s_p := \nabla \log p_t(x)$ and $s_q := \nabla \log q_t(x)$, respectively.

Then, by symmetry and above equation, we have

$$\frac{\partial s_p}{\partial t} = -(\nabla u(x, t))^\top s_p - \nabla s_p u(x, t) - \nabla(\nabla \cdot u(x, t)), \quad (\text{B.9})$$

$$\frac{\partial s_q}{\partial t} = -(\nabla v(x, t))^\top s_q - \nabla s_q v(x, t) - \nabla(\nabla \cdot v(x, t)). \quad (\text{B.10})$$

For simplicity, let $\Delta s := s_p - s_q$ denote the stein score error. Then by (B.9) and (B.10), we have

$$\frac{\partial \Delta s}{\partial t} = -(\nabla u)^\top \Delta s - (\nabla u - \nabla v)^\top s_q - \nabla(\Delta s)u - \nabla s_q(u - v) - \nabla(\nabla \cdot u - \nabla \cdot v). \quad (\text{B.11})$$

By $s_q = s_p - \Delta s$, (B.11) becomes

$$\begin{aligned}\frac{\partial \Delta s}{\partial t} &= -(\nabla u)^\top \Delta s - (\nabla u - \nabla v)^\top (s_p - \Delta s) - \nabla(\Delta s)u - (\nabla s_p - \nabla \Delta s)(u - v) - \nabla(\nabla \cdot u - \nabla \cdot v) \\ &= -(\nabla v)^\top \Delta s - (\nabla u - \nabla v)^\top s_p - \nabla(\Delta s)u - (\nabla s_p - \nabla \Delta s)(u - v) - \nabla(\nabla \cdot u - \nabla \cdot v).\end{aligned} \quad (\text{B.12})$$

Now differentiate the expected mean square Stein score error and get

$$\begin{aligned}&\frac{d}{dt} \mathbb{E}_{x \sim p_t} \left[\frac{1}{2} \|\Delta s\|_2^2 \right] \quad (\text{B.13}) \\ &= \frac{d}{dt} \int_{\Omega} p_t(x) \cdot \frac{1}{2} \|\Delta s\|_2^2 dx \quad (\text{By definition of expectation}) \\ &= \int_{\Omega} p_t(x) \left[\frac{\partial}{\partial t} \frac{1}{2} \|\Delta s\|_2^2 + u(x, t)^\top \nabla \frac{1}{2} \|\Delta s\|_2^2 \right] dx \quad (\text{By replacing } f_t(x) \text{ by } \frac{1}{2} \|\Delta s\|_2^2 \text{ in (B.3)}) \\ &= \int_{\Omega} p_t(x) \left[\frac{\partial \Delta s}{\partial t} + \nabla(\Delta s)u \right]^\top \Delta s dx \quad (\text{By matrix calculus}) \\ &= \underbrace{-\mathbb{E}_{x \sim p_t} [\Delta s^\top (\nabla v) \Delta s]}_{:=(I)} - \underbrace{\mathbb{E}_{x \sim p_t} [(s_p^\top (\nabla u - \nabla v) \Delta s)]}_{:=(II)} - \underbrace{\mathbb{E}_{x \sim p_t} [(u - v)^\top (\nabla s_p) \Delta s]}_{:=(III)} \quad (\text{B.14})\end{aligned}$$

$$\begin{aligned}
& - \underbrace{\mathbb{E}_{x \sim p_t} [(u - v)^\top (-\nabla \Delta s) \Delta s]}_{:= (IV)} - \underbrace{\mathbb{E}_{x \sim p_t} [\nabla (\nabla \cdot u - \nabla \cdot v)^\top \Delta s]}_{:= (V)}. \\
& \hspace{15em} \text{(By substituting the score PDE (B.12))}
\end{aligned}$$

We bound each terms with Lipchitz constants separately.

- For (I), we have

$$- \mathbb{E}_{x \sim p_t} [\Delta s^\top (\nabla u) \Delta s] \leq L(t) \mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2].$$

- For (II), we have

$$\begin{aligned}
- \mathbb{E}_{x \sim p_t} [(s_p^\top (\nabla u - \nabla v) \Delta s)] & \leq 2L(t) \mathbb{E}_{x \sim p_t} [\|s_p\|_2 \cdot \|\Delta s\|_2] \\
& \leq 2L(t) B_p(t) \sqrt{\mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2]}. \\
& \hspace{10em} \text{(By Cauchy-Schwarz Inequality and (3.2))}
\end{aligned}$$

- For (III), we have

$$\begin{aligned}
- \mathbb{E}_{x \sim p_t} [(u - v)^\top (\nabla s_p) \Delta s] & \leq \mathbb{E}_{x \sim p_t} [\|(u - v)^\top (\nabla s_p) \Delta s\|_2] \\
& \leq U_p(t) \mathbb{E}_{x \sim p_t} [\|(u - v)^\top \Delta s\|_2]. \\
& \hspace{10em} \text{(By the definition of } U_p(t) \text{ follows Assumption 3.2)} \\
& \leq U_p(t) \epsilon(t) \sqrt{\mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2]}. \quad \text{(By Cauchy-Schwarz Inequality)}
\end{aligned}$$

- For (IV), we have

$$\begin{aligned}
\mathbb{E}_{x \sim p_t} [(u - v)^\top (\nabla \Delta s) \Delta s] & = \mathbb{E}_{x \sim p_t} [(u - v)^\top \nabla (\frac{1}{2} \|\Delta s\|_2^2)] \quad \text{(By chain rule)} \\
& = \int_{\mathbb{R}^d} p_t(x) (u - v)^\top \nabla (\frac{1}{2} \|\Delta s\|_2^2) dx \\
& \hspace{15em} \text{(By the definition of expectation)} \\
& = - \int_{\mathbb{R}^d} \frac{1}{2} \|\Delta s\|_2^2 \nabla \cdot (p_t(x) (u - v)) dx \\
& \hspace{15em} \text{(By the divergence theorem)} \\
& = - \int_{\mathbb{R}^d} p_t(x) \cdot \frac{1}{2} \|\Delta s\|_2^2 (\nabla \cdot (u - v) + (u - v)^\top s_p) dx \\
& \hspace{15em} \text{(By the product rule)} \\
& = - \mathbb{E}_{x \sim p_t} [\frac{1}{2} \|\Delta s\|_2^2 ((\nabla \cdot u - \nabla \cdot v) + (u - v)^\top s_p)] \\
& \leq K(t) \mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2] + \frac{1}{2} B_p(t) \mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2 \cdot \|u - v\|_2] \\
& \hspace{10em} \text{(By the definition of } K(t) \text{ and } B_p(t) \text{ follow Assumption 3.3)}
\end{aligned}$$

$$\leq K(t) \mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2] + B_p(t)M(t) \mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2].$$

(By definition of $M(t)$ follows [Assumption 3.3](#))

• For (V), we have

$$- \mathbb{E}_{x \sim p_t} [\nabla(\nabla \cdot u - \nabla \cdot v)^\top \Delta s] \leq 2H(t) \sqrt{\mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2]}. \quad (\text{By Cauchy-Schwarz Inequality})$$

Collecting all above into [\(B.13\)](#), dividing both sides by $\sqrt{\mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2]}$, we bound the derivative of $\sqrt{\mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2]}$ as

$$\begin{aligned} & \frac{d}{dt} \sqrt{\mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2]} \\ & \leq (L(t) + K(t) + B_p(t)M(t)) \sqrt{\mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2]} + 2L(t)B_p(t) + U_p(t)\epsilon(t) + 2H(t). \end{aligned}$$

Applying Grönwall's Inequality [Lemma 3.3](#), we have

$$\begin{aligned} \sqrt{\mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2]} & \leq \sqrt{\mathbb{E}_{x \sim p_0} [\|\Delta s\|_2^2]} e^{C_1(t)} + \int_0^t e^{C_1(t)-C_1(\tau)} C_2(\tau) d\tau, \\ & = \int_0^t e^{C_1(t)-C_1(\tau)} C_2(\tau) d\tau, \end{aligned} \quad (\text{By } p_0 = q_0)$$

where $C_1(t) = \int_0^t L(s) + K(s) + B_p(s)M(s) ds$ and $C_2(t) = 2L(t)B_p(t) + U_p(t)\epsilon(t) + 2H(t)$.

Square both sides and take integral over $t \in [0, 1]$,

$$\begin{aligned} \int_0^1 \mathbb{E}_{x \sim p_t} [\|\Delta s\|_2^2] dt & \leq \int_0^1 \left(\int_0^t e^{C_1(t)-C_1(\tau)} C_2(\tau) d\tau \right)^2 dt \\ & \leq \int_0^1 \left(\int_0^t e^{C_1(1)-C_1(\tau)} C_2(\tau) d\tau \right)^2 dt \quad (\text{By } C_1(t) - C_1(\tau) \leq C_1(1)) \\ & \leq \left(\int_0^1 e^{C_1(1)-C_1(\tau)} C_2(\tau) d\tau \right)^2. \quad (\text{By } C_2(t) \geq 0) \end{aligned}$$

Consequently, we bound the KL divergence in terms of Lipschitz Constants as

$$\begin{aligned} \text{KL}(p_1||q_1) & \leq \epsilon \sqrt{\int_0^1 \mathbb{E}_{x \sim p_t} [\|\nabla \log p_t(x) - \nabla \log q_t(x)\|_2^2] dt} \quad (\text{By } \text{[\(B.8\)](#)}) \\ & \leq \epsilon \left(\int_0^1 e^{C_1(1)-C_1(\tau)} C_2(\tau) d\tau \right) \\ & \leq \epsilon \cdot e^{C_1(1)} \left(\int_0^1 2L(t)B_p(t) + U_p(t)\epsilon(t) + 2H(t) dt \right) \quad (\text{By the definition of } C_2(t)) \end{aligned}$$

$$\begin{aligned}
&\leq \epsilon \cdot e^{C_1(1)} \left(\int_0^1 (2L(t)B_p(t) + 2H(t)) dt + \sqrt{\int_0^1 U_p(t)^2 dt} \epsilon \right) \\
&\quad \text{(By the definition of } \epsilon_t \text{ follows [Assumption 3.1](#) and Cauchy-Schwarz Inequality)} \\
&\leq A_1 \epsilon + A_2 \epsilon^2,
\end{aligned}$$

where the last step follows from defining

$$\begin{aligned}
A_1 &:= e^{C_1(1)} \int_0^1 (2L(t)B_p(t) + 2H(t)) dt \\
A_2 &:= e^{C_1(1)} \sqrt{\int_0^1 U_p(t)^2 dt}.
\end{aligned}$$

This completes the proof. \square

B.3 Proof of [Proposition 6.1](#)

Theorem B.2 ([Proposition 6.1](#) Restated: Flow Matching Loss Fails to Control KL (Weak Solution Version)). Fix any constants $M > \epsilon > 0$. There exist probability paths $p_t(x)$ and $q_t(x)$ with $p_0(x) = q_0(x)$, and corresponding velocity fields $u(x, t)$ and $v(x, t)$ such that for almost every $t \in [0, 1]$, [Assumptions 3.1](#) and [3.3](#) hold, and paths satisfy the continuity equations

$$\frac{\partial p_t(x)}{\partial t} + \nabla \cdot (p_t(x)u(x, t)) = 0, \quad \frac{\partial q_t(x)}{\partial t} + \nabla \cdot (q_t(x)v(x, t)) = 0.$$

Moreover,

$$\int_0^1 \mathbb{E}_{x \sim p_t} [\|u(x, t) - v(x, t)\|_2^2] dt \leq \epsilon, \quad \text{but} \quad \text{KL}(p_1 \| q_1) \geq M.$$

Proof. Let $p_t \equiv p_0 = N(0, I)$ with $u \equiv 0$. Then the continuity equation $\frac{\partial p_t}{\partial t} + \nabla \cdot (p_t u_t) = 0$ is satisfied. Define the helping function $\psi_t(x) = a(t)b^\top x$, where $b \in \mathbb{R}^d$ is a constant vector and $a(t)$ is a scalar function satisfy following ordinary differential equation for some $\delta > 0$ to be determined:

$$a'(t) := \frac{d}{dt} a(t) = \delta a(t), \quad \text{almost every } t \in [0, 1]. \tag{B.15}$$

Define estimated probability path q_t via p_t by

$$q_t(x) = Z_t^{-1} p_t(x) \exp(-\psi_t(x)), \quad Z_t = \int p_t e^{-\psi_t} dx \in (0, \infty),$$

which implies that

$$\nabla \log q_t(x) = \nabla \log p_t(x) - \nabla \psi_t(x). \tag{B.16}$$

Next, we define the constant estimated velocity field $v(x, t)$ as

$$v(x, t) := -\delta \nabla \psi_t(x) = -\delta a(t)b. \quad (\text{B.17})$$

Then we check the estimated probability paths q_t and the estimated velocity v satisfy the continuity equation for almost every $t \in [0, 1]$. Notice that $\log q_t = -\log Z_t + \log p_0 - \psi_t(x)$, therefore

$$\frac{\partial}{\partial t} q_t(x) = q_t(x) \cdot \frac{\partial}{\partial t} \log q_t(x) = q_t(x) \cdot \left(-\frac{\partial}{\partial t} \log Z_t - \frac{\partial}{\partial t} \psi_t(x) \right). \quad (\text{B.18})$$

Consequently, for almost everywhere $t \in [0, 1]$, the continuity equation becomes

$$\begin{aligned} & \frac{\partial}{\partial t} q_t(x) + \nabla \cdot (q_t(x)v(x, t)) \\ &= q_t(x) \cdot \left(-\frac{\partial}{\partial t} \log Z_t - \frac{\partial}{\partial t} \psi_t(x) \right) + \nabla \cdot (q_t(x)v(x, t)) \quad (\text{By (B.18)}) \\ &= q_t(x) \cdot \left(-\frac{\partial}{\partial t} \log Z_t - \frac{\partial}{\partial t} \psi_t(x) + \nabla \cdot v(x, t) + \nabla \log q_t(x)^\top v(x, t) \right) \quad (\text{By the product rule}) \\ &= q_t(x) \cdot \left(-\frac{\partial}{\partial t} \log Z_t - \frac{\partial}{\partial t} \psi_t(x) - \delta a(t) \nabla \log q_t(x)^\top b \right) \quad (\text{By (B.17), (B.16) and } \nabla \log p_t(x) = 0) \\ &= q_t(x) \cdot \left(-\frac{\partial}{\partial t} \log Z_t - a'(t)b^\top x - \delta a(t)b^\top (\nabla p_0(x) - \nabla \psi_t(x)) \right) \\ & \quad (\text{By } \psi_t(x) = a(t)b^\top x \text{ and } \log q_t = -\log Z_t + \log p_0 - \psi_t(x)) \\ &= q_t(x) \cdot \left(-\frac{\partial}{\partial t} \log Z_t - a'(t)b^\top x + \delta a(t)b^\top (x + a(t)b) \right) \quad (\text{By } p_0 = N(0, I) \text{ and } \nabla \psi_t(x) = a(t)b) \\ &= q_t(x) \cdot \left(-a'(t)b^\top x + \delta a(t)b^\top x + \delta a(t)^2 \|b\|_2^2 - \frac{\partial}{\partial t} \log Z_t \right). \end{aligned}$$

Following the definition of $a(t)$ (B.15), $-a'(t)b^\top x + \delta a(t)b^\top x = 0$. For another term, recall that $Z_t = \int_{\mathbb{R}^d} p_t(x) e^{-\psi_t(x)} dx$ and $p_t = p_0 = N(0, I)$, which implies that for almost every $t \in [0, 1]$,

$$\begin{aligned} & \delta a(t)^2 \|b\|_2^2 - \frac{\partial}{\partial t} \log Z_t \\ &= \delta a(t)^2 \|b\|_2^2 - \frac{\partial}{\partial t} \log \int_{\mathbb{R}^d} p_0(x) e^{-\psi_t(x)} dx \quad (\text{By the definition of } Z_t) \\ &= \delta a(t)^2 \|b\|_2^2 - \frac{\partial}{\partial t} \log \mathbb{E}_{x \sim p_0} [e^{-\psi_t(x)}] \\ &= \delta a(t)^2 \|b\|_2^2 - \frac{\partial}{\partial t} \log \mathbb{E}_{x \sim p_t} [e^{-a(t)b^\top x}] \quad (\text{By the definition of } \psi_t(x) = a(t)b^\top x) \\ &= \delta a(t)^2 \|b\|_2^2 - \frac{\partial}{\partial t} \frac{1}{2} a(t)^2 \|b\|_2^2 \quad (\text{By } \mathbb{E}_{x \sim N(0, I)} [e^{t^\top x}] = e^{\frac{1}{2} t^\top t}) \\ &= \delta a(t)^2 \|b\|_2^2 - a'(t)a(t) \|b\|_2^2 \quad (\text{By the chain rule}) \\ &= 0. \quad (\text{By } a'(t) = \delta a(t) \text{ follows (B.15)}) \end{aligned}$$

Therefore the continuity equation $\frac{\partial q_t(x)}{\partial t} + \nabla \cdot (q_t(x)v(x, t)) = 0$ holds almost everywhere. Then we calculate the KL bounds between $p_t(x)$ and $q_t(x)$,

$$\begin{aligned}
\frac{d}{dt} \text{KL}(p_t \| q_t) &= \mathbb{E}_{x \sim p_t} [(u(x, t) - v(x, t)) \cdot (\nabla \log p_t(x) - \nabla \log q_t(x))] \\
&\quad \text{(By KL evolution identity \textcolor{red}{Lemma 3.1})} \\
&= \mathbb{E}_{x \sim p_t} [v(x, t) \nabla \log q_t(x)] \quad \text{(By } u \equiv 0\text{)} \\
&= \mathbb{E}_{x \sim p_t} [\delta a(t)^2 \|b\|_2^2] \quad \text{(By } v = -\delta a(t)b \text{ and } \nabla \log q_t(x) = -\nabla \psi_t(x) \text{ follows (\textcolor{red}{B.16})}) \\
&= \delta a(t)^2 \|b\|_2^2. \tag{B.19}
\end{aligned}$$

Following the ordinary differential equation (\textcolor{red}{B.15}), we define $a(t)$ as

$$a(t) = \begin{cases} 0, & 0 \leq t \leq \tau, \\ \eta e^{\delta(t-\tau)} & \tau < t \leq 1. \end{cases} \tag{B.20}$$

Integrating over $t \in [0, 1]$ and using $p_0 = q_0$, we obtain

$$\begin{aligned}
\text{KL}(p_1 \| q_1) &= \int_0^1 \frac{d}{dt} \text{KL}(p_t \| q_t) dt \\
&= \delta \|b\|_2^2 \int_0^1 a(t)^2 dt \quad \text{(By (\textcolor{red}{B.19}))} \\
&= \delta \|b\|_2^2 \int_\tau^1 \eta^2 e^{2\delta(t-\tau)} dt \quad \text{(By the definition of } a(t) \text{ follows (\textcolor{red}{B.20}))} \\
&= \delta \|b\|_2^2 \frac{\eta^2}{2\delta} (e^{2\delta(t-\tau)} - 1).
\end{aligned}$$

Define $J := \int_0^1 a(t)^2 dt = \frac{\eta^2}{2\delta} (e^{2\delta(t-\tau)} - 1)$, then for the flow matching loss

$$\int_0^1 \mathbb{E}_{x \sim p_t} [\|u(x, t) - v(x, t)\|_2^2] dt = \int_0^1 \delta^2 a(t)^2 \|b\|_2^2 dt = \delta^2 \|b\|_2^2 J = \delta \cdot \text{KL}(p_1 \| q_1).$$

Given any $M > \epsilon > 0$, let $\delta = \epsilon/M$, then there exist a unique $\eta > 0$ satisfy $J = \epsilon/\delta^2 \|b\|_2^2$. Then by our construction, we have flow matching loss

$$\int_0^1 \mathbb{E}_{x \sim p_t} [\|u(x, t) - v(x, t)\|_2^2] = \delta^2 \|b\|_2^2 J = \epsilon,$$

and $\text{KL}(p_1 \| q_1) = \epsilon/\delta = M$. This completes the proof. \square

Remark B.1 (Interpretation). The construction aligns a small velocity disparity $u - v$ with a large score gap $\nabla \log p_t - \nabla \log q_t = \nabla \psi_t$. The flow-matching loss controls only $\mathbb{E}_{x \sim p_t} \|u - v\|^2$, whereas the KL evolution depends on their *correlation* $(u - v) \cdot (\nabla \log p_t - \nabla \log q_t)$. Without an independent control on the score gap (e.g., its $L_2(p_t)$ norm), the KL divergence can increase

arbitrarily even when the flow-matching loss remains small.

B.4 Proof of Theorem 4.1

Theorem B.3 (Theorem 4.1 Restated: Distribution Estimation under Total Variation Distance). Let p_1 be the data distribution and q_1 be the flow matching estimated distribution. Let d be the feature dimension. Assume Assumption 4.1 holds. Then we have

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\text{TV}(p_1, q_1)] = O(n^{-\frac{1}{20d}} (\log n)^{5d_x}).$$

Proof. By exploiting the Pinsker's inequality, we bound the total variation distance as

$$\begin{aligned} \text{TV}(p_1, q_1) &\leq \sqrt{\frac{1}{2} \text{KL}(p_1, q_1)} && \text{(By Pinsker's inequality)} \\ &\leq \sqrt{\frac{1}{2} [A_1 \mathcal{R}(\hat{u}_\theta) + A_2 \mathcal{R}(\hat{u}_\theta)^2]}. && \text{(By Theorem 3.1 and let } t_0 = 0, T = 1) \end{aligned}$$

Taking expectation on training data $\{x_i\}_{i=1}^n$,

$$\begin{aligned} \mathbb{E}_{\{x_i\}_{i=1}^n} [\text{TV}(p_1, q_1)] &\lesssim \mathbb{E}_{\{x_i\}_{i=1}^n} [\sqrt{A_1 \mathcal{R}(\hat{u}_\theta) + A_2 \mathcal{R}(\hat{u}_\theta)^2}] \\ &\leq \mathbb{E}_{\{x_i\}_{i=1}^n} [\sqrt{A_1 \mathcal{R}(\hat{u}_\theta)} + \sqrt{A_2 \mathcal{R}(\hat{u}_\theta)^2}] && \text{(By } \sqrt{a+b} \leq \sqrt{a} + \sqrt{b} \text{ for } a, b \geq 0) \\ &\leq \sqrt{\mathbb{E}_{\{x_i\}_{i=1}^n} [A_1 \mathcal{R}(\hat{u}_\theta)]} + \mathbb{E}_{\{x_i\}_{i=1}^n} [\sqrt{A_2 \mathcal{R}(\hat{u}_\theta)}] && \text{(By Cauchy-Schwarz Inequality)} \\ &\lesssim O(n^{-\frac{1}{20d}} (\log n)^{5d_x}) + O(n^{-\frac{1}{10d}} (\log n)^{10d_x}) && \text{(By Lemma 4.1)} \\ &= O(n^{-\frac{1}{20d}} (\log n)^{5d_x}). \end{aligned}$$

This completes the proof. \square

B.5 Proof of Theorem 4.2

Theorem B.4 (Theorem 4.2 Restated: Nearly Minimax Optimality of Flow Matching Transformers). Let C , C_1 and C_2 be positive constants. Assume the data distribution satisfies $p_1(x) = \exp(-C_2 \|x\|_2^2 / 2) \cdot f(x)$, where f belongs to Hölder space $f(x) \in \mathcal{H}^\beta(\mathbb{R}^{d_x}, B)$ (Definition 4.1) and satisfies $C_1 \geq f(x) \geq C$ for all x . Then, within the Hölder distribution class under the Total Variation (TV) metric, the Flow Matching Transformer achieves the minimax-optimal convergence rate when $18d(\beta + 1) = d_x + 2\beta$.

Proof. Notice that the assumption we made in Theorem 4.2 can directly implies Assumption 4.1. Then by Theorem 4.1, we have the distribution estimation rate in Total Variation distance:

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\text{TV}(p_1, q_1)] = O(n^{-\frac{1}{20d}} (\log n)^{5d_x}).$$

Then, by [Lemma 4.2](#), the distribution rates matches the minimax lower bound up to a $\log n$ and Lipschitz constant factors under the setting

$$20d\beta w = d_x + 2\beta.$$

This completes the proof. □

References

- Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*, 2018.
- Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.
- Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.
- Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023.
- Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi 0$: A vision-language-action flow model for general robot control. corr, abs/2410.24164, 2024. doi: 10.48550. *arXiv preprint ARXIV.2410.24164*, 2024.
- Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.
- Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.
- Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.
- Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- 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.
- Kenji Fukumizu, Taiji Suzuki, Noboru Isobe, Kazusato Oko, and Masanori Koyama. Flow matching achieves almost minimax optimal convergence. *arXiv preprint arXiv:2405.20879*, 2024.

- Marta Gentiloni Silveri, Alain Durmus, and Giovanni Conforti. Theoretical guarantees in kl for diffusion flow matching. *Advances in Neural Information Processing Systems*, 37:138432–138473, 2024.
- Thomas Hakon Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919.
- László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.
- Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.
- Jerry Yao-Chieh Hu, Weimin Wu, Yi-Chen Lee, Yu-Chao Huang, Minshuo Chen, and Han Liu. On statistical rates of conditional diffusion transformers: Approximation, estimation and minimax optimality. *arXiv preprint arXiv:2411.17522*, 2024.
- Yuling Jiao, Yanming Lai, Yang Wang, and Bokai Yan. Convergence analysis of flow matching in latent space with transformers. *arXiv preprint arXiv:2404.02538*, 2024.
- Lea Kunkel and Mathias Trabs. On the minimax optimality of flow matching through the connection to kernel density estimation. *arXiv preprint arXiv:2504.13336*, 2025.
- Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034, 2023.
- Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. *The Eleventh International Conference on Learning Representations*, 2022.
- Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.
- Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

- Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *CoRR*, 2024.
- John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Maojiang Su, Jerry Yao-Chieh Hu, Yi-Chen Lee, Ning Zhu, Jui-Hui Chung, Shang Wu, Zhao Song, Minshuo Chen, and Han Liu. High-order flow matching: Unified framework and sharp statistical rates. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025a.
- Maojiang Su, Mingcheng Lu, Jerry Yao-Chieh Hu, Shang Wu, Zhao Song, Alex Reneau, and Han Liu. A theoretical analysis of discrete flow matching generative models. *arXiv preprint arXiv:2509.22623*, 2025b.
- Kunio Takezawa. *Introduction to nonparametric regression*. John Wiley & Sons, 2005.
- Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.
- Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.