# On Kinetic Optimal Probability Paths for Generative Models

Neta Shaul [1]   Ricky T. Q. Chen [2]   Maximilian Nickel [2]   Matt Le [2]   Yaron Lipman [2][1]

## Abstract

Recent successful generative models are trained by fitting a neural network to an a-priori defined tractable probability density path taking noise to training examples. In this paper we investigate the space of Gaussian probability paths, which includes diffusion paths as an instance, and look for an optimal member in some useful sense. In particular, minimizing the Kinetic Energy (KE) of a path is known to make particles' trajectories simple, hence easier to sample, and empirically improve performance in terms of likelihood of unseen data and sample generation quality. We investigate Kinetic Optimal (KO) Gaussian paths and offer the following observations: (i) We show the KE takes a simplified form on the space of Gaussian paths, where the data is incorporated only through a single, one dimensional scalar function, called the *data separation function*. (ii) We characterize the KO solutions with a one dimensional ODE. (iii) We approximate data-dependent KO paths by approximating the data separation function and minimizing the KE. (iv) We prove that the data separation function converges to $1$ in the general case of arbitrary normalized dataset consisting of $n$ samples in $d$ dimension as $n/\sqrt{d} \to 0$. A consequence of this result is that the Conditional Optimal Transport (Cond-OT) path becomes *kinetic optimal* as $n/\sqrt{d} \to 0$. We further support this theory with empirical experiments on ImageNet.

## 1. Introduction

In recent years, deep generative models have become very powerful both in terms of sample quality and density estimation (Rombach et al., 2021; Gafni et al., 2022). The recent revolution is mainly led by a class of time-dependent generative models which make use of predefined probability paths $p_t$, constituting a process interpolating between the target (data) distribution $q$ and the prior (noise) distribution $p$. The training procedure of these models can be generally described as a regression task of a neural network, $v_t(x; \theta)$, to some vector quantity, $u_t$, that allows sampling from the interpolation process:

$$\mathcal{L}(\theta) = \mathbb{E}_{p_t(x)} \ell(v_t(x; \theta), u_t(x)), \tag{1}$$

where $\ell(v, u)$ is some cost function, and $\mathcal{L}$ is the total loss objective that is stochastically estimated and optimized during training. A notable family of algorithms that fits into this framework are diffusion models (Ho et al., 2020; Song et al., 2020) as well as the recent Flow-Matching models, (Lipman et al., 2022; Albergo & Vanden-Eijnden, 2022; Liu et al., 2022; Neklyudov et al., 2022) generalizing the principles used in diffusion models training and apply them to simulation-free continuous normalizing flows (CNFs, (Chen et al., 2018)) training. In particular, (Song et al., 2020; Lipman et al., 2022) show that diffusion models also fall into the simulation-free CNF framework. CNFs parameterize the generation process via a flow vector field $u_t$, which defines the probability path $p_t$.

In order to make the loss in equation 1 tractable, all these methods are confined to a family of *Gaussian probability paths*

$$\mathcal{P} = \left\{ p_t \,\middle|\, p_t(x) = \int p_t(x|x_1) q(x_1) dx_1 \right\} \tag{2}$$

defined by the conditional Gaussian probabilities $p_t(x|x_1) = \mathcal{N}(x|a_t x_1, m_t^2 I)$, on specific training examples $x_1 \sim q$. The space $\mathcal{P}$ allows scalable unbiased approximations of the loss in equation 1 and therefore it is of particular interest. Within this space, diffusion models construct $p_t$ as a diffusion markov process and (Lipman et al., 2022; Liu et al., 2022) draw inspiration from optimal transport to define the conditional probabilities. The different choices of previously explored paths are based on known processes, empirical heuristics (Karras et al., 2022) or learned paths (Nichol & Dhariwal, 2021; Kingma et al., 2021b), but there is no theoretical result analyzing the *optimality* of those paths.

---

[1]Weizmann Institute of Science [2]Meta AI (FAIR). Correspondence to: Neta Shaul <Neta.Shaul@weizmann.ac.il>.

The goal of this paper is to investigate the space of probability paths, $\mathcal{P}$, and single out an *optimal one* in some well defined useful sense. We consider the Kinetic Energy (KE) as the measure of path optimality. The Kinetic energy is directly tied to Optimal Transport solutions in the dynamic formulation (Benamou & Brenier, 2000; Villani, 2009) where the optimal probability path $p_t$ minimizes the KE. Therefore, reducing the KE simplifies the trajectories of individual particles, allowing faster sampling and empirically improved performance. Our key observation is that the KE over $\mathcal{P}$ takes a simplified form, where its dependence on the data distribution $q$ is summarized in a single, one dimensional, scalar function $\lambda : [0, \infty) \to [0, 1]$. Intuitively, $\lambda$ measures the separation of the data samples at different scales, and therefore we call it the *data separation function*. We then characterize the minimizers of KE over $\mathcal{P}$. These observations allow us to approximate Kinetic Optimal (KO) probability paths for different datasets $q$. Note that $\mathcal{P}$ is optimized over the Gaussian paths and therefore the KO path is usually not the Optimal Transport path.

Surprisingly, for high dimensional data we are able to characterize Kinetic Optimal solutions. We prove that for arbitrary normalized datasets consisting of $n$ data points in $\mathbb{R}^d$ the probability path defined by the Conditional Optimal-Transport (Cond-OT) path suggested by (Lipman et al., 2022) becomes kinetic optimal as

$$\frac{n}{\sqrt{d}} \to 0. \tag{3}$$

Empirically, we show evidence that in practice Cond-OT is KO even for lower dimensions than predicted by the above asymptotics, and show that KO paths provides improvement in the KE of trained models over Cond-OT paths in low dimensions and that this improvement is diminished as dimension increases, as predicted by the theory.

## 2. Preliminaries

We consider $\mathbb{R}^d$ as our data domain with data points $x \in \mathbb{R}^d$. Probability densities over $\mathbb{R}^d$ will be denoted by $p, q$. With a slight abuse of notations, but as is usually done, random variables will also be denoted as $x, x_0, x_1$, where a random variable $x$ distributed according to $p$ is denoted $x \sim p(x)$. A time-dependent vector field (VF) is a smooth function $u : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$. A VF defines a *flow*, which is a diffeomorphism $\Phi : \mathbb{R}^d \to \mathbb{R}^d$, defined via a solution $\phi_t(x)$ to the ordinary differential equation (ODE):

$$\frac{d}{dt}\phi_t(x) = u_t(\phi_t(x)) \tag{4}$$

$$\phi_0(x) = x \tag{5}$$

and setting $\Phi(x) = \phi_1(x)$. A probability density path is a time dependent probability density $p_t$, $t \in [0, 1]$, smooth

in $t$. Given a probability density path $p_t$, it is said to be *generated* by $u_t$ from $p$ if

$$p_t = (\phi_t)_{\#}p \tag{6}$$

where the $\#$ is the push-forward operation of probability densities. Continuous Normalizing Flows (CNFs) (Chen et al., 2018) is a general methodology for learning generative models $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ by modeling $u_t$ as a neural network. All the generative models considered in this paper can be seen as instances of CNFs, trained with different losses and probability path supervision.

## 3. Optimal probability paths

Recent generative models learn continuous normalizing flows by regressing a vector field defining some fixed probability density path $p_t$ that is known to take a simple prior normal distribution (noise) $p = \mathcal{N}(0, I)$ to a more complex distribution $q$ (data); equation 1 describes a general loss objective for these models. In practice, the distribution $q$ is constructed (e.g., as a sum of delta distributions or a Gaussian mixture model) from a training set, which is sampled from some unknown distribution. Without limiting the following discussion, we will assume the target distribution $q$ is normalized, i.e.,

$$\mathbb{E}_{q(x_1)}x_1 = 0 \tag{7}$$

$$\frac{1}{d}\mathbb{E}_{q(x_1)}\|x_1\|^2 = 1 \tag{8}$$

The second condition can be understood as setting the average variance (across coordinates of data samples $x_1 \in \mathbb{R}^d$) to be 1.

**Probability paths with affine conditionals.** The probability density path $p_t$ is defined by marginalizing conditional probabilities $p_t(x|x_1)$,

$$p_t(x) = \int p_t(x|x_1)q(x_1)dx_1. \tag{9}$$

Our goal in this paper is to analyze a popular class of such probability paths (or paths, in short), referred here as *Gaussian paths*, defined by the *affine conditional probabilities*

$$p_t(x|x_1) = \mathcal{N}(x|a_t x_1, m_t^2 I), \tag{10}$$

where $(a_t, m_t) \in \mathcal{A}$, and

$$\mathcal{A} = \left\{ (a_t, m_t) \middle| \begin{smallmatrix} a_t, m_t : [0,1] \to [0,\infty] \\ a_0 = 0 = m_1 \\ a_1 = 1 = m_0 \end{smallmatrix} \right\}, \tag{11}$$

where we assume all functions in $\mathcal{A}$ are continuous in $[0, 1]$ and smooth in $(0, 1)$.

Note that any choice of $(a_t, m_t) \in \mathcal{A}$ will make $p_0 = p$ and $p_1 = q$ (in the limit $t \to 1$)[1]. The reason we call such conditionals affine is that they can be seen as affine maps of a standard Gaussian: if $x_0 \sim p = \mathcal{N}(0, I)$ then $x = a_t x_1 + m_t x_0 \sim p_t(x|x_1)$. Affine conditionals have been used in most previous methods, as described next.

*Diffusion models* use affine conditionals of the form (Ho et al., 2020):

$$m_t = \sqrt{1 - \xi_{1-t}^2}, \quad a_t = \xi_{1-t}, \quad (12)$$

where $\xi_t = e^{-\frac{1}{2} \int_0^t \beta(s)ds}$, and $\beta$ is the noise scale function. *Stochastic Interpolants* (Albergo & Vanden-Eijnden, 2022) are suggesting conditionals that are parameterized with trigonometric coefficients, similar also to the so-called cosine scheduling used in Diffusion (Salimans & Ho, 2022)

$$m_t = \cos \frac{\pi}{2} t, \quad a_t = \sin \frac{\pi}{2} t. \quad (13)$$

*Flow Matching* (Lipman et al., 2022; Liu et al., 2022) suggests linear coefficients (in $t$) called Conditional Optimal Transport (Cond-OT),

$$m_t = 1 - t, \quad a_t = t. \quad (14)$$

The different choices of affine conditionals in the probability path were justified mostly heuristically without any principled method for studying *optimality* under some sensible criterion. In this work we will study optimality of affine conditionals from the point of view of *Kinetic Energy*. To define the Kinetic Energy we must attach to the probability path a generating vector field (speed) in the sense of equation 6.

**Generating vector field of paths.** A natural choice for defining a generating vector field $u_t(x)$ for the path $p_t$ in the sense of equation 6 works as follows (see Lipman et al. (2022) for proofs of the following facts): first, the (conditional) vector field $u_t(x|x_1)$ that generates the affine conditional $p_t(x|x_1)$ from $p(x)$ is

$$u_t(x|x_1) = \alpha_t x + \beta_t x_1 \quad (15)$$

where

$$\alpha_t = \frac{\dot{m}_t}{m_t}, \quad \beta_t = \dot{a}_t - a_t \frac{\dot{m}_t}{m_t} \quad (16)$$

where $\dot{a}_t = \frac{d}{dt} a_t$ denotes the time derivative. Further note that $u_t(x|x_1)$ is an affine function in variable $x$, and therefore a gradient in $x$ of some potential making this choice of conditional vector field $u_t(x|x_1)$ unique (Neklyudov et al.,

2022). Secondly, the *marginal* vector field $u_t(x)$ that generates $p_t(x)$ from $p(x)$ is provided via the following formula that aggregates these conditional vector fields

$$u_t(x) = \int u_t(x|x_1) \frac{p_t(x|x_1)q(x_1)}{p_t(x)} dx_1. \quad (17)$$

In the next section we define the Kinetic Energy that uses both the marginal probability path $p_t(x)$ and its generating VF $u_t(x)$.

### 3.1. Kinetic energy of affine paths

Our goal is to study optimal choice of affine conditionals that lead to probability paths with minimal *Kinetic Energy*. The Kinetic Energy (KE) is defined for a pair $(p_t, u_t)$ of a probability path and its generating VF (in the sense of equation 6):

$$\mathcal{E}(a_t, m_t) = \frac{1}{d} \mathbb{E}_{t, p_t(x)} \|u_t(x)\|^2, \quad (18)$$

where $p_t, u_t$ are defined in equations 9 and 17, respectively. The Kinetic Energy measures the *action* of the path, by aggregating the Kinetic Energy of individual particles' paths. The main motivation for studying the Kinetic Energy is that reducing the Kinetic Energy simplifies the overall particles' path. Namely, favors paths that are straight lines with constant speed parameterization. Such paths are in fact globally minimizing the Kinetic Energy and in turn lead to Optimal Transport (OT) interpolants, where the minimal KE value realizes the Wasserstein distance of the source and target probabilities, $\mathcal{T}_2(p, q)$ (McCann, 1997; Villani, 2021). In the context of generative models, simplifying the paths improves sampling efficiency of the model (ideally, OT can be sampled with a single function evaluation), and is empirically shown to improve performance and training speed (Lipman et al., 2022; Albergo & Vanden-Eijnden, 2022; Liu et al., 2022). Note however, that OT paths are generally *outside* the class of probability paths with affine conditionals. Nevertheless, since affine conditionals are of particular interest due to the fact that they are amenable to large scale training, we investigate the kinetic optimal path within this class of paths.

A proxy energy that will be used to investigate the Kinetic Energy is the *Conditional Kinetic Energy* (CKE), that is simply the aggregate kinetic energies of the conditional vector fields,

$$\mathcal{E}_c(a_t, m_t) = \frac{1}{d} \mathbb{E}_{t, q(x_1), p_t(x|x_1)} \|u_t(x|x_1)\|^2 \quad (19)$$

The CKE, for affine conditionals can be expressed as

$$\mathcal{E}_c(a_t, m_t) = \int_0^1 \left( \dot{m}_t^2 + \dot{a}_t^2 \right) dt. \quad (20)$$

Next, the KE takes the form

$$\mathcal{E}(a_t, m_t) = \mathcal{E}_c(a_t, m_t) - \int_0^1 \beta_t^2 \left(1 - \lambda\left(\frac{a_t}{m_t}\right)\right) dt, \quad (21)$$

where $\lambda = \lambda_q : [0, \infty] \to [0, 1]$ is a univariate scalar function defined solely in terms of the target distribution $q$, and bounded by 1, i.e., $\lambda(s) \leq 1$. Computations supporting these derivations can be found in Appendix A.1. The $\lambda$ function measures the separation in the data distribution $q$ and therefore we call it the *data separation function*. Before providing an explicit expression for the data separation function, we note that having access to this function, one can optimize equation 21 to find optimal conditional marginals $(a_t, m_t) \in \mathcal{A}$. In the next section we discuss the data separation function $\lambda = \lambda_q$, followed by characterizing optimal solutions of $\mathcal{E}_c$. Lastly, note that $\lambda(s) \leq 1$ implies that $\mathcal{E}_c \geq \mathcal{E}$.

**The data separation function.** The data separation function $\lambda = \lambda_q$ summarizes the contribution of the data $q$ to the Kinetic Energy (equation 21) in a single, univariate, scalar $[0, \infty] \to [0, 1]$ function. It therefore provides a significant reduction in complexity that is due to the affine conditional probabilities.

To write $\lambda = \lambda_q$ explicitly, we first denote by

$$\rho_s(x) = \int \rho_s(x|x_1) q(x_1) dx_1 \quad (22)$$

the result of adding Gaussian noise, $\rho_s(x|\mu) = \mathcal{N}(x|\mu, s^{-2}I)$, of scale $s^{-1}$, to the data $q$. Now with Bayes' rule we compute the probability of data point $x_1$ given a noisy data point $x$,

$$\rho_s(x_1|x) = \frac{\rho_s(x|x_1) q(x_1)}{\rho_s(x)}.$$

Then the data separation function is

$$\lambda(s) = \frac{1}{d} \mathbb{E}_{\rho_s(x)} \left\| \mathbb{E}_{\rho_s(x_1|x)} x_1 \right\|^2. \quad (23)$$

See Appendix A.1 for detailed derivation of the $\lambda$ function.

To understand the data separation function intuitively, consider discrete data $q$ which consists of data points $x_i, i \in [n]$: if this data is well separated at scale $s^{-1}$, then a noisy sample $x = x_i + \epsilon \sim \rho_s(x)$ will be closer to the data point $x_i$ than any other data point $x_j$. Hence, the softmax term $\rho_s(x_1|x) \approx \delta_{x_i}(x)$, where $\delta_{x_i}(x)$ is the delta distribution centered at $x_i$. Therefore in this case

$$\lambda(s) \approx \frac{1}{d} \mathbb{E}_{\rho_s(x|x_1), q(x_1)} \|x_1\|^2 = \frac{1}{d} \mathbb{E}_{q(x_1)} \|x_1\|^2 = 1,$$

where in the last equality we used our data assumption in equation 8. So, the closer to 1 the function value $\lambda(s)$ is,

$s \in [0, \infty]$, the more separated $q$ at scale $s^{-1}$. Note however that for data with average variance of 1 (see equation 8 again), $\lambda$ is globally bounded by 1. Indeed using Jensen's inequality we have

$$\lambda(s) \leq \frac{1}{d} \mathbb{E}_{\substack{\rho_s(x) \\ \rho_s(x_1|x)}} \|x_1\|^2 = \frac{1}{d} \mathbb{E}_{q(x_1)} \|x_1\|^2 = 1. \quad (24)$$

### 3.2. Kinetic optimal solutions

In this section we characterize kinetic optimal paths within the space of Gaussian paths. These solution will be depending on the data separation function $\lambda$.

To facilitate the variational analysis of the KE functional in equation 21 we perform a change of variables and represent the conditional probability parameters $(a_t, m_t) \in \mathcal{A}$ via

$$\begin{bmatrix} a_t \\ m_t \end{bmatrix} = r_t \begin{bmatrix} \sin(\theta_t) \\ \cos(\theta_t) \end{bmatrix}, \quad (25)$$

where now our optimization space of affine conditionals becomes

$$\mathcal{B} = \left\{ (r_t, \theta_t) \,\middle|\, \begin{matrix} r_t:[0,1]\to[0,\infty) \\ \theta_t:[0,1]\to[0,\frac{\pi}{2}] \\ r_0=1=r_1 \\ \theta_0=0, \theta_1=\frac{\pi}{2} \end{matrix} \right\}. \quad (26)$$

Plugging these coordinates in equation 21 leads to the following form of the KE:

$$\mathcal{E}(r_t, \theta_t) = \int_0^1 \dot{r}_t^2 + r_t^2 \dot{\theta}_t^2 \overbrace{\left[ 1 - \frac{1 - \lambda(\tan(\theta_t))}{\cos^2(\theta_t)} \right]}^{\gamma(\theta_t)} dt. \quad (27)$$

The stationary solutions of equation 27 obey the Euler-Lagrange equations, which is a set of two ordinary differential equations (ODEs) with two unknown functions $r_t, \theta_t$. Surprisingly, it turns out that under the mild extra condition, i.e., $\gamma(\theta_t) > 0$, this system of equations is separable and reduces to two rather simple ODEs, with $r_t$ solvable analytically up to a single unknown parameter, and $\theta_t$ is characterized with a first order ODE:

**Theorem 3.1.** *The minimizer of the Kinetic Energy (equation 27) over all conditional probabilities $(r_t, \theta_t) \in \mathcal{B}$ (equation 26) such that $\gamma(\theta_t) > 0$ for $t > 0$ satisfy*

$$r_t = \sqrt{1 - bt + bt^2}, \quad (28)$$

$$\dot{\theta}_t = \frac{1}{1 - bt + bt^2} \sqrt{\frac{b - \frac{1}{4}b^2}{\gamma(\theta_t)}} \quad (29)$$

*where $b \in [0, 4]$ is determined by the boundary condition on $\theta_t$.*

The proof of this theorem is provided in Appendix A.3. To better grasp the meaning of the condition $\gamma(\theta_t) > 0$ we

show in Appendix A.2 that it is equivalent to

$$\lambda(s) \geq \frac{s^2}{1 + s^2}, \quad s \in [0, \infty]. \tag{30}$$

This implies a condition on the target data distribution $q$ for which our optimality result holds. To get a better sense of this condition in Appendix A.2 we also prove that if $q(x) = \mathcal{N}(0, I)$, namely a standard Gaussian, then $\lambda(s) = \frac{s^2}{1+s^2}$. Hence, the condition in equation 30 requires the separation of $q$ to be at-least that of a standard Gaussian. All the empirical data $q$ we tested in this paper has at-least the separation of a standard Gaussian, namely satisfy equation 30.

**Conditional kinetic optimal solutions.** One case where the optimal solution is known analytically is when $\gamma(\theta_t) \equiv 1$, which happens when $\lambda(s) \equiv 1$. In this case, equation 21 shows that the Kinetic Energy and the Conditional Kinetic Energy are equal, that is $\mathcal{E} = \mathcal{E}_c$. Consequently solving the Euler-Lagrange equations can be done directly for equation 20 and the solutions are the Cond-OT conditionals in equation 14. Alternatively, Theorem 3.1 can be used and equation 29 can be solved to get $\theta_t$ as well as the parameter $b$ leading to the solution

$$r_t = \sqrt{(1-t)^2 + t^2} \quad \theta_t = \arctan\left(\frac{t}{1-t}\right)$$

and using equation 25 to solve for $a_t, m_t$ we get again the Cond-OT conditionals in equation 14.

**Non-analytic solutions.** When an analytic solution of equation 29 cannot be achieved one can approximate the data separation function $\lambda$ numerically, and then optimize the Kinetic Energy $\mathcal{E}$ directly for $r_t, \theta_t$, where for $r_t$ it is convenient to use the one parameter solution family in equation 28. We come back to this procedure in the experiments section.

## 4. Kinetic optimal paths in high dimensions

Finding kinetic optimal conditional paths $(a_t, m_t)$ in closed form for arbitrary data $q$ is usually hard. However, interestingly, as the dimension of the data increases compared to the number of data samples, a kinetic optimal path can be characterized, at least asymptotically, to be paths defined with the Cond-OT conditional (equation 14). To facilitate this discussion consider again equation 21 that provides

$$\mathcal{E}_c(a_t, m_t) - \mathcal{E}(a_t, m_t) = \int_0^1 \beta_t^2 \left(1 - \lambda\left(\frac{a_t}{m_t}\right)\right) dt. \tag{31}$$

As discussed above and can be seen from this equation when the data separation function achieves its upper bound, $\lambda(s) \equiv 1$, the KE coincides with the CKE, and in this case

the optimal path is defined by the Cond-OT conditional (equation 14). However, it is not a-priori clear if $\lambda$ actually approaches 1 for any data $q$, and whether closeness of $\lambda$ to 1 implies Cond-OT is asymptotically kinetic optimal. In this section we show that for *arbitrary* finite normalized dataset consisting of $n$ samples in $\mathbb{R}^d$, indeed $\lambda \to 1$ (in $L^1$ sense) as $n/\sqrt{d} \to 0$, that can be used to show that the KE of the path defined by the Cond-OT conditional is asymptotically optimal. In the experiments section we also verify this property empirically for real-world dataset in high dimensions (i.e., ImageNet).

**Normalized finite data and main results.** We consider an arbitrary *normalized finite data distribution* $q$. That is, $q$ is distributing equal mass (i.e., $n^{-1}$) among $n$ points $x_i \in \mathbb{R}^d$, $i \in [n]$, and is satisfying equations 7 and 8, which in the discrete case means

$$\frac{1}{n} \sum_{i=1}^n x_i = 0, \tag{32}$$

$$\frac{1}{dn} \sum_{i=1}^n \|x_i\|^2 = 1. \tag{33}$$

Our first result considers such data distribution $q$ and shows convergence of $\lambda$ to the constant function 1 as $n/\sqrt{d} \to 0$:

**Theorem 4.1.** *Let $q$ be an arbitrary normalized finite data distribution. Then,*

$$\int_0^\infty |1 - \lambda(s)| \, ds \leq \frac{3n}{\sqrt{d}}. \tag{34}$$

This theorem shows that essentially data separation is inevitable in high dimensions and that we can expect $\lambda$ to approach 1 for finite data in sufficiently high dimension. Using Theorem 4.1 in equation 31 we prove the following theorem, under some mild extra assumption over the family $\mathcal{A}$ of affine conditional probabilities (equation 11):

**Theorem 4.2.** *Let $q$ be an arbitrary normalized finite data distribution. Assume all $(a_t, m_t) \in \mathcal{A}$ in equation 11 satisfy: (i) $\frac{a_t}{m_t}$ is strictly monotonically increasing; and (ii) $a_t, m_t$ have uniform bounded Sobolev $W^{1,\infty}([0,1])$ norm[2], i.e., $\|a_t\|_{1,\infty}, \|m_t\|_{1,\infty} \leq M$. Then, for all $(a_t, m_t) \in \mathcal{A}$*

$$\mathcal{E}_c(a_t, m_t) - 6M^2 \frac{n}{\sqrt{d}} \leq \mathcal{E}(a_t, m_t) \leq \mathcal{E}_c(a_t, m_t). \tag{35}$$

This shows that the KE of any path defined by a conditional $(a_t, m_t) \in \mathcal{A}$ will converge to the CKE as $n/\sqrt{d} \to 0$. Since a probability path defined by Cond-OT is minimizing the CKE, we see that Cond-OT is asymptotically kinetic optimal as $n/\sqrt{d} \to 0$. We summarize:

---

[2]The Sobolev $W^{1,\infty}([0,1])$ norm is defined as $\|f\|_{1,\infty} = \max\left\{\max_{t \in [0,1]} |f(t)|, \max_{t \in [0,1]} |\dot{f}(t)|\right\}$
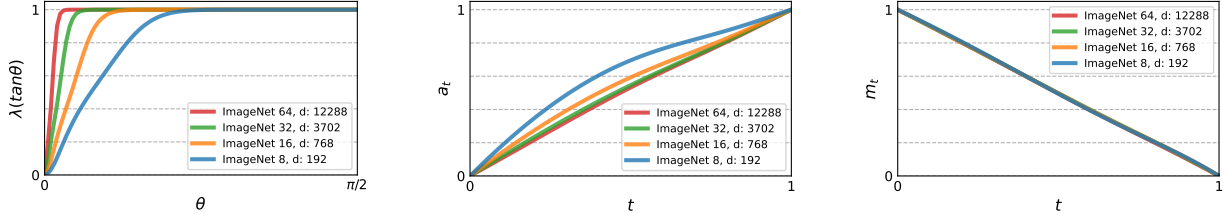
*Figure 1.* (Left) Estimation of the data separation function $\hat{\lambda}$ for ImageNet-8/16/32/64; note the convergence to 1 as data dimension increases, as anticipated by Theorem 4.1. (Middle) depicts the optimal $a_t$; and (right) the optimal $m_t$ for ImageNet-8/16/32/64; note convergence to the linear solution (Cond-OT).

**Corollary 4.3.** *The Gaussian probability path defined by the conditional Cond-OT (equation 14) is asymptotically kinetic optimal as $n/\sqrt{d} \to 0$.*

A comment regrading the first assumption on $\mathcal{A}$: note that this assumption is equivalent to monotonicity of the signal-to-noise-ratio (SNR), $a_t^2/m_t^2$, which is a natural assumption on conditional paths (Kingma et al., 2021a).

**Proof idea.** The complete proofs for Theorems 4.1 and 4.2 are provided in Appendix A.5; here we provide the proof idea for Theorem 4.1. We first formulate $\lambda$ (equation 23) for the finite data case. That is

$$\lambda(s) = \frac{1}{dn} \sum_{j=1}^{n} \mathbb{E}_{\rho_s(x|x_j)} \left\| \sum_{i=1}^{n} x_i \rho_s(x_i|x) \right\|^2, \qquad (36)$$

where using Bayes' rule we have

$$\rho_s(x_i|x) = \frac{\rho_s(x|x_i)q(x_i)}{\sum_{l=1}^{n} q(x_l)\rho_s(x|x_l)} = \frac{e^{-\frac{s^2}{2}\|x_i-x\|^2}}{\sum_{l=1}^{n} e^{-\frac{s^2}{2}\|x_l-x\|^2}}.$$

That is, $\rho_s(x_i|x)$ is the $i$-th entry of a weighted Softmax (with $s^2$ weight) of the squared Euclidean distances between a point $x$ and all data points $x_i$. The key observation is that for $x \sim \rho_s(x|x_j)$, a noisy sample of $x_j$, the probability $\rho_s(x_i|x)$ is bounded above in expectation by a function that depends only on the pairwise distances, i.e., $\|x_i - x_j\|$. Furthermore, this function is decaying exponentially with these distances. That is, if the data is "separated" in the sense that $\|x_i - x_j\|$ is large then the data separation function $\lambda$ will be closer to its maximal value of 1.

**Lemma 4.4.** *Let $q$ be an arbitrary normalized finite data distribution. Then for every $s > 0$,*

$$\mathbb{E}_{\rho_s(x|x_j)} \rho_s(x_i|x) \leq \eta \left( s \|x_i - x_j\| \right), \qquad (37)$$

*where $\eta(t)$ is integrable in $[0, \infty)$ and*

$$\int_0^{\infty} \eta(t) \leq 3. \qquad (38)$$

This lemma is proved in Appendix A.5 as-well. Let us use it to prove our theorem. Using the Cauchy–Schwarz inequality in equation 36 we can show that

$$1 - \lambda(s) \leq \frac{1}{dn} \sum_{i=1}^{n} \|x_i\| \sum_{\substack{j=1 \\ j \neq i}}^{n} \left( \|x_j - x_i\| \, \mathbb{E}_{\rho_s(x|x_i)} \rho_s(x_j|x) \right).$$

Now, the integral w.r.t. $s$ of the term in the parenthesis can be bounded after a change of coordinates $r = s \|x_i - x_j\|$ using Lemma 4.4 by 3. Lastly, using Jensen and equation 33 we have that $\sum_i \|x_i\| \leq n\sqrt{d}$, which provides equation 34.

## 5. Related work

Continuous Normalizing Flows (CNFs) train a flow modeled by learnable vector field $v_t(x)$ (Chen et al., 2018), where traditionally $v_t(x)$ was optimized to reduce the likelihod of the train data. Other CNF methods, have tried to train CNFs by combining the Kinetic Energy as a regularization term alongside a likelihood term to push the resulting path toward kinetic optimal solutions (Finlay et al., 2020; Onken et al., 2021); however, these still require log probabilities during training and the combination of the losses does not guarantee to produce a globally kinetic optimal solution, which is the Optimal Transport (Villani, 2009). Computing the log probabilities and their derivatives during training placed a roadblock on scaling CNFs to high dimensions (Grathwohl et al., 2018). Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020) can be seen as an alternative way to train CNFs with a loss that sidesteps log probabilities and regresses the score of an a-priori defined probability density path $p_t \in \mathcal{P}$ (see equation 2). Recently, (Lipman et al., 2022; Albergo & Vanden-Eijnden, 2022; Liu et al., 2022; Neklyudov et al., 2022) suggest to train a CNF by directly regressing the generating vector field of a probability density paths $p_t \in \mathcal{P}$. The probability paths utilized in these methods belong to the family of paths considered in this paper, namely the probability paths that marginalize affine conditional per-sample probabilities, and from which we characterize the kinetic optimal paths. Lastly, (Albergo & Vanden-Eijnden, 2022) suggested optimizing the kinetic
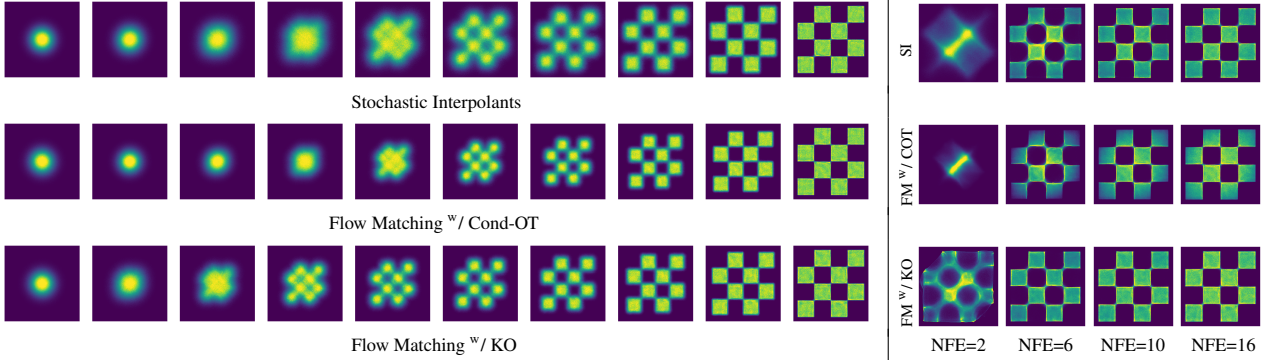
*Figure 2.* Left: Trajectories of trained CNF models with Stochastic Interpolants, Cond-OT, and Kinetic Optimal paths trained with 2D checkerboard data. Right: Results of the same models when sampling with limited number of function evaluations (NFE). This is a case of low dimensional data (i.e., $d = 2$) where Cond-OT is not kinetic optimal.

energy simultaneously to training the CNF, resulting in a challenging high dimensional min-max problem; we consider finding the kinetic optimal path used as supervision in the CNF training in separation.

## 6. Experiments

In this section we: (i) approximate the data separation function $\lambda$ numerically for different real-world datasets, and approximate the corresponding Kinetic Optimal (KO) paths for these datasets. (ii) We validate our theory of the convergence of $\lambda \to 1$ in high dimensions for real-world data. (iii) We empirically test our KO paths compared to paths defined by the conditional probabilities of Cond-OT (Lipman et al., 2022), IS (Albergo & Vanden-Eijnden, 2022), and DDPM (Ho et al., 2020). In terms of datasets, we have been experimenting with a 2D dataset (checkerboard), and the image datasets CIFAR10, ImageNet-32, and Imagenet-64.

### 6.1. Approximation of data separation function

A simple approach to approximate $\lambda$ in equation 23 for a given dataset $x_i \in \mathbb{R}^d$, $i \in [n]$, is to consider normalized finite data $q$ (see Section 4), i.e., giving each $x_i$ the probability $n^{-1}$, normalized in the sense of equation 7 and equation 8. In this case $\lambda$ takes the form of equation 36, and we use Monte-Carlo estimation of the expectation w.r.t. $x \sim \rho_s(x|x_j)$, $j \in [n]$, where we can express $x$ as $x = x_j + s^{-1}z$, and $z \sim \mathcal{N}(0, I)$. We sample $z_l \sim \mathcal{N}(0, I)$, $l \in [k]$ i.i.d. and our estimator becomes

$$\hat{\lambda}(s) = \frac{1}{nd} \sum_{l=1}^{k} \sum_{j=1}^{n} \left\| \sum_{i=1}^{n} x_i \rho_s(x_i | x_j + s^{-1}z_l) \right\|^2. \quad (39)$$

Unfortunately, computing this sum requires the pairwise distances $\|x_i - x_j\|^2$ and therefore direct computation scales quadratically in $n$. To visualize $\hat{\lambda} : [0, \infty) \to [0, 1]$ compactly we instead show $\hat{\lambda} \circ \tan : [0, \pi/2] \to [0, 1]$, so the

$x$-axis is annotated with $\theta$. Figure 4 (left) shows the difference in the estimated lambda (computed with equation 39) for ImageNet-32 using an increasing number of samples from $q$ starting at $n = 10K$ to $n = 100K$ while keeping the number of samples from $\mathcal{N}(0, I)$ constant at $k = 10K$. As can be inspected in this figure, beyond $50K$ the differences are rather small. Note that we display the range $\theta \in [0, 0.3]$ since beyond that all $\hat{\lambda}$ approximations practically reach 1, which is also the upper bound for $\lambda$. Similarly, in Figure 4 (right), we keep the number of samples from $q$ constant at $n = 50K$ and vary the number of samples $k$ from $\mathcal{N}(0, I)$, $k \in [0.3K, 10K]$, again noticing that the differences between the estimators are negligible. We approximate $\hat{\lambda}(\tan(\theta))$ on equispaced grid points $\theta_i = 0.01\frac{\pi}{2}i$, $i \in [0, 100]$ and fit a cubic-spline to this data to be used for estimating $\hat{\lambda}(\tan \theta)$ at arbitrary $\theta \in [0, \frac{\pi}{2}]$.

**Optimizing for kinetic optimal paths.** Given an estimator of $\hat{\lambda}$, we minimize the KE energy in equation 27 as a functional of $(r_t, \theta_t)$. We model $r_t$ with the derived one-parameter solutions family in Theorem 3.1. For $\theta_t$, we use a 10 parameters neural network, and optimize the KE with general purpose gradient descent, see details in Appendix B. Lastly, we apply the coordinate change in equation 25 and transform the optimal $r_t, \theta_t$ back to $a_t, m_t$ that is used for training. Figure 3 shows the kinetic optimal paths computed for all datasets considered in the paper. It is already obvious in this figure that as the dimension of the data increases the kinetic optimal paths get closer to the Cond-OT path, anticipated in Theorem 4.2. We discuss this next.

### 6.2. Kinetic optimal paths in high dimensions

The high dimensional phenomenon presented in Theorem 4.1, namely that $\lambda \to 1$ for discrete data in high dimensions, does in fact manifests in real-world high dimensional datasets such as ImageNet, even though the ratio $\frac{n}{\sqrt{d}}$ is not small for the relevant dimensions. Figure 1 (left)
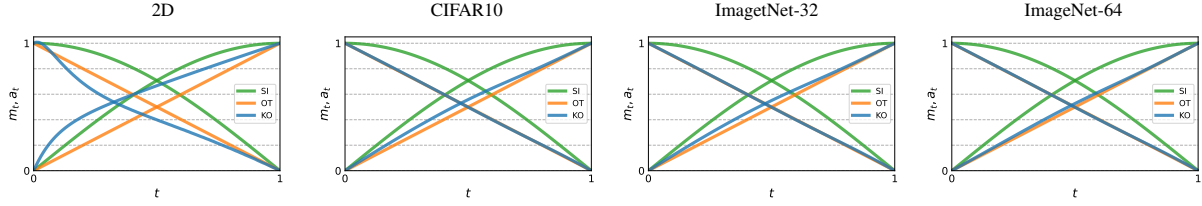
**Figure 3.** The kinetic optimal paths (KO, blue), computed with our algorithm for all datasets considered in the paper, are shown alongside the SI path (green) and Cond-OT path (orange). The KO paths get closer to Cond-OT at high dimensions; for ImageNet-64 the KO and Cond-OT almost coincide.

| | CIFAR-10 | | | | | | ImageNet 32×32 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | NLL↓ | KE↓ | FID$_{20}$↓ | FID$_{50}$↓ | FID$_{100}$↓ | FID$_{Adpt}$↓ | NFE$_{Adpt}$↓ | NLL↓ | KE↓ | FID$_{20}$↓ | FID$_{50}$↓ | FID$_{100}$↓ | FID$_{Adpt}$↓ | NFE$_{Adpt}$↓ |
| *Ablations* | | | | | | | | | | | | | |
| DDPM | 3.11 | 2.30 | $38.5 \pm 0.2$ | $18.0 \pm 0.0$ | $12.5 \pm 0.1$ | $7.4 \pm 0.1$ | $344 \pm 0$ | 3.61 | 2.22 | $66.1 \pm 1.2$ | $15.7 \pm 0.3$ | $8.6 \pm 0.3$ | $5.1 \pm 0.1$ | $333 \pm 2$ |
| SI | 3.00 | 1.30 | $10.4 \pm 0.1$ | $7.4 \pm 0.1$ | $7.4 \pm 0.1$ | $7.5 \pm 0.0$ | $165 \pm 0$ | 3.58 | 1.27 | $23.4 \pm 0.4$ | $9.0 \pm 0.3$ | $6.7 \pm 0.2$ | $5.3 \pm 0.2$ | $158 \pm 2$ |
| FM $^{w/}$ OT path | 2.97 | 1.09 | $9.9 \pm 0.0$ | $6.8 \pm 0.0$ | $6.8 \pm 0.0$ | $6.9 \pm 0.0$ | $146 \pm 0$ | 3.60 | 1.08 | $9.7 \pm 0.1$ | $6.5 \pm 0.1$ | $5.6 \pm 0.0$ | $5.0 \pm 0.1$ | $145 \pm 0$ |
| FM $^{w/}$ KO path | 2.97 | **1.08** | $9.1 \pm 0.0$ | $6.2 \pm 0.1$ | $6.2 \pm 0.1$ | $6.1 \pm 0.0$ | $147 \pm 0$ | 3.57 | **1.07** | $9.8 \pm 0.3$ | $6.6 \pm 0.4$ | $5.8 \pm 0.3$ | $5.1 \pm 0.3$ | $129 \pm 1$ |

**Table 1.** For the datasets of CIFAR-10 and ImageNet-32 we log: Negative log likelihoods in BPD units; Kinetic Energy (KE); quality of generated samples in FID computed with Euler method for 20,50,100 steps and the adaptive method DOPRI5; and Number of Function Evaluation (NFE) with the adaptive solver. The KE, FID and NFE are averaged over three checkpoints (epochs 1990, 1995, 2000 for CIFAR-10; and epochs 390, 395, 400 for ImageNet-32).
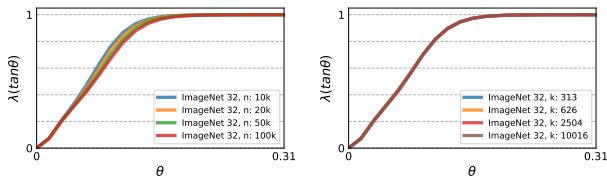


**Figure 4.** Evaluation of data separation estimation: (Left) shows $\hat{\lambda}$ for Imagenet-32 with $n = 10K, 20K, 50K, 100K$ data samples, and $k = 10K$ samples from $\mathcal{N}(0, I)$. (Right) shows $\hat{\lambda}$ for Imagenet-32 with $n = 50k$ data samples and $k = 0.3K, 0.6K, 2.5K, 10K$ samples from $\mathcal{N}(0, I)$. Note the $x$-axis is limited to $[0, 0.31]$ to zoom in on the relevant part.

shows $\hat{\lambda}$ for ImageNet-8/16/32/64, which have dimensions 192/768/3072/12,228 (resp.) computed with $n = 50K$ and $k = 10K$. We provide a detailed run-times discussion in Appendix D. As can be seen in this figure, as the image dimension increases indeed $\lambda \to 1$. In particular, for ImageNet-64, $\hat{\lambda}$ converged to 1 for every $\theta \in [0.08, \frac{\pi}{2}]$. (Middle),(Right) Show the corresponding kinetic optimal paths $a_t$, $m_t$ (resp.); the convergence of the KO paths to Cond-OT as data dimension increases is suggested by theorem 4.2 and the fact that Cond-OT is the minimizer of the CKE. In particular, the KO optimal path for ImageNet-64 is practically identical to the Cond-OT linear path.

### 6.3. Flow Matching with kinetic optimal paths

We also experimented with kinetic optimal (KO) paths for training a continuous normalizing flows. We use the Flow

Matching (FM) approach (Lipman et al., 2022) where (in the notations of equation 1),

$$\mathcal{L}(\theta) = \mathbb{E}_{t, p_t(x)} \|v_t(x; \theta) - u_t(x)\|^2, \quad (40)$$

where $v_t$ is a neural network with learnable parameters $\theta$, $p_t(x)$, $u_t(x)$ are as defined in equations 9, 17 (resp.) and $t$ is uniform in $[0, 1]$. The actual training objective is

$$\mathcal{L}(\theta) = \mathbb{E}_{t, p_t(x|x_1), q(x_1)} \|v_t(x; \theta) - u_t(x|x_1)\|^2, \quad (41)$$

where $p_t(x|x_1)$ is as defined in equation 10 and $u_t(x|x_1)$ as defined in 15, which is shown in (Lipman et al., 2022) to have equivalent gradients to the loss in equation 40. For $a_t, m_t$ we use the optimal solutions of the approximated kinetic energy (see Section 6.1); see Figure 3 for visualizations of these KO paths.

**2D data.** We tested FM-KO first on 2D data of a checkerboard distribution, often used for low dimensional test-bed for CNF models. Figure 2 compares FM-KO, with FM-Cond-OT, and IS: On the left we show the time sequence, $t \in [0, 1]$, of the generated probabilities starting from the standard noise $p_0 = p$. Note that KO is distributed more evenly in time. On the right we show sampling with limited number of function evaluations (NFE) using the Euler solver. Note that KO outperforms the baselines for low NFE counts. In this case the data dimension is low ($d = 2$) and the dataset contains many (more then 1M) data points, which is outside the scope of our Theory that is effective as $n/\sqrt{d}$ is smaller. Indeed, in this case Cond-OT is not kinetic optimal and is being outperformed in terms of NFE counts.
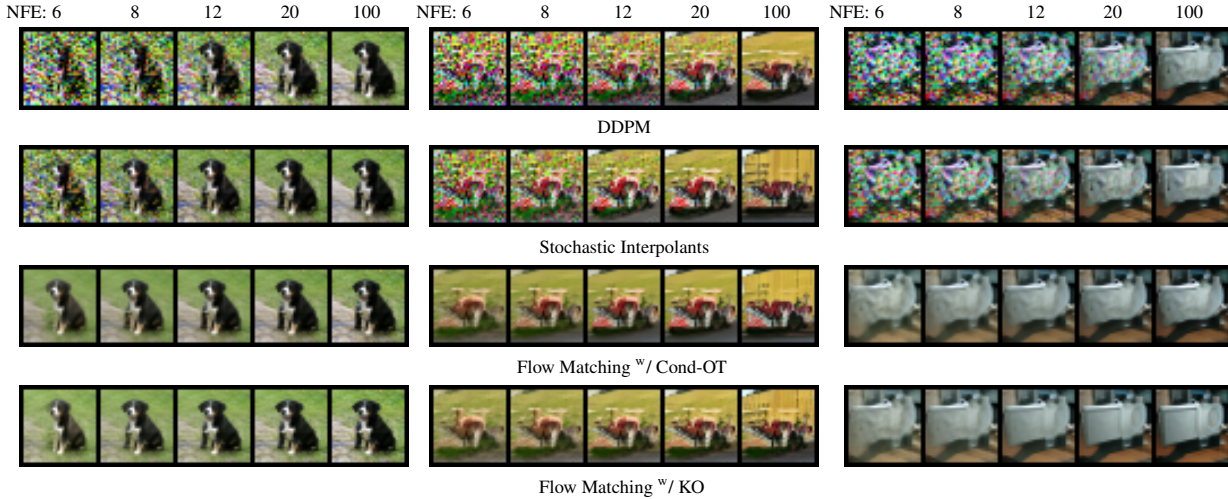
*Figure 5.* Generated images of CNF models trained with DDPM, SI, Cond-OT and Kinetic Optimal paths on *face-blurred* ImageNet-32 data. Generation is done with the Euler method, where in each column all images are generated from the same noise but with a different number of steps (equivalent to NFE for Euler solver) as indicated; 6, 8, 12, 20 and 100 steps.
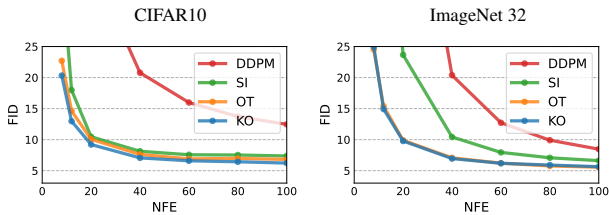


*Figure 6.* FID versus NFE for models trained on the CIFAR-10 (left) and ImageNet-32 (right) datasets. Samples for FID computation are generated with Euler solver.

**Image datasets.** Lastly, we tested KO optimal paths on CIFAR and ImageNet-32 (for ImageNet-64 KO paths practically coincides with Cond-OT). For ImageNet we use the official *face-blurred* ImageNet and downsample to $32 \times 32$ using an open source preprocessing script (Chrabaszcz et al., 2017). Quantitative results are summarizes in Table 1 while qualitative results are depicted in Figure 5; for each method we evaluated Negative Log Likelihoods in Bit-Per-Dimension (BPD); kinetic energy; and FID computed for samples generated by Euler method with 20, 50, 100 steps and the adaptive solver DOPRI5. For FID and NFE (adaptive) we averaged 3 different epochs at the final stage of training. For these datasets, we find that KO and Cond-OT paths already produce similar KE, better than the other baselines considered. Figure 6 depicts curves of generated image quality (FID) as a function of number of function evaluation (NFE) for CIFAR10 ImageNet-32. We provide example of generated images from the KO model in Appendix E.

## 7. Conclusions

In this paper we investigated the space of tractable probability paths, which are used to supervise generative models' training, and searched for an optimal path that minimizes the Kinetic Energy. We started with identifying a more tractable form for the kinetic energy that incorporates the training data using a simple, one dimensional data separation function $\lambda$. We then characterized the kinetic optimal solutions and suggested a method to compute them using a numerical estimation to $\lambda$. The kinetic optimal paths improved performance of the respectively trained generative models mostly in low to medium data dimensions. Lastly, we demonstrated through a theoretical analysis accompanied with empirical experiments that as the dimension of the data increases, the simple Cond-OT probability paths are becoming kinetic optimal. That is, the kinetic energy of paths defined with the Cond-OT conditional is converging to the optimal kinetic energy in a family of probability paths. This result has several implications for future research: First, further searching the class Gaussian paths defined by affine conditionals (equations 9, 10 and 11) for useful probability path might be futile. In fact, in high dimensions, Cond-OT might be a close to optimal choice within this class, for most datasets. Finding probability path with even lower Kinetic Energy would entail abandoning the affine conditional probabilities and searching within a broader class. For example, replacing the affine model (equation 10) with a model that has more complex dependency on $x_1$ might allow further improvement of the kinetic energy and consequently the trained generative model.

# References

Albergo, M. S. and Vanden-Eijnden, E. Building normalizing flows with stochastic interpolants, 2022.

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Chrabaszcz, P., Loshchilov, I., and Hutter, F. A downsampled variant of ImageNet as an alternative to the Cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Finlay, C., Jacobsen, J.-H., Nurbekyan, L., and Oberman, A. How to train your neural ode: the world of jacobian and kinetic regularization. In *International conference on machine learning*, pp. 3154–3164. PMLR, 2020.

Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., and Taigman, Y. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022.

Grathwohl, W., Chen, R. T., Bettencourt, J., Sutskever, I., and Duvenaud, D. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

Kingma, D., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021a.

Kingma, D. P., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021b. URL https://openreview.net/forum?id=2LdBqxc1Yv.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

McCann, R. J. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

Neklyudov, K., Severo, D., and Makhzani, A. Action matching: A variational method for learning stochastic dynamics from samples. *arXiv preprint arXiv:2210.06662*, 2022.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Onken, D., Fung, S. W., Li, X., and Ruthotto, L. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9223–9232, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021. URL https://arxiv.org/abs/2112.10752.

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

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Villani, C. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

# A. Auxiliary computations and proofs

## A.1. Kinetic energy and $\lambda$

*Proof of equation 20.* We compute the Conditional Kinetic Energy (CKE) $\mathcal{E}_c(a_t, m_t)$. By definition, it is

$$\mathcal{E}_c(a_t, m_t) = \frac{1}{d}\mathbb{E}_{t,q(x_1),p_t(x|x_1)} \left\| u_t(x|x_1) \right\|^2, \tag{42}$$

where $u_t(x|x_1)$ is as defined in equation 15. That is

$$u_t(x|x_1) = \alpha_t x + \beta_t x_1 \tag{43}$$

and

$$\alpha_t = \frac{\dot{m}_t}{m_t}, \quad \beta_t = \dot{a}_t - a_t \frac{\dot{m}_t}{m_t}. \tag{44}$$

We remind that $p_t(x|x_1) = \mathcal{N}(x|a_t x_1, m_t^2 I)$ and $p(x) = \mathcal{N}(x|0, I)$. Hence, for $x \sim p(x)$, $m_t x + a_t x_1 \sim p_t(x|x_1)$. We can now compute the expectation w.r.t. $p_t(x|x_1)$ and $q(x_1)$. That is,

$$\frac{1}{d}\mathbb{E}_{t,q(x_1),p_t(x|x_1)} \left\| u_t(x|x_1) \right\|^2 = \frac{1}{d}\mathbb{E}_{t,q(x_1),p_t(x)} \left\| u_t(m_t x + a_t x_1|x_1) \right\|^2 \tag{45}$$

$$= \frac{1}{d}\mathbb{E}_{t,q(x_1),p_t(x)} \left\| \dot{m}_t x + \dot{a}_t x_1 \right\|^2 \tag{46}$$

$$= \frac{1}{d}\mathbb{E}_{t,q(x_1),p_t(x)} \left( \dot{m}_t^2 \left\| x \right\|^2 + 2\dot{m}_t \dot{a}_t x \cdot x_1 + \dot{a}_t^2 \left\| x_1 \right\|^2 \right) \tag{47}$$

$$= \frac{1}{d}\mathbb{E}_{t,q(x_1)} \left( \dot{m}_t^2 d + \dot{a}_t^2 \left\| x_1 \right\|^2 \right) \tag{48}$$

$$= \frac{1}{d}\mathbb{E}_t \left( \dot{m}_t^2 d + \dot{a}_t^2 d \right) \tag{49}$$

$$= \int_0^1 (\dot{m}_t^2 + \dot{a}_t^2) dt, \tag{50}$$

where before the last equality we used equation 8 to compute the expectation on $q(x_1)$.

*Proof of equations 21 and 23.* We compute the Kinetic Energy (KE) $\mathcal{E}(a_t, m_t)$, proving equation 21. By definition, it is

$$\mathcal{E}(a_t, m_t) = \frac{1}{d}\mathbb{E}_{t,p_t(x)} \left\| u_t(x) \right\|^2, \tag{51}$$

where $u_t(x)$ is as defined in equation 17. Equivalently, using Bayes' rule, we can write it as

$$u_t(x) = \mathbb{E}_{p_t(x_1|x)} u_t(x|x_1) \tag{52}$$

and

$$p_t(x_1|x) = \frac{p_t(x|x_1)q(x_1)}{p_t(x)}. \tag{53}$$

Now we can calculate the KE as follows

$$\frac{1}{d}\mathbb{E}_{t,p_t(x)} \left\| u_t(x) \right\|^2 = \frac{1}{d}\mathbb{E}_{t,p_t(x)} \left\| \mathbb{E}_{p_t(x_1|x)} u_t(x|x_1) \right\|^2 \tag{54}$$

$$= \frac{1}{d}\mathbb{E}_{t,p_t(x)} \left\| \alpha_t x + \beta_t \mathbb{E}_{p_t(x_1|x)} x_1 \right\|^2 \tag{55}$$

$$= \frac{1}{d}\mathbb{E}_{t,p_t(x)} \left( \alpha_t^2 \left\| x \right\|^2 + 2\alpha_t \beta_t x \cdot \mathbb{E}_{p_t(x_1|x)} x_1 + \beta_t^2 \left\| \mathbb{E}_{p_t(x_1|x)} x_1 \right\|^2 \right) \tag{56}$$

$$= \frac{1}{d}\mathbb{E}_{t,p_t(x)} \left( \alpha_t^2 \left\| x \right\|^2 + 2\alpha_t \beta_t x \cdot \mathbb{E}_{p_t(x_1|x)} x_1 + \beta_t^2 d - \beta_t^2 d + \beta_t^2 \left\| \mathbb{E}_{p_t(x_1|x)} x_1 \right\|^2 \right) \tag{57}$$

$$= \frac{1}{d}\mathbb{E}_{t,p_t(x)} \left( \alpha_t^2 \left\| x \right\|^2 + 2\alpha_t \beta_t x \cdot \mathbb{E}_{p_t(x_1|x)} x_1 + \beta_t^2 d \right) - \frac{1}{d}\mathbb{E}_{t,p_t(x)} \left( \beta_t^2 d - \beta_t^2 \left\| \mathbb{E}_{p_t(x_1|x)} x_1 \right\|^2 \right) \tag{58}$$

We perform the expectation over $p_t(x)$ for each term on the r.h.s. separately. We invoke Bayes' rule again, which gives

$$p_t(x)p_t(x_1|x) = q(x_1)p_t(x|x_1). \tag{59}$$

Hence, the first term is

$$\frac{1}{d}\mathbb{E}_{t,p_t(x)}\left(\alpha_t^2 \|x\|^2 + 2\alpha_t\beta_t x \cdot \mathbb{E}_{p_t(x_1|x)}x_1 + \beta_t^2 d\right) = \frac{1}{d}\mathbb{E}_{t,p_t(x),p_t(x_1|x)}\left(\alpha_t^2 \|x\|^2 + 2\alpha_t\beta_t x \cdot x_1 + \beta_t^2 d\right) \tag{60}$$

$$= \frac{1}{d}\mathbb{E}_{t,q(x_1),p_t(x|x_1)}\left(\alpha_t^2 \|x\|^2 + 2\alpha_t\beta_t x \cdot x_1 + \beta_t^2 d\right) \tag{61}$$

$$= \frac{1}{d}\mathbb{E}_{t,q(x_1),p_t(x|x_1)}\left(\alpha_t^2 \|x\|^2 + 2\alpha_t\beta_t x \cdot x_1 + \beta_t^2 \|x_1\|^2\right) \tag{62}$$

$$= \frac{1}{d}\mathbb{E}_{t,q(x_1),p_t(x|x_1)}\|\alpha_t x + \beta_t x_1\|^2 \tag{63}$$

$$= \frac{1}{d}\mathbb{E}_{t,q(x_1),p_t(x|x_1)}\|u_t(x|x_1)\|^2 \tag{64}$$

$$= \mathcal{E}_c(a_t, m_t), \tag{65}$$

where we also used equation 8 for the expectation of $\|x_1\|^2$. Next, we define $\lambda$

$$\lambda\left(\frac{a_t}{m_t}\right) = \frac{1}{d}\mathbb{E}_{p_t(x)}\left\|\mathbb{E}_{p_t(x_1|x)}x_1\right\|^2. \tag{66}$$

So the second term is

$$\frac{1}{d}\mathbb{E}_{t,p_t(x)}\left(\beta_t^2 d - \beta_t^2 \left\|\mathbb{E}_{p_t(x_1|x)}x_1\right\|^2\right) = \mathbb{E}_t\left(\beta_t^2 - \beta_t^2\lambda\left(\frac{a_t}{m_t}\right)\right) \tag{67}$$

$$= \int_0^1 \beta_t^2\left(1 - \lambda\left(\frac{a_t}{m_t}\right)\right)dt. \tag{68}$$

Finally, substituting the two terms, gives

$$\mathcal{E}(a_t, m_t) = \mathcal{E}_c(a_t, m_t) - \int_0^1 \beta_t^2\left(1 - \lambda\left(\frac{a_t}{m_t}\right)\right)dt. \tag{69}$$

It is left to show that, indeed, we can write $\lambda$ as a one-variable function. Writing $\lambda$ explicitly gives

$$\frac{1}{d}\mathbb{E}_{p_t(x)}\left\|\mathbb{E}_{p_t(x_1|x)}x_1\right\|^2 = \frac{1}{d}\int\left\|\int\frac{\tilde{x}_1\mathcal{N}\left(x|a_t\tilde{x}_1, m_t^2 I\right)q(\tilde{x}_1)d\tilde{x}_1}{\int\mathcal{N}\left(x|a_t\hat{x}_1, m_t^2 I\right)q(\hat{x}_1)d\hat{x}_1}\right\|^2 \mathcal{N}\left(x|a_t x_1, m_t^2 I\right)q(x_1)dx_1 dx. \tag{70}$$

We will use the following identity: Given $\mu \in \mathbb{R}^d$ and $\Sigma \in S_+^d$,

$$\mathcal{N}(x|\mu, \Sigma) = (\det A)\mathcal{N}(Ax|A\mu, A\Sigma A^T). \tag{71}$$

For $A = \frac{1}{a_t}$ and the change of variable $\frac{x}{a_t} \mapsto x$ we get the r.h.s. below which justifies writing $\lambda$ as a function of $a_t/m_t$:

$$\lambda\left(\frac{a_t}{m_t}\right) = \frac{1}{d}\int\left\|\int\frac{\tilde{x}_1\mathcal{N}\left(x\middle|\tilde{x}_1, \left(\frac{m_t}{a_t}\right)^2 I\right)q(\tilde{x}_1)d\tilde{x}_1}{\int\mathcal{N}\left(x\middle|\hat{x}_1, \left(\frac{m_t}{a_t}\right)^2 I\right)q(\hat{x}_1)d\hat{x}_1}\right\|^2 \mathcal{N}\left(x\middle|x_1, \left(\frac{m_t}{a_t}\right)^2 I\right)q(x_1)dx_1 dx. \tag{72}$$

$\square$

**A.2. The $\gamma(\theta_t) > 0$ condition**

*Proof of $\gamma(\theta_t) > 0$ condition.* We show that $\gamma(\theta_t) > 0$ is equivalent to:

$$\lambda(s) > \frac{s^2}{1+s^2}, \quad s \in [0, \infty], \tag{73}$$

where $\gamma(\theta_t)$ is as defined in equation 27. That is

$$\gamma(\theta_t) = 1 - \frac{1 - \lambda(\tan(\theta_t))}{\cos^2(\theta_t)}. \tag{74}$$

Thus the condition $\gamma(\theta_t) > 0$ holds if and only if

$$\lambda(\tan(\theta_t)) > 1 - \cos^2(\theta_t) \tag{75}$$

$$= \sin^2(\theta_t) \tag{76}$$

$$= \frac{\sin^2(\theta_t)}{\cos^2(\theta_t) + \sin^2(\theta_t)} \tag{77}$$

$$= \frac{\tan^2(\theta_t)}{1 + \tan^2(\theta_t)}. \tag{78}$$

Substituting $\tan(\theta_t) = s$, gives the desired inequality, where for $\theta_t \in [0, \frac{\pi}{2}]$, $\tan(\theta_t) \in [0, \infty]$.

$\square$

*Proof of $\lambda$ of standard Gaussian.* Note, for $q(x_1) = \mathcal{N}(x_1|0, I)$, $\rho_s(x)$ is

$$\rho_s(x) = \int \mathcal{N}(x|x_1, s^{-2}I) \mathcal{N}(x_1|0, I) dx_1 \tag{79}$$

$$= \int \mathcal{N}(x_1|x, s^{-2}I) \mathcal{N}(x_1|0, I) dx_1 \tag{80}$$

$$= \mathcal{N}(x|0, (s^{-2} + 1)I). \tag{81}$$

In addition

$$\mathbb{E}_{\rho_s(x_1|x)} x_1 = \frac{1}{\rho_s(x)} \int x_1 \rho_s(x|x_1) q(x_1) dx_1 \tag{82}$$

$$= \frac{1}{\rho_s(x)} \int x_1 \mathcal{N}(x|x_1, s^{-2}I) \mathcal{N}(x_1|0, I) dx_1 \tag{83}$$

$$= \frac{1}{\rho_s(x)} \int (x + s^{-2}\nabla_x) \mathcal{N}(x|x_1, s^{-2}I) \mathcal{N}(x_1|0, I) dx_1 \tag{84}$$

$$= \frac{1}{\rho_s(x)} (x + s^{-2}\nabla_x) \rho_s(x) \tag{85}$$

$$= \frac{1}{\rho_s(x)} \left(x + s^{-2} \frac{-x}{s^{-2} + 1}\right) \rho_s(x) \tag{86}$$

$$= x - s^{-2} \frac{x}{s^{-2} + 1} \tag{87}$$

$$= \frac{1}{s^{-2} + 1} x. \tag{88}$$

Hence, $\lambda(s)$ is

$$\lambda(s) = \frac{1}{d}\mathbb{E}_{\rho_s(x)}\left\|\mathbb{E}_{\rho_s(x_1|x)}x_1\right\|^2 \tag{89}$$

$$= \frac{1}{d}\mathbb{E}_{\rho_s(x)}\left\|\frac{1}{s^{-2}+1}x\right\|^2 \tag{90}$$

$$= \left(\frac{1}{s^{-2}+1}\right)^2 \frac{1}{d}\mathbb{E}_{\rho_s(x)}\|x\|^2 \tag{91}$$

$$= \frac{1}{s^{-2}+1} \tag{92}$$

$$= \frac{s^2}{1+s^2}. \tag{93}$$

$\square$

## A.3. Optimization of the Kinetic Energy $\mathcal{E}$

We derive the solution for the minimization problem of the Kinetic Energy given in equation 27.

**Theorem 3.1.** The minimizer of the Kinetic Energy (equation 27) over all conditional probabilities $(r_t, \theta_t) \in \mathcal{B}$ (equation 26) such that $\gamma(\theta_t) > 0$ for $t > 0$ satisfy

$$r_t = \sqrt{1 - bt + bt^2}, \tag{94}$$

$$\dot{\theta}_t = \frac{1}{1 - bt + bt^2} \sqrt{\frac{b - \frac{1}{4}b^2}{\gamma(\theta_t)}} \tag{95}$$

where $b \in [0, 4]$ is determined by the boundary condition on $\theta_t$.

*Proof of theorem 3.1.* The Kinetic energy parameterized by $r_t$ and $\theta_t$ given in equation 27 is

$$\mathcal{E}(r_t, \theta_t) = \int_0^1 \dot{r}_t^2 + r_t^2 \dot{\theta}_t^2 \gamma(\theta_t) dt. \tag{96}$$

The Euler-Lagrange equations for the Kinetic Energy are

$$\frac{\partial \mathcal{E}_t}{\partial r_t} - \frac{d}{dt} \frac{\partial \mathcal{E}_t}{\partial \dot{r}_t} = 0 \tag{97}$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta_t} - \frac{d}{dt} \frac{\partial \mathcal{E}_t}{\partial \dot{\theta}_t} = 0, \tag{98}$$

where $\mathcal{E}_t$ is the integrand of $\mathcal{E}(r_t, \theta_t)$. Carrying out the derivation gives

$$r_t \dot{\theta}_t^2 \gamma(\theta_t) - \ddot{r}_t = 0 \tag{99}$$

$$4 r_t \dot{r}_t \dot{\theta}_t \gamma(\theta_t) + 2 r_t^2 \ddot{\theta}_t \gamma(\theta_t) + r_t^2 \dot{\theta}_t^2 \frac{d}{d\theta_t} \gamma(\theta_t) = 0. \tag{100}$$

We notice that the second equation can be separated by dividing by $r_t^2 \dot{\theta}_t \gamma(\theta_t)$, that gives

$$\frac{4\dot{r}_t}{r_t} + 2\frac{\ddot{\theta}_t}{\dot{\theta}_t} + \dot{\theta}_t \frac{\frac{d}{d\theta_t} \gamma(\theta_t)}{\gamma(\theta_t)} = 0 \tag{101}$$

Notice that all terms can be written as derivative in time, that is

$$\frac{4\dot{r}_t}{r_t} = 4 \frac{d}{dt} \log(r_t) \tag{102}$$

$$2\frac{\ddot{\theta}_t}{\dot{\theta}_t} = 2 \frac{d}{dt} \log(\dot{\theta}_t) \tag{103}$$

$$\dot{\theta}_t \frac{\frac{d}{d\theta_t} \gamma(\theta_t)}{\gamma(\theta_t)} = \frac{d}{dt} \log(\gamma(\theta_t)). \tag{104}$$

Preforming the anti-derivative and afterwards exponentiation of both sides, gives

$$r_t^4 \dot{\theta}_t^2 \gamma(\theta_t) = C, \tag{105}$$

where $C > 0$ is a constant depending on the boundary condition. Substitute back to the first Euler-Lagrange equation, that is equation 99, gives

$$\ddot{r}_t = \frac{C}{r^3} \tag{106}$$

that is solved by

$$r_t = \sqrt{c + bt + at^2}, \tag{107}$$

where $a, b, c$ satisfy

$$ac - \frac{1}{4}b^2 = C > 0. \tag{108}$$

Finally imposing boundary conditions on $r_t$ as in A.1 gives

$$1 = \sqrt{c}, \quad 1 = \sqrt{c + b + a}. \tag{109}$$

We solve for $c$ and $a$, and get

$$r_t = \sqrt{1 - bt + bt^2} \tag{110}$$

$$C = b - \frac{1}{4}b^2, \tag{111}$$

where the positivity condition on $C$ gives $b \in [0, 4]$. To get the equation for $\dot{\theta}_t$ we substitute the solution for $r_t$ and $C$ in equation 105, that gives

$$\dot{\theta}_t = \frac{1}{1 - bt + bt^2} \sqrt{\frac{b - \frac{1}{4}b^2}{\gamma(\theta_t)}}. \tag{112}$$

$\square$

## A.4. Boundary condition of polar coordinates

**Lemma A.1.** *For $r_t \in [0, \infty]$, $\theta_t \in [0, \frac{\pi}{2}]$ the boundary conditions on $r_t$ and $\theta_t$ are*

$$r_0 = 1, \qquad\qquad r_1 = 1, \tag{113}$$

$$\theta_0 = 0, \qquad\qquad \theta_1 = \frac{\pi}{2}. \tag{114}$$

*Proof of lemma A.1.* The values of $m_t, a_t$ at $t = 0$ and $t = 1$ are

$$m_0 = 1, \qquad\qquad m_1 = 0, \tag{115}$$

$$a_0 = 0, \qquad\qquad a_1 = 1. \tag{116}$$

The relation between $r_t$, $\theta_t$ and $a_t$, $m_t$ is defined in equation 25, that is

$$\begin{bmatrix} a_t \\ m_t \end{bmatrix} = r_t \begin{bmatrix} \sin(\theta_t) \\ \cos(\theta_t) \end{bmatrix}. \tag{117}$$

Substituting $t = 0$ and $t = 1$ gives

$$1 = r_0 \cos(\theta_0), \qquad\qquad 0 = r_1 \cos(\theta_1), \tag{118}$$

$$0 = r_0 \sin(\theta_0), \qquad\qquad 1 = r_1 \sin(\theta_1). \tag{119}$$

Since neither $r_0 = 0$ is a solution nor $r_1 = 0$, it must be that

$$\theta_0 = 0, \qquad\qquad \theta_1 = \frac{\pi}{2}. \tag{120}$$

Plugging this back gives the boundary conditions also for $r_0$ and $r_1$.

$\square$

## A.5. Data separation in high dimensions

**Theorem 4.2.** Let $q$ be an arbitrary normalized finite data distribution. Assume all $(a_t, m_t) \in \mathcal{A}$ in equation 11 satisfy: (i) $\frac{a_t}{m_t}$ is strictly monotonically increasing; and (ii) $a_t, m_t$ have uniform bounded Sobolev $W^{1,\infty}$ norm, i.e., $\|a_t\|_{1,\infty}, \|m_t\|_{1,\infty} \leq M$. Then, for all $(a_t, m_t) \in \mathcal{A}$

$$\mathcal{E}_c(a_t, m_t) - 6M^2 \frac{n}{\sqrt{d}} \leq \mathcal{E}(a_t, m_t) \leq \mathcal{E}_c(a_t, m_t). \tag{121}$$

*Proof of theorem 4.2.* The KE given in equation 21 is

$$\mathcal{E}(a_t, m_t) = \mathcal{E}_c(a_t, m_t) - \int_0^1 \beta_t^2 \left[ 1 - \lambda \left( \frac{a_t}{m_t} \right) \right] dt. \tag{122}$$

The fact that $\lambda$ is bounded above by 1, shown in equation 142, gives

$$\mathcal{E}(a_t, m_t) \leq \mathcal{E}_c(a_t, m_t). \tag{123}$$

For the second inequality, remember that $\beta_t$ defined in equation 16, is

$$\beta_t = \dot{a}_t - a_t \frac{\dot{m}_t}{m_t}. \tag{124}$$

Further note that

$$\frac{d}{dt} \left( \frac{a_t}{m_t} \right) = \frac{\dot{a}_t}{m_t} - a_t \frac{\dot{m}_t}{m_t^2} = \frac{\beta_t}{m_t}. \tag{125}$$

$\frac{a_t}{m_t}$ is strictly monotonic increasing, $\frac{d}{dt} \frac{a_t}{m_t} > 0$. Hence both $\frac{\beta_t}{m_t} > 0$ and $m_t \beta_t > 0$ for every $t \in (0,1)$. Furthermore, $a_t, m_t$ have uniform bounded Sobolev $W^{1,\infty}([0,1])$ norm by $M > 0$ and thus

$$0 < m_t \beta_t = \dot{a}_t m_t - a_t \dot{m}_t \leq 2M^2. \tag{126}$$

The gap between the CKE and KE is

$$\mathcal{E}_c(a_t, m_t) - \mathcal{E}(a_t, m_t) = \int_0^1 \beta_t^2 \left[ 1 - \lambda \left( \frac{a_t}{m_t} \right) \right] dt \tag{127}$$

$$= \int_0^1 (m_t \beta_t) \left( \frac{\beta_t}{m_t} \right) \left[ 1 - \lambda \left( \frac{a_t}{m_t} \right) \right] dt \tag{128}$$

$$\leq 2M^2 \int_0^1 \frac{d}{dt} \left( \frac{a_t}{m_t} \right) \left[ 1 - \lambda \left( \frac{a_t}{m_t} \right) \right] dt. \tag{129}$$

Since $\frac{a_t}{m_t}$ is strictly monotonic increasing, we can make the change of variable $s = \frac{a_t}{m_t}$ and apply theorem 4.1 that gives

$$2M^2 \int_0^1 \frac{d}{dt} \left( \frac{a_t}{m_t} \right) \left[ 1 - \lambda \left( \frac{a_t}{m_t} \right) \right] dt = 2M^2 \int_0^\infty (1 - \lambda(s)) \, ds \tag{130}$$

$$\leq 6M^2 \frac{n}{\sqrt{d}}. \tag{131}$$

$\square$

**Theorem 4.1.** Let $q$ be an arbitrary normalized finite data distribution. Then,

$$\int_0^\infty |1 - \lambda(s)| \, ds \leq \frac{3n}{\sqrt{d}}. \tag{132}$$

*Proof of theoren 4.1.* In the proof we will use Bayes' Rule multiple times. For convenience, we state it explicitly for $\rho_s(x_i|x)$:

$$\rho_s(x_i|x) = \frac{\rho_s(x|x_i) q(x_i)}{\rho_s(x)}. \tag{133}$$

To bound $\lambda(s)$ from above, we apply Jensen inequality

$$\lambda(s) = \frac{1}{dn} \sum_{j=1}^{n} \mathbb{E}_{\rho_s(x|x_j)} \left\| \sum_{i=1}^{n} x_i \rho_s(x_i|x) \right\|^2 \tag{134}$$

$$\leq \frac{1}{dn} \sum_{j=1}^{n} \mathbb{E}_{\rho_s(x|x_j)} \sum_{i=1}^{n} \|x_i\|^2 \rho_s(x_i|x) \tag{135}$$

$$= \frac{1}{dn} \sum_{i=1}^{n} \|x_i\|^2 \sum_{j=1}^{n} \mathbb{E}_{\rho_s(x|x_j)} \rho_s(x_i|x) \tag{136}$$

$$= \frac{1}{dn} \sum_{i=1}^{n} \|x_i\|^2 \sum_{j=1}^{n} \mathbb{E}_{\rho_s(x|x_j)} \frac{\rho_s(x|x_i)q(x_i)}{\rho_s(x)} \tag{137}$$

$$= \frac{1}{dn} \sum_{i=1}^{n} \|x_i\|^2 \sum_{j=1}^{n} \mathbb{E}_{\rho_s(x|x_j)} \frac{1}{n} \frac{\rho_s(x|x_i)}{\rho_s(x)} \tag{138}$$

$$= \frac{1}{dn} \sum_{i=1}^{n} \|x_i\|^2 \, \mathbb{E}_{\rho_s(x)} \frac{\rho_s(x|x_i)}{\rho_s(x)} \tag{139}$$

$$= \frac{1}{dn} \sum_{i=1}^{n} \|x_i\|^2 \, \mathbb{E}_{\rho_s(x|x_i)} 1 \tag{140}$$

$$= \frac{1}{dn} \sum_{i=1}^{n} \|x_i\|^2 \tag{141}$$

$$= 1. \tag{142}$$

To bound $\lambda(s)$ from below, we first rearrange the expression of $\lambda(s)$

$$\lambda(s) = \frac{1}{dn} \sum_{j=1}^{n} \mathbb{E}_{\rho_s(x|x_j)} \left\| \sum_{i=1}^{n} x_i \rho_s(x_i|x) \right\|^2 \tag{143}$$

$$= \frac{1}{dn} \sum_{j=1}^{n} \mathbb{E}_{\rho_s(x|x_j)} \sum_{i=1}^{n} \sum_{l=1}^{n} x_i \cdot x_l \rho_s(x_i|x) \rho_s(x_l|x) \tag{144}$$

$$= \frac{1}{d} \mathbb{E}_{\rho_s(x)} \sum_{i=1}^{n} \sum_{l=1}^{n} x_i \cdot x_l \rho_s(x_i|x) \rho_s(x_l|x) \tag{145}$$

$$= \frac{1}{d} \sum_{i=1}^{n} \sum_{l=1}^{n} x_i \cdot x_l \mathbb{E}_{\rho_s(x)} \rho_s(x_i|x) \rho_s(x_l|x) \tag{146}$$

$$= \frac{1}{d} \sum_{i=1}^{n} \sum_{l=1}^{n} x_i \cdot x_l \mathbb{E}_{\rho_s(x)} \frac{\rho_s(x|x_i)q(x_i)}{\rho_s(x)} \rho_s(x_l|x) \tag{147}$$

$$= \frac{1}{dn} \sum_{i=1}^{n} \sum_{l=1}^{n} x_i \cdot x_l \mathbb{E}_{\rho_s(x|x_i)} \rho_s(x_l|x) \tag{148}$$

$$= \frac{1}{dn} \sum_{i=1}^{n} \sum_{l=1}^{n} x_i \cdot (x_l - x_i + x_i) \mathbb{E}_{\rho_s(x|x_i)} \rho_s(x_l|x) \tag{149}$$

$$= \frac{1}{dn} \sum_{i=1}^{n} \sum_{l=1}^{n} \|x_i\|^2 \, \mathbb{E}_{\rho_s(x|x_i)} \rho_s(x_l|x) + \frac{1}{dn} \sum_{i=1}^{n} \sum_{l=1}^{n} x_i \cdot (x_l - x_i) \mathbb{E}_{\rho_s(x|x_i)} \rho_s(x_l|x) \tag{150}$$

Remembering that $\rho_s(x_i|x)$ is a discrete probability density on $x_i$, $i \in [n]$, and the normalization as in equation 33, the first

term in 150 is

$$\frac{1}{dn}\sum_{i=1}^{n}\sum_{l=1}^{n}\|x_i\|^2\,\mathbb{E}_{\rho_s(x|x_i)}\rho_s(x_l|x) = \frac{1}{dn}\sum_{i=1}^{n}\|x_i\|^2\,\mathbb{E}_{\rho_s(x|x_i)}\sum_{l=1}^{n}\rho_s(x_l|x) \tag{151}$$

$$= \frac{1}{dn}\sum_{i=1}^{n}\|x_i\|^2\,\mathbb{E}_{\rho_s(x|x_i)}1 \tag{152}$$

$$= \frac{1}{dn}\sum_{i=1}^{n}\|x_i\|^2 \tag{153}$$

$$= 1. \tag{154}$$

We substitute back to the expression of $\lambda(s)$ and apply Cauchy–Schwarz inequality

$$\lambda(s) = 1 + \frac{1}{dn}\sum_{i=1}^{n}\sum_{\substack{l=1\\l\neq i}}^{n}x_i\cdot(x_l-x_i)\mathbb{E}_{\rho_s(x|x_i)}\rho_s(x_l|x) \tag{155}$$

$$\geq 1 - \frac{1}{dn}\sum_{i=1}^{n}\sum_{\substack{l=1\\l\neq i}}^{n}\|x_i\|\,\|x_l-x_i\|\,\mathbb{E}_{\rho_s(x|x_i)}\rho_s(x_l|x). \tag{156}$$

Applying the inequalities in equations 142, 156 and lemma 4.4 gives

$$\int_{0}^{\infty}|1-\lambda(s)|\,ds = \int_{0}^{\infty}\left(1-\lambda\left(s\right)\right)ds \tag{157}$$

$$\leq \int_{0}^{\infty}\frac{1}{dn}\sum_{i=1}^{n}\sum_{\substack{l=1\\l\neq i}}^{n}\|x_i\|\,\|x_l-x_i\|\,\mathbb{E}_{\rho_s(x|x_i)}\rho_s(x_l|x)ds \tag{158}$$

$$\leq \frac{1}{dn}\sum_{i=1}^{n}\sum_{\substack{l=1\\l\neq i}}^{n}\|x_i\|\int_{0}^{\infty}\|x_l-x_i\|\,\eta\left(s\,\|x_l-x_i\|\right)ds \tag{159}$$

$$= \frac{1}{dn}\sum_{i=1}^{n}\sum_{\substack{l=1\\l\neq i}}^{n}\|x_i\|\int_{0}^{\infty}\eta(t)dt \tag{160}$$

$$\leq \frac{3}{dn}\sum_{i=1}^{n}\sum_{\substack{l=1\\l\neq i}}^{n}\|x_i\| \tag{161}$$

$$\leq \frac{3n}{\sqrt{d}}, \tag{162}$$

where in the last equality we use Jensen inequality and equation 8 to bound the sum of norms, $\|x_i\|$, as shown below

$$1 = \frac{1}{dn}\sum_{i=1}^{n}\|x_i\|^2 \geq \frac{1}{d}\left(\frac{1}{n}\sum_{i=1}^{n}\|x_i\|\right)^2. \tag{163}$$

$\square$

**Lemma 4.4.** Let $q$ be an arbitrary normalized finite data distribution. Then for every $s > 0$,

$$\mathbb{E}_{\rho_s(x|x_j)}\rho_s(x_i|x) \leq \eta\left(s\left\|x_i - x_j\right\|\right), \tag{164}$$

where $\eta(t)$ is integrable in $[0, \infty)$ and

$$\int_0^\infty \eta(t) \leq 3. \tag{165}$$

*Proof of lemma 4.4.* Remembering that $p(x) = \mathcal{N}(0, I)$,

$$\mathbb{E}_{\rho_s(x|x_j)}\rho_s(x_i|x) = \mathbb{E}_{\rho_s(x|x_j)} \frac{e^{-\frac{s^2}{2}\|x-x_i\|^2}}{\sum_{l=1}^n e^{-\frac{s^2}{2}\|x-x_l\|^2}} \tag{166}$$

$$= \mathbb{E}_{p(x)} \frac{e^{-\frac{s^2}{2}\left\|xs^{-1}+x_j-x_i\right\|^2}}{\sum_{l=1}^n e^{-\frac{s^2}{2}\|xs^{-1}+x_j-x_l\|^2}} \tag{167}$$

$$\leq \mathbb{E}_{p(x)} \frac{e^{-\frac{s^2}{2}\left\|xs^{-1}+x_j-x_i\right\|^2}}{e^{-\frac{s^2}{2}\|xs^{-1}+x_j-x_i\|^2} + e^{-\frac{s^2}{2}\|xs^{-1}\|^2}} \tag{168}$$

$$= \mathbb{E}_{p(x)} \frac{1}{1 + e^{\frac{s^2}{2}\|x_i-x_j\|^2 - s(x_i-x_j)\cdot x}} \tag{169}$$

Since $x \sim p(x) = \mathcal{N}(0, I)$ is $O(d)$ invariant, the last term depends only on the norm of $x_i - x_j$. That is, we denote by $z \in \mathbb{R}$ the component of $x$ in the direction of $x_i - x_j$

$$z = \frac{x_i - x_j}{\|x_i - x_j\|} \cdot x \tag{170}$$

So $z \sim \mathcal{N}(0, 1)$ and

$$\mathbb{E}_{p(x)} \frac{1}{1 + e^{\frac{s^2}{2}\|x_i-x_j\|^2 - s(x_i-x_j)\cdot x}} = E_{\mathcal{N}(z|0,1)} \frac{1}{1 + e^{\frac{s^2}{2}\|x_i-x_j\|^2 - s\|x_i-x_j\|z}} \tag{171}$$

$$\tag{172}$$

$$= E_{\mathcal{N}(z|0,1)} \frac{1}{1 + e^{\frac{1}{2}(s\|x_i-x_j\|)^2 - (s\|x_i-x_j\|)z}} \tag{173}$$

And we define $\eta(t)$ to be

$$\eta(t) = E_{\mathcal{N}(z|0,1)} \frac{1}{1 + e^{\frac{t^2}{2} - tz}}. \tag{174}$$

So it is left to show that $\eta(t)$ is integrable in $[0, \infty)$ and bound its integral. For every $t > 0$

$$\eta(t) = E_{\mathcal{N}(z|0,1)} \frac{1}{1 + e^{\frac{t^2}{2} - tz}} \tag{175}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \frac{e^{-\frac{z^2}{2}} dz}{1 + e^{\frac{t^2}{2} - tz}} \tag{176}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \frac{dz}{e^{\frac{z^2}{2}} + e^{\frac{1}{2}(z-t)^2}} \tag{177}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \frac{dz}{e^{\frac{1}{2}\left(z+\frac{t}{2}\right)^2} + e^{\frac{1}{2}\left(z-\frac{t}{2}\right)^2}} \tag{178}$$

$$= \sqrt{\frac{2}{\pi}} \int_0^\infty \frac{dz}{e^{\frac{1}{2}\left(z+\frac{t}{2}\right)^2} + e^{\frac{1}{2}\left(z-\frac{t}{2}\right)^2}} \tag{179}$$

$$\leq \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-\frac{1}{2}\left(z+\frac{t}{2}\right)^2} dz \tag{180}$$

$$= \sqrt{\frac{2}{\pi}} \int_{\frac{t}{2}}^\infty e^{-\frac{z^2}{2}} dz, \tag{181}$$

20

where in the third to last equality we used the fact that the integrand is symmetric w.r.t $z \to -z$. Now on the one hand, for every $t > 0$

$$\sqrt{\frac{2}{\pi}} \int_{\frac{t}{2}}^{\infty} e^{-\frac{z^2}{2}} dz \leq \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} e^{-\frac{z^2}{2}} dz = 1 \tag{182}$$

On the other hand, $\frac{2z}{t} > 1$ for every $z > \frac{t}{2}$. Hence

$$\sqrt{\frac{2}{\pi}} \int_{\frac{t}{2}}^{\infty} e^{-\frac{z^2}{2}} dz \leq \sqrt{\frac{2}{\pi}} \int_{\frac{t}{2}}^{\infty} \frac{2z}{t} e^{-\frac{z^2}{2}} dz \tag{183}$$

$$= \sqrt{\frac{2}{\pi}} \int_{\frac{t^2}{4}}^{\infty} \frac{1}{t} e^{-\frac{z'}{2}} dz' \tag{184}$$

$$= \sqrt{\frac{2}{\pi}} \frac{1}{t} e^{-\frac{t^2}{8}} \tag{185}$$

Combining the two equations 182 and 185, we get that for every $t > 0$

$$\eta(t) \leq \min\left\{1, \sqrt{\frac{2}{\pi}} \frac{1}{t} e^{-\frac{t^2}{8}}\right\}, \tag{186}$$

which is integrable in $[0, \infty)$. Furthermore,

$$\int_{0}^{\infty} \eta(t) dt = \int_{0}^{1} \eta(t) dt + \int_{1}^{\infty} \eta(t) dt \tag{187}$$

$$\leq \int_{0}^{1} dt + \int_{1}^{\infty} \sqrt{\frac{2}{\pi}} \frac{1}{t} e^{-\frac{t^2}{8}} dt \tag{188}$$

$$\leq 1 + \int_{0}^{\infty} \sqrt{\frac{2}{\pi}} e^{-\frac{t^2}{8}} dt \tag{189}$$

$$= 3 \tag{190}$$

$$\square$$

## C. Additional tables

|  | CIFAR10 | ImageNet-32 |
|---|---|---|
| Channels | 256 | 256 |
| Depth | 2 | 3 |
| Channels multiple | 1,2,2,2 | 1,2,2,2 |
| Heads | 4 | 4 |
| Heads Channels | 64 | 64 |
| Attention resolution | 16 | 16,8 |
| Dropout | 0.1 | 0.0 |
| Effective Batch size | 256 | 1024 |
| GPUs | 2 | 4 |
| Epochs | 2000 | 400 |
| Iterations | 392k | 500k |
| Learning Rate | 5e-4 | 1e-4 |
| Learning Rate Scheduler | Polynomial Decay | Polynomial Decay |
| Warmup Steps | 19.6k | 20k |

*Table 2.* Hyper-parameters used for training.

| Model | **CIFAR-10** | | | **ImageNet 32×32** | | |
|---|---|---|---|---|---|---|
|  | $K$=1 | $K$=20 | $K$=50 | $K$=1 | $K$=5 | $K$=15 |
| *Ablation* | | | | | | |
| DDPM (Ho et al., 2020) | 3.23 | 3.13 | 3.11 | 3.69 | 3.64 | 3.61 |
| SI (Albergo & Vanden-Eijnden, 2022) | 3.13 | 3.02 | 3.00 | 3.67 | 3.61 | 3.58 |
| FM ᵂ/ OT (Lipman et al., 2022) | 3.10 | 3.00 | 2.98 | 3.69 | 3.64 | 3.60 |
| FM ᵂ/ KO | 3.10 | 3.00 | 2.97 | 3.66 | 3.60 | 3.57 |

*Table 3.* BPD results for different uniform dequantization samples $k$, following the protocol of (Lipman et al., 2022).

## B. Further implementation details

We model $\theta_t$ using the following architecture

$$\theta_t = \frac{\pi}{2} \frac{|m(t) - m(0)|}{|m(1) - m(0)|} \tag{191}$$

where $m$ is an MLP defined as

$$\phi(t) = \texttt{sigmoid}\Big(L_1(t) + L_3\big(2\texttt{sigmoid}\,(L_2(t)) - 1\big)\Big) \tag{192}$$

and $L_i$, $i \in [3]$, are linear layers: $L_1 : \mathbb{R} \to \mathbb{R}$; $L_2 : \mathbb{R} \to \mathbb{R}^2$; and $L_3 : \mathbb{R}^2 \to \mathbb{R}$. In total the model for $\theta_t$ uses 10 learnable parameters. We optimize the KE using ADAM optimizer and Cosine Annealing scheduler.

For the CNF training in Section 6.3 we model $v_t(x)$ as follows: (i) For the 2D data it is an MLP with 5 layers consisting of 512 neurons in each layer. (ii) For the image datasets (CIFAR, ImageNet-32), we used U-Net architecture as in (Dhariwal & Nichol, 2021), where the particular architecture hyper-parameters are detailed in Table 2. In this table we also provide training hyper-parameters. All baselines (IS, DDPM) are implemented in exactly the same setting of architecture and training hyper-parameters.

## D. Runtime Details

First, during deployment (model sampling) there is no change to the original algorithm of sampling from CNFs. Second, during training we use the optimal $a_t, m_t$ found by optimizing equation 27 and equation 39. The solution $a_t, m_t$ is modeled with a -parameter MLP and causes no noticeable change in iteration time compared to Cond-OT. Third, the optimization

| Dataset | Time [s] |
|---|---|
| ImageNet 8 | 30.4 |
| ImageNet 16 | 42.0 |
| ImageNet 32/CIFAR10 | 84.1 |
| ImageNet 64 | 328.6 |

*Table 4.* Runtime of computing $\lambda$ for $k = 1$ and $n = 50,000$ on 1 GPU (Quadro RTX 8000).

of $a_t, m_t$ is done once as a preprocessing step and takes less than a minute. Fourth, the most time consuming part of our algorithm is approximating the data separation function $\lambda$ but it is done once per dataset, so can be reused across training sessions as long as the dataset hasn't changed. The complexity of approximating $\lambda$ can be deduced from equation 39 and it is $O(kn^2)$, where $k$ is the number of samples from $p = \mathcal{N}(0, I)$ and $n$ is the number of samples from $q$. In Table 4 we present the runtime of computing $\lambda$ for $k = 1$ and $n = 50,000$ on 1 GPU (Quadro RTX 8000) for a number of image datasets of different image sizes.
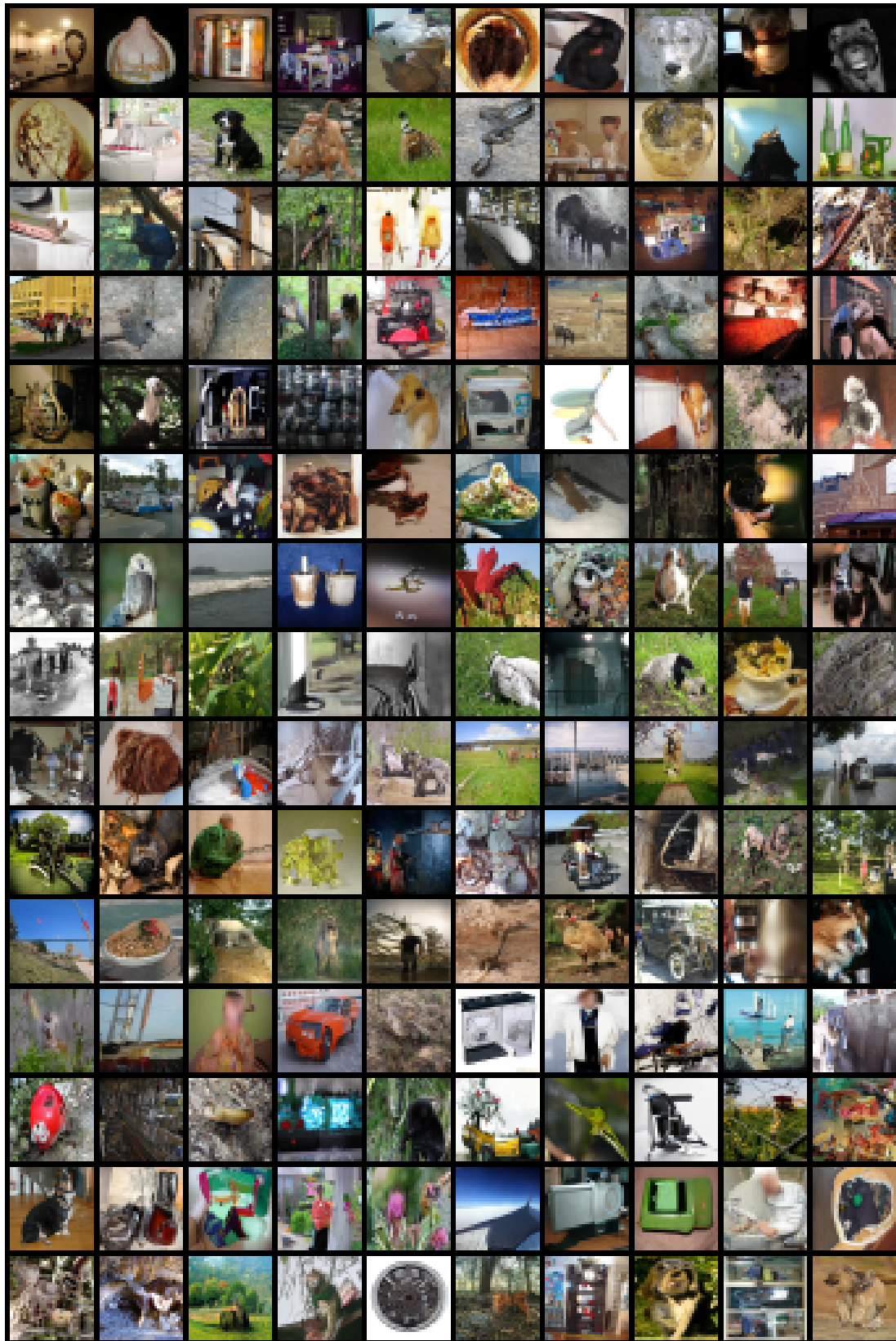
# E. Additional Figures

*Figure 7.* Non-curated unconditional *face-blurred* ImageNet-32 generated images of a CNF model trained with FM-KO.
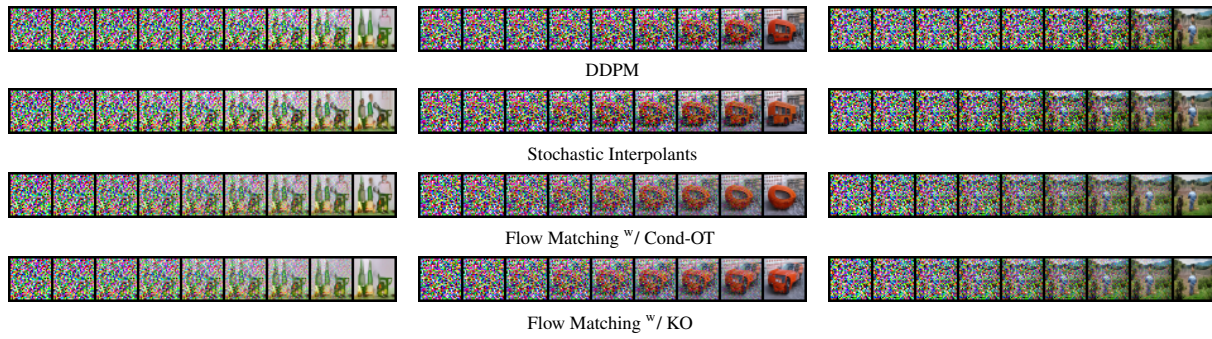
*Figure 8.* Sample trajectories of a CNF models trained with DDPM, SI, Cond-OT and kinetic optimal paths, on *face-blurred* ImageNet 32 data.