Some Notes on the Hidden Biases of Flow Matching Samplers

Soon Hoe Lim[1,2]*

[1]Department of Mathematics, KTH Royal Institute of Technology
[2]Nordita, KTH Royal Institute of Technology and Stockholm University

December 3, 2025

Abstract

We study the implicit bias of empirical Flow Matching (FM) and Conditional Flow Matching (CFM) samplers. Although population FM may produce gradient-field velocities resembling optimal transport (OT), we show that the empirical FM minimizer is almost never a gradient field, even when each conditional flow is. Consequently, empirical FM is intrinsically energetically suboptimal and prone to memorization. We also analyze the kinetic energy of generated samples. With Gaussian sources, both instantaneous and integrated kinetic energies exhibit exponential concentration, while heavy-tailed sources lead to polynomial tails. These behaviors are governed primarily by the choice of source distribution rather than the data. Overall, these notes provide a concise mathematical account of the structural and energetic biases arising in empirical FM.

1 Introduction

The main goal of generative modeling is to use finitely many samples from a distribution to construct a sampling scheme capable of generating new samples from the same distribution. Among the families of existing generative models, flow matching (FM) [28, 29] is notable for its flexibility and simplicity. Given a target probability distribution, FM utilizes a parametric model (e.g., neural network) to learn the velocity vector field that defines a deterministic, continuous transformation (a normalizing flow) and transports a source probability distribution (e.g., standard Gaussian) to the target distribution.

While the population formulation of FM often exhibits appealing structure—sometimes even admitting gradient-field velocities—practical models are trained on finite datasets and therefore optimize empirical objectives. This empirical setting substantially alters the geometry of the learned velocity field and the energetic properties of the resulting sampler. These notes aim to clarify how empirical FM behaves, how it differs from its population counterpart, and what implicit biases arise in the learned sampling dynamics.

From now on, we assume that all the probability distributions/measures (except the empirical distribution) of the random variables considered are absolutely continuous (i.e., they have densities with respect to the Lebesgue measure), in which case we shall abuse the notation and use the same symbol to denote both the distribution and the density. To maintain the flow of the main text, we defer all proofs of the theoretical results to the Appendix.

*Corresponding author: `shlim@kth.se`.

2 Flow Matching (FM) and Conditional Flow Matching (CFM)

Let p_0 be the source distribution and p_1 the target distribution (e.g., the data distribution p^* or a smoothed version) on $\mathbb{R}^d$. We say that T is a transport map if $Z \sim p_0$ implies that $T(Z) \sim p_1$, in which case we write $T_\# p_0 = p_1$, and there exist many such maps. A common generative modeling paradigm aims to learn a transport map T from p_0 to p_1 on $\mathbb{R}^d$ using N i.i.d. samples $x^{(i)} \sim p_1$, where p_1 is typically unknown. One popular approach under this paradigm is flow matching (FM).

FM. The goal of FM is to find a velocity field $v : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, such that, if we solve the ODE:

$$\frac{dz(t)}{dt} = v(t, z(t)), \ z(0) = z_0 \in \mathbb{R}^d,$$

then the law of $z(1)$ when $z_0 \sim p_0$ is p_1 (in which case we say that v drives p_0 to p_1). The law of $z(t)$ for $t \in [0,1]$ is described by a probability path $p : [0,1] \times \mathbb{R}^d \to \mathbb{R}$, denoted $p_t(z)$, that evolves from p_0 at $t = 0$ to p_1 at $t = 1$. If we know v, then we can first sample $z_0 \sim p_0$ and then evolve the ODE from $t = 0$ to $t = 1$ to generate new samples.

The velocity field v generates the flow $\psi : [0,1] \times \mathbb{R}^d$ given as $\psi_t(z) = z(t)$, and the probability path via the push-forward distributions: $p_t = [\psi_t]_\# p_0$, i.e., $\psi_t(Z) \sim p_t$ for $Z \sim p_0$. In particular, $Z \sim p_0$ implies that $\psi_1(Z) \sim p_1$, i.e., ψ_t can be viewed as a dynamical transport map. The ODE corresponds to the Lagrangian description (the v-generated trajectories viewpoint), and a change of variables link it to the Eulerian description (the evolving probability path p_t viewpoint). Indeed, a necessary and sufficient condition for v to generate p_t is given by the continuity equation [2]:

$$\frac{\partial p_t}{\partial t} + \nabla \cdot (p_t v) = 0, \tag{1}$$

where $\nabla \cdot$ denotes the divergence operator. This equation ensures that the flow defined by v conserves the mass (or probability) described by p_t. In general, even for p_t that linear interpolates between p_0 and p_1, the velocity field does not admit a closed-form expression when p_0 and p_1 are known, except in special cases such as Gaussians, mixture of Gaussians and uniform distributions [34].

The above description gives us a population FM model, which we aim to learn using a finite number of samples in practice. Given such a v, it is standard to learn it with a parametric model v_θ (e.g., neural network) by minimizing the FM objective:

$$L_{\mathrm{FM}}[v_\theta] = \mathbb{E}_{t \sim \mathcal{U}[0,1],\ Z_t \sim p_t}[\|v_\theta(t, Z_t) - v(t, Z_t)\|^2]. \tag{2}$$

CFM. In CFM [28, 41], we consider a probability path in the mixture form:

$$p_t(z) = \int p_t(z|x) p_1(x) dx, \tag{3}$$

where $p_t(\cdot|x) : \mathbb{R}^d \to \mathbb{R}^+$ is a conditional probability path generated by some vector field $v(t, \cdot|x) : \mathbb{R}^d \to \mathbb{R}^d$ for $x \in \mathbb{R}^d$. Moreover, consider the vector field:

$$v(t, z) = \int v(t, z|x) \frac{p_t(z|x) p_1(x)}{p_t(z)} dx. \tag{4}$$

In this setting, it can be shown in [28] that minimizing the FM objective L_{FM} is equivalent to minimizing the CFM objective:

$$L_{\mathrm{CFM}}[v_\theta] = \mathbb{E}_{t \sim \mathcal{U}[0,1],\ X \sim p_1,\ Z_t \sim p_t(\cdot|X)}[\|v_\theta(t, Z_t) - v(t, Z_t|X)\|^2]. \tag{5}$$

In order to apply CFM, we need to specify the boundary distributions p_0 and p_1, and the conditional probability path $p_t(z|x)$. Below are some examples.

Example 1 (Rectified Flow)**.** *A canonical choice [31] is* $p_0 = \mathcal{N}(0, I_d)$*,* $p_1 = p^*$*, and*

$$p_t(z|X = x_1) = \mathcal{N}(z; tx_1, (1-t)^2 I_d), \tag{6}$$

which corresponds to the conditional velocity field $v(t, z|X = x_1) = \frac{x_1 - z}{1-t}$*. This conditional probability path realizes linear interpolating paths of the form* $Z_t = (1-t)x_0 + tx_1$ *between a (reference) Gaussian sample* x_0 *and a data sample* x_1*. In practice, regularized versions of rectified flow are preferred for numerical stability (since* v *blows up as* $t \to 1$*). A simple version is to modify the conditional probability path to* $p_t(\cdot|X = x_1) = \mathcal{N}(tx_1, (1-(1-\sigma_{min})t)^2 I_d)$ *for some small* $\sigma_{min} > 0$*, which corresponds to the regularized conditional velocity field* $v(t, z|X = x_1) = \frac{x_1 - (1-\sigma_{min})z}{1-(1-\sigma_{min})t}$*. Another version is to consider a smoothed version of the data distribution* p^**; e.g.,* $p_1 = p^* \star \mathcal{N}(0, \sigma_{min}^2 I_d)$*, where* $\star$ *denotes convolution.*

Example 2 (Affine Flows)**.** *More generally, consider a latent variable* $Z \sim \mathbb{Q}$ *with probability density function (PDF)* K *(not necessarily Gaussian) and, for* $t \in [0, 1]$*, the affine conditional flow defined by* $\psi_t(Z|X) = m_t(X) + \sigma_t(X)Z$ *for some time-differentiable functions* $m : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ *and* $\sigma : [0,1] \times \mathbb{R}^d \to \mathbb{R}^+$*. Since* ψ_t *is linear in* Z*, we can obtain its density via the change of variables:*

$$p_t(z|X) = \frac{1}{\sigma_t^d(X)} K\left(\frac{z - m_t(X)}{\sigma_t(X)}\right). \tag{7}$$

Then, as in Theorem 3 in [28], we can show that the unique vector field that defines $\psi_t(\cdot|X)$ *via the ODE* $\frac{d}{dt}\psi_t(z|X) = v(t, \psi_t(z|X)|X)$ *has the form:*

$$v(t, z|X) = a_t(X)z + b_t(X), \tag{8}$$

where

$$a_t(X) = \frac{\frac{\partial \sigma_t}{\partial t}(X)}{\sigma_t(X)}, \quad b_t(X) = \frac{\partial m_t}{\partial t}(X) - m_t(X)a_t(X). \tag{9}$$

The rectified flow in previous example is a special case of this family of conditional flows (with $K = \mathcal{N}(0, I_d)$*,* $m_t(X) = tX$ *and* $\sigma_t(X) = 1 - t$*). The Gaussian flows considered in [28, 41, 1] are also special cases.*

All the formulations thus far are in the idealized continuous-time setting. In practice, we work with Monte Carlo estimates of the objective and use the optimized v_θ to generate new samples by simulating the ODE with a numerical scheme. Note, however, that the training of CFM is simulation-free: the dynamics are only simulated at inference time and not when training the parametric model. In practice, affine flows are most widely used, and thus we will focus on them here, using the rectified flow model as a canonical example.

3 Empirical Flow Matching

Suppose that we are given a p_0 and N i.i.d. samples $x^{(i)} \sim p_1$, i.e., we only have access to p_1 via a finite number of i.i.d. samples. The target distribution p_1 needed to compute L_{FM} and L_{CFM} can be approximated by the empirical distribution $\hat{p}_1 := \frac{1}{N}\sum_{i=1}^N \delta_{x^{(i)}}$. We shall call the FM and CFM for the case when $p_1 = \hat{p}_1$ the empirical FM and empirical CFM respectively. The empirical counterparts of $p_t(z)$ and $v(t, z)$ are given by:

$$\hat{p}_t(z) = \frac{1}{N}\sum_{i=1}^N p_t(z|x^{(i)}), \tag{10}$$

$$\hat{v}(t, z) = \sum_{i=1}^N v(t, z|x^{(i)}) \frac{p_t(z|x^{(i)})}{\sum_{j=1}^N p_t(z|x^{(j)})} \tag{11}$$

respectively. The objectives that the empirical FM and empirical CFM minimize are then given by, respectively:

$$\hat{L}_{\mathrm{FM}}[v'] = \mathbb{E}_{t\sim\mathcal{U}[0,1],\ Z_t\sim\hat{p}_t}[\|v'(t,Z_t) - \hat{v}(t,Z_t)\|^2], \tag{12}$$

$$\begin{aligned}\hat{L}_{\mathrm{CFM}}[v'] &= \mathbb{E}_{t\sim\mathcal{U}[0,1],\ X\sim\hat{p}_1,\ Z_t\sim p_t(\cdot|X)}[\|v'(t,Z_t) - v(t,Z_t|X)\|^2] \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{t\sim\mathcal{U}[0,1],\ Z_t\sim p_t(\cdot|x^{(i)})}[\|v'(t,Z_t) - v(t,Z_t|x^{(i)})\|^2],\end{aligned} \tag{13}$$

where $p_t(\cdot|x^{(i)})$ is the conditional probability path (given by, e.g., (7) or (6)).

One can show that if $v(t,\cdot|x^{(i)})$ generates $p_t(\cdot|x^{(i)})$ for all $i \in [N]$, then $\hat{v}(t,\cdot)$ generates $\hat{p}_t$ (see Lemma 2.1 in [24]). Just as before, the equivalence (with respect to the optimizing arguments) between FM and CFM carries over to empirical FM and empirical CFM naturally (see Theorem 2.2 in [24]). Moreover, for the examples of conditional probability path considered earlier, we can derive a closed-form expression for the minimizer $\hat{v}^* \in \mathrm{argmin}_v \hat{L}_{CFM}[v] = \mathrm{argmin}_v \hat{L}_{FM}[v]$, giving us a training-free model for generating new samples. This sampler is described by the ODE:

$$\frac{d\hat{z}^*(t)}{dt} = \hat{v}^*(t,\hat{z}^*(t)), \quad \hat{z}^*(0) \sim p_0, \tag{14}$$

which we evolve to $t = 1$ to obtain new samples.

Example 3 (Empirical Rectified Flow)**.** *For the rectified flow example in Example 1, the minimizer $\hat{v}^*$ has a closed-form formula [7]:*

$$\hat{v}^*(t,z) = \sum_{i=1}^{N} w_i(t,z)\frac{x^{(i)} - z}{1-t}, \tag{15}$$

where $w_i(t,z) = softmax_i\left(\left(-\frac{1}{2(1-t)^2}\|z - tx^{(j)}\|^2\|\right)_{j\in[N]}\right)$, with $softmax_i$ denoting the ith component of the vector obtained after applying the softmax operation. This optimal velocity field is thus a time-dependent weighted average of the N different directions towards the $x^{(i)}$. Similar formula can also be obtained for regularized versions of rectified flow.

Example 4 (Empirical Affine Flows)**.** *To construct a flow from $p_0(z) = K(z)$ to $\tilde{p}_1(z) = \frac{1}{N\sigma_{min}^d}\sum_{i=1}^{N} K\left(\frac{z-x^{(i)}}{\sigma_{min}}\right)$ (a smoothed version of $\hat{p}_1$), we can choose any m_t and σ_t such that $m_0(X) = 0$, $m_1(X) = X$, $\sigma_0(X) = 1$, $\sigma_1(X) = \sigma_{min}$. The final marginal distribution, obtained by averaging the evolved conditional distribution $p_1(z|X = x^{(i)})$ across the empirical distribution $\hat{p}_1$, coincides with the (Nadaraya-Watson) kernel density estimator (KDE) $\tilde{p}_1$. Heuristically, if K is the standard Gaussian PDF, then we recover the rectified flow model as $\sigma_{min} \to 0$.*

Moreover, similar to the empirical rectified flow, we can obtain the following result for the empirical affine flows.

Proposition 1. *For the family of affine flows, the minimizer $\hat{v}^*$ of the empirical FM objective admits a closed-form formula:*

$$\hat{v}^*(t,z) = \sum_{i=1}^{N} w_i(t,z)\cdot(a_t(x^{(i)})z + b_t(x^{(i)})), \tag{16}$$

where a_t and b_t are given in (9), and $w_i(t,z)$ is a kernel-dependent weighting function given by:

$$w_i(t,z) = \frac{p_t(z|x^{(i)})}{\sum_{j=1}^{N} p_t(z|x^{(j)})} \tag{17}$$

with

$$p_t(z|x^{(i)}) = \frac{1}{\sigma_t^d(x^{(i)})} K\left(\frac{z - m_t(x^{(i)})}{\sigma_t(x^{(i)})}\right). \tag{18}$$

Intuitively, $\hat{v}^*$ is a convex combination of the individual conditional velocity fields $v(t, z|x^{(i)})$, weighted by $w_i(t, z)$ which tells us how likely the observed point z at time t is to belong to the flow path originating from the sample $x^{(i)}$.

4 FM Through the Lens of Energetics

In this section, we study the optimal empirical FM model defined by the $\hat{v}^*$ given in (15) and Proposition 1, and the associated energetics (such as kinetic energy). First, we need to introduce optimal transport, its dynamical representation, and the Wasserstein distance.

Optimal Transport (OT). OT is the problem of efficiently moving probability mass from a source distribution p_0 to a target distribution p_1 such that a given cost function has minimal expected value. More precisely, we aim to find a coupling (Z_0, Z_1) of random variables $Z_0 \sim p_0$ and $Z_1 \sim p_1$ such that the expected cost $\mathbb{E}[c(Z_0, Z_1)]$ is minimal, where c is a cost function, typically chosen as $c_1(z_0, z_1) = \|z_0 - z_1\|$ or $c_2(z_0, z_1) = \|z_0 - z_1\|^2$.

The Monge map (or OT map) T_0 is the transport map that minimizes $\mathbb{E}_{p_0}[c_2(Z_0, T(Z_0))]$. The squared 2-Wasserstein distance $W_2^2(p_0, p_1)$ is defined by the minimum expected squared distance over all couplings:

$$W_2^2(p_0, p_1) := \inf_{\gamma \in \Pi(p_0, p_1)} \mathbb{E}_{(Z_0, Z_1) \sim \gamma}[\|Z_0 - Z_1\|^2] = \inf_{\gamma \in \Pi(p_0, p_1)} \int \|x - y\|^2 d\gamma(x, y),$$

where $\Pi(p_0, p_1)$ is the set of all joint probability distributions with marginals p_0 and p_1. For the squared cost, this minimum is achieved by the Monge map T_0, such that $W_2^2(p_0, p_1) = \mathbb{E}_{Z_0 \sim p_0}[\|Z_0 - T_0(Z_0)\|^2]$. The Wasserstein distance W_2 defines a metric on $\mathcal{P}_2(\mathbb{R}^d)$, the space of probability measures on $\mathbb{R}^d$ with finite second moment.

Let $\mathcal{T}(p_0, p_1) := \{T : \mathbb{R}^d \to \mathbb{R}^d : T_{\#} p_0 = p_1\}$. The following is a key result in OT theory due to Brenier (see, e.g., Chapter 3 in [42], [33]): there exists a unique (up to a p_0-negligible set) minimizer T_0 to the Monge problem:

$$d(p_0, p_1)^2 := \inf_{T \in \mathcal{T}(p_0, p_1)} \int \|x - T(x)\|^2 dp_0(x)$$

such that $d(p_0, p_1)^2 = W_2^2(p_0, p_1)$. Moreover, T_0 can be represented (p_0-almost everywhere) as $T_0 = \nabla \Phi$ for some convex function $\Phi : \mathbb{R}^d \to \mathbb{R}$ (this T_0 is the optimal transport map).

Dynamical Representation (Benamou-Brenier Formulation). Just like any transport map, OT map can be expressed in a dynamic form as a continuous flow from the source distribution p_0 to the target distribution p_1 [6]. Consider a flow $\psi_t(z)$ defined by the ODE:

$$\frac{\partial}{\partial t} \psi_t(z) = v(t, \psi_t(z)), \quad \text{for all } t \in [0, 1],$$

for a velocity field $v(t, z)$, with the initial condition $\psi_0(z) = z$. The flow ψ_t induces a probability path, $p_t = [\psi_t]_{\#} p_0$, in the Wasserstein space.

Let $\mathcal{U}$ be the collection of all velocity fields v such that the flow $\psi_t(z)$ is uniquely defined and transports p_0 to p_1 over the unit time interval. The OT map $T_0(z)$ is given by the end-point of the optimal flow: $T_0(z) = \psi_1^{\text{OT}}(z)$, where the associated optimal velocity field $v^{\text{OT}}(\cdot, \cdot)$ is the minimizer of the expected kinetic energy[1]:

$$\mathbb{E}\left[\int_0^1 \|v(t, \psi_t(Z_0))\|^2 dt\right]$$

over all $v \in \mathcal{U}$. This minimal expected energy is equal to the squared 2-Wasserstein distance $W_2^2(p_0, p_1)$. Importantly, the W_2 optimal velocity field v^{OT} must be irrotational (curl-free), meaning that $v^{\text{OT}}(t, z) = -\nabla_z \Phi(t, z)$ for some scalar potential Φ (otherwise, the curl component would introduce unnecessary looping or rotational motion, which would increase the total cost); see also Theorem 8.3.1 in [3].

If p_t denotes the density of the distribution at time t (i.e., the law of $\psi_t(Z_0)$), the optimal solution must satisfy the continuity equation (which ensures mass conservation):

$$\partial_t p_t + \nabla \cdot (v^{\text{OT}} p_t) = 0.$$

Hence, the optimization problem (Benamou-Brenier formulation) can be written in its Eulerian form, and minimizes the total kinetic energy:

$$\inf_{v,p} \int_0^1 \int_{\mathbb{R}^d} \frac{1}{2} \|v(t, z)\|^2 p_t(z)\, dz\, dt$$
$$\text{subject to} \quad \partial_t p_t + \nabla \cdot (v_t p_t) = 0,$$

with the boundary conditions p_0 (at $t = 0$) and p_1 (at $t = 1$).

Empirical Continuity Equation. Now, the empirical counterpart of the continuity equation (1) is:

$$\frac{\partial \hat{p}_t^*}{\partial t} + \nabla \cdot (\hat{v}^* \hat{p}_t^*) = 0. \tag{19}$$

Remark 1. *By the Helmholtz-Hodge decomposition, any sufficiently smooth vector field in $\mathbb{R}^d$ (decaying sufficiently fast at infinity) can be uniquely decomposed into the sum of a gradient field (irrotational) and a divergence-free field (solenoidal) [8]:*

$$\hat{v}^*(t, z) = \underbrace{-\nabla \Phi(t, z)}_{\text{gradient field}} + \underbrace{u(t, z)}_{\text{divergence-free}},$$

for some scalar potential function Φ and some u such that $\nabla \cdot u = 0$. This is analogous to the fact that any matrix can be uniquely decomposed into a symmetric and an antisymmetric part. In the context of OT, the optimality condition requires the velocity to be a gradient field ($u = 0$). The presence of a non-zero divergence-free component u implies that the mass transport includes rotational or "looping" movements that increase the kinetic energy cost without contributing to the net displacement of mass.

It is natural to ask if the $\hat{v}^*$ (the velocity field that a trainable CFM model is really optimizing for) in (15) and Proposition 1 corresponds to an optimal velocity field in the OT sense.

In fact, except for special cases, even the velocity fields v_t arising from the population FM framework are generally not gradient functions [44, 30], thus not optimal in the OT sense. Indeed, OT paths are generally outside the class of probability paths with affine conditionals. Since affine conditionals are of particular interest due to the fact that they enable scalable training, [39] studied the kinetic optimal path within this class of paths using a proxy for the kinetic energy.

The following example gives a special case in which we have velocity fields which can be represented as gradient fields. We will look at the empirical case later.

[1]This is also, up to a multiplicative constant involving d, the kinetic energy considered in [39].

Example 5 (Optimal Velocity of Rectified Flow Can Be a Gradient Field)**.** *If the joint distribution of the source and target is a product distribution, i.e., $p_{0,1} = p_0 \times p_1$ (independent coupling), then for the interpolating path of the rectified flow $Z_t = (1-t)x_0 + tx_1$, $x_0 \sim \mathcal{N}(0, I_d)$, and $p_1 \in \mathcal{P}_2(\mathbb{R}^d)$, the optimal velocity can be shown to be the conditional expectation [44, 51]:*

$$v(t,z) = \mathbb{E}_{x_0 \sim p_0,\ x_1 \sim p_1}[x_1 - x_0 | Z_t = z] = -\nabla_z \Phi(t,z), \tag{20}$$

where

$$\Phi(t,z) = -\frac{1}{2t}\|z\|^2 - \frac{1-t}{t}\log p_t(z). \tag{21}$$

We also see that the optimal score function is related to the velocity by: $\nabla_z \log p_t(z) = \frac{t}{1-t}v(t,z) - \frac{1}{1-t}z$. Analogous formula can also be derived had we consider a slightly more general flow with $Z_t = \alpha_t x_0 + \beta_t x_1$ for some time-differentiable α_t, β_t such that $\alpha_0 = \beta_1 = 0$ and $\alpha_1 = \beta_0 = 1$. This tells us that the rectified flow's optimal velocity, under the independent coupling, is a gradient field (but does not generally give us an OT map due to the independent coupling assumption; being a gradient field is necessary but not sufficient for OT).

Let us consider Gaussian distributions for p_0 and p_1, in which case the OT map can be computed explicitly [15].

Example 6 (Explicit Examples [34])**.** *Take $p_0 \sim \mathcal{N}(0, \Sigma_0)$, $p_1 \sim \mathcal{N}(m_1, \Sigma_1)$ and consider the rectified flow (RF) map, denoted $R(x) := x + \int_0^1 v(t, \psi_t(x))dt$ with $v = \dot{\psi}_t$, where $\psi_t(x) = (1-t)x + tR(x)$ is the displacement interpolation between the independent Gaussians $X_0 \sim p_0$ and $X_1 \sim p_1$. If $\Sigma_0 = I_d$, then the Monge's OT map and the RF map between X_0 and X_1 coincide: $T_0(x) = m_1 + \Sigma_1^{1/2}x = R(x)$. However, if $\Sigma_0 \neq I_d$, then the two maps are not equivalent. In practice, p_0 is often chosen to be $\mathcal{N}(0, I_d)$ and so the RF map is the OT map as well.*

Optimal Velocity of the Empirical Flows Generally Fails to be a Gradient Field. A crucial observation is that even if the optimal velocity is a gradient field, the empirical version is generally not a gradient field. This is the main message of the following proposition.

Proposition 2. *Let the empirical target distribution be $\hat{p}_1 = \frac{1}{N}\sum_{i=1}^N \delta_{x^{(i)}}$. Consider the family of empirical affine flows defined by the conditional probability paths $p_t(z|x^{(i)})$ and their corresponding conditional velocity fields $v_i(t,z) := v(t,z|x^{(i)}) = a_t(x^{(i)})z + b_t(x^{(i)})$ from Proposition 1. The minimizer $\hat{v}^*(t,z)$ of the empirical FM/CFM objective is not a gradient field if and only if the following data-dependent condition holds for all $t \in [0,1]$:*

$$\sum_{i=1}^N \left(v_i(t,z)\nabla_z w_i(t,z)^\top - \nabla_z w_i(t,z) v_i(t,z)^\top \right) = 0.$$

In general, this sum does not vanish as this is not true for generic datasets, implying that $\hat{v}^*$ is generally not a gradient field and thus not the true OT solution. Intuitively, this says that even if every individual conditional flow is a straight line (gradient field), their weighted sum is not generally a gradient field because the weights $w_i(t,z)$ vary spatially (dependent on z).

An important consequence of Proposition 2, together with Proposition 1, is that the velocity field (even if it is originally formulated so that the idealized model is a gradient field) that a trainable CFM model is optimizing for is generally not a gradient field, and thus does not learn the OT solution. Combining these results with a recent line of studies on memorization in empirical FM [7], we obtain the following interpretation:

"Neural network trained CFM models are implicitly optimizing for a sampler that is not only energetically inefficient (not minimizing the kinetic energy), but also leads to memorization (producing samples that are close enough to those from the training set)."

Quantifying the difference in the generative path generated by the idealized gradient field model of the population FM and the closed-form velocity field model of the empirical FM is a natural direction to consider in order to understand how likely samples with high energy are generated and the generation paths that lead to them.

First, we focus on the Gaussian rectified flow (RF) example in Example 6, which is tractable enough to allow for precise analysis. The following result shows that the probability of a generated sample under the population RF model that has high kinetic energy decays exponentially. Since this is the OT map and velocity is constant along straight paths, this bound applies simultaneously to the instantaneous kinetic energy at any time t and the integrated total energy.

Proposition 3 (Population setting, OT case)**.** *Let $p_0 = \mathcal{N}(0, I_d)$ and $p_1 = \mathcal{N}(m_1, \Sigma_1)$, where Σ_1 is positive definite. Let $R(x) = m_1 + \Sigma_1^{1/2} x$ be the Rectified Flow map from Example 6. For a generated sample $Y \sim p_1$, let $E(Y) = \int_0^1 \|v(t, R^{-1}(Y))\|^2 dt = \|Y - R^{-1}(Y)\|^2$ be the random variable representing the kinetic energy (integrated or instantaneous).*

(a) For all $y \in \mathbb{R}^d$, $\frac{1}{2}E(y) = -\log p_1(y) + C(y)$, where

$$C(y) = \frac{1}{2} y^T (I_d - 2\Sigma_1^{-1/2}) y + m_1^T \Sigma_1^{-1/2} y - \frac{1}{2} \log \det(2\pi \Sigma_1). \tag{22}$$

(b) Assume $\Sigma_1 \neq I_d$. Let $\lambda_i(\Sigma_1)$ denote the eigenvalues of Σ_1, and define

$$\rho := \max_{i=1,\dots,d} \left(\sqrt{\lambda_i(\Sigma_1)} - 1 \right)^2 > 0.$$

Then, for any energy threshold $u > 0$, we have the concentration bound:

$$\mathbb{P}_{Y \sim p_1} \big(E(Y) \geq u \big) \leq C \cdot \exp \left(-\frac{u}{4\rho} \right),$$

where the constant C is given explicitly by $C = 2^{d/2} \exp\left(\frac{\|m_1\|^2}{2\rho}\right)$.

The result in (a) above connects the kinetic energy to negative log density point-wise, saying that the kinetic energy is essentially the negative log density plus a correction term (which is not necessarily positive for all $y \in \mathbb{R}^d$). Meanwhile, the exponential concentration of kinetic energy in (b) implies that samples with high kinetic energy are predominantly found in the low-density tail regions of the target distribution p_1 (i.e., random sample Y drawn from p_1 is exponentially unlikely to have required high energy). Importantly, this phenomenon arises purely from the design of the Gaussian RF model itself and the assumption that p_1 is Gaussian. Similar bound also holds for the case when p_1 is sub-Gaussian.

Remark 2. *We could also possibly extend the analysis for the mixture of Gaussian case at the cost of more complicated statements and blurring the key message. It could also be interesting to look at the cases where p_1 has heavy tails, in which case we expect the decay to be slower than exponential.*

It turns out that we can also obtain a similar bound for the empirical RF model $\hat{v}^*(t, z)$ (which is nonlinear in z) under the same assumption for p_1, despite the fact that it does not give rise to an OT map.

Theorem 1 (Empirical setting, Gaussian source)**.** *Let $X_0 \sim \mathcal{N}(0, I_d)$ and suppose that we are given a fixed dataset $\mathcal{D}_N = \{x^{(i)}\}_{i \in [N]}$, $x^{(i)} \in \mathbb{R}^d$, with $M := \max_i \|x^{(i)}\| < \infty$. Let $T \in [0, 1)$ and define the instantaneous kinetic energy, $K_t = \|\hat{v}^*(t, \psi_t(X_0))\|^2$, and the corresponding time-integrated kinetic energy, $E_T = \int_0^T K_t \, dt$, where $\hat{v}^*$ is given in (15) and ψ_t solves $\dot{\psi}_t(X) = \hat{v}^*(t, \psi_t(X))$, $\psi_0(X) = X_0$, for $t \in [0, 1)$. Assume that there exists a unique solution to the latter ODE on $[0, T]$.*

(a) For each $t \in [0, T]$, there exist constants $C > 0$ (depending only on M) and $c_t > 0$ (depending only on t) such that for all sufficiently large energy thresholds U_t:

$$\mathbb{P}(K_t \geq U_t \mid \mathcal{D}_N) \leq C\, e^{-c_t U_t}.$$

(b) There exist constants $C > 0$ (depending only on M) and $c_T > 0$ (depending only on t) such that for all sufficiently large energy thresholds U_T:

$$\mathbb{P}(E_T \geq U_T \mid \mathcal{D}_N) \leq C\, e^{-c_T U_T}.$$

Theorem 1 implies that, just as in the population case, both instantaneous and integrated empirical kinetic energy have exponential tails in the energy level. Note that such result arises due to the property of the Gaussian distribution, and holds regardless of whether the velocity fields give an OT map.

The above results do not assume specific distribution from which the $x^{(i)}$ (treated as fixed) are sampled from, giving conditional bounds for K_t and E_T (the probability is over only the random draw of $X_0 \sim \mathcal{N}(0, I_d)$). Interestingly, this implies that even if the underlying samples $x^{(i)}$ come from a heavy-tailed distribution (in which case we expect the decay to be polynomial and thus slower than exponential decay in the population setting), the empirical model still exhibits exponential decay of kinetic energy conditional on the realized dataset. In practice, empirical FM/CFM models therefore always display exponential concentration of kinetic energy (regardless of whether the true data distribution is heavy-tailed) because all observable sampling randomness is conditioned on the fixed finite dataset used during training. The exponential concentration comes purely from the Gaussian source distribution.

From these results, we obtain the following interpretation:

> *"Neural network trained CFM models are implicitly optimizing for a sampler that generates samples with exponentially concentrated kinetic energy, where high-energy samples are predominantly found in low-density tail regions of the target distribution."*

Combining these two interpretations, we see that neural network trained CFM models are implicitly optimizing for a sampler that is suboptimal in expected kinetic energy (using more energy on average than the OT solution due to lack of gradient structure), and also leads to memorization. However, regardless of optimality, both OT and CFM models generate samples with exponentially concentrated kinetic energy around typical values, with high-energy samples (which are rare) predominantly found in low-density tail regions.

Going back to our earlier observation that, for the standard Gaussian p_0, even if the underlying samples $x^{(i)}$ come from a heavy-tailed distribution (in which case we expect the decay to be polynomial and thus slower than exponential decay in the population setting), the empirical model still exhibits exponential decay of kinetic energy conditional on the realized dataset. Therefore, simply adding heavy-tailed noise to data samples would not achieve the polynomial decay. In order to achieve polynomial decay, one solution is to work with a heavy-tailed p_0 instead.

Indeed, while Theorem 1 establishes exponential concentration due to the Gaussian source, the empirical framework allows for heavy-tailed modeling if consider instead a smoothed model from Example 4 and choosing the source kernel K to be heavy-tailed. Specifically, if $X_0 \sim K$ satisfies $\mathbb{P}(\|X_0\| > s) \propto s^{-\alpha}$ (e.g., Empirical Affine Flow with a Student-t source), the linear growth of the vector field $\hat{v}^*$ preserves this tail index. Consequently, the kinetic energy decays polynomially, which is the main message of the following theorem.

Theorem 2 (Empirical setting, heavy-tailed source)**.** *Let $D_N = \{x^{(i)}\}_{i\in[N]}$ be a fixed dataset with $M := \max_{i\in[N]} \|x^{(i)}\| < \infty$. Let $T \in [0, 1)$. Suppose the source distribution $p_0(z) = K(z)$ is heavy-tailed, in the sense that the initial norm $\|X_0\|$ satisfies*

$$\mathbb{P}(\|X_0\| \geq s) \;\leq\; \frac{C_\alpha}{s^\alpha} \quad \textit{for all } s \geq 1,$$

for some constants $C_\alpha > 0$ and tail index $\alpha > 0$.

For the velocity field $\hat{v}^*$ defined in Proposition 1, let

$$A_{\max} := \sup_{t\in[0,T],\, i\in[N]} |a_t(x^{(i)})|, \qquad B_{\max} := \sup_{t\in[0,T],\, i\in[N]} \|b_t(x^{(i)})\|,$$

and assume that there exists a unique solution to the ODE driven by $\hat{v}^*$ on $[0, T]$. Then there exist constants $C_{\text{poly}} > 0$ and $\gamma = \alpha/2$ such that, for all sufficiently large thresholds U_t, U_T,

$$\mathbb{P}(K_t \geq U_t \mid D_N) \;\leq\; \frac{C_{\text{poly}}}{U_t^\gamma}, \qquad \mathbb{P}(E_T \geq U_T \mid D_N) \;\leq\; \frac{C_{\text{poly}}}{U_T^\gamma}.$$

Moreover, C_{poly} depends only on $T, A_{\max}, B_{\max}, C_\alpha$.

These results demonstrate that the tail behavior of the generated energy profile by the empirical model is strictly controlled by the choice of the source distribution p_0.

5 Conclusion

In these notes, we show that empirical FM optimizes for velocity fields that generally lack a gradient structure and therefore cannot reproduce OT maps or minimize kinetic energy. This explains both the energetic inefficiency and the memorization tendencies observed in practice. Despite this, the generated kinetic energy concentrates sharply: exponentially for Gaussian sources and polynomially for heavy-tailed ones. This shows that the tail behavior of FM samplers is governed mainly by the source distribution, not the underlying data. Understanding these structural biases clarifies the limitations of empirical FM and may guide improved sampler designs, which we leave for future work.

Acknowledgment. SHL would like to acknowledge support from the Wallenberg Initiative on Networks and Quantum Information (WINQ) and the Swedish Research Council (VR/2021-03648).

References

[1] 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.

[2] Michael S Albergo and Eric Vanden-Eijnden. Learning to sample better. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104014, 2024.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures.* Springer, 2005.

[4] Jacob Bamberger, Iolo Jones, Dennis Duncan, Michael M Bronstein, Pierre Vandergheynst, and Adam Gosztolai. Carr\'e du champ flow matching: better quality-generalisation tradeoff in generative models. *arXiv preprint arXiv:2510.05930*, 2025.

[5] Ricardo Baptista, Agnimitra Dasgupta, Nikola B Kovachki, Assad Oberai, and Andrew M Stuart. Memorization and regularization in generative diffusion models. *arXiv preprint arXiv:2501.15785*, 2025.

[6] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[7] Quentin Bertrand, Anne Gagneux, Mathurin Massias, and Rémi Emonet. On the closed-form of flow matching: Generalization does not arise from target stochasticity. *arXiv preprint arXiv:2506.03719*, 2025.

[8] Harsh Bhatia, Gregory Norgard, Valerio Pascucci, and Peer-Timo Bremer. The Helmholtz-Hodge decomposition—a survey. *IEEE Transactions on visualization and computer graphics*, 19(8):1386–1404, 2012.

[9] Zachary Charles and Keith Rush. Iterated vector fields and conservatism, with applications to federated learning. In *International Conference on Algorithmic Learning Theory*, pages 130–147. PMLR, 2022.

[10] Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, and Andrew M Stuart. Sampling via gradient flows in the space of probability measures. *arXiv preprint arXiv:2310.03597*, 2023.

[11] Yifan Chen, Eric Vanden-Eijnden, and Jiawei Xu. Lipschitz-guided design of interpolation schedules in generative models. *arXiv preprint arXiv:2509.01629*, 2025.

[12] Yongxin Chen, Tryphon Georgiou, and Michele Pavon. Gradient flows as optimal controlled evolutions: From rn to wasserstein product spaces. *arXiv preprint arXiv:2510.27120*, 2025.

[13] Zhengdao Chen. On the interpolation effect of score smoothing. *arXiv preprint arXiv:2502.19499*, 2025.

[14] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 3, 2024.

[15] Julie Delon and Agnes Desolneux. A wasserstein-type distance in the space of gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

[16] Ruiqi Feng, Chenglei Yu, Wenhao Deng, Peiyan Hu, and Tailin Wu. On the guidance of flow matching. *arXiv preprint arXiv:2502.02150*, 2025.

[17] Anne Gagneux, Ségolène Martin, Rémi Gribonval, and Mathurin Massias. The generation phases of flow matching: a denoising perspective. *arXiv preprint arXiv:2510.24830*, 2025.

[18] Weiguo Gao and Ming Li. How do flow matching models memorize and generalize in sample data subspaces? *arXiv preprint arXiv:2410.23594*, 2024.

[19] Johannes Hertrich, Antonin Chambolle, and Julie Delon. On the relation between rectified flows and optimal transport. *arXiv preprint arXiv:2505.19712*, 2025.

[20] Christian Horvat and Jean-Pascal Pfister. On Gauge freedom, conservativity and intrinsic dimensionality estimation in diffusion models. *arXiv preprint arXiv:2402.03845*, 2024.

[21] Samuel Hurault, Matthieu Terris, Thomas Moreau, and Gabriel Peyré. From score matching to diffusion: A fine-grained error analysis in the gaussian setting. *arXiv preprint arXiv:2503.11615*, 2025.

[22] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[23] Lea Kunkel. Distribution estimation via flow matching with Lipschitz guarantees. *arXiv preprint arXiv:2509.02337*, 2025.

[24] 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.

[25] Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via Wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35:14434–14447, 2022.

[26] Sixu Li, Shi Chen, and Qin Li. A good score does not lead to a good generative model. *arXiv preprint arXiv:2401.04856*, 2024.

[27] Yunchen Li, Shaohui Lin, and Zhou Yu. Generation properties of stochastic interpolation under finite training set. *arXiv preprint arXiv:2509.21925*, 2025.

[28] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[29] 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.

[30] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

[31] 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.

[32] Yang Lyu, Tan Minh Nguyen, Yuchun Qian, and Xin T Tong. Resolving memorization in empirical diffusion model for manifold data in high-dimensional spaces. *arXiv preprint arXiv:2505.02508*, 2025.

[33] Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966–998, 2024.

[34] Gonzalo Mena, Arun Kumar Kuchibhotla, and Larry Wasserman. Statistical properties of rectified flow. *arXiv preprint arXiv:2511.03193*, 2025.

[35] Jakiw Pidstrigach. Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, 35:35852–35865, 2022.

[36] Teodora Reu, Sixtine Dromigny, Michael Bronstein, and Francisco Vargas. Gradient variance reveals failure modes in flow-based generative models. *arXiv preprint arXiv:2510.18118*, 2025.

[37] Christopher Scarvelis, Haitz Sáez de Ocáriz Borde, and Justin Solomon. Closed-form diffusion models. *arXiv preprint arXiv:2310.12395*, 2023.

[38] Louis Sharrock and Christopher Nemeth. Tuning-free sampling via optimization on the space of probability measures. *arXiv preprint arXiv:2510.25315*, 2025.

[39] Neta Shaul, Ricky TQ Chen, Maximilian Nickel, Matthew Le, and Yaron Lipman. On kinetic optimal probability paths for generative models. In *International Conference on Machine Learning*, pages 30883–30907. PMLR, 2023.

[40] Dejan Stancevic, Florian Handke, and Luca Ambrogioni. Entropic time schedulers for generative diffusion models. *arXiv preprint arXiv:2504.13612*, 2025.

[41] Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

[42] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[43] An B Vuong, Michael T McCann, Javier E Santos, and Yen Ting Lin. Are we really learning the score function? reinterpreting diffusion models through Wasserstein gradient flow matching. *arXiv preprint arXiv:2509.00336*, 2025.

[44] Christian Wald and Gabriele Steidl. Flow matching: Markov kernels, stochastic processes and transport plans. *Variational and Information Flows in Machine Learning and Optimal Transport*, pages 185–254, 2025.

[45] Zhengchao Wan, Qingsong Wang, Gal Mishne, and Yusu Wang. Elucidating flow matching ode dynamics via data geometry and denoisers. In *Forty-second International Conference on Machine Learning*.

[46] Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on learning theory*, pages 2093–3027. PMLR, 2018.

[47] Yewei Xu and Qin Li. Forward-euler time-discretization for Wasserstein gradient flows can be wrong. *arXiv preprint arXiv:2406.08209*, 2024.

[48] Zeqi Ye, Qijie Zhu, Molei Tao, and Minshuo Chen. Provable separations between memorization and generalization in diffusion models. *arXiv preprint arXiv:2511.03202*, 2025.

[49] Donggeun Yoon, Minseok Seo, Doyi Kim, Yeji Choi, and Donghyeon Cho. Deterministic guidance diffusion model for probabilistic weather forecasting. *arXiv preprint arXiv:2312.02819*, 2023.

[50] Huijie Zhang, Zijian Huang, Siyi Chen, Jinfan Zhou, Zekai Zhang, Peng Wang, and Qing Qu. Understanding generalization in diffusion models via probability flow distance. *arXiv preprint arXiv:2505.20123*, 2025.

[51] Yasi Zhang, Peiyu Yu, Yaxuan Zhu, Yingshan Chang, Feng Gao, Ying Nian Wu, and Oscar Leong. Flow priors for linear inverse problems via iterative corrupted trajectory matching. *Advances in Neural Information Processing Systems*, 37:57389–57417, 2024.

[52] Jia-Jie Zhu and Alexander Mielke. Approximation, kernelization, and entropy-dissipation of gradient flows: from wasserstein to fisher-rao. 2022.

Appendix

A Related Work

Flow Matching and Related Models. Flow Matching (FM) and Conditional Flow Matching (CFM) have been developed as scalable alternatives to diffusion-based generative models. Recent work has analyzed their statistical, geometric, and algorithmic foundations, including distributional properties of FM [23], particle and bridge-based interpretations [4], and geometric structure and gauge freedom in learned flows [45, 20]. Extensions include guided generation [16], statistical efficiency analyses [34], and rigorous comparisons between FM and optimal transport [19, 44]. The kinetic and energetic behavior of flow-based samplers has also been examined in [39].

Memorization vs. Generalization in Generative Models. A large body of recent work studies memorization, generalization, and interpolation phenomena in modern generative models. For diffusion models, prior work has analyzed identifiability, overfitting, and deterministic sampling behavior [35, 49]. Further studies provide theoretical and empirical characterizations of interpolation, dataset coverage, and memorization tendencies [32, 37, 5, 7, 13]. Broader perspectives on generalization in flow-based and likelihood-based models appear in [50, 18, 26, 48, 36, 27]. Our work contributes to this line by showing that empirical FM induces a structural bias—non-conservative velocities—that naturally leads to memorization-like behavior even without model approximation error.

Understanding and Improving the Sampling Process. A complementary literature studies the dynamics and stability of generative sampling. This includes analyses of Lipschitz regularity and robustness [11], entropic or kinetic regularization [40], and methods aimed at accelerating or stabilizing the generation process [17]. For diffusion and score-based models, [21] examines how score estimation affects sampling quality. Our work adds to this view by characterizing the kinetic energy and tail behavior induced by empirical FM.

B Sampling as Optimization in Measure Space

There are rich connections between sampling and optimization, which we briefly discuss here. It is often useful to study sampling as an optimization problem in the space of probability measures [46]. The main idea is to find an objective function that is minimized exactly at a given target measure p_∞. One such special objective function is the relative entropy (or free energy, or KL divergence):

$$F[p] = \int p \log(p/p_\infty), \tag{23}$$

where p is a probability measure on $\mathbb{R}^d$ absolutely continuous with respect to p_∞. It is easy to see that $F[p] \geq 0$ and is minimized at the target measure, i.e., $H[p] = 0$ if and only if $p = p_\infty$, and p_∞ is the only stationary point of $F[p]$. Now, consider target measures of the form $p_\infty(z) = \exp(-f(z))$ (i.e., $f = -\log p_\infty$). Then we can write $F[p] = \mathbb{E}_p[f] - H[p]$, where $H[p] = -\mathbb{E}_p[\log p]$ is the entropy of p.

Therefore, if we can minimize $F[p]$, then we can sample from p_∞. This means that we need to solve an optimization problem over the space of probability measures in which we aim to minimize the KL divergence from the target p_∞. The standard trick is to minimize $F[p]$ by running the gradient flow dynamics in the space of measures over $\mathbb{R}^d$ endowed with the Wasserstein-2 metric. In this metric space, the gradient flow of $F[p]$ is given by the Fokker-Planck equation

(FPE):

$$\frac{\partial p_t}{\partial t} = \nabla \cdot \left(p_t \nabla \log \left(\frac{p_t}{p_\infty} \right) \right) = \nabla \cdot (p_t \nabla f) + \Delta p_t, \tag{24}$$

where p_t is a smooth positive density evolving over time. If we can follow the flow of this FPE then p_t converges to the target measure p_∞ as $t \to \infty$. The convergence is exponentially fast if p_∞ satisfies the log Sobolev inequality; e.g., this is true when $p_\infty = \mathcal{N}(m_\infty, \Sigma_\infty)$ where Σ_∞ is positive definite. The FPE is the continuity equation of the (Probability Flow) ODE:

$$\frac{dZ_t}{dt} = -\nabla_{Z_t} \log \frac{p_t(Z_t)}{p_\infty(Z_t)}, \tag{25}$$

as well as the SDE (Langevin dynamics):

$$dZ_t = -\nabla_{Z_t} f(Z_t)dt + \sqrt{2}dW_t = \nabla_{Z_t} \log p_\infty(Z_t)dt + \sqrt{2}dW_t, \tag{26}$$

where W_t is the standard Brownian motion in $\mathbb{R}^d$. This means that if $Z_t \sim p_t$ evolves via the above ODE or the SDE in space, then p_t evolves according to the FPE in the space of measures. As an example, if $p_\infty = \mathcal{N}(m_\infty, \Sigma_\infty)$, then we can use the above SDE with $f(z) = \frac{1}{2}(z - m_\infty)^T \Sigma_\infty^{-1}(z - m_\infty) + \frac{1}{2} \log \det(2\pi\Sigma_\infty)$ (quadratic) and run the SDE to obtain samples from p_∞ (since we know, by design, $Z_t \sim p_t$ converges to $e^{-f} = \mathcal{N}(m_\infty, \Sigma_\infty)$ exponentially fast). Equivalently, we can use the ODE $\dot{Z}_t = v(Z_t) := -Z_t - \nabla_Z p_t(Z_t)$ instead for the same purpose.

One then asks what are the connections between sampling via measure transport (e.g., the rectified flow model) and sampling via optimization. We can try to map the rectified flow example in Example 5 to the above optimization perspective but the mapping is not clean. The main differences are that here we consider finite-time dynamics ($t \in [0, 1]$) instead and there are time-dependent coefficients in $\Phi(t, z)$ (see Eq. (21)). Nevertheless, under the same setting as Example 5, if we consider the (generalized) free energy functional

$$F[p_t] = \int \Phi(t, z)p_t(z)dz = \frac{1}{t} \left[\int \frac{1}{2}\|z\|^2 p_t(z)dz + (1 - t) \int p_t(z) \log p_t(z)dz \right] \tag{27}$$

$$=: \frac{1}{t} \left(\mathbb{E}_{z \sim p_t} \left[\frac{\|z\|^2}{2} \right] - \beta(t)H[p_t] \right), \tag{28}$$

with $\beta(t) = 1 - t$, then we can interpret it as a time-dependent[2] weighted sum of average quadratic potential energy and Shannon entropy. Then, in order to minimize the free energy functional (over the probability measures in the Wasserstein space) via the gradient flows, one should take the velocity field v_{WGF} that corresponds to the steepest descent [43] of $F[p_t]$ in the Wasserstein space, and this can be shown to be precisely

$$v_{\mathrm{WGF}}(t, z) = -\nabla_z \Phi(t, z) = -\frac{1}{t}z - \frac{1 - t}{t} \nabla_z \log p_t(z)$$

(note the similarity of this with $v(z)$ up to the presence of time-dependent coefficients). This is the Wasserstein gradient flow (WGF) perspective for the rectified flow model. More details on WGFs are in [43], [12], [47], [25], [22], [3], [52], [10], [38].

When the metric is chosen to be the Fisher-Rao metric instead, we have the Fisher-Rao gradient flows perspective [10, 14]. In [6] (see Section 3), sampling in unit time was considered with the kernel Fisher-Rao gradient flow.

[2]The time-dependent nature of the weights causes a loss of direct analogy with the free energy functional encountered in statistical mechanics, which are concerned with the study of thermodynamic systems in the large time (equilibrium) limit.

C Proof of Theoretical Results

C.1 Proof of Proposition 1

Proof of Proposition 1. Let $\hat{\mathcal{L}}_{\mathrm{CFM}}[v'] = \mathbb{E}_{t,X,Z_t}\|v'(t,Z_t) - v(t,Z_t \mid X)\|^2$. Since $X \sim \hat{p}_1$, the expectation over X can be written as:

$$\hat{\mathcal{L}}_{\mathrm{CFM}}[v'] = \mathbb{E}_t\left[\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{Z_t\sim p_t(\cdot|X^{(j)})}\|v'(t,Z_t) - v(t,Z_t \mid X^{(j)})\|^2\right].$$

The optimal field $v'(t,z)$ that minimizes $\mathbb{E}_t\left[\int_{\mathbb{R}^d}\|v'(t,z) - \hat{v}(t,z)\|^2\hat{p}_t(z)\,dz\right]$ is the conditional expectation of $v(t,Z_t|X)$ given $Z_t = z$:

$$\hat{v}^*(t,z) = \mathbb{E}_{X\sim\hat{p}_1}\left[v(t,z \mid X) \mid Z_t = z\right].$$

Using Bayes' theorem:

$$P(X = X^{(j)} \mid Z_t = z) = \frac{p_t(z \mid X^{(j)})\hat{p}_1(X^{(j)})}{\hat{p}_t(z)} = \frac{p_t(z \mid X^{(j)})}{\sum_{k=1}^{N} p_t(z \mid X^{(k)})} =: w_j(t,z).$$

Substituting the conditional velocity $v(t,z \mid X^{(j)}) = a_t(X^{(j)})z + b_t(X^{(j)})$:

$$\hat{v}^*(t,z) = \sum_{j=1}^{N} P(X = X^{(j)} \mid Z_t = z)\cdot v(t,z \mid X^{(j)}) = \sum_{j=1}^{N} w_j(t,z)\left(a_t(X^{(j)})\,z + b_t(X^{(j)})\right),$$

which is the weighted sum over all possible samples $X^{(j)}$ that we aimed to show. □

C.2 Proof of Proposition 2

Proof of Proposition 2. By Poincaré lemma [9], a continuously differentiable vector field $F : \mathbb{R}^d \to \mathbb{R}^d$ on a simply connected domain in $\mathbb{R}^d$ (e.g., the whole $\mathbb{R}^d$) is a gradient field if and only if its Jacobian matrix $J_F := (\frac{\partial F_j}{\partial z_k})_{j,k}$ is symmetric everywhere. Therefore, it suffices to identify the condition under which the Jacobian matrix of $\hat{v}^*(t,z)$ is symmetric on $\mathbb{R}^d$ for all $t \in [0,1]$.

Using the product rule $\nabla_z(cu) = cJ_u + u(\nabla_z c)^T$ for a scalar-valued function c and a vector-valued function u, we obtain, for any $t \in [0,1]$,

$$J_{\hat{v}^*} = \sum_{i=1}^{N}[w_i(t,z)J_{v_i}(t,z) + v_i(\nabla_z w_i)^T], \tag{29}$$

which is symmetric if and only if its skew-symmetric part is zero, i.e., $J_{\hat{v}^*} - J_{\hat{v}^*}^T = 0$. Since $J_{v_i} = J_{v_i}^T$ for all i (as the conditional velocity fields v_i themselves are gradient fields), the latter condition simplifies to

$$\sum_{i=1}^{N}[v_i(\nabla_z w_i)^T - (\nabla_z w_i)v_i^T] = 0,$$

which is exactly the condition stated in the proposition. □

C.3 Proof of Proposition 3

Proof of Proposition 3 (a). Since $p_1 = \mathcal{N}(m_1, \Sigma_1)$, we have, for all $y \in \mathbb{R}^d$,

$$-\log p_1(y) = \frac{1}{2}(y - m_1)^T \Sigma_1^{-1}(y - m_1) + \frac{1}{2}\log\det(2\pi\Sigma_1) \tag{30}$$

$$= \frac{1}{2}y^T\Sigma_1^{-1}y - y^T\Sigma_1^{-1}m_1 + \frac{1}{2}m_1^T\Sigma_1^{-1}m_1 + \frac{1}{2}\log\det(2\pi\Sigma_1). \tag{31}$$

Meanwhile, $E(y) = \|y - R^{-1}(y)\|^2 = \|y - \Sigma_1^{-1/2}(y - m_1)\|^2 = \|(I_d - \Sigma_1^{-1/2})y + \Sigma_1^{-1/2}m_1\|^2$. Expanding the term and then regrouping the resulting terms, we obtain, for all $y \in \mathbb{R}^d$,

$$\frac{1}{2}E(y) = \frac{1}{2}y^T(I - 2\Sigma_1^{-1/2} + \Sigma_1^{-1})y + m_1^T\Sigma_1^{-1/2}y - m_1^T\Sigma_1^{-1}y + \frac{1}{2}m_1^T\Sigma_1^{-1}m_1.$$

The desired result then follows from the above formula for $-\log p_1(y)$ and $\frac{1}{2}E(y)$. □

Before proving Proposition 3 (b), we need the following auxiliary result.

Lemma 1. *Let $W \sim \mathcal{N}(0, 1)$ be a scalar standard Gaussian random variable. Let $a, b \in \mathbb{R}$ be constants. If $b < \frac{1}{2}$, then:*

$$\mathbb{E}\left[e^{aW + bW^2}\right] = \frac{1}{\sqrt{1 - 2b}}\exp\left(\frac{a^2}{2(1 - 2b)}\right).$$

Proof. Compute the integral explicitly, complete the square, and simplify. □

With this lemma in place, we can now prove part (b) in Proposition 3.

Proof of Proposition 3 (b). The RF map is given as $R(x) = m_1 + \Sigma_1^{1/2}x$, and the inverse map is given by $R^{-1}(y) = \Sigma_1^{-1/2}(y - m_1)$. We analyze the random variable $E(Y)$ where $Y \sim p_1$. Since p_1 is the pushforward of $p_0 = \mathcal{N}(0, I_d)$ through R, we can parameterize Y using $X \sim \mathcal{N}(0, I_d)$ via $Y = R(X)$.

Substituting this into the energy definition:

$$E(Y) = \|Y - R^{-1}(Y)\|^2$$

Since $R^{-1}(R(X)) = X$ by definition of the inverse, this simplifies to:

$$E = \|R(X) - X\|^2.$$

Using the definition of $R(X)$:

$$E = \|(m_1 + \Sigma_1^{1/2}X) - X\|^2 = \|m_1 + (\Sigma_1^{1/2} - I_d)X\|^2.$$

Let $A = \Sigma_1^{1/2} - I_d$. Note that A is symmetric and we can consider the eigen-decomposition $A = UDU^T$, where U is orthogonal and D is diagonal with elements d_i. The eigenvalues of $\Sigma_1^{1/2}$ are $\sqrt{\lambda_i(\Sigma_1)}$. Thus, the eigenvalues of A are:

$$d_i = \sqrt{\lambda_i(\Sigma_1)} - 1.$$

The kinetic energy can then be written as:

$$E = \|m_1 + UDU^T X\|^2.$$

Since the Euclidean norm is rotation-invariant, $\|v\|^2 = \|U^T v\|^2$ for any orthogonal matrix U, we obtain:

$$E = \|U^T m_1 + D(U^T X)\|^2$$

Let $\tilde{m} = U^T m_1$ (note $\|\tilde{m}\|^2 = \|m_1\|^2$) and $Z = U^T X$. Since $X \sim \mathcal{N}(0, I_d)$ and U is orthogonal, $Z \sim \mathcal{N}(0, I_d)$. The energy decomposes into a sum of independent terms:

$$E = \sum_{i=1}^{d} (\tilde{m}_i + d_i Z_i)^2.$$

Let $u > 0$ be given. Applying the Chernoff bound gives $\mathbb{P}(E \geq u) \leq e^{-tu}\mathbb{E}[e^{tE}]$ for any $t > 0$. Using the independence of Z_i, we have:

$$\mathbb{E}[e^{tE}] = \prod_{i=1}^{d} \mathbb{E}\left[\exp\left(t(\tilde{m}_i + d_i Z_i)^2\right)\right] =: \prod_{i=1}^{d} M_i.$$

Expanding the term in the exponents, we see that:

$$t(\tilde{m}_i^2 + 2\tilde{m}_i d_i Z_i + d_i^2 Z_i^2) = (t\tilde{m}_i^2) + (2t\tilde{m}_i d_i)Z_i + (td_i^2)Z_i^2.$$

Now, we apply Lemma 1 for $\mathbb{E}[e^{aW+bW^2}]$ with $W = Z_i$, $a = 2t\tilde{m}_i d_i$ and $b = td_i^2$, for $b < 1/2$. Let $\rho = \max_i(\sqrt{\lambda_i(\Sigma_1)} - 1)^2 = \max_i d_i^2$ (which is positive since we assume $\Sigma_1 \neq I_d$) and choose $t = \frac{1}{4\rho}$. Then $b = \frac{d_i^2}{4\rho} \leq \frac{1}{4} < \frac{1}{2}$, and so the condition needed to apply the lemma is satisfied.

Applying the lemma to the M_i, we have:

$$M_i = \frac{1}{\sqrt{1 - 2td_i^2}} \cdot \exp\left(t\tilde{m}_i^2 + \frac{(2t\tilde{m}_i d_i)^2}{2(1 - 2td_i^2)}\right).$$

Now, we bound the terms: $2td_i^2 = \frac{d_i^2}{2\rho} \leq \frac{1}{2}$. Thus $\sqrt{1 - 2td_i^2} \geq \sqrt{1/2}$, and $\frac{1}{\sqrt{1-2td_i^2}} \leq \sqrt{2}$. Thus, for the term in the exponent of M_i:

$$t\tilde{m}_i^2 + \frac{4t^2\tilde{m}_i^2 d_i^2}{2(1 - 2td_i^2)} = t\tilde{m}_i^2\left(1 + \frac{2td_i^2}{1 - 2td_i^2}\right) = \frac{t\tilde{m}_i^2}{1 - 2td_i^2}.$$

Since $1 - 2td_i^2 \geq 1/2$,

$$\frac{t\tilde{m}_i^2}{1 - 2td_i^2} \leq 2t\tilde{m}_i^2 = \frac{\tilde{m}_i^2}{2\rho}.$$

Combining these, we have:

$$M_i \leq \sqrt{2}\exp\left(\frac{\tilde{m}_i^2}{2\rho}\right).$$

Therefore,

$$\mathbb{E}[e^{tE}] \leq \prod_{i=1}^{d}\left(\sqrt{2}e^{\frac{\tilde{m}_i^2}{2\rho}}\right) = 2^{d/2}\exp\left(\frac{\sum \tilde{m}_i^2}{2\rho}\right) = 2^{d/2}\exp\left(\frac{\|m_1\|^2}{2\rho}\right) =: C.$$

Finally, substituting this into the earlier Chernoff bound:

$$\mathbb{P}(E \geq u) \leq e^{-tu}C = C\exp\left(-\frac{u}{4\rho}\right).$$

$\square$

C.4 Proof of Theorem 1

Lemma 2. *Let $X_0 \sim \mathcal{N}(0, I_d)$. Define $U = \frac{\|X_0\|^2}{d}$. For all $s \geq 2$, we have:*

$$\mathbb{P}(U \geq s) \leq \exp\left(-\frac{sd}{16}\right).$$

Proof. First, we claim that for all $s \geq 1$,

$$\mathbb{P}\left(U \geq s\right) \leq \exp\left(-\frac{d}{2} f(s)\right), \tag{32}$$

where $f(s) = s - 1 - \ln(s)$.

To verify this claim, let $S := \|X_0\|^2$ and compute, for $\lambda > 0$,

$$\mathbb{P}(S \geq ds) = \mathbb{P}(e^{\lambda S} \geq e^{\lambda ds}) \tag{33}$$

$$\leq e^{-\lambda ds}\mathbb{E}[e^{\lambda S}] = \frac{e^{-\lambda ds}}{(1-2\lambda)^{d/2}}, \tag{34}$$

where we have used the fact that $\|X_0\|^2 \sim \chi_d^2$ (chi-squared distributed) and the formula for its moment generating function in the last line. Choosing $\lambda = \frac{s-1}{2s} \in (0, 1/2)$ minimizes the upper bound. Plugging this minimizer back into the upper bound, we obtain the result as claimed.

Now, observe that for $s \geq 2$, $f(s) \geq s/8$. Therefore, using (32) and this observation, we have, for all $s \geq 2$,

$$\mathbb{P}\left(U \geq s\right) \leq \exp\left(-\frac{sd}{16}\right), \tag{35}$$

which is the result that we wanted to show. □

With this lemma in place, we can now prove Theorem 1.

Proof of Theorem 1. Let $T \in [0, 1)$ and $\mathcal{D}_N$ be given. For all $t \in [0, T]$ and $z \in \mathbb{R}^d$,

$$\|\hat{v}^*(t, z)\| \leq \frac{1}{1-t}\sum_{i=1}^N w_i(t, z)\|x^{(i)} - z\| \tag{36}$$

$$\leq \frac{1}{1-t}\sum_{i=1}^N w_i(t, z)(\|x^{(i)}\| + \|z\|) \tag{37}$$

$$\leq \frac{1}{1-t}(M + \|z\|), \tag{38}$$

where we have used the fact that $\sum_i w_i(t, z) = 1$ and the notation $M := \max_i \|x^{(i)}\|$.

Let $r_t := \|\psi_t(X_0)\|$. For all t with $r_t > 0$,

$$\dot{r}_t := \frac{dr_t}{dt} = \frac{\psi_t(X_0) \cdot \dot{\psi}_t(X_0)}{\|\psi_t(X_0)\|} \leq \frac{|\psi_t(X_0) \cdot \dot{\psi}_t(X_0)|}{\|\psi_t(X_0)\|} \leq \|\dot{\psi}_t(X_0)\| = \|\hat{v}^*(t, \psi_t(X_0))\|,$$

where we have used the chain rule for differentiation and Cauchy-Schwarz inequality.

Then, using (38):

$$\dot{r}_t \leq \frac{1}{1-t}(M + r_t)$$

and so $(1-t)\dot{r}_t - r_t \leq M$. Now,

$$\frac{d}{dt}\left((1-t)r_t\right) = (1-t)\dot{r}_t - r_t \leq M.$$

Integrating both sides from 0 to t gives (and noting that $r_0 = \|X_0\|$):

$$(1-t)r_t - r_0 \leq Mt \tag{39}$$
$$(1-t)r_t \leq \|X_0\| + Mt \tag{40}$$
$$\|\psi_t(X_0)\| \leq \frac{\|X_0\| + Mt}{1-t} =: c_1(t)\|X_0\| + c_2(t)M, \tag{41}$$

where $c_1(t) = 1/(1-t)$ and $c_2(t) = t/(1-t)$.

Let $\hat{V}_t := \hat{v}^*(t, \psi_t(X_0))$. Using (38) and (41), we have:

$$\begin{aligned}\|\hat{V}_t\| &\leq \frac{1}{1-t}(M + \|\psi_t(X_0)\|) &(42)\\ &\leq \frac{1}{1-t}(M + c_1(t)\|X_0\| + c_2(t)M) &(43)\\ &\leq c_1^2(t)(M + \|X_0\|). &(44)\end{aligned}$$

Therefore,

$$\begin{aligned}K_t := \|\hat{V}_t\|^2 &\leq c_1^4(t)(\|X_0\| + M)^2 &(45)\\ &\leq 2c_1^4(t)(\|X_0\|^2 + M^2), &(46)\end{aligned}$$

where we have used the inequality $(x+y)^2 \leq 2(x^2+y^2)$ for $x, y \in \mathbb{R}$.

Integrating from 0 to T on both sides gives:

$$E_T = \int_0^T K_t dt \leq c_3(T)(\|X_0\|^2 + M^2),$$

where $c_3(T) = 2\int_0^T c_1^4(t)dt = \frac{2}{3}((1-T)^{-3} - 1)$.

Now, for any $u > 0$, since $\{K_t \geq u\} \subset \left\{\|X_0\|^2 \geq \frac{u}{2c_1^4(t)} - M^2\right\}$, we have:

$$\mathbb{P}[K_t \geq u \mid \mathcal{D}_N] \leq \mathbb{P}[\|X_0\|^2/d \geq s \mid \mathcal{D}_N], \tag{47}$$

where $s := u/(2dc_1^4(t)) - M^2/d$.

Since this holds for any $u > 0$, we can choose $u \geq 2c_1^4(t)(2d + M^2) =: U_t$ so that $s \geq 2$ and apply Lemma 2 to obtain $\mathbb{P}[K_t \geq U_t \mid \mathcal{D}_N] \leq C_t \exp(-c_t U_t)$ with $C_t = e^{M^2/16}$ and $c_t = (1-t)^4/32$. This shows part (a).

For part (b), we can choose $u \geq c_3(T)(2d + M^2) =: U_T$ and proceed analogously to obtain $\mathbb{P}[E_T \geq U_T \mid \mathcal{D}_N] \leq C_T \exp(-c_T U_T)$ with $C_T = e^{M^2/16}$ and $c_T = 1/(16c_3(T)) = \frac{3}{32((1-T)^{-3}-1)}$.

□

C.5 Proof of Theorem 2

Proof of Theorem 2. The proof is analogous to that of Theorem 1, with the Gaussian tail bound replaced by the assumed power-law tail.

Recall from Proposition 1 that the empirical affine-flow minimizer has the form

$$\hat{v}^*(t,z) = \sum_{i=1}^{N} w_i(t,z)\,\big(a_t(x^{(i)})z + b_t(x^{(i)})\big),$$

where the weights $w_i(t,z)$ are nonnegative and sum to one. By the definition of

$$A_{\max} := \sup_{t\in[0,T],\, i\in[N]} |a_t(x^{(i)})|, \qquad B_{\max} := \sup_{t\in[0,T],\, i\in[N]} \|b_t(x^{(i)})\|,$$

we have, for all $t \in [0,T]$ and all $z \in \mathbb{R}^d$,

$$\|\hat{v}^*(t,z)\| \;\le\; \sum_{i=1}^{N} w_i(t,z)\big(|a_t(x^{(i)})|\,\|z\| + \|b_t(x^{(i)})\|\big) \;\le\; A_{\max}\|z\| + B_{\max}. \tag{48}$$

Let ψ_t denote the flow driven by $\hat{v}^*$, i.e.,

$$\dot{\psi}_t(X_0) = \hat{v}^*(t,\psi_t(X_0)), \qquad \psi_0(X_0) = X_0,$$

and define $r_t := \|\psi_t(X_0)\|$. Whenever $r_t > 0$, we have, by the chain rule and Cauchy–Schwarz,

$$\dot{r}_t = \frac{\psi_t(X_0)}{\|\psi_t(X_0)\|}\cdot\dot{\psi}_t(X_0) \le \|\hat{v}^*(t,\psi_t(X_0))\|.$$

Using (48) at $z = \psi_t(X_0)$ gives

$$\dot{r}_t \;\le\; A_{\max} r_t + B_{\max}.$$

By Grönwall's lemma, there exist constants $C_1(T), C_2(T) > 0$, depending only on $T, A_{\max}, B_{\max}$, such that for all $t \in [0,T]$,

$$r_t = \|\psi_t(X_0)\| \;\le\; C_1(T)\,\|X_0\| + C_2(T). \tag{49}$$

Define $V_t := \hat{v}^*(t,\psi_t(X_0))$ and the instantaneous kinetic energy $K_t := \|V_t\|^2$. Combining (48) and (49), we obtain

$$\|V_t\| \;\le\; A_{\max} r_t + B_{\max} \;\le\; A_{\max}\big(C_1(T)\|X_0\| + C_2(T)\big) + B_{\max} \;\le\; C_3(T)\,\|X_0\| + C_4(T),$$

for suitable constants $C_3(T), C_4(T) > 0$ depending only on $T, A_{\max}, B_{\max}$. Hence, by the inequality $(x+y)^2 \le 2(x^2+y^2)$,

$$K_t = \|V_t\|^2 \;\le\; 2C_3(T)^2\|X_0\|^2 + 2C_4(T)^2 \;\le\; C_K(T)\,\big(\|X_0\|^2 + 1\big), \tag{50}$$

where we may take $C_K(T) := 2\max\{C_3(T)^2, C_4(T)^2\}$. Integrating (50) over $t \in [0,T]$ yields the same type of bound for the integrated kinetic energy

$$E_T := \int_0^T K_t\,dt,$$

i.e.,

$$E_T \;\le\; C_E(T)\,\big(\|X_0\|^2 + 1\big), \tag{51}$$

for some constant $C_E(T) := TC_K(T) > 0$ depending only on $T, A_{\max}, B_{\max}$.

Tail bounds. From (50), for any $u > 0$,

$$\{K_t \ge u\} \;\subseteq\; \Big\{\|X_0\|^2 \;\ge\; \frac{u}{C_K(T)} - 1\Big\}.$$

Fix U_t large enough so that for all $u \geq U_t$, $\frac{u}{C_K(T)} - 1 \geq 1$. Writing $s := \sqrt{\frac{u}{C_K(T)} - 1}$, we obtain

$$\mathbb{P}\big(K_t \geq u \,\big|\, D_N\big) \;\leq\; \mathbb{P}\big(\|X_0\| \geq s \,\big|\, D_N\big) = \mathbb{P}\big(\|X_0\| \geq s\big),$$

since X_0 is independent of D_N. By the heavy-tailed assumption on p_0, for all $s \geq 1$,

$$\mathbb{P}\big(\|X_0\| \geq s\big) \;\leq\; \frac{C_\alpha}{s^\alpha}.$$

For $u \geq U_t$ large enough so that $s^2 = \frac{u}{C_K(T)} - 1 \geq \frac{u}{2C_K(T)}$, we have

$$\frac{1}{s^\alpha} \;\leq\; \left(\frac{2C_K(T)}{u}\right)^{\alpha/2},$$

and hence

$$\mathbb{P}\big(K_t \geq u \,\big|\, D_N\big) \;\leq\; \frac{C'_\alpha}{u^{\alpha/2}},$$

where $C'_\alpha := C_\alpha(2C_K(T))^{\alpha/2}$. This proves the first inequality in (2) with $\gamma = \alpha/2$.

The argument for E_T is identical, using (51) in place of (50). For any $u > 0$,

$$\{E_T \geq u\} \;\subseteq\; \left\{\|X_0\|^2 \;\geq\; \frac{u}{C_E(T)} - 1\right\},$$

and the same substitution $s = \sqrt{\frac{u}{C_E(T)} - 1}$ together with the heavy-tailed bound on $\|X_0\|$ yields

$$\mathbb{P}\big(E_T \geq u \,\big|\, D_N\big) \;\leq\; \frac{C'_{\alpha,T}}{u^{\alpha/2}}$$

for all sufficiently large u, for some constant $C'_{\alpha,T} > 0$ depending on C_α and $C_E(T)$.

Collecting all constants into a single C_{poly} that depends only on $T, A_{\max}, B_{\max}, C_\alpha$ gives the desired polynomial decay with exponent $\gamma = \alpha/2$, which completes the proof.

□