Let us Flow Together

Qiang Liu

December 24, 2024

(Working Draft)

This is a collection of notes on rectified flow. It is still under development. The references are not complete.

It is accompanied with a code base and blog:

Code: `https://github.com/lqiang67/rectified-flow`.

Blog: `https://rectifiedflow.github.io/`

Please let us know (`rectifiedflow@gmail.com`) if you have any suggestions, comments, or notice any typos.

Contents

CHAPTER ONE

Rectified Flow

1.1 Overview

Generative modeling can be formulated as finding a computational procedure that transforms a noise distribution, denoted by π_0, into the unknown data distribution π_1. In flow models, this procedure is represented by a ordinary differential equation (ODE):

$$\dot{Z}_t = v_t(Z_t), \quad \forall t \in [0,1], \quad \text{starting from} \quad Z_0 \sim \pi_0, \tag{1.1}$$

where $\dot{Z}_t = \mathrm{d}Z_t/\mathrm{d}t$ denotes the time derivative, and the velocity field $v_t(x) = v(x,t)$ is a learnable function to be estimated to ensure that Z_1 follows the target distribution π_1 when starting from $Z_0 \sim \pi_0$. In this case, we say that the stochastic process $Z = \{Z_t\}$ provides an (ODE) transport from π_0 to π_1.

It is important to note that, in all but trivial cases, there exist *infinitely many* ODE transports from π_0 to π_1, provided that at least one such process exists. Thus, it is essential to be clear about which types of ODEs we should prefer.

One option is to favor ODEs that are *easy* to solve at inference time. In practice, the ODEs are approximated using numerical methods, which typically construct *piecewise linear* approximations of the ODE trajectories. For instance, a common choice is Euler's method:

$$\hat{Z}_{t+\varepsilon} = \hat{Z}_t + \varepsilon v_t(\hat{Z}_t), \quad \forall t \in \{0, \varepsilon, 2\varepsilon, \ldots, 1\}, \tag{1.2}$$

where $\varepsilon > 0$ is a step size. Varying the step size ε introduces a trade-off between accuracy and computational cost: smaller ε yields high accuracy, but incurs larger number of calculation steps. Therefore, we should seek ODEs that can be approximated accurately even with large step sizes.

The ideal scenario arises when the ODE follows straight-line trajectories, in which case Euler approximation yields *zero discretization error* regardless of the choice step sizes. In such cases, the ODE, up to time reparameterization, should satisfy:

$$Z_t = tZ_1 + (1-t)Z_0, \quad \Rightarrow \quad \dot{Z}_t = Z_1 - Z_0.$$

These ODEs, known as *straight transports*, enable *fast* generative models that can be simulated in a single step. We refer to the resulting pair (Z_0, Z_1) as a *straight coupling* of π_0 and π_1. In practice, we may not achieve perfect

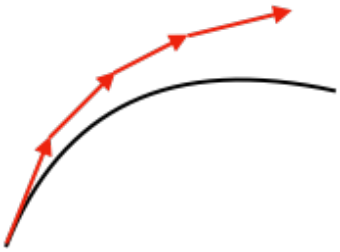


Figure 1.1: Curved trajectories suffer from discretization error when approximated by Euler's method.

straightness but can aim to make the ODE trajectories as straight as possible to maximize computational efficiency.

It is possible to discuss generalized notions of straightness when solvers other than Euler's method are used.

Rectified Flow

To construct a flow transporting π_0 to π_1, let us assume that we are given an arbitrary coupling (X_0, X_1) of π_0 and π_1, from which we can obtain empirical draws. This can be simply the *independent coupling* with law $\pi_0 \times \pi_1$, as is common in practice when we have access to independent samples from π_0 and π_1. The idea is that we are going to take (X_0, X_1) and convert it to a better coupling generated by an ODE model, and optionally, we can go further to iteratively repeat this process to further enhance desired properties, such as straightness.

The independent coupling $(X_0, X_1) \sim \pi_0 \times \pi_1$ serves as a special starting point because it is what we can observe empirically even without any meaningful pairing relations between data from π_0 and π_1. But the algorithm works for arbitrary couplings, if available.

Rectified flow works in the following ways:

1. Build Interpolation: We build an interpolation process $\{X_t\} = \{X_t : t \in [0,1]\}$ that smoothly interpolate between X_0 and X_1. Although general choices are possible, let us consider the canonical choice of straight-line interpolation:

$$X_t = tX_1 + (1-t)X_0.$$

Here $\{X_t\}$ is a stochastic process generated in a special way: we first sample the endpoints X_0 and X_1 and then sample the intermediate trajectory connecting them. Such processes are also known as *bridge processes*, where the intermediate values of X_t smoothly "bridge" the distribution between X_0 and X_1.

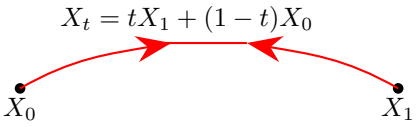


Figure 1.2: Interpolation process X: We first draw the random end points (X_0, X_1), and then build the intermediate path.

2. Marginal Matching: By construction, the marginal distributions of X_0 and X_1 match the target distributions π_0 and π_1 through the interpolation process $\{X_t\}$. However, $\{X_t\}$ is not a *causal* ODE process like $\dot{Z}_t = v_t(Z_t)$, which evolves forward from Z_0. Instead, generating X_t requires knowledge of both X_0 and X_1, rather than evolving solely from X_0 as t increases.

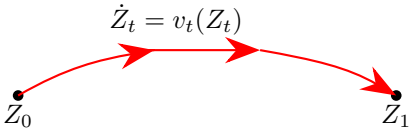


Figure 1.3: The ODE (a.k.a flow) process Z_t, which generates Z_t causally with increasing t starting from the initialization Z_0.

This issue can be resolved if we can convert $\{X_t\}$ somehow into a causal ODE process, while preserving the marginal distributions of X_t at each time t. Perhaps surprisingly, this can be achieved by simply training the ODE model $\dot{Z}_t = v_t(Z_t)$ to match the slope $\dot{X}_t$ of the interpolation process via:

$$\min_v \int_0^1 \mathbb{E}\left[\left\|\dot{X}_t - v_t(X_t)\right\|^2\right] \mathrm{d}t. \tag{1.3}$$

The theoretical minimum is achieved by

$$v_t^*(x) = \mathbb{E}\left[\dot{X}_t \mid X_t = x\right],$$

which denotes the expectation of the slope $\dot{X}_t$ of the interpolation process passing through a given point $X_t = x$. We have $v_t^*(x) = \dot{X}_t$ only one trajectory passing $X_t = x$. If multiple trajectories pass point $X_t = x$, the velocity $v_t^*(x)$ is the average of $\dot{X}_t$ for these trajectories.

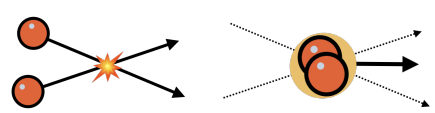


Figure 1.4: Imagine emitting particles along the trajectories of the interpolation paths $\{X_t\}$. When these trajectories intersect, the particles collide and merge into a larger particle, which then continues moving in the average direction.

With the canonical straight interpolation, we have $\dot{X}_t = X_1 - X_0$ by taking derivative of X_t w.r.t. t. It yields

$$\min_v \int_0^1 \mathbb{E}\left[\|X_1 - X_0 - v_t(X_t)\|^2\right] \mathrm{d}t, \qquad X_t = tX_1 + (1-t)X_0.$$

In practice, the optimization in (1.3) can be efficiently solved *even for large AI models* when v is parameterized as modern deep neural networks. This is achieved by leveraging off-the-shelf optimizers with stochastic gradients, computed by drawing pairs (X_0, X_1) from data, sampling t uniformly in $[0, 1]$, and then computing the corresponding $(X_t, \dot{X}_t)$ using the interpolation formula.

Background (Random Variables and Expectation). To put it simply, a random variable $X = X(\omega)$ is a measurable function of a "random seed" ω following a baseline distribution $\mathbb{P}$. A stochastic process $X_t = X(t, \omega)$ is a time-dependent random variable. We use uppercase letters like X, Y to represent random variables (RVs).

Viewing the interpolation process above, the random seed is given by the endpoints, i.e., $\omega = (X_0, X_1)$. The slope is defined as $\dot{X}_t = \partial_t X(t, \omega)$, which is another function of the same random seed.

The expectation in the loss (1.3), written in full, is

$$\mathbb{E}_{\omega\sim\mathbb{P}}\left[\|\partial_t X(t,\omega) - v_t(X(t,\omega))\|^2\right],$$

though we often omit the random seed in writing.

Background (Conditional Expectation). For any joint random variable (X, Y), the conditional expectation $\mathbb{E}[Y|X]$ is a function f^* of X that yields the best prediction of Y given X,

$$f^* = \arg\min_f \mathbb{E}\left[\|Y - f(X)\|^2\right],$$

that is, $\mathbb{E}[Y|X] = f^*(X)$. This can be seen from the bias-variance decomposition,

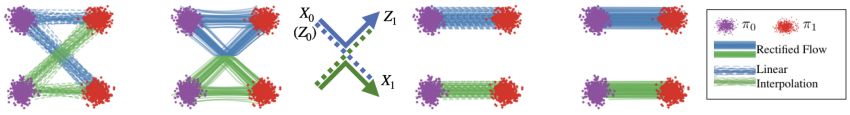
$$\begin{aligned}\mathbb{E}\left[\|Y - f(X)\|^2\right] &= \mathbb{E}\left[\|Y - \mathbb{E}[Y|X]\|^2 + \|\mathbb{E}[Y|X] - f(X)\|^2\right] \\ &= \underbrace{\mathbb{E}\left[\mathrm{Var}(Y \mid X)\right]}_{variance} + \underbrace{\mathbb{E}\left[\|\mathbb{E}[Y|X] - f(X)\|^2\right]}_{bias},\end{aligned}$$

where the first term represents the variance of Y given X, which is independent of f. The second term is the bias, which is zero when $f(X) = \mathbb{E}[Y \mid X]$. Thus, the optimal choice for f is $f^*(X) = \mathbb{E}[Y \mid X]$.

So $\mathbb{E}[Y \mid X]$ is a random variable, as it is a function of X. We can "instantiate" it at a fixed value $X = x$, and get the deterministic quantity $\mathbb{E}[Y \mid X = x] = f^*(x)$.

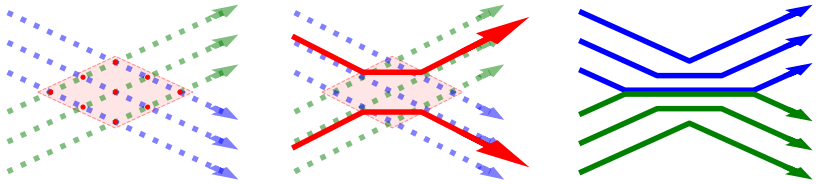
Figure 1.5 illustrates the intuition:

1. In the interpolation process $\{X_t\}$, different trajectories may have intersecting points, resulting in multiple possible values of $\dot{X}_t$ associ-



(a) Linear interpolation $X_t = tX_1 + (1-t)X_0$ (b) Rectified flow Z_t induced by (X_0, X_1) (c) Linear interpolation $Z_t = tZ_1 + (1-t)Z_0$ (d) Rectified flow Z'_t induced by (Z_0, Z_1)

Figure 1.5: Intuition of rectified flow and reflow. (a) The interpolation $\{X_t\}$ constructed from independent coupling (X_0, X_1), which has intersecting points in the middle. (b) The rectified flow $\{Z_t\}$ induced from $\{X_t\}$, which rewires the trajectories at the intersecting points, while preserving the marginal distributions. (c) The new interpolation build from (Z_0, Z_1) of the rectified from in (b), which has less intersecting points. (d) The new rectified flow build from the interpolation from (c), which is now almost straight.



(a) (b) (c)

Figure 1.6: A close-up view of how rectification "rewires" interpolation trajectories. (a) Interpolation trajectories with intersections. (b) Averaged velocity directions at intersection points (red arrows). (c) Trajectories of the resulting rectified flow.

ated with a same point X_t due to uncertainty about which trajectory it was drawn from (see Figure 1.5(a)).

2. On the other hand, by the definition of an ODE $\dot{Z}_t = v_t^*(Z_t)$, the update direction $\dot{Z}_t$ at each point Z_t is uniquely determined by Z_t, making it impossible for different trajectories of $\{Z_t\}$ to intersect and then diverge along different directions.

3. At these intersection points of $\{X_t\}$, where $\dot{X}_t$ is stochastic and non-unique, Z_t "derandomizes" the update direction by following the conditional expectation $v_t^*(X_t) = \mathbb{E}[\dot{X}_t \mid X_t]$, thus providing the unique update direction required by ODEs.

4. Since ODE trajectories $\{X_t\}$ cannot intersect, they must curve at potential intersection points to "rewire" the original interpolation paths and avoid crossing.

Remark 1. Figure 1.6 illustrates a close-up view of how rectification "rewires" interpolation trajectories.

Consider two "beams" of interpolation trajectories intersecting to form the "region ofz confusion" (shaded area in the middle). Within this region, a particle moving along the rectified flow follows the averaged direction v_t^*. Upon exiting, the particle joins one of the original interpolation streams based on its exit side and continues moving. Since rectified flow trajectories do not intersect within the region, they remain separated and exit from their respective sides, effectively "rewiring" the original interpolation trajectories.

The example described above results in a velocity field that is

discontinuous at the boundary of the region of confusion. However, when the coupling is randomized, this intersection process can be viewed as occurring infinitely many times, yielding a smooth velocity field.

Definition 1. For any time-differential stochastic process $\{X_t\} = \{X_t : t \in [0,1]\}$, We call the ODE process $\dot{Z}_t = v_t^*(Z_t)$ with $v_t^*(z) = \mathbb{E}\left[\dot{X}_t \mid X_t = z\right]$, and $Z_0 = X_0$ the *rectified flow* induced by $\{X_t\}$. We denote it as

$$\{Z_t\} = \texttt{Rectify}(\{X_t\}).$$

Remark 2. Although $\{X_t\}$ is an interpolation process in the algorithm, the definition of $\texttt{Rectify}(\cdot)$ applies to general time-differential stochastic processes.

What makes rectified flow $\{Z_t\}$ useful is that it preserves the marginal distributions of $\{X_t\}$ at each point, while resulting in a "better" coupling (Z_0, Z_1) in terms of optimal transport:

[Marginal Preservation]

The $\{X_t\}$ and its rectified flow $\{Z_t\}$ share the same marginal distributions at each time $t \in [0,1]$, that is,

$$\text{Law}(Z_t) = \text{Law}(X_t), \quad \forall t \in [0,1].$$

[Transport Cost]

The start-end pairs (Z_0, Z_1) from the rectified flow $\{Z_t\}$ guarantees to yield no larger transport cost than (X_0, X_1), simultaneously for *all* convex cost functions c:

$$\mathbb{E}\left[c(Z_1 - Z_0)\right] \le \mathbb{E}\left[c(X_1 - X_0)\right], \quad \forall \text{convex } c : \mathbb{R}^d \to \mathbb{R}.$$

1.2 Reflow

While rectified flows tend to favor straight trajectories, they are not perfectly straight. As shown in Figure 1.5(b), the flow makes turns at intersection points of the interpolation trajectories $\{X_t\}$. How can we further improve the flow to achieve straighter trajectories and hence speed up inference?

A key insight is that the start-end pairs (Z_0, Z_1) generated by rectified flow, called the *rectified coupling* of (X_0, X_1), form a better and "straighter" coupling compared to (X_0, X_1). This is because if we connect Z_0 and Z_1 with a new straight-line interpolation, it would yield less intersection points. Hence, training a new rectified flow based on this new interpolation would result in straighter trajectories, leading to faster inference.

Formally, we apply the $\texttt{Rectify}(\cdot)$ procedure recursively, yielding a sequence of rectified flows starting from $(Z_0^0, Z_1^0) := (X_0, X_1)$:

$$\text{Reflow:} \qquad \{Z_t^{k+1}\} = \texttt{Rectify}(\texttt{Interp}(Z_0^k, Z_1^k)), \tag{1.4}$$

where $\texttt{Interp}(Z_0^k, Z_1^k)$ denotes an interpolation process given (Z_0^k, Z_1^k) as the endpoints. We call $\{Z_t^k\}$ the k-th rectified flow, or simply *k-rectified flow*, induced from (X_0, X_1).

This *reflow* procedure is proved to "straightening" the paths of rectified flows in the following sense: Define the following measure of straightness of $\{Z_t\}$:

$$\text{Define} \quad S(\{Z_t\}) = \int_0^1 \mathbb{E}[\|Z_1 - Z_0 - \dot{Z}_t\|^2]dt,$$

where $S(\{Z_t\})$ is a measure of the straightness of $\{Z_t\}$, with $S(\{Z_t\}) = 0$ corresponding to straight paths. Then we have

$$\mathbb{E}_{k\sim\text{Unif}(\{1,\dots,K\})}[S(\{Z_t^k\})] = \mathcal{O}(1/K), \tag{1.5}$$

which says that the average of $S(\{Z_t^k\})$ of the first K steps decrease with an $\mathcal{O}(1/K)$ rate. Hence, we would obtain perfectly straight-line dynamics with $(S(\{Z_t^k\}) \to 0)$ in the limit of $k \to +\infty$. Note that reflow can begin from any coupling (X_0, X_1), so it provides a *general procedure* for straightening and thus speeding up *any* given dynamics while preserving the marginals.

1.3 Interpolations

The algorithm is not limited to the straight-line interpolation. In general, we can consider any smooth interpolation process of form:

$$X_t = \texttt{I}_t(X_0, X_1), \quad \forall t \in [0, 1] \tag{1.6}$$

where $\texttt{I}$ is a function that satisfies the boundary conditions of $X_0 = \texttt{I}_0(X_0, X_1)$, $X_1 = \texttt{I}_1(X_0, X_1)$ to ensure that interpolation process is valid.

Definition 2. A function $\texttt{I}\colon [0,1]\times\mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}^d$, denoted as $\texttt{I}_t(x_0, x_1)$, is said to be an *interpolation*, or an *interpolation function*, if it satisfies

$$\texttt{I}_0(x_0, x_1) = x_0, \qquad \texttt{I}_1(x_0, x_1) = x_1, \qquad \text{for any} \quad x_0, x_1 \in \mathbb{R}^d.$$

We call $\{X_t\}$ with $X_t = \texttt{I}_t(X_0, X_1)$ the interpolation process constructed from $\texttt{I}$ and coupling (X_0, X_1).

For now, we assume that X_t is a time-differentiable process, that is, the derivative $\dot{X}_t := \partial\texttt{I}_t(X_0, X_1)$ exists pointwisely. Hence, the loss function in (1.3) reduces to

$$\min_v \int_0^1 \mathbb{E}\|\partial_t\texttt{I}_t(X_0, X_1) - v_t(\texttt{I}_t(X_0, X_1))\|^2,$$

which yields

$$v_t^*(x) = \mathbb{E}\left[\partial_t\texttt{I}_t(X_0, X_1) \mid \texttt{I}_t(X_0, X_1) = x\right].$$

Here we first find the pairs (X_0, X_1) that yields $X_t = \texttt{I}_t(X_0, X_1) = x$, and then calculating the corresponding derivative $\dot{X}_t = \partial_t\texttt{I}_t(X_0, X_1)$. If (X_0, X_1) is not fully determined by $\texttt{I}_t(X_0, X_1) = x$, then the derivatives

are averaged across all solutions of (X_0, X_1). If (X_0, X_1) can be fully determined by X_t, corresponding to the case when no interpolation trajectories intersect, then conditional expectation reduces to calculating an inverse function: first invert function $\mathtt{I}$ to get (X_0, X_1) from X_t, and then calculate the derivative $\partial \mathtt{I}_t(X_0, X_1)$.

Example 1 (Affine Interpolations). In the literature, the class of *affine interpolations* is mostly studied:

$$X_t = \alpha_t X_1 + \beta_t X_0,$$

where α_t, β_t are sequences satisfying

$$\alpha_0 = \beta_1 = 0, \qquad \alpha_1 = \beta_0 = 1.$$

In addition, we may want α_t to be monotonically increasing, and β_t monotonically decreasing, even though this is not strictly required by the theory.

Example 2 (Straight Interpolation). With $\alpha_t = t, \beta_t = 1 - t$, we obtain the time-uniform straight interpolation:

$$X_t = tX_1 + (1-t)X_0, \qquad \dot{X}_t = X_1 - X_0. \tag{1.7}$$

In general, affine interpolations satisfying $\alpha_t + \beta_t = 1$ yields a straight interpolation trajectories. In this case, different choices of α_t introduces a time scaling.

Example 3 (The DDPM Interpolation). The DDPM [Ho et al., 2020], DDIM [Song et al., 2020a], and the VP ODE of [Song et al., 2020b] use α_t and β_t satisfying $\alpha_t^2 + \beta_t^2 = 1$, which corresponds to a spherical curve. In particular, DDPM&DDIM use a special non-uniform speed:

$$\alpha_t = \exp\left(-\frac{1}{4}a(1-t)^2 - \frac{1}{2}b(1-t)\right),$$

where the suggested default values are $a = 19.9$, $b = 0.1$.

The time-uniform variant of this is

$$X_t = \sin(\frac{\pi}{2}t)X_1 + \cos(\frac{\pi}{2}t)X_0, \quad \forall t \in [0,1].$$

Example 4 (General Spherical Interpolations). The more general spherical interpolation, a.k.a. Slerp, is

$$X_t = \frac{\sin(\omega t)}{\sin(\omega)}X_1 + \frac{\sin(\omega(1-t))}{\sin(\omega)}X_0, \quad \forall t \in [0,1],$$

where $\omega \in [-\pi, \pi)$ is a parameter. This reduces to the straight interpolation with $\omega \to 0$, and it satisfies $\alpha_t^2 + \beta_t^2 = 1$ only with $\omega = \pm\pi/2$.

Remark 3. The interpolation process can be further generalized in at least two ways:

1) The interpolation function $\mathtt{I}$ can be randomized:

$$X_t = \mathtt{I}_t(X_0, X_1, \omega), \qquad \omega \sim \pi_\omega,$$

which depends on a random seed ω drawn form distribution π_ω.

2) Further, $\mathtt{I}_t(X_0, X_1)$ may not be differentiable w.r.t. t, such as the case when X_t is a diffusion process [Song et al., 2020b, Liu et al., 2022c, Peluchetti, 2021, Albergo et al., 2023]. We will discuss these possibilities.

Impacts of Different Interpolations

Understanding the impact of different interpolation processes is a key question of both theoretical and practical significance. Section 3 elaborates on this issue, that all interpolation processes that are pointwise transformable in a suitable sense yield essentially equivalent rectified flow dynamics and rectified couplings. Notably, all affine interpolation processes are pointwise transformable and therefore "essentially equivalent" [Kingma et al., 2021, Karras et al., 2022b, Shaul et al., 2023, Gao et al., 2024]. Consequently, it suffices to adopt a simple form, such as the straight $X_t = tX_1 + (1-t)X_0$, while maintaining the flexibility to recover all affine interpolations through adjustments in time parameterization and inference algorithms.

1.4 Models and Loss Functions

Beyond the standard quadratic loss (1.3), a variety of alternative loss functions have been explored. A notable example is the time-weighted loss function:

$$\int_0^1 \eta_t \mathbb{E}\left[\left\|\dot{X}_t - v_t(X_t)\right\|^2\right] \mathrm{d}t,$$

where η_t is a positive time weight. The non-uniform weights have been found useful and used in training large models such as Stable Diffusion 3, Flux and MovieGen.

In addition, the training procedure may vary depending on the choice of the target function that the neural network is designed to estimate. For example, rather than estimating v_t, many studies suggest training neural networks to approximate the condition expectation of noise X_0 or target X_1 given X_t:

$$\hat{x}_{0|t}(x) = \mathbb{E}\left[X_0 \mid X_t = x\right], \qquad \hat{x}_{1|t}(x) = \mathbb{E}\left[X_1 \mid X_t = x\right]. \tag{1.8}$$

With $X_t = tX_1 + (1-t)X_0$, we can see that $\hat{x}_{0|t}$ and $\hat{x}_{1|t}$ are related to v_t via linear relations:

$$\hat{x}_{1|t}(x) = x + (1-t)v_t(x), \qquad \hat{x}_{0|t}(x) = x - tv_t(x).$$

As a result, the different model formulations can be converted into one another by adjusting the time weightings in the training loss. For example,

if we set the problem as predicting $\hat{x}_{1|t}(x)$ with a time weighting η, the loss becomes

$$\int_0^1 \eta_t \mathbb{E}\left[\left\|X_1 - \hat{x}_{1|t}(x)\right\|^2\right] \mathrm{d}t = \int_0^1 \eta_t (1-t)^2 \mathbb{E}\left[\left\|X_1 - X_0 - v_t(X_t)\right\|^2\right] \mathrm{d}t,$$

which is equivalent to learning v_t with a weighting of $\tilde{\eta}_t = (1-t)^2 \eta_t$. Similarly, using different affine interpolation can be shown to correspond to training with different time weights (Section 3.3).

Beyond time weighting with square losses, it is possible to explore more general loss functions. A key property is that the minimum of the loss should enforce $v_t^X(x) = \mathbb{E}\left[\dot{X}_t | X_t = x\right]$. This can be ensured in general with Bregman divergence (Section 2.5).

1.5 Samplers

At inference time, we need to numerically solve the ODE $\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t$ to obtain samples. In addition to using off-the-shelf numerical ODE solvers, specialized algorithms should be developed by exploiting the intrinsic properties of rectified flow.

Euler's method is default algorithm for solving the ODEs of rectified flow. However, it has become common to 1) use non-uniform step sizes for better performance, and 2) in the case of curved interpolations, apply specialized natural Euler discretization update rules that align with the underlying interpolation. As discussed in Section 1.3, these two choices are interconnected, as using curved interpolations is equivalent to using straight interpolations with non-uniform time steps. See Section 3 for further details.

Another direction concerns the distinction between flow and diffusion models. Although rectified flow is introduced as a method for learning an ODE, it is possible to introduce stochasticity into the sampling process, yielding an stochastic differential equation (SDE) for sampling at inference time. Specifically, we may consider

$$\mathrm{d}Z_t = \underbrace{v_t(Z_t)\mathrm{d}t}_{\text{Rectified flow}} + \underbrace{\sigma_t^2 \nabla \log \rho_t(Z_t) + \sqrt{2\sigma_t^2}\mathrm{d}W_t}_{\text{Langevin dynamics on } \rho_t = \mathrm{Law}(Z_t)},$$

where ρ_t is the density function of Z_t. Here, we introduce an additional Langevin dynamics component on the density of the current random variable Z_t, which is always in an equilibrium state due to the rectified flow, and hence does not contribute to the change of the density in the ideal case when Z_t follows ρ_t exactly. However, when approximated in practice, the Langevin dynamics acts as a negative feedback mechainsm to help bring Z_t back towards ρ_t.

In general, we need to learn the score function $\nabla \log \rho_t$ in addition to v_t. In the special case of independent coupling and Gaussian X_0, however, ρ_t can be obtained from v_t using an explicit formula, which allows us to freely switch between ODE and SDE samplers without retraining models [Song et al., 2020b,a, Karras et al., 2022b].

1.6 Literature

We follow the order of first introducing flow and then presenting diffusion as an added option at inference time. This is in contrast to the historical development of diffusion models, which follows the reversed order:

1. **Diffusion.** Denoising diffusion models were initially developed with diffusion processes playing a fundamental role. Their development began from the perspective of hierarchical variational inference [Sohl-Dickstein et al., 2015, Ho et al., 2020], the learning of score functions for energy-based models [Song and Ermon, 2019, 2020], and time-reversal SDEs as introduced by Song et al. [2020b]. Complementary perspectives were offered through Schrödinger bridges [De Bortoli et al., 2021b, Shi et al., 2024], mixtures of diffusion [Peluchetti, 2021, 2023], and lazy EMs and h-transforms [Liu et al., 2022c, Ye et al., 2022]. All these works established diffusion processes and SDE concepts as central and indispensable building blocks.

2. **Diffusion $\Rightarrow$ Flow.** It was discovered that at inference time, the learned SDE models can be converted to deterministic ODE models. This means that one can switch between SDE and ODE samplers at inference time without re-training the model. This insight led to the development of denoising diffusion implicit models (DDIM) [Song et al., 2020a] and probability-flow ODEs (PF-ODEs) [Song et al., 2020b]. ODE-based procedures are simpler and faster than SDE-based inference, making them more appealing when computational speed is a concern. However, these methods still rely on the SDE as an intermediate step, which is somewhat counterintuitive, given that SDEs are more sophisticated than ODEs.

3. **Flow.** It was later observed that ODE models can be introduced directly, bypassing the need for SDEs altogether, whether for theoretical or practical reasons. The works include rectified flow [Liu et al., 2022a], flow matching [Lipman et al., 2022], stochastic interpolation [Albergo et al., 2023], and alpha-blending [Heitz et al., 2023]. Also related is related the action matching proposed by Neklyudov et al. [2023].

> Remark 4. The term *diffusion models* has frequently been used to refer specifically to the DDIM and DDPM approaches. Here, we use it to refer to general models that learn the diffusion process as a generative mechanism.

CHAPTER TWO

Marginals and Errors

We show that any time-differential stochastic process $\{X_t\}$ shares the same marginal distributions with the rectified flow $\{Z_t\}$ induced from $\{X_t\}$, that is,

$$\text{Law}(Z_t) = \text{Law}(X_t) \quad \text{for} \quad \forall t.$$

We provide the proof and also establish error bounds when the rectified flow is learned exactly.

2.1 Marginals are Determined by $\mathbb{E}[\dot{X}_t|X_t]$

In the following, we establish the marginal preserving property, drawing connection to the continuity equation.

Definition 3. For a path-wise continuously differentiable random process $\{X_t\} = \{X_t : t \in [0,1]\}$, its expected velocity v^X, also called its rectified flow (RF) velocity, is defined as

$$v_t^X(x) = \mathbb{E}\left[\dot{X}_t \mid X_t = x\right], \quad \forall x \in \text{supp}(X_t).$$

For $x \notin \text{supp}(X_t)$, the conditional expectation is not defined and we set v^X arbitrarily, say $v_t^X(x) = 0$.

Definition 4. We call that $\{X_t\}$ is *rectifiable* if v^X is locally bounded and the solution of the integral equation below exists and is unique initialized from $Z_0 = X_0$ almost surely:

$$Z_t = Z_0 + \int_0^t v_s^X(Z_s)\mathrm{d}s, \quad \forall t \in [0,1], \quad Z_0 = X_0. \tag{2.1}$$

In this case, $\{Z_t\} = \{Z_t : t \in [0,1]\}$ is called the *rectified flow* induced from $\{X_t\}$.

The integral equation (2.1) reduces to the ODE $\frac{\mathrm{d}}{\mathrm{d}t}Z_t = v_t^X(Z_t)$ if Z_t is time-differentiable for all t.

Theorem 1. Assume $\{X_t\}$ is rectifiable and $\{Z_t\}$ is its rectified flow. Then $\text{Law}(Z_t) = \text{Law}(X_t)$ for $\forall t \in [0,1]$.

Remark 5. The marginal distributions of X_t in a time-differential stochastic process $\{X_t\}$ are fully determined by the initial condition and the expected velocity $v^X(x) = \mathbb{E}\left[\dot{X}_t \mid X_t = x\right]$, which depends solely on the conditional expectation of $\dot{X}_t$ given X_t at each point.

Higher-order moments of the conditional distribution $\dot{X}_t|X_t$ *do not* influence the marginal distributions of X_t at each individual time t, but they do affect the joint distributions, such as (X_{t_1}, X_{t_2}) for different time points t_1 and t_2.

Proof. Denote by $\mathcal{C}_c^1(\mathbb{R}^d;\mathbb{R})$ the set of compactly supported continuously differentiable functions. For each $h \in \mathcal{C}_c^1(\mathbb{R}^d;\mathbb{R})$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[h(X_t)\right] = \mathbb{E}\left[\nabla h(X_t)^\top \dot{X}_t\right] = \mathbb{E}\left[\nabla h(X_t)^\top v_t^X(X_t)\right], \tag{2.2}$$

where we used $v_t^X(X_t) = \mathbb{E}\left[\dot{X}_t|X_t\right]$. This readily establishes a series of equations on the probability measures $\pi_t := \mathrm{Law}(X_t)$:

$$\int h\mathrm{d}\pi_t - \int h\pi_0 = \int_0^t \int \nabla h^\top v_s^X \mathrm{d}\pi_s, \quad \forall h \in \mathcal{C}_c^1(\mathbb{R}^d;\mathbb{R}). \tag{2.3}$$

It turns out these (infinite number of) equations identifies a unique solution of the measures $\{\pi_t\}$ given an initial π_0, if and only if Equation (2.1) admits an unique solution initialized from $Z_0 \sim \pi_0$. (See Corollary 1.3 of Kurtz [2011] or Theorem 4.1 of Ambrosio and Crippa [2008]).

Because Z_t is driven by the same velocity field v^X, its marginal law $\mathrm{Law}(Z_t)$ solves the very same equation (2.3) with the same initial condition ($Z_0 = X_0$). Hence, we have $\mathrm{Law}(Z_t) = \mathrm{Law}(X_t)$ by the uniqueness of the solutions. □

An Elementary Proof

The key part of the proof is the uniqueness of the solutions of (2.3). We give an elementary proof to illustrate the idea here.

Let $\Phi_{t|s}$ be the transfer map from Z_s to Z_t following the ODE $\mathrm{d}Z_t = v_t^X(Z_t)\mathrm{d}t$, such that $Z_t = \Phi_{t|s}(Z_s)$. Note that

$$0 = \frac{\mathrm{d}}{\mathrm{d}s}Z_t = \frac{\mathrm{d}}{\mathrm{d}s}\Phi_{t|s}(Z_s) = \partial_s\Phi_{t|s}(Z_s) + \nabla\Phi_{t|s}(Z_s)^\top v_s^X(Z_s),$$

where we write $\nabla\Phi = [\partial_i\Phi_j]_{ij}$, which is the transpose of the Jacobian equation.

Assume the solution of the ODE exists and is unique starting from any point $Z_s = z$ on $\mathbb{R}^d$. We can replace Z_s with any z and get the so-called Kolmogorov backward equation:

$$\partial_s\Phi_{t|s}(z) + \nabla\Phi_{t|s}(z)^\top v_s^X(z) = 0, \quad \forall z \in \mathbb{R}^d. \tag{2.4}$$

As illustrated in Figure 2.1, define

$$\hat{Z}_{t|s} = \Phi_{t|s}^Z(X_s),$$

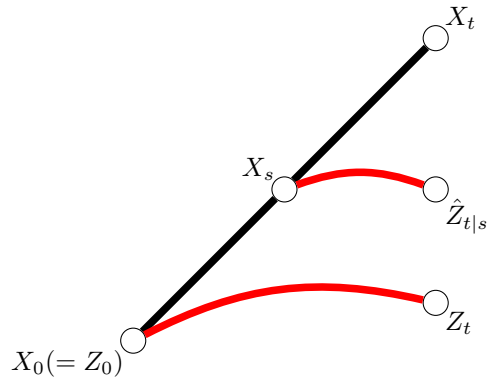


Figure 2.1: Illustrating $\hat{Z}_{t|s} = \Phi_{t|s}^Z(X_s)$, which yields $\hat{Z}_{t|0} = Z_t$ and $\hat{Z}_{t|t} = X_t$.

which is the result we get by following X from X_0 to X_s, and then switch to the ODE to get $\hat{Z}_{t|s}$. Obviously, we have $\hat{Z}_{t|t} = X_t$ and $\hat{Z}_{t|0} = Z_t$.

For any test function $h \in \mathcal{C}_c^1(\mathbb{R}^d; \mathbb{R})$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\left[h(\hat{Z}_{t|s})\right] &= \frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\left[h(\Phi_{t|s}(X_s))\right] \\
&= \mathbb{E}\left[\nabla h(\hat{Z}_{t|s})^\top(\partial_s\Phi_{t|s}(X_s) + \nabla\Phi_{t|s}(X_s)^\top\dot{X}_s)\right] \\
&= \mathbb{E}\left[\nabla h(\hat{Z}_{t|s})^\top(\nabla\Phi_{t|s}(X_s)^\top(\dot{X}_s - v_s^X(X_s))\right] \qquad \text{//using (2.4)} \\
&= \mathbb{E}\left[\nabla h(\hat{Z}_{t|s})^\top(\nabla\Phi_{t|s}(X_s)^\top(\mathbb{E}\left[\dot{X}_s|X_s\right] - v_s^X(X_s))\right] \\
&= 0.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[h(X_t)\right] - \mathbb{E}\left[h(Z_t)\right] = \mathbb{E}\left[h(\hat{Z}_{t|t})\right] - \mathbb{E}\left[h(\hat{Z}_{t|0})\right] = \int_0^t \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[h(\hat{Z}_{t|s})\right]\mathrm{d}s = 0.$$

Because $\mathbb{E}\left[h(X_t)\right] = \mathbb{E}\left[h(Z_t)\right]$ for any test function h, we conclude that X_t and Z_t have the same distributions.

2.2 Continuity Equations

If $\pi_t = \mathrm{Law}(X_t)$ admits a smooth density function ρ_t, then, as we show below, Equation (2.3) can be reduced to the *continuity equation*:

$$\partial_t\rho_t(x) = -\nabla\cdot(v_t^X(x)\rho_t(x)), \tag{2.5}$$

which explicitly characterizes the evolution of the densities ρ_t of X_t based on the expected velocity $v_t^X(x) = \mathbb{E}\left[\dot{X}_t|X_t = x\right]$. Here, for a velocity field $v\colon \mathbb{R}^d \to \mathbb{R}^d$, we denote by $\nabla\cdot v$ the *divergence operator* of velocity field v, defined as

$$\nabla\cdot v(x) = \sum_i \partial_{x_i} v(x)_i = \mathrm{Trace}(\nabla v(x)).$$

It is the trace of the Jacobian matrix of $v(x)$.

Derivation by Integration by Parts

The key of the derivation of (2.5) from 2.3 is using integration by parts. Writing the equation in terms of the density function:

$$\int h(x)\partial_t\rho_t(x)\mathrm{d}x = \int \nabla h(x)^\top v_t^X(x)\rho_t(x)\mathrm{d}x.$$

Applying integration by parts on the right hand side: Note that left and right hand side of (2.3) are

$$\int \nabla h(x)^\top v_t^X(x)\rho_t(x)\mathrm{d}x = -\int h(x)\,\nabla\cdot(v_t^X(x)\rho_t(x))\mathrm{d}x \qquad \text{//integration by parts.}$$

This shows that

$$\int h(x)(\partial_t\rho_t(x) + \nabla\cdot(v_t^X(x)\rho_t(x)))\mathrm{d}x = 0,$$

for any test function h, which implies that $\partial_t \rho_t + \nabla\cdot(v_t^X \rho_t) = 0$.

If π_t does not admit density functions, Equation (2.3) is simply defined as the continuity equation in the weak sense, which is written formally as

$$\dot{\pi}_t + \nabla\cdot(v_t^X \pi_t) = 0. \tag{2.6}$$

2.2.1 The Divergence Operator

In physics, the divergence of a vector field measures how much the field acts as a source or sink at a specific point. It quantifies the "outflow" of field vectors from an infinitesimal region. As shown in Figure 2.2, positive divergence ($\nabla\cdot v(x) > 0$) indicates a source, while negative divergence ($\nabla\cdot v(x) < 0$) indicates a sink.

To see why this is the case, note that with integration by parts, we have for any velocity field v and density function ρ,

$$\int \nabla\cdot v(x)\rho(x)\mathrm{d}x = -\int v(x)^\top \nabla\rho(x)\mathrm{d}x.$$

This is rewritten into

$$\mathbb{E}_{X\sim\rho}[\nabla\cdot v(X)] = -\mathbb{E}_{X\sim\rho}[v(X)^\top \nabla \log \rho(X)],$$

which is also known as *Stein's identity*. It shows that the expectation of divergence $\nabla\cdot v(X)$ under any distribution ρ equals the negative of expected inner product of $v(X)$ and $\nabla \log \rho(X)$.

Now, let $\rho \sim \texttt{Normal}(x_0, \sigma^2 I)$ be a Gaussian distribution centered a point x_0. We have $\nabla \log \rho(X) = -(X - x_0)/\sigma^2$. Hence,

$$\mathbb{E}\left[\nabla\cdot v(X)\right] = \frac{1}{\sigma^2}\mathbb{E}\left[v(X)^\top (X - x_0)\right]. \tag{2.7}$$

Thus, the expected divergence around x_0 is proportional to the expected inner product of $v(X)$ with the lines $(X - x_0)$ radiating outward from x_0 to X. Taking the limit as $\sigma \to 0$, this confirms that $\nabla\cdot v(x_0)$ serves as a local measure of "sourceness" at x_0.

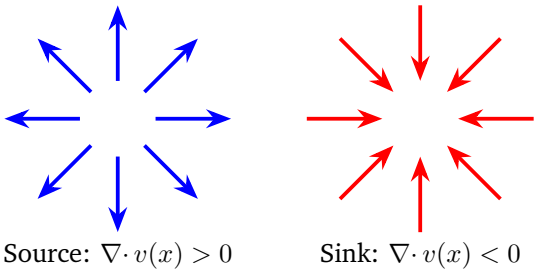


Figure 2.2: Illustrating divergence.

> Remark 6. Back to the continuity equation $\partial_t \rho_t = -\nabla\cdot(v_t \rho_t)$. It shows that the rate of change of ρ_t at each point equals the negative divergence (which measures the amount of inward flow) of the flux $v_t\rho_t$. Here, $v_t\rho_t$ is the product of the velocity field v_t and the density ρ_t, which together define the flow of the density ρ_t driven by the velocity field v_t.
>
> $\partial_t \rho_t = -\nabla\cdot(v_t\rho_t)$: *Change rate of* ρ_t *= negative divergence of flux.*

2.2.2 Numerical Approximation of $\nabla\cdot v$

Calculating the divergence requires to sum over the partial derivatives $\partial_{x_i} v(x)_i$. Because each of these terms takes derivative of a different objective $v(x)_i$ w.r.t. a different input x_i, there exists no efficient vectorized operators that calculate all the terms simultaneously, they have to be calculated separately with a for loop, which is highly inefficient.

The Stein's identity in (2.7) provides a convenient approximation of divergence. Taking a small $\sigma \approx 0$, we have

$$\begin{aligned}\nabla\cdot v(x_0) &\approx \mathbb{E}\left[\nabla\cdot v(x_0+\sigma\xi)\right] \\ &= \frac{1}{\sigma}\mathbb{E}\left[v(x_0+\sigma\xi)^\top \xi\right] \\ &= \frac{1}{\sigma}\mathbb{E}\left[(v(x_0+\sigma\xi)-v(x_0))^\top \xi\right], \end{aligned} \tag{2.8}$$

where the right hand side can be approximated by Monte Carlo sampling of $\xi \sim \texttt{Normal}(0, I)$.

In the limit of $\sigma \to 0$, we obtain an unbiased estimation.

Proposition 1. Assume $\mathbb{E}\left[\xi\xi^\top\right] = I$, we have

$$\nabla\cdot v(x) = \frac{\mathrm{d}}{\mathrm{d}\sigma}\mathbb{E}\left[v(x+\sigma\xi)^\top\xi\right]\Big|_{\sigma=0}. \tag{2.9}$$

Remark 7. It does not necessarily require a Gaussian ξ. The finite difference approximation of (2.9) yields (2.8).

Proof.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathbb{E}\left[v(x+\sigma\xi)^\top\xi\right]\Big|_{\sigma=0} &= \mathbb{E}\left[\xi^\top\nabla v(x)\xi\right] \\ &= \mathrm{Trace}(\nabla v(x)\mathbb{E}\left[\xi\xi^\top\right]) \\ &= \mathrm{Trace}(\nabla v(x)) = \nabla\cdot v(x).\end{aligned}$$

□

2.3 Wasserstein Bounds

Due to learning and optimization errors, we may only yield an approximation $\hat{v}_t$ of the expected velocity $v_t^X(x) = \mathbb{E}\left[\dot{X}_t | X_t = x\right]$. Let $\pi_t^{\hat{Z}} = \mathrm{Law}(\hat{Z}_t)$ be the distribution of $\hat{Z}_t$ following the ODE $\frac{\mathrm{d}}{\mathrm{d}t}\hat{Z}_t = \hat{v}_t(Z_t)$ with $\hat{Z}_0 = X_0$, and $\pi_t^X = \mathrm{Law}(X_t)$ the marginal distribution of the interpolation process. We show in the following that

$$\begin{aligned}W_{1,q}(\pi_t^X, \pi_t^{\hat{Z}}) &\le \int_0^t \mathbb{E}\left[\left\|\nabla\Phi_{t|s}^{\hat{Z}}(X_s)^\top(v_s^X(X_s)-\hat{v}_s(X_s))\right\|_{q^*}\right]\mathrm{d}s \\ &\le \int_0^t \mathbb{E}\left[\left\|\nabla\Phi_{t|s}^{\hat{Z}}(X_s)^\top(\dot{X}_s-\hat{v}_s(X_s))\right\|_{q^*}\right]\mathrm{d}s, \end{aligned} \tag{2.10}$$

where $W_{1,q}(\cdot,\cdot)$ is the 1-Wasserstein distance under a norm $\|\cdot\|_q$, and $\|\cdot\|_{q^*}$ is the dual norm of $\|\cdot\|_q$, and $\Phi_{t|s}^{\hat{Z}}$ is the transfer map from $\hat{Z}_s$ to $\hat{Z}_t$. We can see that this bound mimics the training loss (1.3), but depends on the Jacobian $\nabla\Phi_{t|s}^{\hat{Z}}$ and does not square the loss. It is possible to use this bound as the training loss, if the Jacobian can be approximated properly.

Lemma 1 (Wasserstein Distances). 1) Let $\frac{\mathrm{d}}{\mathrm{d}t} Z_t = v_t^Z(Z_t)$ be an ODE that yields an unique solution passing $Z_t = z$ for $\forall t \in [0,1]$ and $z \in \mathbb{R}^d$. Let $\Phi_{t|s}^Z$ be its transfer map, such that $Z_t = \Phi_{t|s}(Z_s)$.

2) Let $\{X_t\}$ be a time-differentiable process with RF velocity field $v_t^X(x) = \mathbb{E}\left[\dot{X}_t | X_t = x\right]$. Assume $X_0 = Z_0$.

3) Let $\pi_t^X = \mathrm{Law}(X_t)$ and $\pi_t^Z = \mathrm{Law}(Z_t)$ be the marginal laws. Let $\|\cdot\|_q$ be any norm on $\mathbb{R}^d$. Consider the 1-Wasserstein distance w.r.t. a norm $\|\cdot\|_q$:

$$W_{1,q}(\pi_t^X, \pi_t^Z) = \sup_h \{\mathbb{E}\left[h(X_t)\right] - \mathbb{E}\left[h(Z_t)\right] \quad s.t. \quad \sup_{x \in \mathbb{R}^d} \|\nabla h(x)\|_q \leq 1\}.$$

Then, we have

$$\begin{aligned} W_{1,q}(\pi_t^X, \pi_t^Z) &\leq \int_0^t \mathbb{E}\left[\left\|\nabla \Phi_{t|s}^Z(X_s)^\top (v_s^X(X_s) - v_s^Z(X_s))\right\|_{q^*}\right] \mathrm{d}s \\ &\leq \int_0^t \mathbb{E}\left[\left\|\nabla \Phi_{t|s}^Z(X_s)^\top (\dot{X}_s - v_s^Z(X_s))\right\|_{q^*}\right] \mathrm{d}s, \end{aligned}$$

where $\|\cdot\|_{q^*}$ is the dual norm of $\|\cdot\|_q$, given by $\|x\|_{q^*} = \sup_y \{x^\top y \colon \|y\| \leq 1\}$.

Proof. Let $\hat{Z}_{t|s} = \Phi_{t|s}^Z(X_s)$, so that we have $\hat{Z}_{t|t} = X_t$ and $\hat{Z}_{t|0} = Z_t$.

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\left[h(\hat{Z}_{t|s})\right] &= \frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\left[h(\Phi_{t|s}(X_s))\right] \\ &= \mathbb{E}\left[\nabla h(\hat{Z}_{t|s})^\top (\partial_s \Phi_{t|s}(X_s) + \nabla \Phi_{t|s}(X_s)^\top \dot{X}_s)\right] \\ &= \mathbb{E}\left[\nabla h(\hat{Z}_{t|s})^\top (\nabla \Phi_{t|s}(X_s)^\top (\dot{X}_s - v_s^Z(X_s))\right] \\ &= \mathbb{E}\left[\nabla h(\hat{Z}_{t|s})^\top (\nabla \Phi_{t|s}(X_s)^\top (v_s^X(X_s) - v_s^Z(X_s))\right], \end{aligned}$$

where we used the backward equation $\partial_s \Phi_{t|s}(z) + \nabla \Phi_{t|s}(z)^\top v_s^Z(z) = 0$. Hence,

$$\begin{aligned} &\mathbb{E}\left[h(X_t)\right] - \mathbb{E}\left[h(Z_t)\right] \\ &= \mathbb{E}\left[h(\hat{Z}_{t|t})\right] - \mathbb{E}\left[h(\hat{Z}_{t|0})\right] \\ &= \int_0^t \frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\left[h(\hat{Z}_{t|s})\right] \mathrm{d}s \\ &= \int_0^t \mathbb{E}\left[\nabla h(\hat{Z}_{t|s})^\top (\nabla \Phi_{t|s}(X_s)^\top (v_s^X(X_s) - v_s^Z(X_s))\right] \mathrm{d}s \\ &\leq \int_0^t \mathbb{E}\left[\left\|\nabla h(\hat{Z}_{t|s})\right\|_q \left\|\nabla \Phi_{t|s}(X_s)^\top (v_s^X(X_s) - v_s^Z(X_s))\right\|_{q^*}\right] \mathrm{d}s. \end{aligned}$$

This yields the result by the definition of the 1-Wasserstein distance:

$$W_{1,q}(\pi_t^X, \pi_t^Z) \le \int_0^t \mathbb{E}\left[\left\|\nabla \Phi_{t|s}^Z(X_s)^\top (v_s^X(X_s) - v_s^Z(X_s))\right\|_{q^*}\right] \mathrm{d}s$$
$$\le \int_0^t \mathbb{E}\left[\left\|\nabla \Phi_{t|s}^Z(X_s)^\top (\dot{X}_s - v_s^Z(X_s))\right\|_{q^*}\right] \mathrm{d}s,$$

where the last inequality follows Jensen's inequality given that every norm is convex. □

The following is an upper bound of KL divergence between the marginal distributions traded by an ODE and time-differential process.

2.4 KL Divergence

Besides Wasserstein distances, we often need to keep track of the KL divergence of the marginal distributions driven by two time-differentiable processes. The follow result plays a key role in many places.

Lemma 2 (KL Divergence Between ODEs). Let $\{X_t\}$ and $\{X_t'\}$ be two time-differentiable stochastic processes with RF velocity fields v_t and v_t', and smooth log densities ρ_t, ρ_t', respectively. We have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\rho_t \,||\, \rho_t') = \mathbb{E}_{X_t\sim\rho_t}[(\nabla \log \rho_t(X_t) - \nabla \log \rho_t'(X_t))^\top (v_t(X_t) - v_t'(X_t))].$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{H}(\rho_t) = -\mathbb{E}_{X_t\sim\rho_t}[\log \rho_t(X_t)^\top v_t(X_t)],$$

where $\mathrm{KL}(\cdot \,||\, \cdot)$ and $\mathrm{H}(\cdot)$ denotes KL divergence and entropy, respectively.

Remark 8. This shows that the change rate of KL divergence equals the expected inner product of the score difference $\nabla \log \rho_t - \nabla \log \rho_t'$ and velocity difference $v_t - v_t'$. The change rate of entropy $\mathrm{H}(\rho_t)$ equals the negative expected inner product of $\nabla \log \rho_t$ and v_t.

Remark 9. We have by Cauchy–Schwarz inequality,

$$\mathrm{KL}(\rho_1 \,||\, \rho_1') \le \int_0^1 \mathbb{E}_{X_t\sim\rho_t}[\|\nabla \log \rho_t(X_t) - \nabla \log \rho_t'(Z_t)\| \, \|v_t(X_t) - v_t'(X_t)\|]\mathrm{d}t.$$

The bound depends on both ρ_t and v_t.

Remark 10. As shown in Section 4.1, if $X_t = \alpha_t X_1 + \beta_t X_0$ is an affine interpolation with $X_0 \perp\!\!\!\perp X_1$ and $X_0 \sim \texttt{Normal}(0, I)$, then $\nabla \log \rho_t$ is related to v_t in closed form via

$$\nabla \log \rho_t(x) = \eta_t \left(v_t(x) - \frac{\dot{\alpha}_t}{\alpha_t} x\right), \quad \text{with} \quad \eta_t = \frac{1}{\beta_t^2}\left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right)^{-1}.$$

Assume the same relation also holds for $\nabla \log \rho_t$ and v_t. We have

$$\mathrm{KL}(\rho_1 \mid\mid \rho_1') = \int_0^1 \mathbb{E}_{X_t \sim \rho_t}[\eta_t \left\| v_t(X_t) - v_t'(X_t) \right\|^2] \mathrm{d}t,$$

which reduces to a time-weighted square loss function.

Proof. Let $\dot{Z}_t = v_t(Z_t)$ and $\dot{Z}_t' = v_t'(Z_t')$ be the rectified flow of $\{X_t\}$ and $\{X_t'\}$, respectively. Note that

$$\mathrm{KL}(\rho_t \mid\mid \rho_t') = \mathbb{E}_{Z_t \sim \rho_t}[\log(\rho_t(Z_t)/\rho_t'(Z_t)].$$

Its time derivative is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\rho_t \mid\mid \rho_t') = \underbrace{\mathbb{E}\left[\nabla(\log \rho_t(Z_t)/\rho_t'(Z_t))\dot{Z}_t\right]}_{I} + \underbrace{\mathbb{E}\left[\partial_t \log(\rho_t(Z_t)/\rho_t'(Z_t))\right]}_{II},$$

where we used $\frac{\mathrm{d}}{\mathrm{d}t} r_t(Z_t) = \partial_t r_t(Z_t) + \nabla r_t(Z_t)^\top \dot{Z}_t$ with $r_t = \log(\rho_t/\rho_t')$; the first term here differentiate through the random variable Z_t, and the second term through the log density ratio r_t.

Because $\dot{Z}_t = v_t(Z_t)$, the first term equals

$$I = \mathbb{E}\left[(\nabla \log \rho_t(Z_t) - \nabla \log \rho_t'(Z_t))^\top v_t(Z_t)\right]. \tag{2.11}$$

The second term can be separated into

$$II = \underbrace{\int \rho_t(z) \partial_t \log \rho_t(z) \mathrm{d}z}_{II_1} - \underbrace{\int \rho_t(z) \partial_t \log \rho_t'(z) \mathrm{d}z}_{II_2},$$

where the first part II_1 equals zero, because

$$II_1 = \int \rho_t(z) \partial_t \log \rho_t(z) \mathrm{d}z = \int \partial_t \rho_t(z) \mathrm{d}z = \partial_t \int \rho_t(z) \mathrm{d}z = \partial_t 1 = 0. \tag{2.12}$$

For the second part II_2, note the continuity equation $\partial_t \rho_t'(z) = -\nabla \cdot (v_t'(z)\rho_t'(z))$. By dividing both sides with $\rho_t'(z)$, it can be rewritten into

$$\partial_t \log \rho_t'(z) = -(\nabla \cdot v_t'(z) + \nabla \log \rho_t'(z)^\top v_t'(z)).$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[\partial_t \log \rho_t(Z_t)\right] &= \mathbb{E}\left[-\nabla \cdot v_t'(Z_t) - \nabla \log \rho_t'(Z_t)^\top v_t'(Z_t)\right] \\
&= \mathbb{E}\left[\nabla \log \rho_t(Z_t)^\top v_t'(Z_t) - \nabla \log \rho_t'(Z_t)^\top v_t'(Z_t)\right], \\
&= \mathbb{E}\left[(\nabla \log \rho_t(Z_t) - \nabla \log \rho_t'(Z_t))^\top v_t'(Z_t)\right], \quad (2.13)
\end{aligned}$$

where we used the integration by parts formula for divergence: $\int \rho_t(z) \nabla \cdot v_t(z) \mathrm{d}z = -\int \nabla \rho_t(z)^\top v_t(z) \mathrm{d}z$.

Combing Equation 2.11, 2.12 and 2.13 yields the result for $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\rho_t \mid\mid \rho_t')$. The result for the entropy $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{H}(\rho_t)$ follows similarly except simpler. □

2.5 Bregman Divergence

The use of the quadratic loss (1.3) is critical in ensuring that $v_t^*(X_t) = \mathbb{E}\left[\dot{X}_t | X_t\right]$ for $\forall t \in [0,1]$ at the optimum. The same property can be achieved by other loss functions. A general loss that we may consider is of form

$$\int_0^1 \psi_t \left(\mathbb{E}\left[\ell_t(\dot{X}_t, v_t(X_t)) \right] \right) \mathrm{d}t,$$

where $\ell_t(\cdot,\cdot)$ is a loss function measuring the difference between $\dot{X}_t$ and $v_t(X_t)$ at each time t, and $\psi_t(\cdot)$ decides how the losses across different time are aggregated. We should choose ψ_t and ℓ_t to satisfy the following conditions to ensure that minimum of the loss is attained by $v_t^*(X_t) = \mathbb{E}\left[\dot{X}_t | X_t\right]$ for $\forall t \in [0,1]$:

1. $\psi_t(x) \geq 0$ for $\forall x$, and $\psi_t(x) = 0$ iff $x = 0$.
2. $\mathbb{E}\left[\ell_t(Y,x)\right] \geq 0$, and $\mathbb{E}\left[\ell_t(Y,x)\right] = 0$ iff $x = \mathbb{E}\left[Y\right]$, for any random variable x.

Here, the property of ℓ_t, referred to as the *mean-as-minimizer*, is critical. It is obviously ensured by the squared loss $\ell_t(y,x) = (y-x)^2$. Bregman divergence represents a more general class of losses that satisfy this property [e.g., Banerjee et al., 2005].

Definition 5. Let $h\colon \Omega \to \mathbb{R}$ be a continuously-differentiable, strictly convex function defined on a convex set Ω. The Bregman divergence associated with h is defined as

$$B_h(y,x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle, \quad \forall x, y \in \Omega.$$

It is the difference between $h(y)$, and the first-order Taylor expansion $h(x) + \langle \nabla h(x), y - x \rangle$ around point x evaluated at point y.

We have $B_h(y,x) \geq 0$ for all x, y as a consequence of the convexity of h, and $B_h(y,x) = 0$ iff $x = y$, which is true when h is *strictly* convex. More importantly, Bregman divergence admits a generalized bias–variance decomposition similar to that of mean square errors (MSEs), which ensures that the minimum of $\mathbb{E}\left[B_h(Y, f(X))\right]$ is $f^*(X) = \mathbb{E}\left[Y|X\right]$.

Lemma 3. Bregman divergence yields the following decomposition:

$$\mathbb{E}\left[B_h(Y,x)\right] = \underbrace{\mathbb{E}\left[B_h(Y,\ \mathbb{E}\left[Y\right])\right]}_{variance} + \underbrace{\mathbb{E}\left[B_h(\mathbb{E}\left[Y\right],\ x)\right]}_{bias}.$$

Hence, the solution of $\min_x \mathbb{E}\left[B_h(Y,x)\right]$ is achieved by $x^* = \mathbb{E}\left[Y\right]$.

Proof.

$$\begin{aligned}\mathbb{E}\left[B_h(Y,x)\right] &= \mathbb{E}\left[h(Y) - h(x) - \nabla h(x)^\top (Y - x)\right] \\ &= \mathbb{E}\left[h(Y) - h(x) - \nabla h(x)^\top (\mathbb{E}\left[Y\right] - x)\right] \\ &= B_h(\mathbb{E}\left[Y\right],\, x) + \mathbb{E}\left[h(Y)\right] - \mathbb{E}\left[h(\mathbb{E}\left[Y\right])\right].\end{aligned}$$

The result follows if we note that

$$\begin{aligned}\mathbb{E}\left[B_h(Y,\, \mathbb{E}\left[Y\right])\right] &\overset{*}{=} \mathbb{E}\left[h(Y) - h(\mathbb{E}\left[Y\right]) - \nabla h(\mathbb{E}\left[Y\right])^\top (Y - \mathbb{E}\left[Y\right])\right] \\ &= \mathbb{E}\left[h(Y)\right] - \mathbb{E}\left[h(\mathbb{E}\left[Y\right])\right],\end{aligned}$$

where the last term in $\overset{*}{=}$ is canceled because $\mathbb{E}[Y - \mathbb{E}[Y]] = 0$. □

Remark 11. As an alternative proof, note that

$$\mathbb{E}\left[B_h(Y,x)\right] = \mathbb{E}\left[h(Y)\right] - h(x) - \nabla h(x)^\top (\mathbb{E}\left[Y\right] - x).$$

Taking the derivative w.r.t. x:

$$\begin{aligned}\nabla_x \mathbb{E}\left[B_h(Y,x)\right] &= -\nabla h(x) - \nabla^2 h(x)(\mathbb{E}\left[Y\right] - x) + \nabla h(x) \\ &= -\nabla^2 h(x)(\mathbb{E}\left[Y\right] - x).\end{aligned}$$

Because $\nabla^2 h(x) \succ 0$ as h is strictly convex, solving $\nabla_x \mathbb{E}\left[B_h(Y,x)\right] = -\nabla^2 h(x)(\mathbb{E}\left[Y\right] - x) = 0$ yields $x = \mathbb{E}\left[Y\right]$.

2.5.1 Semantic Losses

In practice, it can be beneficial to impose the loss on a semantic space. Let $f(x)$ be feature mapping, one can consider

$$\int_0^1 \mathbb{E}\left[\|f(X_1) - f(X_t + (1-t)v_t(X_t))\|^2\right] \mathrm{d}t, \tag{2.14}$$

where we consider $\hat{X}_{1|t} = X_t + (1-t)v_t(X_t)$ as a predict of X_1. An example of this is the LPIPS loss [Zhang et al., 2018]; see Lee et al. [2024]. However, for nonlinear f, the loss above does not generally satisfy the *mean-as-minimizer* property, and hence may yield biased estimation.

Tangent Loss

Tangent loss is one approach to incorporate loss to avoid introducing bias while incorporating the information of nonlinear feature maps.

Assume that $f_t(X_t)$ is a useful representation of X_t at each time t. The transformed interpolation and ODEs are:

$$X_t^f = f_t(X_t), \qquad Z_t^f = f_t(Z_t).$$

By chain rule, the slopes of the induced curves are

$$\dot{X}_t^f = \nabla f_t(X_t)^\top \dot{X}_t + \partial_t f_t(X_t), \qquad \dot{Z}_t^f = \nabla f_t(Z_t)^\top v_t(Z_t) + \partial_t f_t(Z_t).$$

Matching these two slopes yield

$$\min_v \int_0^1 \mathbb{E}\left[\left\|\nabla f_t(X_t)^\top(\dot{X}_t - v_t(X_t))\right\|^2\right] \mathrm{d}t. \tag{2.15}$$

This allows us to place higher weight on the directions are important w.r.t. the feature network f_t. In practice, the space spanned by $\nabla f(X_t)$ may be degenerate. In this case, we can take the linear combination of the loss above with the standard L2 loss. See Liu et al. [2022a].

Remark 12. Note that (2.15) coincides with the Wasserstein bound in (2.10) if we take $f_t(x) = \Phi_{1|t}(x)$ to be the transfer mapping of $\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t$.

To avoid calculating the derivative, we can use Taylor approximation:

$$\min_v \int_0^1 \mathbb{E}\left[\left\|\frac{1}{\varepsilon_t}(f_{t+\varepsilon_t}(X_t + \varepsilon_t\dot{X}_t) - f_{t+\varepsilon_t}(X_t + \varepsilon_t v_t(X_t)))\right\|^2\right] \mathrm{d}t, \tag{2.16}$$

which converges to (2.15) when $\varepsilon_t \approx 0$. It reduces to (2.14) when $\varepsilon = 1-t$ when $X_t = tX_1 + (1-t)X_0$.

What is the Effect of Using Biased Losses?

Using losses that do not satisfy the mean-as-minimizer property is equivalent to impose a loss-dependent reweighing on the data points. Let us consider, for example,

$$\min_f \mathbb{E}\left[\ell(Y, f(X))\right].$$

Assume we minimize f in the set of all functions, using calculus of variations, the optimal solution f^* should satisfy

$$\mathbb{E}\left[\nabla_f \ell(Y, f^*(X)) \mid X\right] = 0,$$

which can be writing into

$$f^*(X) = \frac{\mathbb{E}\left[w(X,Y)Y \mid X\right]}{\mathbb{E}\left[w(X,Y) \mid X\right]}.$$

where the weighting function is $w(X,Y) := \nabla_f \ell(Y, f^*(X))/(Y - f^*(X))$, if they can be defined properly. Hence, these loss functions induce a non-uniform weighting of data points, failing to accurately capture the original distribution. While such biases might be hard to detect when training models on very large datasets, they can potentially lead to unexpected effects, such as bias amplification or mode collapse.

Remark 13. The magnitude of the bias depends on the variance of Y conditioned on X. If X is fully deterministic given X, then no bias is induced even if ℓ does not have the mean-as-minimizer property. As suggested in Lee et al. [2024], the label $Y = \dot{X}_t$ can be highly deterministic given X_t, using a nonlinear loss such as LPIPS may suffer from the bias issue.

CHAPTER THREE

Interpolations and Equivariance

The choice of the interpolation process can have a significant impact on inference performance and speed, and it appears to be a decision that must be made during the pre-training phase. However, for interpolation processes that are pointwise transformable one-to-one in a suitable sense, the trajectories of their induced rectified flow can also be transformed one-to-one, indicating that $\texttt{Rectify}(\cdot)$ is equivariant under pointwise transformations. This result holds not only for continuous-time ODEs but also for discretized trajectories of the ODEs if a particular "natural Euler" method is used, which takes the underlying interpolation into account.

Notably, all affine interpolation processes are pointwise transformable through a simple rescaling of the time t and the input x, and hence they satisfy the equivariance results mentioned above. In particular, the discretized DDIM algorithm is a natural Euler method for spherical interpolation, which is therefore equivalent to straight interpolation with the vanilla Euler method. Straight interpolation is particularly simple because its natural Euler method coincides with the vanilla Euler method. Consequently, it suffices to adopt a simple form, such as the straight interpolation $X_t = tX_1 + (1-t)X_0$, while still maintaining the flexibility to recover all affine interpolations through adjustments in time parameterization and inference algorithms. Similar observations was made in various settings and from different perspectives [Kingma et al., 2021, Karras et al., 2022b, Shaul et al., 2023, Gao et al., 2024].

3.1 Point-wisely Transformable Interpolations

We show that if two processes $\{X_t\}$ and $\{X'_t\}$ are related pointwise by

$$X'_t = \phi_t(X_{\tau_t}),$$

for some differentiable and invertible maps $\phi\colon (t,x) \mapsto \phi_t(x)$ and $\tau\colon t \mapsto \tau_t$, then their corresponding rectified flows, $\{Z_t\}$ and $\{Z'_t\}$, satisfy the same relation:

$$Z'_t = \phi_t(Z_{\tau_t}),$$

provided the relation holds at initialization, that is, $Z'_0 = \phi_t(Z_0)$.

This result suggests that rectified flows of pointwisely transformable interpolations are essentially the same, upto the same pointwise transform. Furthermore, if $X_t = \texttt{I}_t(X_0, X_1)$ and $X'_t = \texttt{I}_t(X_0, X_1)$ are constructed

from the same coupling (X_0, X_1), then they yield the same rectified coupling $(Z_0', Z_1') = (Z_0, Z_1)$.

Write $\{X_t'\} = \texttt{Transform}(\{X_t\})$ the pointwise transform above. The results suggest that $\texttt{Rectify}(\cdot)$ is an equivariant map under the transforms:

$$\texttt{Rectify}(\texttt{Transform}(\{X_t\})) = \texttt{Transform}(\texttt{Rectify}(\{X_t\})).$$

Why is this true? The intuition is illustrated in Figure 3.1. The trajectories of the rectified flow (RF) are simply a "rewiring" of the interpolation trajectories at their intersection points to avoid crossings. As a result, they occupy the same "trace" as the interpolation process, even though they switch between different trajectories at intersection points. Consequently, any deformation applied to the interpolation trajectories is inherited by the rectified flow trajectories. The deformation must be point-to-point here to make it insensitive to the rewiring of the trajectories.

This is a general and fundamental property of the rectification process and is not restricted to specific distributions, couplings, or interpolations.

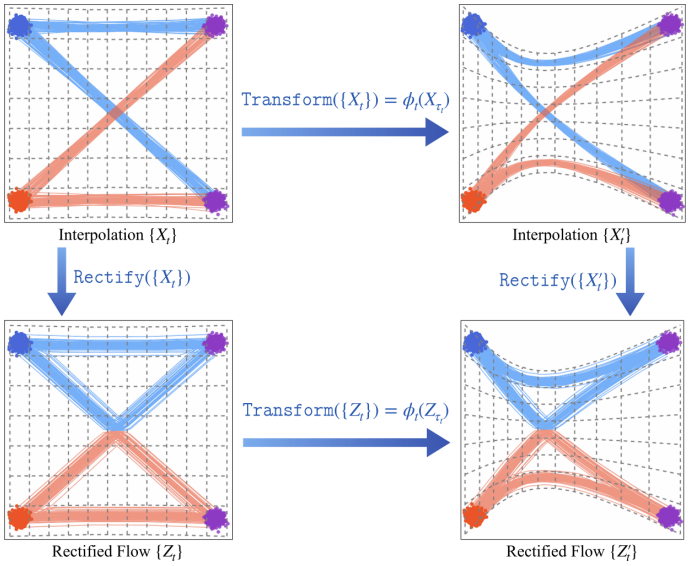


Figure 3.1: Rectified flow rewires the interpolation trajectories to avoid crossing. When a deformation is applied on the interpolation trajectories, the trajectories of the corresponding rectified flow is deformed in the same way.

We now introduce and prove the results.

Definition 6. Two stochastic processes $\{X_t\}$ and $\{X_t'\}$ is said to be *pointwisely transformable* if

$$X_t' = \phi_t(X_{\tau_t}), \quad \forall t \in [0,1],$$

where $\tau\colon [0,1] \to [0,1]$ and $\phi\colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ for $t \in [0,1]$ are differentiable maps, and ϕ_t is invertible for $t \in [0,1]$.

Theorem 2. Assume that $\{X_t\}$ and $\{X_t'\}$ are pointwise transformable as per Definition 6. Let v_t and v_t' be their respective RF velocity fields.

Then we have

$$v_t'(x) = \partial_t \phi_t(\phi_t^{-1}(x)) + \nabla \phi_t(\phi_t^{-1}(x))^\top v_{\tau_t}(\phi_t^{-1}(x))\dot{\tau}_t.$$

In addition, let $\{z_t\}$ be a trajectory of the rectified flow of $\{X_t\}$, satisfying $\frac{\mathrm{d}}{\mathrm{d}t} z_t = v_t(z_t)$. Then a curve $\{z_t'\}$ satisfies $z_t' = \phi_t(z_{\tau_t})$, $\forall t \in [0,1]$ if and only if it is the trajectory of the rectified flow of $\{X_t'\}$ initialized from $z_0' = \phi_0(z_{\tau_0})$, that is,

$$z_t' = \phi_t(z_{\tau_t}),\ \forall t \in [0,1] \quad \Leftrightarrow \quad z_0' = \phi_0(z_{\tau_0}),\ \text{and}\ \frac{\mathrm{d}}{\mathrm{d}t} z_t' = v_t'(z_t'),\ \forall t \in [0,1].$$

Proof. 1) By definition of v_t', we have

$$\begin{aligned}
v_t'(x) &= \mathbb{E}\left[\dot{X}_t' \mid X_t' = x\right] \\
&= \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}t}\phi_t(X_{\tau_t}) \mid \phi_t(X_{\tau_t}) = x\right] \\
&= \mathbb{E}\left[\partial_t \phi_t(X_{\tau_t}) + \nabla \phi_t(X_{\tau_t})^\top \dot{X}_{\tau_t}\dot{\tau}_t \mid X_{\tau_t} = \phi_t^{-1}(x)\right] \\
&= \partial_t \phi_t(\phi_t^{-1}(x)) + \nabla \phi_t(\phi_t^{-1}(x))^\top v_{\tau_t}(\phi_t^{-1}(x))\dot{\tau}_t,
\end{aligned}$$

where we used $v_t(x) = \mathbb{E}\left[\dot{X}_t | X_t = x\right]$.

2) Assume $z_t' = \phi_t(z_{\tau_t})$, then it implies $\frac{\mathrm{d}}{\mathrm{d}t} z_t' = v_t(z_t')$, because

$$\begin{aligned}
\dot{z}_t' &= \frac{\mathrm{d}}{\mathrm{d}t}\phi_t(z_{\tau_t}) \\
&= \partial_t \phi_t(z_{\tau_t}) + \nabla \phi_t(z_{\tau_t})^\top \dot{z}_{\tau_t}\dot{\tau}_t \\
&= \partial_t \phi_t(z_{\tau_t}) + \nabla \phi_t(z_{\tau_t})^\top v_{\tau_t}(z_{\tau_t})\dot{\tau}_t \\
&= \partial_t \phi_t(\phi_t^{-1}(z_t')) + \nabla \phi_t(\phi_t^{-1}(z_t'))^\top v_{\tau_t}(\phi_t^{-1}(z_t'))\dot{\tau}_t \\
&= v_t'(z_t').
\end{aligned}$$

where we used $z_{\tau_t} = \phi_t^{-1}(z_t')$ and $\frac{\mathrm{d}}{\mathrm{d}t} z_t = v_t(z_t)$.

Conversely, if $\frac{\mathrm{d}}{\mathrm{d}t} z_t' = v_t(z_t')$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(z_t' - \phi_t(z_{\tau_t})) = v_t'(z_t') - \frac{\mathrm{d}}{\mathrm{d}t}\phi_t(z_{\tau_t}) = 0.$$

Hence, $z_t' - \phi_t(z_{\tau_t}) = z_0' - \phi_0(z_{\tau_0}) = 0$, which shows that $z_t' = \phi_t(z_{\tau_t})$. □

Remark 14. Let $\frac{\mathrm{d}}{\mathrm{d}t} Z_t = v_t(Z_t)$ be the rectified flow of $\{X_t\}$, which is initialized with $Z_0 = X_0$ by default. Then $Z_t' = \phi_t(Z_{\tau_t})$ is the rectified flow of $\{X_t'\}$ with a specific initialization

$$\frac{\mathrm{d}}{\mathrm{d}t} Z_t' = v_t'(Z_t'), \quad \forall t \in [0,1], \quad \text{and} \quad Z_0' = \phi_0(Z_{\tau_0}).$$

Note that the initialization Z_0' has the same distribution as X_0', even though we may not have $Z_0' = X_0'$ in random variables. It is because

$$Z_0' = \phi_0(Z_{\tau_0}) \overset{\text{law}}{=} \phi_0(X_{\tau_0}) = X_0',$$

where we use the marginal preservation property ($X_{\tau_0} \overset{\text{law}}{=} Z_{\tau_0}$) of rectified flow.

If we further assume $\tau_0 = 0$, which is a natural condition, then we have $Z_{\tau_0} = Z_0 = X_0 = X_{\tau_0}$ (not just equal in law), and hence $Z'_0 = X'_0$, and $\{Z'_t\}$ is the rectified flow with the default initialization of $Z'_0 = X'_0$.

Corollary 1. Assume the same conditions as Theorem 2, with the additional assumption that $\tau(0) = 0$. Let $\{Z_t\}$ and $\{Z'_t\}$ be the rectified flows of $\{X_t\}$ and $\{X'_t\}$, respectively. Then

$$Z'_t = \phi_t(Z_{\tau_t}) \quad \text{for all} \quad t \in [0, 1].$$

Equivalence of Rectified Couplings

If the two pointwisely transformable interpolation processes are constructed from the same coupling, then, they yield the same rectified coupling.

Corollary 2. If $\{X_t\}$ and $\{X'_t\}$ share the same coupling, that is,

$$(X_0, X_1) = (X'_0, X'_1),$$

and they satisfy the condition in Theorem 2 with $\tau(0) = 0$ and $\tau(1) = 1$, then their rectified flow yields the same coupling, that is,

$$(Z_0, Z_1) = (Z'_0, Z'_1).$$

Proof. Corollary 1 gives

$$(Z'_0, Z'_1) = (\phi_0(Z_{\tau_0}), \phi_1(Z_{\tau_1})).$$

One the other hand, we have $X_0 = X'_0 = \phi_0(X_{\tau_0}) = \phi_0(X_0)$, which suggests that ϕ_0 is the identity mapping, and hence $Z_0 = \phi_0(Z_0) = \phi_0(Z_{\tau_0})$. Similarly, $Z_1 = \phi_1(Z_1) = \phi_1(Z_{\tau_1})$. Therefore,

$$(Z'_0, Z'_1) = (\phi_0(Z_{\tau_0}), \phi_1(Z_{\tau_1}) = (Z_0, Z_1).$$

□

3.2 Equivalence of Affine Interpolations

We now show that all affine interpolations can be pointwise transformed to each other by appropriately scaling both time and the variable. Then by Corollary 1 and Corollary 2, their rectified flows can be transformed pointwisely with same maps, and they yield the identical rectified couplings.

Lemma 4. Let $X_t = \alpha_t X_1 + \beta_t X_0$ and $X'_t = \alpha'_t X_1 + \beta'_t X_0$ be two affine interpolation processes constructed from a common coupling

(X_0, X_1), where $\alpha_0 = \beta_1 = \alpha_0' = \beta_1' = 0$ and $\alpha_1 = \beta_1 = \alpha_1' = \beta_1' = 1$.

Then we have

$$X_t' = \frac{1}{\omega_t} X_{\tau_t}, \quad \forall t \in [0,1],$$

where τ_t and ω_t are found by solving:

$$\frac{\alpha_{\tau_t}}{\beta_{\tau_t}} = \frac{\alpha_t'}{\beta_t'}, \qquad \omega_t = \frac{\alpha_{\tau_t}}{\alpha_t'} = \frac{\beta_{\tau_t}}{\beta_t'}, \ \forall t \in (0,1), \tag{3.1}$$

with the boundary condition of

$$\omega_0 = \omega_1 = 1, \quad \tau_0 = 0, \quad \tau_1 = 1.$$

In addition, there is at least one solution of (τ_t, ω_t) in (3.1) if $\alpha_t'/\beta_t' \geq 0$ for $t \in [0,1]$, and α_t/β_t is continuous w.r.t. t. The solution is unique if α_t/β_t is strictly increasing w.r.t. t.

Proof. 1) Write $\omega = \omega_t$ and $\tau = \tau_t$. When (3.1) holds, we have $\omega_t \alpha_t' = \alpha_\tau$ and $\omega_t \beta_t' = \beta_{\tau_t}$. Hence, $\omega_t X_t' = \omega_t \alpha_t' X_1 + \omega_t \beta_t' X_0 = \alpha_{\tau_t} X_1 + \beta_{\tau_t} X_0 = X_{\tau_t}$.

2) For the existence of solution, note that by the boundary conditions of the interpolation processes, we must have $\alpha_0 = \alpha_0' = \beta_1 = \beta_1' = 0$, and $\alpha_1 = \alpha_1' = \beta_0 = \beta_0' = 1$. Hence,

$$\frac{\alpha_0}{\beta_0} = \frac{\alpha_0'}{\beta_0'} = 0, \qquad \frac{\alpha_1}{\beta_1} = \frac{\alpha_1'}{\beta_1'} = +\infty. \tag{3.2}$$

Therefore, there is at least a solution of τ_t for each t once $\frac{\alpha_t'}{\beta_t'} \geq 0$ and $\frac{\alpha_\tau}{\beta_\tau}$ is continuous w.r.t. τ.

In addition, for the boundary conditions, note that $\tau_0 = 0$ and $\tau_1 = 1$ can be achieved due to (3.2), and with it we have $\omega_0 = \beta_0/\beta_0' = 1$, and $\omega_1 = \alpha_1/\alpha_1' = 1$. □

Theorem 3. Assume $\{X_t\}$ and $\{X_t'\}$ are two affine interpolations in Lemma 4.

1) Their respective rectified flows $\{Z_t\}$ and $\{Z_t'\}$ satisfy:

$$Z_t' = {\omega_t}^{-1} Z_{\tau_t}, \quad \forall t \in [0,1].$$

2) Their rectified couplings are equivalent:

$$(Z_0, Z_1) = (Z_0', Z_1').$$

3) Their RF velocity fields v_t and v_t' satisfy

$$v_t'(x) = \frac{1}{\omega_t} \left(\dot{\tau}_t v_{\tau_t}(\omega_t x) - \dot{\omega}_t x \right).$$

Proof. It is the direct implication of Lemma 4, Corollary 1, and Corollary 2.

But let us give a direct derivation of $v_t'(x) = \mathbb{E}\left[\dot{X}_t' | X_t = x\right]$. Tak-

ing derivative of $X'_t = \frac{1}{\omega_t} X_{\tau_t}$,

$$\dot{X}'_t = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{\omega_t} X_{\tau_t}\right) = \frac{1}{\omega_t}\dot{X}_{\tau_t}\dot{\tau}_t - \frac{\dot{\omega}_t}{{\omega_t}^2} X_{\tau_t}.$$

Hence,

$$\begin{aligned} v'_t(x) &= \mathbb{E}\left[\dot{X}'_t \mid X'_t = x\right] \\ &= \mathbb{E}\left[\frac{1}{\omega_t}\dot{X}_{\tau_t}\dot{\tau}_t - \frac{\dot{\omega}_t}{{\omega_t}^2} X_{\tau_t} \mid \frac{1}{\omega_t} X_{\tau_t} = x\right] \\ &= \mathbb{E}\left[\frac{1}{\omega_t}\dot{X}_{\tau_t}\dot{\tau}_t - \frac{\dot{\omega}_t}{{\omega_t}^2} X_{\tau_t} \mid X_{\tau_t} = \omega_t x\right] \\ &= \frac{1}{\omega_t}\dot{\tau}_t v_{\tau_t}(\omega_t x) - \frac{\dot{\omega}_t}{\omega_t} x. \end{aligned}$$

□

Remark 15. The pointwise transform of affine interpolations was discussed in [Kingma et al., 2021, Karras et al., 2022b]. A proof of the equivariance of RF ODEs with Gaussian X_0 and independent couplings was provided in Shaul et al. [2023]. See Gao et al. [2024] for an exploratory introduction to the topic.

Remark 16. Further derivation shows that

$$\begin{aligned} v'_t(x) &= \frac{\alpha_t \kappa_t}{\alpha_{\tau_t}\kappa_{\tau_t}} v_{\tau_t}(\omega_t x) + \left(\frac{\dot{\beta}_t}{\beta_t} - \frac{\alpha_t\kappa_t}{\alpha_{\tau_t}\kappa_{\tau_t}}\frac{\dot{\beta}_{\tau_t}}{\beta_{\tau_t}}\omega_t\right) x \\ &= \frac{\kappa'_t}{\omega_t \kappa_{\tau_t}} v_{\tau_t}(\omega_t x) + \left(\frac{\dot{\beta}'_t}{\beta'_t} - \frac{\kappa'_t}{\kappa_{\tau_t}}\frac{\dot{\beta}_{\tau_t}}{\beta_{\tau_t}}\right) x, \end{aligned}$$

where $\kappa_t = \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right)$.

Example 5. Consider the straight interpolation $X_t = tX_1 + (1-t)X_0$ with $\alpha_t = t$ and $\beta_t = 1 - t$, and we want to transfer it into another interpolation $X'_t = \alpha'_t X_1 + \beta'_t X_0$. We need to solve

$$\omega_t = \frac{\tau_t}{\alpha'_t} = \frac{1-\tau_t}{\beta'_t}.$$

This gives

$$\tau_t = \frac{\alpha'_t}{\alpha'_t + \beta'_t}, \qquad \omega_t = \frac{1}{\alpha'_t + \beta'_t}$$

The velocity field is converted by

$$v'_t(x) = \frac{\dot{\alpha}'_t\beta'_t - \alpha'_t\dot{\beta}'_t}{\alpha'_t + \beta'_t} v_{\tau_t}(\omega_t x) + \frac{\dot{\alpha}'_t + \dot{\beta}'_t}{\alpha'_t + \beta'_t} x. \tag{3.3}$$

Proof.

$$
\begin{aligned}
v'_t(x) &= \frac{1}{\omega}\left(\dot\tau_t v_{\tau_t}(\omega x) - \dot\omega_t x\right) \\
&= (\alpha'_t + \beta'_t)\left(\frac{\dot\alpha'_t(\alpha'_t+\beta'_t) - \alpha'_t(\dot\alpha'_t+\dot\beta'_t)}{(\alpha'_t+\beta'_t)^2} v_{\tau_t}(\omega_t x) + \frac{\dot\alpha'_t+\dot\beta'_t}{(\alpha'_t+\beta'_t)^2} x\right) \\
&= \frac{\dot\alpha'_t\beta'_t - \alpha'_t\dot\beta'_t}{\alpha'_t+\beta'_t} v_{\tau_t}(\omega_t x) + \frac{\dot\alpha'_t+\dot\beta'_t}{\alpha'_t+\beta'_t} x.
\end{aligned}
$$

Example 6. In particular, consider converting the straight interpolation $X_t = tX_1 + (1-t)X_0$ to the spherical interpolation $X'_t = \sin(\frac{\pi}{2}t)X_1 + \cos(\frac{\pi}{2}t)X_0$ with $\alpha'_t = \sin(\frac{\pi}{2}t)$ and $\beta'_t = \cos(\frac{\pi}{2}t)$. We need to solve

$$\omega_t = \frac{\tau_t}{\sin(\frac{\pi}{2}t)} = \frac{1-\tau_t}{\cos(\frac{\pi}{2}t)}.$$

This gives

$$\tau_t = \frac{\sin(\frac{\pi}{2}t)}{\sin(\frac{\pi}{2}t)+\cos(\frac{\pi}{2}t)}, \qquad \omega_t = \frac{1}{\sin(\frac{\pi}{2}t)+\cos(\frac{\pi}{2}t)}.$$

The velocity field is converted by

$$v'_t(x) = \frac{1}{\omega}\left(\dot\tau_t v_{\tau_t}(\omega_t x) - \dot\omega_t x\right).$$

Calculation shows that

$$\dot\tau_t = \frac{\pi}{2}{\omega_t}^2, \qquad \dot\omega_t = -\frac{\pi}{2}{\omega_t}^2\left(\cos(\frac{\pi}{2}t) - \sin(\frac{\pi}{2}t)\right).$$

Hence,

$$v'_t(x) = \frac{\pi\omega_t}{2}\left(v_{\tau_t}(\omega_t x) + \left(\cos(\frac{\pi}{2}t) - \sin(\frac{\pi}{2}t)\right)x\right). \tag{3.4}$$

Figure 3.2 shows the plot of τ_t and ω_t. We can see from the plot that τ_t is a monotonic function of t and is similar to the identity mapping, and the scaling factor ω_t lies in interval $[1/\sqrt{2}, 1]$. Hence, the transform between v_t and v'_t is relatively smooth.

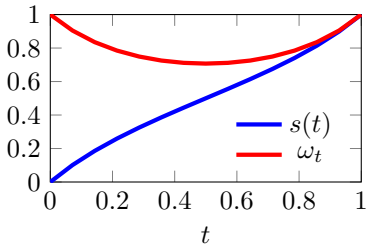


Figure 3.2: Plot of $s(t)$ and ω_t for $t \in [0, 1]$.

Relations of Other Quantities

In the following, we demonstrate how other quantities, such as marginal densities and score functions, can be converted between affine interpolations. Their transformations are simpler than those between v_t and v'_t, and they can sometimes be used to simplify derivations.

Theorem 4. Assume the conditions in Lemma 4. Let ρ_t be the density function of X_t, and $\hat x_{1|t}(x) = \mathbb{E}[X_1 \mid X_t = x]$ and $\hat x_{0|t}(x) = \mathbb{E}[X_0 \mid X_t = x]$ the expected target and noise. Let ρ'_t, $\hat x'_{1|t}$, $\hat x'_{0|t}$ be the

analogous quantities of X'_t. Let d be the dimension of X_t. We have

Density Function: $\rho'_t(x) = (\omega_t)^d \rho_{\tau_t}(\omega_t x)$,

Score Function: $\nabla \log \rho'_t(x) = \omega_t \nabla \log \rho_{\tau_t}(\omega_t x)$,

Expected X_0, X_1**:** $\hat{x}'_{1|t}(x) = \hat{x}_{1|\tau_t}(\omega_t x), \quad \hat{x}'_{0|t}(x) = \hat{x}_{0|\tau_t}(\omega_t x)$.

Proof. The result on ρ_t and ρ'_t directly follows from the change-of-variables formula for transform $X'_t = {\omega_t}^{-1} X_{\tau_t}$. For $\hat{x}_{1|t}$, we have

$$\begin{aligned}\hat{x}'_{1|t}(x) &= \mathbb{E}\left[X'_1 \mid X'_t = x\right] \\ &= \mathbb{E}\left[\frac{1}{\omega_1} X_{\tau_1} \mid \frac{1}{\omega_t} X_{\tau_t} = x\right] \\ &= \mathbb{E}\left[X_1 \mid X_{\tau_t} = \omega_t x\right] \\ &= \hat{x}_{1|\tau_t}(\omega_t x),\end{aligned}$$

where we used $X_1 = \frac{1}{\omega_1} X_{\tau_1}$. The result for $\hat{x}_{0|t}$ follows similarly. □

There are some quantities that are invariant if we transform the process together with the time.

Proposition 2. Let $\{X_t\}$ and $\{X'_t\}$ be the two affine interpolations satisfying the conditions in Lemma 4, and let $t = \tau(t')$. We have

$$\frac{\alpha'_{t'}}{\beta'_{t'}} = \frac{\alpha_t}{\beta_t}, \qquad \frac{X'_{t'}}{\alpha'_{t'}} = \frac{X_t}{\alpha_t}, \qquad \frac{X'_{t'}}{\beta'_{t'}} = \frac{X_t}{\beta_t},$$

and

$$\hat{x}'_{1|t'}(X'_{t'}) = \hat{x}_{1|t}(X_t), \qquad \hat{x}'_{0|t'}(X'_{t'}) = \hat{x}_{0|t}(X_t).$$

Proof. $\frac{\alpha'_{t'}}{\beta'_{t'}} = \frac{\alpha_t}{\beta_t}$ is directly from (3.1). From $X'_{t'} = \frac{1}{\omega_{t'}} X_t$ and $\omega_{t'} = \frac{\alpha_t}{\alpha'_{t'}} = \frac{\beta_t}{\beta'_{t'}}$, we have

$$X'_{t'} = \frac{\alpha'_{t'}}{\alpha_t} X_t = \frac{\beta'_{t'}}{\beta_t} X_t.$$

The proves that $\frac{X'_{t'}}{\alpha'_{t'}} = \frac{X_t}{\alpha_t}$, and $\frac{X'_{t'}}{\beta'_{t'}} = \frac{X_t}{\beta_t}$.

Finally, we have $\hat{x}'_{1|t'}(X'_{t'}) = \hat{x}_{1|t}(\omega_{t'} X'_{t'}) = \hat{x}_{1|t}(X_t)$, and the same holds for $\hat{x}'_{0|t}(X'_{t'})$. □

Practical Implications

Despite the theoretical equivalences discussed above, using different interpolation methods can still have practical impacts on performance due to the following reasons:

1) Training: Training with different interpolation schemes effectively applies different time-weighting to the training loss and results in a differ-

ent parameterization of the model.

2) Inference: Although different interpolation methods yield the same rectified coupling (Z_0, Z_1), their flows have different trajectories $\{Z_t\}$ and are subject to different discretization errors during inference, especially when large step sizes are used.

Straighter trajectories of $\{Z_t\}$ are generally preferred. Thanks to the transformation relations outlined above, it is possible to convert the interpolation scheme of a pre-trained model without retraining, allowing one to identify the scheme that yields straighter trajectories of $\{Z_t\}$.

3.3 Implications on Loss Functions

Assume that we have trained a model $\hat{v}_t$ for the RF velocity field v_t under an affine interpolation. Using the formulas from the previous section, we can convert it to an approximation $\hat{v}'_t$ for the RF velocity v'_t corresponding to a different interpolation scheme at the post-training phase. This raises the question of what properties the converted model $\hat{v}'_t$ may have compared to the models trained directly on the same interpolation, and whether it suffers from performance degradation due to the conversion.

This section investigates this question. We show that the effect of using different affine interpolation schemes during training is equivalent to applying different time-weighting in the loss, and an affine transform on the parametric model. Unless ω_t and τ_t are highly singular, the conversion does not necessarily degrade performance.

Specifically, assume we have trained a parametric model $v_t(x;\ \theta)$ to approximate the RF velocity v_t of interpolation $X_t = \alpha_t X_1 + \beta_t X_0$, using the mean square loss:

$$L(\theta) = \int_0^1 \mathbb{E}\left[\eta_t \left\|\dot{X}_t - v_t(X_t;\ \theta)\right\|^2\right] \mathrm{d}t. \tag{3.5}$$

After training, we may convert the obtained model $v_t(x;\ \theta)$ to an approximation of v'_t of a different interpolation $X'_t = \alpha_t X_1 + \beta_t X_0$ via

$$v'_t(x;\ \theta) = \frac{\dot{\tau}_t}{\omega_t} v_{\tau_t}(\omega_t x;\ \theta) - \frac{\dot{\omega}_t}{\omega_t} x, \tag{3.6}$$

with ω and τ defined in (3.1).

On the other hand, if we train $v'_t(x;\theta)$ directly to approximate v'_t of interpolation $X'_t = \alpha'_t X_1 + \beta'_t X_0$, the loss function with a time weight η' is

$$L'(\theta) = \int_0^1 \mathbb{E}\left[\eta'_t \left\|\dot{X}'_t - v'_t(X'_t;\ \theta)\right\|^2\right] \mathrm{d}t. \tag{3.7}$$

Plugging (3.6) and $X_{\tau_t} = \omega X'_t$, and $\dot{X}'_t = \frac{1}{\omega_t}\left(\dot{\tau}_t \dot{X}_{\tau_t} - \dot{\omega}_t X'_t\right)$, we have

$$
\begin{aligned}
L'(\theta) &= \int_0^1 \mathbb{E}\left[\eta'_t \left\| \dot{X}'_t - v'_t(X'_t;\, \theta)\right\|^2\right] \mathrm{d}t \\
&= \int_0^1 \mathbb{E}\left[\eta'_t \left\| \frac{\dot{\tau}_t}{\omega_t}\left(\dot{X}_{\tau_t} - v_{\tau_t}(X_{\tau_t};\, \theta)\right)\right\|^2\right] \mathrm{d}t \\
&= \int_0^1 \mathbb{E}\left[\eta'_t \frac{\dot{\tau}_t^2}{\omega_t^2} \left\| \dot{X}_{\tau_t} - v_{\tau_t}(X_{\tau_t};\, \theta)\right\|^2\right] \mathrm{d}t \\
&= \int_0^1 \mathbb{E}\left[\eta'_t \frac{\dot{\tau}_t}{\omega_t^2} \left\| \dot{X}_{\tau_t} - v_{\tau_t}(X_{\tau_t};\, \theta)\right\|^2\right] \mathrm{d}\tau_t \qquad //\mathrm{d}\tau_t = \dot{\tau}_t \mathrm{d}t \\
&= \int_0^1 \mathbb{E}\left[\eta'_{t_\tau} \frac{\dot{\tau}(t_\tau)}{\omega(t_\tau)^2} \left\| \dot{X}_{\tau} - v_{\tau}(X_{\tau};\, \theta)\right\|^2\right] \mathrm{d}\tau. \qquad //\text{rename } \tau_t \text{ to } \tau
\end{aligned}
$$

where we denote t_τ as the inverse of the map τ_t. To match the loss in (3.5), we should set $\eta'_{t_\tau} \frac{\dot{\tau}(t_\tau)}{\omega(t_\tau)^2} = \eta_\tau$, which gives

$$\eta'_t = \frac{\omega_t^2}{\dot{\tau}_t} \eta_{\tau_t}.$$

Proposition 3. Training $v_t(x;\theta)$ with the loss in (3.5) is equivalent to training $v'_t(x;\theta)$ with the loss in (3.5), but with the following time-weighting and parameterization:

$$\eta'_t = \frac{\omega_t^2}{\dot{\tau}_t}\eta_{\tau_t}, \qquad v'_t(x;\, \theta) = \frac{\dot{\tau}_t}{\omega_t} v_{\tau_t}(\omega_t x;\, \theta) - \frac{\dot{\omega}_t}{\omega_t} x. \tag{3.8}$$

Remark 17. Notably, since the loss functions are equivalent, the equivalence described above holds even when the models are approximately optimized using gradient-based optimizers, as is common in practice, provided that the random seeds for initialization and mini-batch sampling are matched exactly.

Example 7 (Loss from Straight to Affine). Consider the straight interpolation $X_t = tX_1 + (1-t)X_0$ with $\alpha_t = t$ and $\beta_t = 1 - t$, and another affine interpolation $X'_t = \alpha'_t X_1 + \beta'_t X_0$. Assume we train the v_t for X_t with time weight η_t, then v'_t converted from v_t is equivalent to be trained with the reparametrization in (3.3), and time weight

$$\eta'_t = \frac{{\omega_t}^2}{\dot{\tau}_t}\eta_{\tau_t} = \frac{1}{\dot{\alpha}'_t\beta'_t - \alpha'_t\dot{\beta}'_t}\eta_{\tau_t},$$

where we used $\omega_t = \frac{1}{\alpha'_t + \beta'_t}$, $\tau_t = \frac{\alpha'_t}{\alpha'_t + \beta'_t}$,, and $\dot{\tau}_t = \frac{\dot{\alpha}'_t\beta'_t - \alpha'_t\dot{\beta}'_t}{(\alpha'_t + \beta'_t)^2}$ from Example 5.

Example 8 (Losses of Straight vs Spherical). Continuing from Example 7, interesting cases occur when

$$\dot{\alpha}'_t\beta'_t - \alpha_t\dot{\beta}'_t = const,$$

in which case we have $\eta'_t \propto \eta_{\tau_t}$. For instance, this holds for spherical interpolation $X'_t = \sin(\frac{\pi}{2}t)X_1 + \cos(\frac{\pi}{2}t)X_0$, which has

$$\dot{\alpha}'_t\beta'_t - \alpha_t\dot{\beta}'_t = \frac{\pi}{2},$$

and hence

$$\eta'_t = \frac{2}{\pi}\eta_{\tau_t}, \qquad \text{where} \quad \tau_t = \tan(\frac{\pi}{2}t)/(\tan(\frac{\pi}{2}t) + 1).$$

This can be also seen by noting that $\dot{\tau}_t = \frac{\pi}{2}{\omega_t}^2$ in Example 6.

Therefore, using the straight and spherical interpolations corresponds to using equivalent time weights in the training loss, up to the time scaling with τ_t.

In particular, *training v_t with straight interpolation using a uniform weight $\eta_t = 1$ is equivalent to training v'_t with spherical interpolation, also using a uniform weight $\eta'_t = 1$.* In this case, the only difference lies in the model parameterization.

From Equation (3.4), the model reparameterization is

$$v'_t(x,\theta) = \frac{\pi\omega_t}{2}\left(v_{\tau_t}(\omega_t x,\theta) + \left(\cos(\frac{\pi}{2}t) - \sin(\frac{\pi}{2}t)\right)x\right).$$

Given that the variable scaling factor $\omega_t = (\sin(\frac{\pi}{2}t) + \cos(\frac{\pi}{2}t))^{-1}$ is bounded in $[1/\sqrt{2}, 1]$ (see Example 6 and Figure 3.2), This reparameteriztation may not impact the performance significantly. Overall, the choice of using straight or spherical might have limited impact in terms of the training performance.

3.4 Equivariance of Natural Euler Samplers

The Euler method can be seen as a piecewise linear approximation of the ODE trajectory, where the curve is locally approximated using straight lines tangent to the curve. This local straight-line approximation is a natural choice for rectified flows induced by straight interpolation. However, if the rectified flow is induced by a curved interpolation scheme, it may be more appropriate to use curves derived from the underlying interpolation scheme as the local approximation.

We refer to such approximation schemes as *natural Euler samplers*, as they are similar to *natural gradient descent* in terms of the invariance of trajectories under re-parameterizations. Popular methods that employ curved interpolations, such as DDIM, DDPM, EDM, and DMP solvers, all utilize natural Euler samplers.

In this section, we first introduce the concept of *natural Euler samplers* and then establishe that the trajectories of natural Euler samplers are equivariant for pointwise transformable interpolations:

When two interpolation processes are pointwise transformable, the trajectories of their corresponding natural Euler samplers are also pointwise transformable under the same maps, provided their time grids are mapped using the corresponding time-scaling function τ_t.

This result implies that, when natural Euler samplers are used, employing different affine interpolations corresponds to using different time

grids.

3.4.1 Natural Euler Samplers

Let $X_t = \mathtt{I}_t(X_0, X_1)$ be an interpolation process whose RF velocity field is v_t. Recall that in the vanilla Euler method, the trajectories of the flow are approximated on a time grid $\{t_i\}_i$ via

$$\hat{z}_{t_{i+1}} = \hat{z}_{t_i} + (t_{i+1} - t_i)\, v_t(\hat{z}_{t_i}),$$

where the solution is locally approximated by the straight line tangent to the rectified flow curve at the point $\hat{z}_{t_i}$.

In a curved Euler method with an interpolation scheme $\mathtt{I}_t$, we replace the straight line with an interpolation curve that is tangent at $\hat{z}_{t_i}$. The update rule becomes

$$\hat{z}_{t_{i+1}} = \mathtt{I}_{t_{i+1}}(\hat{x}_{0|t_i}, \hat{x}_{1|t_i}),$$

where $\hat{x}_{0|t_i}$ and $\hat{x}_{1|t_i}$ are determined by solving

$$\hat{z}_{t_i} = \mathtt{I}_{t_i}(\hat{x}_{0|t_i}, \hat{x}_{1|t_i}), \qquad v_{t_i}(\hat{z}_{t_i}) = \partial_{t_i} \mathtt{I}_t(\hat{x}_{0|t_i}, \hat{x}_{1|t_i}). \tag{3.9}$$

This identifies the endpoints $\hat{x}_{0|t_i}$ and $\hat{x}_{1|t_i}$ of the interpolation curve that passes through the point $\hat{z}_{t_i}$ with slope $\partial \hat{z}_{t_i} = v_{t_i}(\hat{z}_{t_i})$ at time t_i, ensuring that the interpolation curve is tangent to the rectified flow at $\hat{z}_{t_i}$ at time t_i. We assume that the solution of Equation (3.9) exists and is unique. For affine interpolation, it admits a simple closed-form solution.

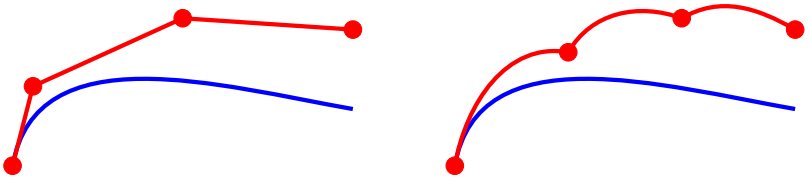


Figure 3.3: Vanilla Euler solver yields piecewise linear approximation,while curved Euler solvers yield piecewise curved approximation.

The curved Euler method defines a general family of numerical methods for solving ODEs. When the step size approaches zero, and provided that the interpolation is differentiable, the approximation converges to the ODE trajectory, as in the case of the vanilla Euler method. The idea is to employ a better interpolation scheme that better approximates the true ODE trajectories.

For rectified flows, it is natural to use the interpolation scheme used to construct the flow itself. In this context, we refer to the method as the *natural Euler method*.

> Definition 7. For the rectified flow ODE induced by an interpolation $X_t = \mathtt{I}_t(X_0, X_1)$, its natural Euler sampler is the curved Euler sampler that employs the same interpolation $\mathtt{I}_t$.

Example 9. For affine interpolation $X_t = \alpha_t X_1 + \beta_t X_0$, solving (3.9) yields

$$\hat{x}_{0|t}(\hat{z}_t) = \frac{-\alpha_t v_t(\hat{z}_t) + \dot{\alpha}_t \hat{z}_t}{\dot{\alpha}_t \beta_t - \alpha_t \dot{\beta}_t}, \qquad \hat{x}_{1|t}(\hat{z}_t) = \frac{\beta_t v_t(\hat{z}_t) - \dot{\beta}_t \hat{z}_t}{\dot{\alpha}_t \beta_t - \alpha_t \dot{\beta}_t}.$$

Hence, the update from $t_i = t$ and $t_{i+1} = t + \varepsilon$ is

$$\begin{aligned}
\hat{z}_{t+\varepsilon} &= \alpha_{t+\varepsilon} \hat{x}_{1|t}(\hat{z}_t) + \beta_{t+\varepsilon} \hat{x}_{0|t}(\hat{z}_t) \\
&= \alpha_{t+\varepsilon} \frac{\beta_t v_t(\hat{z}_t) - \dot{\beta}_t \hat{z}_t}{\dot{\alpha}_t \beta_t - \alpha_t \dot{\beta}_t} + \beta_{t+\varepsilon} \frac{-\alpha_t v_t(\hat{z}_t) + \dot{\alpha}_t \hat{z}_t}{\dot{\alpha}_t \beta_t - \alpha_t \dot{\beta}_t} \\
&= \frac{\dot{\alpha}_t \beta_{t+\varepsilon} - \alpha_{t+\varepsilon} \dot{\beta}_t}{\dot{\alpha}_t \beta_t - \alpha_t \dot{\beta}_t} \hat{z}_t + \frac{\alpha_{t+\varepsilon} \beta_t - \alpha_t \beta_{t+\varepsilon}}{\dot{\alpha}_t \beta_t - \alpha_t \dot{\beta}_t} v_t(\hat{z}_t).
\end{aligned}$$

For straight interpolation $X_t = tX_1 + (1-t)X_0$, this reduces to the standard Euler scheme. In general cases, however, it is a different update that is nonlinear on the step size ε.

For the spherical interpolation $X_t = \sin(\frac{\pi}{2}t)X_1 + \cos(\frac{\pi}{2}t)X_0$, the update reduces to

$$\hat{z}_{t+\varepsilon} = \cos(\tfrac{\pi}{2}\varepsilon)\hat{z}_t + \tfrac{2}{\pi}\sin(\tfrac{\pi}{2}\varepsilon)v_t(\hat{z}_t).$$

where we used the trigonometric identities:

$$\begin{aligned}
\dot{\alpha}_t \beta_{t'} - \alpha_{t'} \dot{\beta}_t &= \tfrac{\pi}{2}(\cos\left(\tfrac{\pi}{2}t\right)\cos\left(\tfrac{\pi}{2}t'\right) + \sin\left(\tfrac{\pi}{2}t'\right)\sin\left(\tfrac{\pi}{2}t\right)) = \tfrac{\pi}{2}\cos\left(\tfrac{\pi}{2}(t'-t)\right), \\
\alpha_{t'}\beta_t - \alpha_t \beta_{t'} &= \sin\left(\tfrac{\pi}{2}t'\right)\cos\left(\tfrac{\pi}{2}t\right) - \sin\left(\tfrac{\pi}{2}t\right)\cos\left(\tfrac{\pi}{2}t'\right) = \sin\left(\tfrac{\pi}{2}(t'-t)\right), \\
\dot{\alpha}_t \beta_t - \alpha_t \dot{\beta}_t &= \tfrac{\pi}{2}\cos^2\left(\tfrac{\pi}{2}t\right) + \tfrac{\pi}{2}\sin^2\left(\tfrac{\pi}{2}t\right) = \tfrac{\pi}{2}.
\end{aligned}$$

Remark 18. The discretized inference scheme of DDIM is an instance of natural Euler sampler. To see this, note that the inference update of DDIM is written in terms of the expected noise $\hat{x}_{0|t}(x) = \mathbb{E}[X_0 | X_t = x]$. Hence, we write the update of $\hat{z}_t$ in terms of $\hat{x}_{0|t}(x)$:

$$\begin{aligned}
\hat{z}_{t+\varepsilon} &= \alpha_{t+\varepsilon} \hat{x}_{1|t}(\hat{z}_t) + \beta_{t+\varepsilon} \hat{x}_{0|t}(\hat{z}_t) \\
&\overset{*}{=} \alpha_{t+\varepsilon} \left(\frac{\hat{z}_t - \beta_t \hat{x}_{0|t}(\hat{z}_t)}{\alpha_t} \right) + \beta_{t+\varepsilon} \hat{x}_{0|t}(\hat{z}_t) \\
&= \frac{\alpha_{t+\varepsilon}}{\alpha_t} \hat{z}_t + \left(\beta_{t+\varepsilon} - \frac{\alpha_{t+\varepsilon}\beta_t}{\alpha_t} \right) \hat{x}_{0|t}(\hat{z}_t)
\end{aligned}$$

where in $\overset{*}{=}$ we used $\alpha_t \hat{x}_{1|t}(\hat{z}_t) + \beta_t \hat{x}_{0|t}(\hat{z}_t) = \hat{z}_t$. We can slightly rewrite the update as

$$\frac{\hat{z}_{t+\varepsilon}}{\alpha_{t+\varepsilon}} = \frac{\hat{z}_t}{\alpha_t} + \left(\frac{\beta_{t+\varepsilon}}{\alpha_{t+\varepsilon}} - \frac{\beta_t}{\alpha_t} \right) \hat{x}_{0|t}(\hat{z}_t), \tag{3.10}$$

which matches the DDIM update in Equation 13 of Song et al. [2020a].

3.4.2 Equivalence of Natural Euler Trajectories

We show that the trajectories of natural Euler samplers are equivariant under point-wise transforms. Let $\{X_t\}$ and $\{X'_t\}$ be two interpolation processes that are point-wisely transformable via $X'_t = \phi_t(X_{\tau_t})$, and let $\{\hat{z}_{t_i}\}_i$ and $\{\hat{z}'_{t'_i}\}_i$ be the trajectories returned by the natural Euler method of the rectified flow under each interpolation on time grid $\{t_i\}$ and $\{t'_i\}$, respectively. Assume that time grids satisfies $\tau(t'_i) = t_i$ for $\forall i$, and $\hat{z}'_{t'_0} = \phi(\hat{z}_{\tau(t'_0)})$. Then, we can show

$$\hat{z}'_{t'_i} = \phi_{t'_i}(\hat{z}_{t_i}), \quad \forall i = 0, 1, \dots. \tag{3.11}$$

Moreover, if $X_1 = X'_1$ and $\tau(1) = 1$, and the time grid ends at $t_i = t'_i = 1$, then $\hat{z}'_1 = \hat{z}_1$, that is, they provide the same final output, even though the intermediate trajectories can be different.

In particular, applying natural Euler method on an affine interpolation X'_t on uniform time grid $t'_i = \frac{i}{n}$ is equivalent to applying standard Euler method on the straight interpolation with on a non-uniform time grid $t_i = \tau(\frac{i}{n})$.

Note that this strengthens the equivariance result of the RF ODEs in Theorem 2, which arises as the infinitesimal step size limit of the natural Euler trajectories.

Let $\{X'_t\} = \texttt{Transform}(\{X_t\})$ denote the pointwise transform, and $\hat{Z} = \texttt{NaturalEulerRF}(\{X_t\})$ the mapping from $\{X_t\}$ to the discretized natural Euler trajectories. The result is the following equivariant property:

$$\texttt{NaturalEulerRF}(\texttt{Transform}(\{X_t\})) = \texttt{Transform}(\texttt{NaturalEulerRF}(\{X_t\})).$$

Example 10 (Equivalence of Straight Euler and DDIM). From Example 5, The pointwise transform from straight interpolation $X_t = tX_1 + (1-t)X_0$ to a general affine interpolation $X'_t = \alpha'_t X_1 + \beta'_t X_0$ is

$$X'_t = \frac{1}{\omega_t} X_{\tau_t}, \quad \text{with} \quad \tau_t = \frac{\alpha'_t}{\alpha'_t + \beta'_t}, \quad \omega_t = \frac{1}{\alpha'_t + \beta'_t}. \tag{3.12}$$

The vanilla Euler method under the straight interpolation is

$$\hat{z}_{t_{i+1}} = \hat{z}_{t_i} + (t_{i+1} - t_i) v_{t_i}(\hat{z}_{t_i}). \tag{3.13}$$

As predicted by the theory, transforming the trajectories in (3.13) with the mapping in (3.12) would yield the natural Euler trajectories of $X'_t = \alpha'_t X_1 + \beta'_t X_0$, which coincides with the DDIM inference (see Remark 18), on time grid $\{t'_i\}$ that solves the equation:

$$t_i = \frac{\alpha'_{t'_i}}{\alpha'_{t'_i} + \beta'_{t'_i}}. \tag{3.14}$$

Conversely, if we run natural Euler sampler of $X'_t = \alpha'_t X_1 + \beta'_t X_0$ with a uniform time grid $t'_i = \frac{i}{n}$, then it corresponds to running

vanilla Euler of straight interpolation with a non-uniform time grid $t_i = \frac{\alpha'_{i/n}}{\alpha'_{i/n}+\beta'_{i/n}}$.

Proof. As a (tedious) exercise, let us manually apply the pointwise transform in (3.12) to the straight Euler trajectory in (3.13), and show that it coincides with the natural Euler samplers corresponding to DDIM in Remark 18.

First, we rewrite the straight Euler update w.r.t. the predicted noise $\hat{x}_{0|t}(x) = \mathbb{E}[X_0|X_t = x]$:

$$\hat{z}_{t_{i+1}} = \hat{z}_{t_i} + (t_{i+1} - t_i)v_{t_i}(\hat{z}_{t_i}) = \frac{t_{i+1}}{t_i}\hat{z}_{t_i} - \frac{t_{i+1} - t_i}{t_i}\hat{x}_{0|t}(\hat{z}_{t_i}), \tag{3.15}$$

where we used $v_t(x) = (x - \hat{x}_{0|t}(x))/t$. Our goal is to show that applying the transform $\hat{z}'_{t'_i} = \frac{1}{\omega_{t'_i}}\hat{z}_{t_i}$ with $t_i = \tau(t'_i)$ to (3.15) yields

$$\frac{\hat{z}'_{t'_{i+1}}}{\alpha'_{t'_{i+1}}} = \frac{\hat{z}'_{t'_i}}{\alpha'_{t'_i}} + \left(\frac{\beta'_{t'_{i+1}}}{\alpha'_{t'_{i+1}}} - \frac{\beta'_{t'_i}}{\alpha'_{t'_i}}\right)\hat{x}'_{0|t'_i}(z'_{t'_i}), \tag{3.16}$$

which coincides with the DDIM inference update in (3.10).

To do so, applying the transform $\hat{z}'_{t'_i} = \frac{1}{\omega_{t'_i}}\hat{z}_{t_i}$ and $\hat{z}'_{t'_{i+1}} = \frac{1}{\omega_{t'_{i+1}}}\hat{z}_{t_{i+1}}$ to (3.15) yields

$$\begin{aligned}
\hat{z}'_{t'_{i+1}} &= \frac{1}{\omega_{t'_{i+1}}}\hat{z}_{t_{i+1}} \\
&= \frac{1}{\omega_{t'_{i+1}}}\left(\frac{t_{i+1}}{t_i}\hat{z}_{t_i} - \frac{t_{i+1} - t_i}{t_i}\hat{x}_{0|t_i}(\hat{z}_{t_i})\right) \\
&= \frac{\omega_{t'_i}}{\omega_{t'_{i+1}}}\frac{t_{i+1}}{t_i}\hat{z}'_{t'_i} - \frac{1}{\omega_{t'_{i+1}}}\frac{t_{i+1} - t_i}{t_i}\hat{x}_{0|t_i}(\omega_{t'_i}\hat{z}'_{t'_i}) \\
&= \frac{\omega_{t'_i}}{\omega_{t'_{i+1}}}\frac{t_{i+1}}{t_i}\hat{z}'_{t'_i} - \frac{1}{\omega_{t'_{i+1}}}\frac{t_{i+1} - t_i}{t_i}\hat{x}_{0|t'_i}(\hat{z}'_{t'_i}),
\end{aligned}$$

where we used $\hat{x}_{0|t_i}(\omega_{t'_i}z'_{t'_i}) = \hat{x}'_{0|t'_i}(z'_{t'_i})$ by Theorem 4.

We just need to clean up the coefficients. By (3.12) and (3.14), we have $t_i = \alpha'_{t'_i}\omega_{t'_i}$, and hence

$$\frac{\omega_{t_i}}{\omega_{t'_{i+1}}}\frac{t_{i+1}}{t_i} = \frac{\alpha'_{t'_{i+1}}}{\alpha'_{t'_i}}.$$

$$\frac{1}{\omega_{t'_{i+1}}}\frac{t_{i+1}-t_i}{t_i} = (\alpha'_{t'_{i+1}} + \beta'_{t'_{i+1}})\frac{\frac{\alpha'_{t'_{i+1}}}{\alpha'_{t'_{i+1}}+\beta'_{t'_{i+1}}} - \frac{\alpha'_{t'_i}}{\alpha'_{t'_i}+\beta'_{t'_i}}}{\frac{\alpha'_{t'_i}}{\alpha'_{t'_i}+\beta'_{t'_i}}}$$

$$= \frac{\alpha'_{t'_{i+1}}\beta'_{t'_i} - \alpha'_{t'_i}\beta'_{t'_{i+1}}}{\alpha'_{t'_i}}$$

$$= \alpha'_{t'_{i+1}}\left(\frac{\beta'_{t'_i}}{\alpha'_{t'_i}} - \frac{\beta'_{t'_{i+1}}}{\alpha'_{t'_{i+1}}}\right).$$

Rearranging the terms yields (3.16).

Proofs

In the following, we prove the general equivariance results of natural Euler samplers in (3.11).

Definition 8. A function $\mathtt{I}\colon [0,1]\times\mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}^d$, denoted as $\mathtt{I}_t(x_0, x_1)$, is said to be an *interpolation*, or interpolation function, if it satisfies $\mathtt{I}_0(x_0, x_1) = x_0$ and $\mathtt{I}_1(x_0, x_1) = x_1$ for any $x_0, x_1 \in \mathbb{R}^d$.

An interpolation $\mathtt{I}$ is said to be *invertible* if, for any $x \in \mathbb{R}^d$, $v \in \mathbb{R}^d$, and $t_0 \in [0,1]$, there exists a unique interpolation curve $x_t = \mathtt{I}_t(\hat{x}_0, \hat{x}_1)$ that passes through the point $x_{t_0} = x$ with slope $\dot{x}_{t_0} = v$.

In other words, there exists a unique solution for $\hat{x}_0$ and $\hat{x}_1$ such that

$$x = \mathtt{I}_{t_0}(\hat{x}_0, \hat{x}_1), \quad v = \partial_t \mathtt{I}_t(\hat{x}_0, \hat{x}_1)\big|_{t=t_0}.$$

With an abuse of notation, we write $x_t = \mathtt{I}_t(x_{t_0} = x, \dot{x}_{t_0} = v)$.

Theorem 5. 1) Let $\{X_t\}$ and $\{X'_t\}$ be two interpolation processes constructed from the same coupling via

$$X_t = \mathtt{I}_t(X_0, X_1), \qquad X'_t = \mathtt{I}'_t(X_0, X_1),$$

where $\mathtt{I}$ and $\mathtt{I}'$ are two invertible interpolations. Let v_t, v'_t be the RF velocity fields of $\{X_t\}$ and $\{X'_t\}$, respectively.

2) Assume $\mathtt{I}$ and $\mathtt{I}'$ are point-wisely transformable via

$$\mathtt{I}'_t(x_0, x_1) = \phi_t(\mathtt{I}_{\tau_t}(x_0, x_1)), \quad \forall t \in [0,1], x_0, x_1 \in \mathbb{R}^d,$$

where $\phi_t(x)$ and $\tau_t = \tau(t)$ are differentiable functions and ϕ_t is invertible for $\forall t \in [0,1]$.

3) Let $\{\hat{z}_{t_i}\}_i$ and $\{\hat{z}'_{t'_i}\}_i$ be the trajectories returned by the natural Euler method of the rectified flow under each interpolation on time grid $\{t_i\}$ and $\{t'_i\}$, respectively.

Assume that the time grids satisfy $\tau(t'_i) = t_i$ for all i, and that the initial condition satisfies $\hat{z}'_{t'_0} = \phi_{t'_0}(\hat{z}_{t_0})$. Then, the two trajectories can be pointwise mapped via:

$$\hat{z}'_{t'_i} = \phi_{t'_i}(\hat{z}_{t_i}), \quad \forall i = 0, 1, \ldots.$$

Proof. Assume that $\hat{z}'_{t'_i} = \phi_{t'_i}(\hat{z}_{t_i})$ holds, we just need to prove that $\hat{z}'_{t'_{i+1}} = \phi_{t'_{i+1}}(z_{t_{i+1}})$ for the next time point. Note that the next points are fetched on the interpolation curve that is tangent to the rectified flow curves at the current time point:

$$\begin{aligned}\hat{z}_{t_{i+1}} &= \mathtt{I}_{t_{i+1}}(z_{t_i} = \hat{z}_{t_i}, \dot{z}_{t_i} = v_t(\hat{z}_{t_i})),\\ \hat{z}'_{t'_{i+1}} &= \mathtt{I}_{t'_{i+1}}(z'_{t'_i} = \hat{z}'_{t'_i}, \dot{z}'_{t'_i} = v'_t(\hat{z}'_{t'_i})),\end{aligned}$$

where we have $z'_{t'_i} = \phi_{t'_i}(z_{t_i})$ and $t_i = \tau(t'_i)$. Following Lemma 5, we have $z'_{t'_{i+1}} = \phi_{t'_{i+1}}(z_{t_{i+1}})$.

□

Lemma 5. 1) Let $\{X_t\}$ and $\{X'_t\}$ be two interpolation processes constructed from the same coupling via

$$X_t = \mathtt{I}_t(X_0, X_1), \qquad X'_t = \mathtt{I}'_t(X_0, X_1),$$

where $\mathtt{I}$ and $\mathtt{I}'$ are two invertible interpolations. Let v_t, v'_t be the RF velocity fields of $\{X_t\}$ and $\{X'_t\}$, respectively.

2) Assume $\mathtt{I}$ and $\mathtt{I}'$ are point-wisely transformable via

$$\mathtt{I}'_t(x_0, x_1) = \phi_t(\mathtt{I}_{\tau_t}(x_0, x_1)), \quad \forall t \in [0,1], x_0, x_1 \in \mathbb{R}^d,$$

where $\phi_t(x)$ and $\tau_t = \tau(t)$ are differentiable functions and ϕ_t is invertible for $\forall t \in [0,1]$.

3) Let $\frac{\mathrm{d}}{\mathrm{d}t} z_t = v_t(z_t)$ be a trajectory of RF of $\{X_t\}$ and $\{y_t\}$ the interpolation curve that is tangent to $\{z_t\}$ at time t_0. Similarly, let $\frac{\mathrm{d}}{\mathrm{d}t} z'_t = v'_t(z'_t)$ be a trajectory of RF of $\{X'_t\}$, and $\{y'_t\}$ the interpolation curve that is tangent to $\{z'_t\}$ at time t'_0. In other words:

$$\begin{aligned}y_t &= \mathtt{I}_t(y_{t_0} = z_{t_0}, \dot{y}_{t_0} = v_t(z_{t_0}))\\ y'_t &= \mathtt{I}_t(y'_{t_0} = z'_{t'_0}, \dot{y}'_{t'_0} = v'_t(z'_{t'_0})).\end{aligned}$$

We assume that $z'_0 = \phi_0(z_{\tau_0})$ and $\tau(t'_0) = t_0$.

Then $\{y_t\}$ and $\{y'_t\}$ can be mapped point-wisely via

$$y'_t = \phi_t(y_{\tau_t}), \quad \forall t \in [0,1].$$

Proof. To prove the result, we define $\{y'_t\}$ by the mapping $y'_t = \phi_t(y_{\tau_t})$, and then we have:

1) $\{y'_t\}$ and $\{z'_t\}$ are tangent at time t'_0. This follows Lemma 6 because $\{y_t\}$ and $\{z_t\}$ are tangent at $t_0 = \tau(t'_0)$, and $y'_t = \phi_t(y_{\tau_t})$, and $z'_t = \phi_t(z_{\tau_t})$ (following Theorem 2).

2) $\{y'_t\}$ is an interpolation curve under interpolation I'. This is because

$$\begin{aligned}y'_t &= \phi_t(\mathtt{I}_{\tau_t}(y_0, y_1))\\ &= \mathtt{I}'_t(y_0, y_1).\end{aligned}$$

This shows that $\{y'_t\}$ is an interpolation curve that is tangent to $\{z'_t\}$

at t_0', which is unique because the interpolation $\mathtt{I}'$ is invertible. The proves the result. □

Lemma 6. Let $\{x_t\}$ and $\{y_t\}$ are be time-differentiable curves that are tangent at a time point t_0, that is,

$$x_{t_0} = y_{t_0}, \quad \dot{x}_{t_0} = \dot{y}_{t_0}.$$

Assume the $x_t' = \phi_t(x_{\tau_t})$ and $y_t' = \phi_t(y_{\tau_t})$, where $\phi\colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ and $\tau\colon [0,1] \to [0,1]$ are differentiable maps.

Then $\{x_t'\}$ and $\{y_t'\}$ are tangent at time t_0', that is,

$$x'_{t_0'} = y'_{t_0'}, \quad \dot{x}'_{t_0'} = \dot{y}'_{t_0'},$$

where t_0' satisfies $\tau(t_0') = t_0$.

Remark 19. In other words, when $x_t' = \phi_t(x_{\tau_t})$, there exists a function F, such that $\dot{x}_t' = F(x_{\tau_t}, \dot{x}_{\tau_t})$, so that $\dot{x}_t'$ is completely determined by x_τ and $\dot{x}_{\tau_t}$.

Proof. Obviously, we have $x'_{t_0'} = y'_{t_0'}$ by definition. For the slope, taking derivative of x_t' and y_t':

$$\dot{x}_t' = \partial_t \phi_t(x_{\tau_t}) + \nabla \phi_t(x_{\tau_t}) \dot{\tau}_t \dot{x}_{\tau_t}, \quad \dot{y}_t' = \partial_t \phi_t(y_{\tau_t}) + \nabla \phi_t(y_{\tau_t}) \dot{\tau}_t \dot{y}_{\tau_t}.$$

Plugging $t = t_0'$ and $\tau_t = \tau(t_0') = t_0$, we get

$$\dot{x}'_{t_0'} = \partial_t \phi_{t_0'}(x_{t_0}) + \nabla \phi_{t_0'}(x_{t_0}) \dot{\tau}(t_0') \dot{x}_{t_0},$$
$$\dot{y}'_{t_0'} = \partial_t \phi_{t_0'}(y_{t_0}) + \nabla \phi_{t_0'}(y_{t_0}) \dot{\tau}(t_0') \dot{y}_{t_0}.$$

Hence, $x_{t_0} = y_{t_0}$ and $\dot{x}_{t_0} = \dot{y}_{t_0}$ imply that $\dot{x}'_{t_0'} = \dot{y}'_{t_0'}$. □

3.5 Stochastic Smooth Interpolations

Although we have focused on deterministic interpolation processes $X_t = \mathtt{I}_t(X_0, X_1)$ in which X_t is deteministic given (X_0, X_1), the same algorithm extends naturally to stochastic interpolations of form

$$X_t = \mathtt{I}_t(X_0, X_1, \omega), \quad \omega \sim \pi_\xi,$$

where X_t depends on an extra random variable ω. It should still satisfy the correct boundary condition $X_i = \mathtt{I}_i(X_0, X_1, \omega)$ for $i \in \{0, 1\}$.

Assume the trajectory-wise time derivative $\dot{X}_t = \partial_t \mathtt{I}_t(X_0, X_1, \omega)$, then the rectified flow $\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t$ is defined as $v_t(x) = \mathbb{E}\left[\dot{X}_t \mid X_t = x\right]$ as usual, which solves

$$\min_\mu \int_0^1 \mathbb{E}\left[\|\partial_t \mathtt{I}_t(X_0, X_1, \omega) - \mu_t(\mathtt{I}_t(X_0, X_1, \omega),\ t)\|^2\right] \mathrm{d}t.$$

A natural example of this is the following randomized affine interpolation:

$$X_t = \alpha_t X_1 + \beta_t X_0 + \gamma_t \xi, \tag{3.17}$$

where ξ is an independent noise, and $\alpha_t, \beta_t, \gamma_t$ are differentiable time sequences satisfying $\alpha_0 = 1 - \alpha_1 = \beta_1 = 1 - \beta_0 = \gamma_0 = \gamma_1 = 0$. It yields a loss of form

$$\min_{\mu} \int_0^1 \mathbb{E}\left[\left\|\dot{\alpha}_t X_1 + \dot{\beta}_t X_0 + \dot{\gamma}_t \xi - \mu_t(X_t,\, t)\right\|^2\right] \mathrm{d}t.$$

It introduces the extra flexibility of choosing γ_t and the distribution of ξ beyond the standard affine interpolation.

The question is, however, when and how stochastic interpolations are useful beyond deterministic interpolations. We have the following results:

For a time-differential process $\{X_t\}$, one can de-randomize it into a deterministic interpolation $X_t^{\mathtt{d}} = \mathtt{I}_t^{\mathtt{d}}(X_0, X_1)$, satisfying $\dot{X}_t^{\mathtt{d}} = \mathbb{E}\left[\dot{X}_t \mid X_t, X_1\right]$, such that

1. $\{X_t\}$ and $\{X_t^{\mathtt{d}}\}$ share the same marginal distributions: $\mathrm{Law}(X_t) = \mathrm{Law}(X^{\mathtt{d}})$, $\forall t \in [0,1]$.
2. They yield the same rectified flow, that is,
$$\mathtt{Rectify}(\{X_t^{\mathtt{d}}\}) = \mathtt{Rectify}(\{X_t\}).$$
3. RF loss of $\{X_t^{\mathtt{d}}\}$ yields a Rao–Blackwellization of the RF loss of $\{X_t\}$, and hence has smaller variance.

This suggests that there is no motivation to use a stochastic interpolation if we can use its de-randomized interpolation. In particular, for the randomized affine interpolation in (3.17), when (X_0, X_1, ξ) are mutually independent, and $X_0, \xi \sim \mathtt{Normal}(0, I)$, as the case of 1-rectified flow, the de-randomized interpolation is an affine interpolation $X_t = \alpha_t^{\mathtt{d}} X_1 + \beta_t^{\mathtt{d}} X_0$, where $\alpha_t^{\mathtt{d}}, \beta_t^{\mathtt{d}}$ are determined by $\alpha_t, \beta_t, \gamma_t$. In this case, there may be no clear motivation to use the randomized interpolation (3.17).

On the other hand, when it is computationally intractable to calculate the slope $\dot{X}_t^{\mathtt{d}} = \mathbb{E}\left[\dot{X}_t | X_t, X_1\right]$ of the de-randomized interpolation, it may still be useful to employ randomized interpolations. This is especially true when (X_0, X_1) does not form an independent coupling, as in the case of the reflow step. In such scenarios, introducing randomness enables greater flexibility in the algorithm design space and may induce better regularity conditions on the resulting rectified flow.

3.5.1 De-randomized Interpolation

Let $\{X_t\}$ be a time-differentiable stochastic process that corresponds to a randomized interpolation, Its de-randomized interpolation $\{X_t^{\mathtt{d}}\}$ can be constructed as the rectified flow of the conditioned process $\{X_t\} \mid X_1$.

Definition 9. $\{X_t^{\mathtt{d}}\}$ is said to be the X_1-conditioned rectified flow of $\{X_t\}$ if $(X_0^{\mathtt{d}}, X_1^{\mathtt{d}}) = (X_0, X_1)$, and conditioned on $(X_0^{\mathtt{d}}, X_1^{\mathtt{d}}) = (x_0, x_1)$,

the path $X_t^{\mathtt{d}}$ for $t \in [0,1]$ follows the ODE:

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t^{\mathtt{d}} = v_t^{X|X_1}(X_t^{\mathtt{d}} \mid x_1), \qquad X_0^{\mathtt{d}} = x_0, \tag{3.18}$$

where

$$v_t^{X|X_1}(x_t \mid x_1) = \mathbb{E}\left[\dot{X}_t \mid X_t = x_t, X_1 = x_1\right].$$

We assume the solutions of the ODEs exist and are unique.

Proposition 4. Assume $\{X_t^{\mathtt{d}}\}$ is the X_1-conditioned rectified flow of $\{X_t\}$. We have

1. $\{X_t\}$ and $\{X_t^{\mathtt{d}}\}$ share the same marginal distributions:

$$\mathrm{Law}(X_t) = \mathrm{Law}(X^{\mathtt{d}}), \quad \forall t \in [0,1].$$

2. They yield the same rectified flow, that is,

$$\mathtt{Rectify}(\{X_t^{\mathtt{d}}\}) = \mathtt{Rectify}(\{X_t\}).$$

3. It is easier to predict $\dot{X}_t^{\mathtt{d}}$ from $X_t^{\mathtt{d}}$ than $\dot{X}_t$ from X_t:

$$\begin{aligned}\mathrm{Cov}(\dot{X}_t^{\mathtt{d}} \mid X_t^{\mathtt{d}}) &= \mathrm{Cov}(\dot{X}_t \mid X_t) - \mathbb{E}\left[\mathrm{Cov}(\dot{X}_t | X_t, X_1)\right] \\ &\preceq \mathrm{Cov}(\dot{X}_t \mid X_t).\end{aligned}$$

Proof. 1) From the definition, $(X_0, X_1) = (X_0^{\mathtt{d}}, X_1^{\mathtt{d}})$, and the conditioned process $X_t^{\mathtt{d}}|X_1 = x_1$ is the rectified flow of the conditioned process $X_t|X_1 = x_1$. Therefore, the conditional marginal distribution of $X_t^{\mathtt{d}}|X_1 = x_1$ equals that of $X_t|X_1 = x_1$. Hence, $(X_t^{\mathtt{d}}, X_1^{\mathtt{d}}) \overset{\text{law}}{=} (X_t, X_1)$.

2) Because $X_t^{\mathtt{d}}|X_1 = x_1$ is the rectified flow of $X_t|X_1 = x_1$, we have

$$\mathbb{E}\left[\dot{X}_t^{\mathtt{d}} \mid X_t^{\mathtt{d}} = x_t, X_1 = x_1\right] = \mathbb{E}\left[\dot{X}_t \mid X_t = x_t, X_1 = x_1\right], \quad \forall x_t, x_1.$$

Marginalizing x_1 out:

$$\mathbb{E}\left[\dot{X}_t^{\mathtt{d}} \mid X_t^{\mathtt{d}}\right] = \mathbb{E}\left[\dot{X}_t \mid X_t\right].$$

This suggests that $\{X_t^{\mathtt{d}}\}$ shares the same rectified flow as $\{X_t\}$.

3) For the conditional variance, we have

$$\begin{aligned}\mathrm{Cov}(\dot{X}_t^{\mathtt{d}} \mid X_t^{\mathtt{d}}) &= \mathrm{Cov}(v_t^{X_t|X_1}(X_t^{\mathtt{d}}, X_1) \mid X_t^{\mathtt{d}}) && \text{//by (3.18)} \\ &= \mathrm{Cov}(v_t^{X_t|X_1}(X_t, X_1) \mid X_t) && \text{//}(X_t^{\mathtt{d}}, X_1) \overset{\text{law}}{=} (X_t, X_1) \\ &= \mathrm{Cov}\left(\mathbb{E}\left[\dot{X}_t \mid X_t, X_1\right] \mid X_t\right) \\ &= \mathrm{Cov}(\dot{X}_t \mid X_t) - \mathbb{E}\left[\mathrm{Cov}(\dot{X}_t \mid X_t, X_1)\right] \\ &\preceq \mathrm{Cov}(\dot{X}_t \mid X_t),\end{aligned}$$

where we use the law of total variance.

□

3.5.2 De-randomizing Affine Interpolations

Lemma 7. Assume $X_t = \alpha_t X_1 + \beta_t X_0 + \gamma_t \xi$, where (X_0, X_1, ξ) are mutually independent and $X_0, \xi \sim \texttt{Normal}(0, I)$. The de-randomized interpolation $\{X_t^{\mathtt{d}}\}$ of $\{X_t\}$ satisfies

$$X_t^{\mathtt{d}} = \alpha_t^{\mathtt{d}} X_1 + \beta_t^{\mathtt{d}} X_0,$$

where $\alpha_t^{\mathtt{d}}$, $\beta_t^{\mathtt{d}}$ satisfies

$$\dot{\alpha}_t^{\mathtt{d}} - \frac{\dot{\beta}_t^{\mathtt{d}}}{\beta_t^{\mathtt{d}}}\alpha_t^{\mathtt{d}} = \dot{\alpha}_t - \frac{\dot{\beta}_t}{\beta_t}\alpha_t - \left(\dot{\gamma}_t - \frac{\dot{\beta}_t}{\beta_t}\gamma_t\right)\gamma_t(\beta_t^2 + \gamma_t^2)^{-1}\alpha_t,$$

$$\frac{\dot{\beta}_t^{\mathtt{d}}}{\beta_t^{\mathtt{d}}} = \frac{\dot{\beta}_t}{\beta_t} + \left(\dot{\gamma}_t - \frac{\dot{\beta}_t}{\beta_t}\gamma_t\right)\gamma_t(\beta_t^2 + \gamma_t^2)^{-1}.$$

Lemma 8. Assume $\{X_t\}$ satisfies $X_t = \alpha_t X_1 + \beta_t X_0 + \gamma_t \xi$, where $\alpha_t, \beta_t, \gamma_t$ are time-differential sequences. Define

$$\hat{\xi}_t^{X|X_t}(x_t \mid x_1) = \mathbb{E}\left[\xi \mid X_t = x_t, X_1 = x_1\right].$$

1) We have

$$v_t^{X|X_1}(x_t \mid x_1) = \left(\dot{\alpha}_t - \frac{\dot{\beta}_t}{\beta_t}\alpha_t\right)x_1 + \frac{\dot{\beta}_t}{\beta_t}x_t + \left(\dot{\gamma}_t - \frac{\dot{\beta}_t}{\beta_t}\gamma_t\right)\hat{\xi}_t^{X|X_t}(x_t \mid x_1).$$

2) Further, assume $\xi \sim \texttt{Normal}(0, I)$, and $(X_0, X_1) \perp\!\!\!\perp \xi$. Let ρ_t be the density function of X_t. We have

$$\hat{\xi}_t^{X|X_t}(x_t \mid x_1) = -\gamma_t \nabla \log \rho_{X_t|X_1}(x_t|x_1).$$

3) Further, assume $X_0 \perp\!\!\!\perp X_1$ and $X_0 \sim \texttt{Normal}(\mu_0, \Sigma_0)$. Then

$$\hat{\xi}_t^{X|X_t}(x_t \mid x_1) = \gamma_t \Sigma_t^{-1}(x_t - \mu_t(x_1)),$$

where $\mu_t(x_1) = \alpha_t x_1 + \beta_t \mu_0$, and $\Sigma_t = \beta_t^2 \Sigma_0 + \gamma_t^2 I$.

4) Further, assume $\mu_0 = 0$ and $\Sigma_0 = I$. We have

$$v_t^{X|X_1}(x_t \mid x_1) = a_t^{\mathtt{d}} x_1 + b_t^{\mathtt{d}} x_t.$$

where

$$a_t^{\mathtt{d}} = \dot{\alpha}_t - \frac{\dot{\beta}_t}{\beta_t}\alpha_t - \left(\dot{\gamma}_t - \frac{\dot{\beta}_t}{\beta_t}\gamma_t\right)\gamma_t(\beta_t^2 + \gamma_t^2)^{-1}\alpha_t,$$

$$b_t^{\mathtt{d}} = \frac{\dot{\beta}_t}{\beta_t} + \left(\dot{\gamma}_t - \frac{\dot{\beta}_t}{\beta_t}\gamma_t\right)\gamma_t(\beta_t^2 + \gamma_t^2)^{-1}.$$

Hence, $X_t^{\mathtt{d}}$ satisfies

$$X_t^{\mathtt{d}} = \alpha_t^{\mathtt{d}} X_1 + \beta_t^{\mathtt{d}} X_0,$$

where $\alpha_t^{\mathtt{d}}$ and $\beta_t^{\mathtt{d}}$ should satisfy

$$\dot{\alpha}_t^{\mathtt{d}} - \frac{\dot{\beta}_t^{\mathtt{d}}}{\beta_t^{\mathtt{d}}}\alpha_t^{\mathtt{d}} = \dot{\alpha}_t - \frac{\dot{\beta}_t}{\beta_t}\alpha_t - \left(\dot{\gamma}_t - \frac{\dot{\beta}_t}{\beta_t}\gamma_t\right)\gamma_t(\beta_t^2 + \gamma_t^2)^{-1}\alpha_t$$

$$\frac{\dot{\beta}_t^{\mathtt{d}}}{\beta_t^{\mathtt{d}}} = \frac{\dot{\beta}_t}{\beta_t} + \left(\dot{\gamma}_t - \frac{\dot{\beta}_t}{\beta_t}\gamma_t\right)\gamma_t(\beta_t^2 + \gamma_t^2)^{-1}.$$

Proof. For 1), we have

$$\begin{aligned}
v_t^{X|X_1}(x_t \mid x_1) &:= \mathbb{E}\left[\dot{X}_t \mid X_t = x_t, X_1 = x_1\right] \\
&= \mathbb{E}\left[\dot{\alpha}_t X_1 + \dot{\beta}_t X_0 + \dot{\gamma}_t \xi \mid X_t = x_t, X_1 = x_1\right] \\
&= \mathbb{E}\left[\dot{\alpha}_t x_1 + \dot{\beta}_t\left(\frac{x_t - \alpha_t x_1 - \gamma_t \xi}{\beta_t}\right) + \dot{\gamma}_t \xi \mid X_t = x_t, X_1 = x_1\right] \\
&= \left(\dot{\alpha}_t - \frac{\dot{\beta}_t}{\beta_t}\alpha_t\right) x_1 + \frac{\dot{\beta}_t}{\beta_t} x_t + \left(\dot{\gamma}_t - \frac{\dot{\beta}_t}{\beta_t}\gamma_t\right)\mathbb{E}\left[\xi \mid X_t = x_t, X_1 = x_1\right].
\end{aligned}$$

For 2), using the generalized Tweedie's formula in Theorem 6, we have

$$\nabla \log \rho_{X_t|X_1}(x_t \mid x_1) = \gamma_t^{-1}\mathbb{E}\left[\nabla_\xi \log \rho_\xi(\xi) \mid X_t = x_t, X_1 = x_1\right].$$

The results follow given that $\nabla \log \rho_\xi(\xi) = -\xi$.

For 3), we just use 2) and note that $X_t \mid X_1 = x_1 \sim \texttt{Normal}(\mu_t(x_1), \Sigma_t)$.

For 4), the form of $v_t^{X|X)t}$ is obtained by combining the results above.

□

3.6 Affine Interpolation Identities

In many derivations and algorithms, it is often necessary to solve equations related to interpolation functions. For instance, given X_t and $\dot{X}_t$, we often need to to find X_0 and X_1 that satisfy $X_t = \mathtt{I}_t(X_0, X_1)$ and $\dot{X}_t = \dot{\mathtt{I}}_t(X_0, X_1)$.

For affine interpolations, this problem reduces to solving a simple 2×2 linear system, which yields a closed-form solution. Additionally, the conditional expectation counterparts, $\hat{x}_{i|t}(X_t) = \mathbb{E}[X_i \mid X_t]$ and $v_t(X_t) = \mathbb{E}[\dot{X}_t \mid X_t]$, satisfy the same relations due to the linearity of expectation.

For convenience, we collect these formulas here for easy reference.

Lemma 9. Let $X_t = \alpha_t X_1 + \beta_t X_0$, and $\dot{X}_t = \dot{\alpha}_t X_1 + \dot{\beta}_t X_0$, and

$$\begin{aligned}
&\text{RF velocity field :} && v_t(x) = \mathbb{E}\left[\dot{X}_t \mid X_t = x\right] \\
&\text{Expected noise } X_0 : && \hat{x}_{0|t}(x) = \mathbb{E}\left[X_1 \mid X_t = x\right] \\
&\text{Expected target } X_1 : && \hat{x}_{1|t}(x) = \mathbb{E}\left[X_1 \mid X_t = x\right].
\end{aligned}$$

We have

$$\begin{aligned}
v_t(x) &= \frac{\dot{\alpha}_t}{\alpha_t} x - \beta_t \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right) \hat{x}_{0|t}(x), \\
v_t(x) &= \frac{\dot{\beta}_t}{\beta_t} x + \alpha_t \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right) \hat{x}_{1|t}(x),
\end{aligned}$$

and

$$\begin{aligned}
\hat{x}_{0|t}(x) &= \frac{-\dot{\alpha}_t x + \alpha_t v_t(x)}{\alpha_t \dot{\beta}_t - \dot{\alpha}_t \beta_t} = -\frac{1}{\beta_t} \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right)^{-1} \left(v_t(x) - \frac{\dot{\alpha}_t}{\alpha_t} x\right), \\
\hat{x}_{1|t}(x) &= \frac{\dot{\beta}_t x - \beta_t v_t(x)}{\alpha_t \dot{\beta}_t - \dot{\alpha}_t \beta_t} = \frac{1}{\alpha_t} \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right)^{-1} \left(v_t(x) - \frac{\dot{\beta}_t}{\beta_t} x\right),
\end{aligned}$$

and

$$\hat{x}_{1|t}(x) = \frac{x - \beta_t \hat{x}_{0|t}(x)}{\alpha_t}, \qquad \hat{x}_{0|t}(x) = \frac{x - \alpha_t \hat{x}_{1|t}(x)}{\beta_t}.$$

Proof.

$$\begin{aligned}
v_t(x) &= \mathbb{E}\left[\dot{X}_t \mid X_t = x\right] \\
&= \mathbb{E}\left[\dot{\alpha}_t X_1 + \dot{\beta}_t X_0 \mid X_t = x\right] \\
&= \mathbb{E}\left[\dot{\alpha}_t X_1 + \dot{\beta}_t \frac{(X_t - \alpha_t X_1)}{\beta_t} \mid X_t = x\right] \\
&= \frac{\dot{\beta}_t}{\beta_t} x + \alpha_t \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right) \hat{x}_{1|t}(x).
\end{aligned}$$

The other relations are derived similarly using the relations in Lemma 10. □

Lemma 10. Let $X_t = \alpha_t X_1 + \beta_t X_0$, and $\dot{X}_t = \dot{\alpha}_t X_1 + \dot{\beta}_t X_0$, then
1) X_0, X_1 from $X_t, \dot{X}_t$:

$$\begin{aligned}
X_1 &= \frac{\dot{\beta}_t X_t - \beta_t \dot{X}_t}{\alpha_t \dot{\beta}_t - \dot{\alpha}_t \beta_t} = \frac{1}{\alpha_t} \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right)^{-1} \left(\dot{X}_t - \frac{\dot{\beta}_t}{\beta_t} X_t\right), \\
X_0 &= \frac{-\dot{\alpha}_t X_t + \alpha_t \dot{X}_t}{\alpha_t \dot{\beta}_t - \dot{\alpha}_t \beta_t} = -\frac{1}{\beta_t} \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right)^{-1} \left(\dot{X}_t - \frac{\dot{\alpha}_t}{\alpha_t} X_t\right).
\end{aligned}$$

2) $\dot{X}_t$ from X_t, X_0 (or X_1):

$$\dot{X}_t = \frac{\dot{\alpha}_t}{\alpha_t} X_t - \beta_t (\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}) X_0,$$
$$\dot{X}_t = \frac{\dot{\beta}_t}{\beta_t} X_t + \alpha_t (\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}) X_1.$$

3) X_t from $\dot{X}_t$ and X_0 (or X_1):

$$X_t = \frac{\alpha_t}{\dot{\alpha}_t} \left(\dot{X}_t + \beta_t (\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}) X_0 \right),$$
$$X_t = \frac{\beta_t}{\dot{\beta}_t} \left(\dot{X}_t - \alpha_t (\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}) X_1 \right).$$

4) X_0 (resp. X_1) from X_t and X_1 (resp. X_0):

$$X_1 = \frac{X_t - \beta_t X_0}{\alpha_t}, \qquad X_0 = \frac{X_t - \alpha_t X_0}{\beta_t}.$$

Lemma 11. Let $X_t = \alpha_t X_1 + \beta_t X_0$, and $\dot{X}_t = \dot{\alpha}_t X_1 + \dot{\beta}_t X_0$. We have
1) X_0, X_1 from $\dot{X}_t$:

$$X_1 - \mathbb{E}[X_1|X_t] = \frac{1}{\alpha_t} \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t} \right)^{-1} \left(\dot{X}_t - \mathbb{E}\left[\dot{X}_t | X_t \right] \right),$$
$$X_0 - \mathbb{E}[X_0|X_t] = -\frac{1}{\beta_t} \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t} \right)^{-1} \left(\dot{X}_t - \mathbb{E}\left[\dot{X}_t | X_t \right] \right).$$

2) $\dot{X}_t$ from X_0 (or X_1):

$$\dot{X}_t - \mathbb{E}\left[\dot{X}_t | X_t \right] = -\beta_t (\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})(X_0 - \mathbb{E}[X_0|X_t]),$$
$$\dot{X}_t - \mathbb{E}\left[\dot{X}_t | X_t \right] = \alpha_t (\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})(X_1 - \mathbb{E}[X_1|X_t]).$$

3) X_0 and X_1:

$$X_1 - \mathbb{E}[X_1|X_t] = -\frac{\beta_t}{\alpha_t}(X_0 - \mathbb{E}[X_0|X_t]),$$
$$X_0 - \mathbb{E}[X_0|X_t] = -\frac{\alpha_t}{\beta_t}(X_1 - \mathbb{E}[X_1|X_t]).$$

CHAPTER FOUR

Identities

We present some important identities related to rectified flow, particularly in the case when (X_0, X_1) is an independent coupling, which corresponds to the 1-rectified flow derived from the data. Furthermore, more appealing properties can be established if π_0 is a Gaussian distribution.

4.1 Score Identities

Assume that ρ_t is the density function of X_t for $X_t = \alpha_t X_1 + \beta_t X_0$. In this section, we provide a set of formulas regarding the score function $\nabla \log \rho_t$.

In particular, when $X_0 \perp\!\!\!\perp X_1$ and X_0 is Gaussian, the score function is connected with the RF velocity field via

$$\nabla \log \rho_t(x) = \frac{1}{\beta_t^2}\left(\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t}\right)^{-1}\left(v_t(x) - \frac{\dot\alpha_t}{\alpha_t}x\right). \tag{4.1}$$

This is a well known result that have been heavily explored in the literature.

We start with an array of basic identities regarding score functions involving marginalizing latent variables. Note the subtle differences between the three results below and their conditions.

Lemma 12 (Score Functions and Latent Variables). Assume all log densities involved below are continuously differentiable.

1) For any continuous random variable (X, Y),

$$\begin{aligned}\nabla_x \log \rho_X(x) &= \mathbb{E}\left[\nabla_X \log \rho_{X,Z}(X, Z) \mid X = x\right] \\ &= \mathbb{E}\left[\nabla_X \log \rho_{X|Z}(X|Z) \mid X = x\right],\end{aligned}$$

where $\rho_{X|Z}, \rho_{X,Z}$ is the conditional and joint density functions.

2) Moreover, if $X = Y + Z$, we have

$$\begin{aligned}\nabla_x \log \rho_X(x) &= \mathbb{E}\left[\nabla_Z \log \rho_{Z|Y}(Z|Y) \mid X = x\right] \\ &= \mathbb{E}\left[\nabla_Y \log \rho_{Y|Z}(Y|Z) \mid X = x\right].\end{aligned}$$

3) Further, if $X = Y + Z$ and $Y \perp\!\!\!\perp Z$, we have

$$\begin{aligned}\nabla_x \log \rho_X(x) &= \mathbb{E}\left[\nabla_Z \log \rho_Z(Z) \mid X = x\right] \\ &= \mathbb{E}\left[\nabla_Y \log \rho_Y(Y) \mid X = x\right].\end{aligned}$$

Proof. 1) Directly taking the derivative:

$$\begin{aligned}
\nabla_x \log \rho_X(x) &= \frac{\nabla_x \rho_X(x)}{\rho_X(x)} \\
&\overset{*}{=} \frac{\nabla_x \int_z \rho_{X,Z}(x,z)dz}{\rho_X(x)} \\
&= \frac{\int \nabla_x \log \rho_{X,Z}(x,z)\rho_{X,Z}(x,z)dz}{\rho_X(x)} \\
&= \int \nabla_x \log \rho_{X,Z}(x,z)\rho_{Z|X}(x,z)dz \\
&= \mathbb{E}[\nabla_x \log \rho_{X,Z}(X,Z) \mid X = x] \\
&\overset{**}{=} \mathbb{E}[\nabla_x \log \rho_{X|Z}(X|Z) \mid X = x].
\end{aligned}$$

The major step here is to exchange the order of derivative and integral operators in $\overset{*}{=}$. In the last step $\overset{**}{=}$, we used

$$\nabla_x \log \rho_{X,Z}(x,z) = \nabla_x \log \rho_Z(z) + \nabla_x \log \rho_{X|Z}(x|z) = \nabla_x \log \rho_{X|Z}(x|z).$$

2) If $X = Y + Z$, we have

$$\rho_{X|Z}(x|z) = \rho_{Y|Z}(x - z|z).$$

Plugging it into the result above yields

$$\begin{aligned}
\nabla_x \log \rho_X(x) &= \mathbb{E}[\nabla_X \log \rho_{Y|Z}(X - Z|Z) \mid X = x] \\
&= \mathbb{E}[\nabla_Y \log \rho_{Y|Z}(Y|Z) \mid X = x],
\end{aligned}$$

where we used $Y = X - Z$ in the last step.

3) Further, if $Y \perp\!\!\!\perp Z$, we have $\rho_{Y|Z}(y|z) = \rho_Y(y)$, and hence the results follow directly.

□

Theorem 6. Assume $X_t = \alpha_t X_1 + \beta_t X_0$, where $X_0 \perp\!\!\!\perp X_1$. We have

$$\begin{aligned}
\nabla \log \rho_t(x) &= \mathbb{E}\left[\beta_t^{-1} \nabla \log \rho_0(X_0) \mid X_t = x\right] \\
&= \mathbb{E}\left[\alpha_t^{-1} \nabla \log \rho_1(X_1) \mid X_t = x\right],
\end{aligned} \tag{4.2}$$

where ρ_0, ρ_1 are the density functions of X_0 and X_1, respectively.

Equation (4.2) expresses $\nabla \log \rho_X$ using either $\nabla \log \rho_Y$ or $\nabla \log \rho_Z$. We should decide which one to use based on the problems and our needs.

Proof. Let $Y = \beta_t X_0$ and $Z = \alpha_t X_1$, we have $\rho_Y(y) = \rho_{X_0}(\beta_t^{-1} y)\beta_t^{-1}$, and hence $\nabla_y \log \rho_Y(y) = \beta_t^{-1} \nabla \log \rho_{X_0}(\beta_t^{-1} y)$.

Using Lemma 12, we have

$$\begin{aligned}
\nabla \log \rho_t(x) &= \mathbb{E}\left[\nabla \log \rho_Y(Y) \mid X_t = x\right] \\
&= \mathbb{E}\left[\beta_t^{-1} \nabla \log \rho_{X_0}(\beta_t^{-1} Y) \mid X_t = x\right] \\
&= \mathbb{E}\left[\beta_t^{-1} \nabla \log \rho_{X_0}(X_0) \mid X_t = x\right],
\end{aligned}$$

where we substitute $X_0 = \beta_t^{-1} Y$ in the last step. The other form follows by symmetry. □

The case when X_0 is Gaussian, say $X_0 \sim \texttt{Normal}(\mu_0, \Sigma_0)$, is particularly interesting because the score function of the noise $\nabla \log \rho_0(X_0) = \Sigma_0^{-1}(X_0 - \mu)$ is linear on X_0, which allows for a simple relation between $\nabla \log \rho_t$, the expected noise $\hat{x}_{0|t}(x) = \mathbb{E}[X_0 | X_t = x]$, and the velocity v_t.

Corollary 3 (Tweedie's Formula). Let $X_t = \alpha_t X_1 + \beta_t X_0$ and $X_0 \perp\!\!\!\perp X_1$. If $X_0 \sim \texttt{Normal}(\mu_0, \Sigma_0)$, we have $\nabla \log \rho_0(x) = -\Sigma_0^{-1}(x - \mu_0)$, and hence

$$\begin{aligned}\nabla \log \rho_t(x) &= -\frac{1}{\beta_t} \mathbb{E}\left[\Sigma_0^{-1}(X_0 - \mu_0) \mid X_t = x\right] \\ &= -\frac{1}{\beta_t} \Sigma_0^{-1}(\hat{x}_{0|t}(x) - \mu_0).\end{aligned}$$

Using the relation of $\hat{x}_{0|t}(x)$ and $v_t(x)$ in Lemma 9, we have

$$\begin{aligned}\nabla \log \rho_t(x) &= -\frac{1}{\beta_t} \Sigma_0^{-1}(\hat{x}_{0|t}(x) - \mu_0) \\ &= \frac{1}{\beta_t^2} \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right)^{-1} \Sigma_0^{-1} \left(v_t(x) - \frac{\dot{\alpha}_t}{\alpha_t} x\right) + \frac{1}{\beta_t} \Sigma_0^{-1} \mu_0.\end{aligned}$$

Hence,

$$v_t(x) = \frac{\dot{\alpha}_t}{\alpha_t} x + \beta_t \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right) \Sigma_0 (\beta_t \nabla \log \rho_t(x) - \mu_0).$$

Corollary 4 (Tweedie's Formula). In particular, if $X_0 \sim \texttt{Normal}(0, I)$ is the standard Gaussian noise, we have

$$\nabla \log \rho_t(x) = -\frac{1}{\beta_t} \mathbb{E}\left[X_0 \mid X_t = x\right].$$

Hence,

$$v_t(x) = \frac{\dot{\alpha}_t}{\alpha_t} x + \beta_t^2 \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right) \nabla \log \rho_t(x),$$

and

$$\nabla \log \rho_t(x) = \frac{1}{\beta_t^2} \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right)^{-1} \left(v_t(x) - \frac{\dot{\alpha}_t}{\alpha_t} x\right).$$

Example 11 (Score Function of Straight Interpolation). For straight interpolation $X_t = tX_1 + (1-t)X_0$ with $X_0 \perp\!\!\!\perp X_1$ and $X_0 \sim \texttt{Normal}(0, I)$,

the score function $\nabla \log \rho_t$ and velocity field v_t is related via

$$\begin{aligned} \nabla \log \rho_t(x) &= \frac{t v_t(x) - x}{1-t}, \\ v_t(x) &= \frac{(1-t)\nabla \log \rho_t(x) + x}{t}. \end{aligned} \tag{4.3}$$

Remark 20. If X_0 is not Gaussian, the score function of the noise $\nabla \log \rho_0(X_0)$ is not linear on X_0, and hence $\nabla \log \rho_t$ can not be expressed with v_t. In this case, we may need to learn an extra model $g_t(x, \theta)$ to approximate $\nabla \log \rho_t(x)$. Using Equation (4.2), a loss function can be

$$\min_\theta \int \eta_t \mathbb{E}\left[\|\beta_t g_t(X_t, \theta) - \nabla \log \rho_0(X_0)\|^2\right] \mathrm{d}t.$$

4.2 Covariance Identities

For $X_t = \alpha_t X_1 + \beta_t X_0$ with $X_0 \perp\!\!\!\perp X_1$, and $X_0 \sim \texttt{Normal}(0, I)$, we provide some identities related to the derivative matrix ∇v_t and the Hessian matrix $\nabla^2 \log \rho_t$:

$$\begin{aligned} \nabla^2 \log \rho_t(x) &= \beta_t^{-2}(\mathrm{Var}(X_0 \mid X_t = x) - I) \\ \nabla v_t(x) &= \frac{\dot\beta_t}{\beta_t} I + \frac{1}{\beta_t^2}\left(\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t}\right)^{-1} \mathrm{Var}(\dot X_t \mid X_t = x), \end{aligned} \tag{4.4}$$

where $\mathrm{Var}(\cdot|\cdot)$ denotes the conditional covariance matrix. These formulas may find various applications. We provide two examples here.

Divergence and Straightness

Taking the sum of diagonals yields a formula for the divergence $\nabla \cdot v_t$:

$$\nabla \cdot v_t(x) = \frac{\dot\beta_t}{\beta_t} d + \frac{1}{\beta_t^2}\left(\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t}\right)^{-1} \mathrm{Trace}(\mathrm{Var}(\dot X_t \mid X_t = x)), \tag{4.5}$$

where d is the dimension of X_t.

On the other hand, note that the trace of conditional variance $\mathrm{Trace}(\mathrm{Var}(\dot X_t \mid X_t = x))$ coincides with the minimum square loss:

$$\ell_t^* := \min_{v_t} \mathbb{E}\left[\left\|\dot X_t - v_t(X_t)\right\|^2\right] = \mathbb{E}\left[\mathrm{Trace}(\mathrm{Var}(\dot X_t \mid X_t))\right], \tag{4.6}$$

where $\ell_t^* = 0$ implies that $\dot X_t$ is determined by X_t, indicating no different trajectories of $\{X_t\}$ intersect at t. Hence ℓ_t^* is an indication of straightness of the rectified flow. Combining (4.5) and (4.6) yields

$$\ell_t^* = \beta_t^2\left(\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t}\right)\left(\mathbb{E}\left[\nabla \cdot v_t(X_t)\right] - \frac{\dot\beta_t}{\beta_t} d\right), \tag{4.7}$$

This suggests that the divergence $\mathbb{E}\left[\nabla \cdot v_t(X_t)\right]$ reveals information regarding the straightness of the rectified flow. This also provides an approach to empirically estimate ℓ_t^*, since the right hand side can be estimated given v_t.

Monotonicity of One-Step Euler Steps

Equation (4.4) suggest that $\nabla v_t(x) - \frac{\dot{\beta}_t}{\beta_t} I$ is a positive semi-definite matrix. This can be used to show that the one-step Euler update $\Phi_{s|t}^{\texttt{Euler}}(x) = x + (s-t)v_t(x)$ is always a monotonic map and corresponds to the gradient of a convex function. We discuss this result in detail in Section 4.4, and use it to prove a bound in relation to L2 optimal transport in (4.5)

Proofs

We now provide the proofs of Equation (4.4) and more related results.

Definition 10. The cross-covariance matrix of a joint random vector (X, Y) is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top\right],$$

The covariance matrix of X is

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\right].$$

Lemma 13. Let $X_t = \alpha_t X_1 + \beta_t X_0$ where $X_0 \perp\!\!\!\perp X_1$, and $X_0 \sim \texttt{Normal}(0, I)$.
For $\nabla_x \hat{x}_{1|t}(x) := \mathbb{E}[X_1 \mid X_t = x]$, we have

$$\begin{aligned}
\nabla_x \hat{x}_{1|t}(x) &:= \mathbb{E}[X_1 \mid X_t = x] \\
&= -\frac{1}{\beta_t}\mathrm{Cov}(X_0,\ X_1 \mid X_t = x) \\
&= \frac{\alpha_t}{\beta_t^2}\mathrm{Var}(X_1 \mid X_t = x) \\
&= \frac{1}{\alpha_t}\mathrm{Var}(X_0 \mid X_t = x).
\end{aligned} \tag{4.8}$$

For $v_t(x) = \mathbb{E}\left[\dot{X}_t | X_t = x\right]$, we have

$$\begin{aligned}
\nabla_x v_t(x) &= \frac{\dot{\beta}_t}{\beta_t} I - \frac{\alpha_t}{\beta_t}\left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right)\mathrm{Cov}(X_0, X_1 \mid X_t = x) \\
&= \frac{\dot{\beta}_t}{\beta_t} I + \frac{\alpha_t^2}{\beta_t^2}\left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right)\mathrm{Var}(X_1 \mid X_t = x) \\
&= \frac{\dot{\beta}_t}{\beta_t} I + \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right)\mathrm{Var}(X_0 \mid X_t = x) \\
&= \frac{\dot{\beta}_t}{\beta_t} I + \frac{1}{\beta_t^2}\left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right)^{-1}\mathrm{Var}(\dot{X}_t \mid X_t = x)
\end{aligned} \tag{4.9}$$

Let ρ_t be the density function of X_t, the Hessian of $\log \rho_t$ is

$$\nabla^2 \log \rho_t(x) = \beta_t^{-2}(\mathrm{Var}(X_0 \mid X_t = x) - I). \tag{4.10}$$

Proof. Write $X_t = \alpha_t X_1 + Y$, where $Y = \beta_t X_0 \sim \texttt{Normal}(0, \beta_t^2 I)$.

Using Lemma 16 on $\hat{x}_{1|t}(x) = \mathbb{E}[X_1|X_t = x]$, we have

$$\nabla \hat{x}_{1|t}(x) = \mathrm{Cov}((-\frac{Y}{\beta_t^2}), X_1 \mid X_t = x) = -\frac{1}{\beta_t}\mathrm{Cov}(X_0,\ X_1 \mid X_t = x).$$

We then have (4.8) using Lemma 15.

For $v_t(x)$, Eq (4.9) follows by noting that $\hat{x}_{1|t}(x)$ and $v_t(x)$ are relate via

$$\begin{aligned} v_t(x) &= \mathbb{E}\left[\dot{\alpha}_t X_1 + \dot{\beta}_t X_0 \mid X_t = x\right] \\ &= \mathbb{E}\left[\dot{\alpha}_t X_1 + \dot{\beta}_t \frac{(X_t - \alpha_t X_1)}{\beta_t} \mid X_t = x\right] \\ &= \frac{\dot{\beta}_t}{\beta_t} x + \alpha_t(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})\hat{x}_{1|t}(x). \end{aligned}$$

For $\nabla \log \rho_t(x)$, note that

$$\begin{aligned} \nabla \log \rho_t(x) &= \mathbb{E}\left[-\frac{X_0}{\beta_t} \mid X_t = x\right] \\ &= \mathbb{E}\left[-\frac{X_t - \alpha_t X_1}{\beta_t^2} \mid X_t = x\right] \\ &= -\frac{x}{\beta_t^2} + \frac{\alpha_t}{\beta_t^2}\hat{x}_{1|t}(x). \end{aligned}$$

Hence, plugging (4.8) gives

$$\begin{aligned} \nabla^2 \log \rho_t(x) &= \frac{\alpha_t}{\beta_t^2}\nabla \hat{x}_{1|t}(x) - \frac{1}{\beta_t^2} I \\ &= -\frac{\alpha_t}{\beta_t^3}\mathrm{Cov}(X_1,\ X_0 \mid X_t = x) - \frac{1}{\beta_t^2} I \\ &= \frac{\alpha_t^2}{\beta_t^4}\mathrm{Var}(X_1 \mid X_t = x) - \frac{1}{\beta_t^2} I \\ &= \frac{1}{\beta_t^2}(\mathrm{Var}(X_0 \mid X_t = x) - I). \end{aligned}$$

□

Lemma 14. For any random variables (X, Y, Z), we have

$$\mathrm{Cov}(X,\ (aY + bZ)) = a\mathrm{Cov}(X, Y) + b\mathrm{Cov}(X, Z).$$

Proof.

$$\begin{aligned} &\mathrm{Cov}(X,\ (aY + bZ)) \\ &= \mathbb{E}\left[X(aY + bZ)^\top\right] - \mathbb{E}[X]\,\mathbb{E}\left[(aY + bZ)^\top\right] \\ &= a\mathbb{E}\left[XY^\top\right] + b\mathbb{E}\left[XZ^\top\right] - a\mathbb{E}[X]\,\mathbb{E}\left[Y^\top\right] - b\mathbb{E}[X]\,\mathbb{E}\left[Z^\top\right] \\ &= a\mathrm{Cov}(X, Y) + b\mathrm{Cov}(X, Z). \end{aligned}$$

□

Lemma 15. For $X_t = \alpha_t X_1 + \beta_t X_0$ and $\dot{X}_t = \dot{\alpha}_t X_1 + \dot{\beta}_t X_0$, we have

$$\begin{aligned}\mathrm{Cov}(X_1,\ X_0 \mid X_t) &= -\frac{\beta_t}{\alpha_t}\mathrm{Var}(X_0 \mid X_t) \\ &= -\frac{\alpha_t}{\beta_t}\mathrm{Var}(X_1 \mid X_t) \\ &= -\frac{1}{\alpha_t\beta_t(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})^2}\mathrm{Var}(\dot{X}_t \mid X_t = x).\end{aligned}$$

This suggests that X_0 and X_1 are negatively correlated conditioned on X_t if $\alpha_t, \beta_t \geq 0$, regardless of the unconditioned coupling (X_0, X_1).

Proof. Plugging $X_0 = (X_t - \alpha_t X_1)/\beta_t$, we get

$$\begin{aligned}&\mathrm{Cov}(X_1,\ X_0 \mid X_t) \\ &= \frac{1}{\beta_t}\mathrm{Cov}(X_1,\ (X_t - \alpha_t X_1) \mid X_t = x) \\ &= \frac{1}{\beta_t}(\mathrm{Cov}(X_1,\ X_t | X_t = x) - \alpha_t \mathrm{Cov}(X_1,\ X_1 \mid X_t = x)) \\ &= -\frac{\alpha_t}{\beta_t}\mathrm{Cov}(X_1,\ X_1 \mid X_t = x),\end{aligned}$$

where we used $\mathrm{Cov}(X_1, X_t | X_t = x) = \mathrm{Cov}(X_1, x | X_t = x) = 0$, that the covariance of any random variable with a constant equals zero.

Similarly, using $X_1 = (X_t - \beta_t X_0)/\alpha_t$, we get

$$\begin{aligned}&\mathrm{Cov}(X_1,\ X_0 \mid X_t) \\ &= \mathrm{Cov}((X_t - \beta_t X_0)/\alpha_t,\ X_0 \mid X_t = x) \\ &= -\frac{\beta_t}{\alpha_t}\mathrm{Cov}(X_0,\ X_0 \mid X_t = x).\end{aligned}$$

Finally, solving $X_t = \alpha_t X_1 + \beta_t X_0$ and $\dot{X}_t = \dot{\alpha}_t X_1 + \dot{\beta}_t X_0$ for X_0, and X_1, we get

$$X_0 = \frac{\dot{\alpha}_t X_t - \alpha_t \dot{X}_t}{\dot{\alpha}_t\beta_t - \alpha_t\dot{\beta}_t}, \qquad X_1 = \frac{\dot{\beta}_t X_t - \beta_t \dot{X}_t}{\dot{\beta}_t\alpha_t - \beta_t\dot{\alpha}_t}.$$

Hence,

$$\begin{aligned}\mathrm{Cov}(X_0, X_1 \mid X_t) &= \mathrm{Cov}\left(\frac{\dot{\alpha}_t X_t - \alpha_t \dot{X}_t}{\dot{\alpha}_t\beta_t - \alpha_t\dot{\beta}_t},\ \frac{\dot{\beta}_t X_t - \beta_t \dot{X}_t}{\dot{\beta}_t\alpha_t - \beta_t\dot{\alpha}_t} \,\middle|\, X_t\right) \\ &= \mathrm{Cov}\left(\frac{-\alpha_t \dot{X}_t}{\dot{\alpha}_t\beta_t - \alpha_t\dot{\beta}_t},\ \frac{-\beta_t \dot{X}_t}{\dot{\beta}_t\alpha_t - \beta_t\dot{\alpha}_t} \,\middle|\, X_t\right) \qquad //X_t \text{ are viewed as constants} \\ &= -\frac{1}{\alpha_t\beta_t\kappa_t^2}\mathrm{Cov}(\dot{X}_t, \dot{X}_t \mid X_t),\end{aligned}$$

where we write $\kappa_t = \frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\alpha}_t}{\alpha_t}$. □

Lemma 16 (Second Order Score Identities). Let $\mu_h(X) = \mathbb{E}[h(Z) \mid X]$

where (X, Z) is a joint random variable and h is a function. Let $\rho_{X,Z}$, $\rho_{X|Z}$, and ρ_X denote the joint, conditional, and marginal density functions, respectively. Assume all relevant derivatives exist and continuous.
1) For any (X, Z), we have

$$\begin{aligned}\nabla_X \mu_h(X) &= \mathrm{Cov}(\nabla_X \log \rho_{X,Z}(X,\ Z),\ h(Z) \mid X) \\ &= \mathrm{Cov}(\nabla_X \log \rho_{X|Z}(X \mid Z),\ h(Z) \mid X).\end{aligned}$$

2) If $X = Y + Z$, we have

$$\nabla_X \mu_h(X) = \mathrm{Cov}(\nabla_Y \log \rho_{Y|Z}(Y \mid Z),\ h(Z) \mid X).$$

3) If $X = Y + Z$ and $Y \perp\!\!\!\perp Z$, we have

$$\nabla_X \mu_h(X) = \mathrm{Cov}(\nabla_Y \log \rho_Y(Y),\ h(Z) \mid X).$$

4) If $X = Y + Z$, $Y \perp\!\!\!\perp Z$ and $Y \sim \texttt{Normal}(\mu, \Sigma)$, we have

$$\nabla_X \mu_h(X) = \mathrm{Cov}(\Sigma^{-1}(\mu - Y),\ h(Z) \mid X).$$

Proof.

$$\mu_h(x) = \frac{\int h(z)\rho_{X,Z}(x,z)\mathrm{d}z}{\rho_X(x)}.$$

Taking derivative, we have

$$\begin{aligned}\nabla\mu_h(x) &= \frac{\int h(z)\nabla_x\rho_{X,Z}(x,z)\mathrm{d}z}{\rho_X(x)} - \frac{\int h(z)\rho_{X,Z}(x,z)\mathrm{d}z}{\rho_X(x)}\frac{\int \nabla_x\rho_{X,Z}(x,z)\mathrm{d}z}{\rho_X(x)} \\ &= \mathbb{E}\left[\nabla_X \log \rho_{X,Z}(X,Z)h(Z)^\top \mid X = x\right] - \mathbb{E}\left[\nabla_X \log \rho_{X,Z}(X,Z) \mid X = x\right]\mathbb{E}\left[h(Z)^\top \mid X = x\right] \\ &= \mathrm{Cov}(\nabla_X \log \rho_{X,Z}(X,Z),\ h(Z) \mid X = x) \\ &= \mathrm{Cov}(\nabla_X \log \rho_{X|Z}(X|Z),\ h(Z) \mid X = x),\end{aligned}$$

If $X = Y + Z$, we have $\rho_{X|Z}(x \mid z) = \rho_{Y|Z}(x - z \mid z)$, and hence

$$\begin{aligned}\nabla\mu_h(x) &= \mathrm{Cov}(\nabla_X \log \rho_{Y|Z}(X - Z \mid Z),\ h(Z) \mid X = x) \\ &= \mathrm{Cov}(\nabla_Y \log \rho_{Y|Z}(Y \mid Z),\ h(Z) \mid X = x).\end{aligned}$$

If $Y \perp\!\!\!\perp Z$, we have $\nabla_Y \log \rho_{Y|Z}(Y \mid Z) = \nabla_Y \log \rho_Y(Y)$ and hence the third result follows. The fourth result follows $\nabla_y \log \rho_Y(y) = \Sigma^{-1}(y - \mu)$ for $Y \sim \texttt{Normal}(\mu, \Sigma)$. □

4.3 Curvature Identities

For the rectified flow $\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t$, it is of interest to estimate the curvature $\ddot{Z}_t = \frac{\mathrm{d}}{\mathrm{d}t}\dot{Z}_t$ of the trajectories. Small curvatures means straighter trajectories and hence it incurs smaller discretization error when solved with numerical algorithms. As we discussed in Section 4.2, the conditional variance $\mathrm{Var}(\dot{X}_t \mid X_t)$, which is related to $\nabla v_t(x)$, provides a measure of how much the trajectories of $\{X_t\}$ intersect, which is is related to the straightness of the $\{Z_t\}$ trajectories. However, $\mathrm{Var}(\dot{X}_t \mid X_t)$ does not exactly equal the curvature $\ddot{Z}_t$.

We provide an explicit formula for the curvature $\ddot{Z}_t$ for rectified flows. Note that the curvature of $\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t$ is

$$\ddot{Z}_t = \frac{\mathrm{d}}{\mathrm{d}t} v_t(Z_t) = \partial_t v_t(Z_t) + \nabla v_t(Z_t)^\top v_t(Z_t).$$

We have the following formula:

Theorem 7. Let $X_t = \alpha_t X_1 + \beta_t X_0$ with $X_0 \perp\!\!\!\perp X_1$ and $X_0 \sim \texttt{Normal}(0, I)$. Let v_t the RF velocity field and ρ_t the density function of X_t. We have

$$\partial_t v_t(X_t) + \nabla v_t^\top v_t(X_t) \quad = \mathbb{E}\left[\ddot{X}_t | X_t\right] - \mathrm{Var}(\dot{X}_t | X_t) \nabla \log \rho_t(X_t) + I,$$

where $\ddot{X}_t = \ddot{\alpha}_t X_1 + \ddot{\beta}_t X_0$, and I is a term involving third order central moment:

$$\begin{aligned} I &= \frac{1}{\beta_t} \mathbb{E}\left[(\dot{X}_t^c)(\dot{X}_t^c)^\top (X_0^c) \mid X_t\right] \\ &= -\frac{1}{\beta_t^2}(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})^{-1} \mathbb{E}\left[(\dot{X}_t^c)(\dot{X}_t^c)^\top (\dot{X}_t^c) \mid X_t\right] \\ &= -\frac{\alpha_t^3}{\beta_t^2}(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})^2 \mathbb{E}\left[(X_1^c)(X_1^c)^\top (X_1^c) \mid X_t\right] \\ &= \beta_t (\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})^2 \mathbb{E}\left[(X_0^c)(\dot{X}_0^c)^\top (\dot{X}_0^c) \mid X_t\right], \end{aligned}$$

where $X_0^c = X_0 - \mathbb{E}[X_0|X_t]$, $X_1^c = X_1 - \mathbb{E}[X_1|X_t]$ and $\dot{X}_t = \dot{X}_t - \mathbb{E}[\dot{X}_t|X_t]$ are the centered random variables.

Proof. It is rewritten from Lemma 17 by noting that $\mathbb{E}[X_0|X_t] = -\beta_t \log \rho_t(X_t)$. □

For the straight interpolation $X_t = tX_1 + (1-t)X_0$, we have $\ddot{X}_t = 0$, and hence

$$\begin{aligned} &\partial_t v_t(X_t) + \nabla v_t^\top v_t(X_t) \\ &= -\mathrm{Var}(\dot{X}_t | X_t) \nabla \log \rho_t(X_t) - \frac{t}{1-t} \mathbb{E}\left[(\dot{X}_t^c)(\dot{X}_t^c)^\top (\dot{X}_t^c) \mid X_t\right]. \end{aligned}$$

Theorem 8. Let $X_t = \alpha_t X_1 + \beta_t X_0$ with $X_0 \perp\!\!\!\perp X_1$ and $X_0 \sim \texttt{Normal}(0, I)$. Let v_t the RF velocity field and ρ_t the density function of X_t.

Then, if $\mathrm{Var}(\dot{X}_t | X_t = x) = 0$ and $\beta_t > 0$, we have

$$\partial_t v_t(x) + \nabla v_t(x)^\top v_t(x) = \mathbb{E}\left[\ddot{X}_t \mid X_t = x\right].$$

Proof. If $\mathrm{Var}(\dot{X}_t | X_t = x)$, then it means $\dot{X}_t$ is deterministic given $X_t = x$, hence $\mathbb{E}\left[(\dot{X}_t^c)(\dot{X}_t^c)^\top (\dot{X}_t^c) \mid X_t = x\right] = 0$ and result follows. □

Remark 21. We observe that curvature involves third-order moments, while $\mathrm{Var}(\dot{X}_t \mid X_t = x)$ is second-order. It is possible to bound third-order moments in terms of second-order moments using sub-Gaussian

inequalities.

Proofs

Lemma 17. Let $X_t = \alpha_t X_1 + \beta_t X_0$ with $X_0 \perp\!\!\!\perp X_1$ and $X_0 \sim \texttt{Normal}(0, I)$. For $\hat{x}_{1|t}(x) = \mathbb{E}\left[X_1 \mid X_t = x\right]$, we have

$$\partial_t \hat{x}_{1|t}(x) = \frac{1}{\beta_t} \mathrm{Cov}(X_1, X_0^\top \dot{X}_t \mid X_t),$$

and

$$\partial_t v_t(X_t) = \mathbb{E}\left[\ddot{X}_t | X_t\right] + \frac{\alpha_t}{\beta_t}\left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right) \mathrm{Cov}(X_1, X_0^\top \dot{X}_t \mid X_t) - \frac{\dot{\beta}_t}{\beta_t} v_t(x),$$

and

$$\begin{aligned}
&\partial_t v_t(X_t) + \nabla v_t^\top v_t(X_t) \\
&= \mathbb{E}\left[\ddot{X}_t | X_t\right] + \frac{\alpha_t}{\beta_t}(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})\mathrm{Cov}(X_1, X_0^\top(\dot{X}_t - \mathbb{E}[\dot{X}_t|X_t]) \mid X_t) \\
&= \mathbb{E}\left[\ddot{X}_t | X_t\right] + \beta_t(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})^2 \mathbb{E}\left[(X_0 - \mathbb{E}\left[X_0|X_t\right])(X_0 - \mathbb{E}[X_0|X_t])^\top X_0 \mid X_t\right] \\
&= \mathbb{E}\left[\ddot{X}_t | X_t\right] + \frac{1}{\beta_t}\mathbb{E}\left[(\dot{X}_t - \mathbb{E}[\dot{X}_t|X_t])(\dot{X}_t - \mathbb{E}[\dot{X}_t|X_t])^\top X_0 \mid X_t\right],
\end{aligned}$$

where $\ddot{X}_t = \ddot{\alpha}_t X_1 + \ddot{\beta}_t X_0$.

Proof. The formula of $\partial_t \hat{x}_{1|t}(x)$ is obtained by applying Lemma 18 with $X_t = X$ and $X_1 = Z$, $t = \theta$. Note that

$$\rho_{X_t, X_1}(x, x_1) \propto \rho_{X_1}(x_1) \exp\left(-\frac{\left\|x - \alpha_t x_1\right\|^2}{2\beta_t^2}\right) \frac{1}{\beta_t^d}.$$

Taking the log, we get

$$\log \rho_{X_t, X_1}(x, x_1) = \log \rho_{X_1}(x_1) - \frac{\left\|x - \alpha_t x_1\right\|^2}{2\beta_t^2} - d \log \beta_t,$$

where d is the dimension of X_t. Taking derivative w.r.t. t,

$$\partial_t \log \rho_{X_t, X_1}(x, x_1) = \frac{(x - \alpha_t x_1)^\top x_1 \dot{\alpha}_t}{\beta_t^2} + \frac{\left\|x - \alpha_t x_1\right\|^2 \dot{\beta}_t}{\beta_t^3} - \frac{d\dot{\beta}_t}{\beta_t}.$$

Using Lemma 18, we have

$$
\begin{aligned}
&\partial_t \hat{x}_{1|t}(X_t) \\
&= \mathrm{Cov}(X_1, \partial_t \log \rho_{X_t, X_1}(X, X_1) \mid X_t) \\
&= \mathrm{Cov}(X_1, \frac{(X_t - \alpha_t X_1)^\top X_1 \dot{\alpha}_t}{\beta_t^2} + \frac{\|X_t - \alpha_t X_1\|^2 \dot{\beta}_t}{\beta_t^3} - \frac{d\dot{\beta}_t}{\beta_t} \mid X_t) \\
&= \mathrm{Cov}(X_1, \frac{X_0^\top X_1 \dot{\alpha}_t}{\beta_t} + \frac{\|X_0\|^2 \dot{\beta}_t}{\beta_t} - \frac{d\dot{\beta}_t}{\beta_t} \mid X_t) \\
&= \frac{\dot{\alpha}_t}{\beta_t} \mathrm{Cov}(X_1, X_0^\top X_1 \mid X_t) + \frac{\dot{\beta}_t}{\beta_t} \mathrm{Cov}(X_1, \|X_0\|^2 \mid X_t) \\
&= \frac{1}{\beta_t} \mathrm{Cov}(X_1, \dot{\alpha}_t X_1 + \dot{\beta}_t X_0)^\top X_0 \mid X_t) \\
&= \frac{1}{\beta_t} \mathrm{Cov}(X_1, \dot{X}_t^\top X_0 \mid X_t).
\end{aligned}
$$

Next, we need to convert the formula to that of $\partial_t v_t$. From $X_t = \alpha_t X_1 + \beta_t X_0$ and $\dot{X}_t = \dot{\alpha}_t X_1 + \dot{\beta}_t X_0$, we have by taking the conditional expectation $\mathbb{E}[\cdot \mid X_t]$:

$$x = \alpha_t \hat{x}_{1|t}(x) + \beta_t \hat{x}_{0|t}(x), \tag{4.11}$$

$$v_t(x) = \dot{\alpha}_t \hat{x}_{1|t}(x) + \dot{\beta}_t \hat{x}_{0|t}(x). \tag{4.12}$$

Taking derivatives of Equation (4.11) w.r.t. time yields $\dot{\alpha}_t \hat{x}_{1|t}(x) + \dot{\beta}_t \hat{x}_{0|t}(x) + \alpha_t \partial_t \hat{x}_{1|t}(x) + \beta_t \partial_t \hat{x}_{0|t}(x) = 0$. Combining it with Equation (4.12) yields

$$v_t(x) + \alpha_t \partial_t \hat{x}_{1|t}(x) + \beta_t \partial_t \hat{x}_{0|t}(x) = 0.$$

Hence,

$$\partial_t \hat{x}_{0|t}(x) = -\frac{1}{\beta_t}(v_t(x) + \alpha_t \partial_t \hat{x}_{1|t}(x)). \tag{4.13}$$

Taking the derivative of Equation (4.12) w.r.t. t:

$$
\begin{aligned}
\partial_t v_t(x) &= \ddot{\alpha}_t \hat{x}_{1|t}(x) + \ddot{\beta}_t \hat{x}_{0|t}(x) + \dot{\alpha}_t \partial_t \hat{x}_{1|t}(x) + \dot{\beta}_t \partial_t \hat{x}_{0|t}(x) \\
&= \mathbb{E}\left[\ddot{X}_t | X_t = x\right] + \dot{\alpha}_t \hat{x}_{1|t}(x) - \frac{\dot{\beta}_t}{\beta_t}(v_t(x) + \alpha_t \partial_t \hat{x}_{1|t}(x)) \qquad \text{//by (4.13)} \\
&= \mathbb{E}\left[\ddot{X}_t | X_t = x\right] + (\dot{\alpha}_t - \frac{\alpha_t \dot{\beta}_t}{\beta_t}) \partial_t \hat{x}_{1|t}(x) - \frac{\dot{\beta}_t}{\beta_t} v_t(x) \\
&= \mathbb{E}\left[\ddot{X}_t | X_t = x\right] + \frac{\alpha_t}{\beta_t}\left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right) \mathrm{Cov}(X_1, \dot{X}_t^\top X_0 \mid X_t) - \frac{\dot{\beta}_t}{\beta_t} v_t(x).
\end{aligned}
$$

One the other hand, we have $\nabla_x v_t(x) = \frac{\dot{\beta}_t}{\beta_t} I - \frac{\alpha_t}{\beta_t}(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}) \mathrm{Cov}(X_0, X_1 \mid X_t = x)$ from Equation (4.9), which gives

$$(\nabla v_t(x))^\top v_t(x) = \frac{\dot{\beta}_t}{\beta_t} v_t(x) - \frac{\alpha_t}{\beta_t}(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}) \mathrm{Cov}(X_1, X_0 \mid X_t = x) v_t(x).$$

Sum them together:

$$\begin{aligned}
&\partial_t v_t(X_t) + \nabla v_t^\top v_t(X_t) \\
&= \mathbb{E}\left[\ddot{X}_t | X_t\right] + \frac{\alpha_t}{\beta_t}(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})(\mathrm{Cov}(X_1, \dot{X}_t^\top X_0 \mid X_t) - \mathrm{Cov}(X_1, X_0 |\, X_t) v_t(X_t)) \\
&= \mathbb{E}\left[\ddot{X}_t | X_t\right] + \frac{\alpha_t}{\beta_t}(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})\mathrm{Cov}(X_1, X_0^\top(\dot{X}_t - \mathbb{E}[\dot{X}_t | X_t]) \mid X_t) \\
&= \mathbb{E}\left[\ddot{X}_t | X_t\right] + \frac{\alpha_t}{\beta_t}(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})\mathbb{E}\left[(X_1 - \mathbb{E}\left[X_1 | X_t\right])(\dot{X}_t - \mathbb{E}[\dot{X}_t | X_t])^\top X_0 \mid X_t\right].
\end{aligned}$$

Note that $X_1 - \mathbb{E}\left[X_1 | X_t\right] = \frac{1}{\alpha_t}(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})^{-1}(\dot{X}_t - \mathbb{E}\left[\dot{X}_t | X_t\right])$. We have

$$\begin{aligned}
&\partial_t v_t(X_t) + \nabla v_t^\top v_t(X_t) \\
&= \mathbb{E}\left[\ddot{X}_t | X_t\right] + \frac{1}{\beta_t}\mathbb{E}\left[(\dot{X}_t - \mathbb{E}[\dot{X}_t | X_t])(\dot{X}_t - \mathbb{E}[\dot{X}_t | X_t])^\top X_0 \mid X_t\right].
\end{aligned}$$

Note that $\dot{X}_t - \mathbb{E}[\dot{X}_t | X_t] = -\beta_t(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})(X_0 - \mathbb{E}[X_0 | X_t])$ and $X_1 - \mathbb{E}[X_1 | X_t] = -\frac{\beta_t}{\alpha_t}(X_0 - \mathbb{E}[X_0 | X_t])$, we get

$$\begin{aligned}
&\partial_t v_t(X_t) + \nabla v_t^\top v_t(X_t) \\
&= \mathbb{E}\left[\ddot{X}_t | X_t\right] + \beta_t(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})^2\mathbb{E}\left[(X_0 - \mathbb{E}\left[X_0 | X_t\right])(X_0 - \mathbb{E}[X_0 | X_t])^\top X_0 \mid X_t\right].
\end{aligned}$$

□

Lemma 18 (Parametric Second Order Score Identities). Let $\mu_h(X) = \mathbb{E}\left[h(Z) \mid X\right]$ where (X, Z) is a joint random variable and h is a function. Assume the density functions $\rho_{X,Z} = \rho^\theta_{X,Z}$ depend on a parameter. Assume all relevant derivatives exist and are continuous.
1) For any (X, Z), we have

$$\nabla_\theta \mu_h(X) = \mathrm{Cov}(\nabla_\theta \log \rho_{X,Z}(X,\ Z),\ h(Z) \mid X).$$

2) If $X = Y + Z$, we have

$$\nabla_\theta \mu_h(X) = \mathrm{Cov}(\nabla_\theta \log \rho_Z(Z) + \nabla_\theta \log \rho_{Y|Z}(Y \mid Z),\ h(Z) \mid X).$$

3) If $X = Y + Z$ and $Y \perp\!\!\!\perp Z$, we have

$$\nabla_\theta \mu_h(X) = \mathrm{Cov}(\nabla_\theta \log \rho_Z(Z) + \nabla_\theta \log \rho_Y(Y),\ h(Z) \mid X).$$

Proof.

$$\mu_h(x) = \frac{\int h(z)\rho_{X,Z}(x, z)\mathrm{d}z}{\rho_X(x)}.$$

Taking derivative

$$\begin{aligned}\nabla_\theta \mu_h(x) &= \frac{\int h(z)\nabla_\theta \rho_{X,Z}(x,z)\mathrm{d}z}{\rho_X(x)} - \frac{\int h(z)\rho_{X,Z}(x,z)\mathrm{d}z}{\rho_X(x)}\frac{\int \nabla_\theta \rho_{X,Z}(x,z)\mathrm{d}z}{\rho_X(x)} \\ &= \mathbb{E}\left[\nabla_\theta \log \rho_{X,Z}(X,Z)h(Z)^\top \mid X = x\right] - \mathbb{E}\left[\nabla_\theta \log \rho_{X,Z}(X,Z) \mid X = x\right]\mathbb{E}\left[h(Z)^\top \mid X = x\right] \\ &= \mathrm{Cov}(\nabla_\theta \log \rho_{X,Z}(X,Z), h(Z) \mid X = x).\end{aligned}$$

If $X = Y + Z$, we have $\rho_{X,Z}(x,\ z) = \rho_Z(z)\rho_{Y|Z}(x - z \mid z)$. Hence

$$\begin{aligned}\nabla_\theta \mu_h(x) &= \mathrm{Cov}(\nabla_\theta \log \rho_Z(Z) + \nabla_\theta \log \rho_{Y|Z}(X - Z \mid Z),\ h(Z) \mid X = x) \\ &= \mathrm{Cov}(\nabla_\theta \log \rho_Z(Z) + \nabla_\theta \log \rho_{Y|Z}(Y \mid Z),\ h(Z) \mid X = x).\end{aligned}$$

If $Y \perp\!\!\!\perp Z$, we have $\nabla_Y \log \rho_{Y|Z}(Y \mid Z) = \nabla_Y \log \rho_Y(Y)$ and hence the third result. □

4.4 Monotonicity of the Euler Updates

As an application of Lemma 13, we show that the update rule of Euler schemes on rectified flow are *monotonic maps* for affine interpolations with independent Gaussian noises. Specifically, assume we solve the rectified flow $\frac{\mathrm{d}}{\mathrm{d}t}Z_t = v_t(Z_t)$ with Euler updates,

$$\hat{\Phi}^{\texttt{Euler}}_{s|t}(z) = z + (s - t)v_t(z),$$

which advances from $\hat{Z}_t$ to $\hat{Z}_s$ following one step of Euler update. Assume we use the straight interpolation $X_t = tX_1 + (1 - t)X_0$, then we show that $\hat{\Phi}^{\texttt{Euler}}_{s|t}(z)$ is a monotonic map for any $0 \le t \le s \le 1$, in the sense that

$$(\hat{\Phi}^{\texttt{Euler}}_{s|t}(z) - \hat{\Phi}^{\texttt{Euler}}_{s|t}(z'))^\top (z - z') \ge 0, \quad \forall z, z'.$$

In fact, $\hat{\Phi}^{\texttt{Euler}}_{s|t}$ is the gradient of a convex function, that is, there exists a convex function $f\colon \mathbb{R}^d \to \mathbb{R}$, such that $\hat{\Phi}^{\texttt{Euler}}_{s|t}(z) = \nabla f(z)$.

Definition 11. A mapping $\Phi\colon \mathbb{R}^d \to \mathbb{R}^d$ is said to be monotonic if

$$(\Phi(x) - \Phi(x'))^\top (x - x') \ge 0, \quad \forall x, x' \in \mathbb{R}^d.$$

It is called strictly monotonic if $(\Phi(x) - \Phi(x'))^\top (x - x') > 0$ unless $x = x'$.

Lemma 19. Assume Φ is continuously differentiable, then it is monotonic iff $\nabla\Phi$ is positive definite (even though it may not be symmetric) in the sense of

$$u^\top \nabla\Phi(x)u \ge 0, \quad \forall x, u \in \mathbb{R}^d. \tag{4.14}$$

It is strictly monotonic if $u^\top \nabla\Phi(x)u > 0$ unless $u = 0$.

Eq. (4.14) is equivalent to the symmetric matrix $\nabla\Phi(x) + \nabla\Phi(x)^\top$ is positive definite in the typical sense.

Proof. Let $u = x - x'$ and $x_t = x + t(x' - x)$ for $t \in [0, 1]$. We have

$$(\Phi(x) - \Phi(x'))^\top (x - x') = \int_0^1 u^\top \nabla\Phi(x_t) u \mathrm{d}t.$$

Hence, if $u^\top \Phi(x_t) u \geq 0$ everywhere, then Φ is monotonic.

For the other direction, if there exists u, z such that $u^\top \Phi(z) u < 0$, we can choose $x = z + \varepsilon u$ and $x' = z - \varepsilon u$ with very small ε, such that $(\Phi(x) - \Phi(x'))^\top (x - x') = \int_0^1 u^\top \nabla\Phi(x_t) u < 0$. □

Theorem 9. Let $\frac{\mathrm{d}}{\mathrm{d}t} Z_t = v_t(Z_t)$ be the rectified flow induced by $X_t = \alpha_t X_1 + \beta_t X_0$, with $X_0 \perp\!\!\!\perp X_1$ and $X_0 \sim \texttt{Normal}(0, I)$, and $\frac{\dot\alpha_t}{\alpha_t} \geq 0 \geq \frac{\dot\beta_t}{\beta_t}$.

Assume we solve $\frac{\mathrm{d}}{\mathrm{d}t} Z_t = v_t(Z_t)$ with Euler updates,

$$\hat\Phi^{\texttt{Euler}}_{s|t}(z) = z + (s - t) v_t(z),$$

which maps $\hat Z_t$ to $\hat Z_s$ following one Euler update.

Then $\hat\Phi^{\texttt{Euler}}_{s|t}$ is a monotonic mapping if $0 \leq (s - t) \leq -\beta_t / \dot\beta_t$. Moreover, there exists a convex function $f(x)$, such that $\hat\Phi^{\texttt{Euler}}_{s|t}(z) = \nabla f(z)$.

Proof. Taking the Jacobian matrix using Lemma 13,

$$\begin{aligned}\nabla\hat\Phi^{\texttt{Euler}}_{s|t}(z) &= I + (s - t)\nabla v_t(z) \\ &= (1 + (s - t)\frac{\dot\beta_t}{\beta_t}) I + (s - t)(\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t})\mathrm{Var}(X_0 \mid X_t = x).\end{aligned}$$

Hence, $\nabla\hat\Phi^{\texttt{Euler}}_{s|t}(z)$ is positive semi-definite if $(s - t)\frac{\dot\beta_t}{\beta_t} + 1 \geq 0$ and $(s - t)(\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t}) \geq 0$. Because $\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t} \geq 0$, the condition reduces to $0 \leq (s - t) \leq -\beta_t / \dot\beta_t$. □

Remark 22. For straight interpolation $X_t = tX_1 + (1 - t)X_0$, we have $-\frac{\beta_t}{\dot\beta_t} = 1 - t$, and hence Theorem 9 holds for all Euler updates with $0 \leq t \leq s \leq 1$.

Remark 23. Unfortunately, the composition of multiple Euler steps is not guaranteed to be monotonic, except in one-dimensional cases. It is because compositions of monotonic maps are not necessarily monotonic in multi-dimensional cases. The fundamental difficulty here is that the product AB of two positive definite matrices A, B is not necessarily positive definite, unless A and B are commutative.

Remark 24. As a result, the exact transport mapping $\Phi_{s|t}(z)$, which can be viewed as composing infinitely many small Euler steps, is not monotonic in general as well. It is well known that $\Phi_{t|s}$ of a general

ODE is orientation-preserving in that $\det(\nabla\Phi_{t|s}) > 0$. In the one-dimensional case of $Z_t \in \mathbb{R}$, this implies that $\Phi_{t|s}$ is a monotonic map. In high-dimensional cases ($Z_t \in \mathbb{R}^d$), orientation-preserving maps are not necessarily monotonic in the typical sense.

Remark 25. This result can be leveraged to avoid calculating the Jacobian matrix in control or inverse problems. Assume we are interested in adjusting z_t to minimize an objective defined on $\hat{z}_1 = \Phi_{1|t}^{\texttt{Euler}}(z_t)$:

$$\min_z \hat{\ell}(z) := \ell(\Phi_{1|t}^{\texttt{Euler}}(z_t)).$$

The exact gradient is

$$\nabla_z \hat{\ell}(z) = \nabla\Phi_{1|t}^{\texttt{Euler}}(z_t)\nabla\ell(\hat{z}_1).$$

However, because $\nabla\Phi_{1|t}^{\texttt{Euler}}(z_t)$ is positive definite, we can use $\nabla\ell(\hat{z}_1)$ in place of $\nabla_z\hat{\ell}(z)$ during optimization, because they have positive inner products: $\nabla_z\ell(\hat{z}_1)^\top\nabla\ell(\hat{z}_1) \geq 0$.

4.5 An Error Bound w.r.t. L2 Optimal Transport

With Theorem 9, we show that the accuracy of the one-step Euler update controls how optimal the rectified coupling (Z_0, Z_1) in terms of sovling the L2 optimal transport problem.

Theorem 10. Let $\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t$ be the rectified flow of $\{X_t\}$ with $X_t = tX_1 + (1-t)X_0$ with $X_0 \perp\!\!\!\perp X_1$ and $X_0 \sim \texttt{Normal}(0, I)$. We have for $0 \leq t \leq s \leq 1$,

$$\mathbb{E}\left[\|Z_s - Z_t\|^2\right]^{1/2} - W_2(\pi_t, \pi_s) \leq 2\mathbb{E}\left[\|Z_s - Z_t - v_t(Z_t)\|^2\right]^{1/2}.$$

In particular, at $t = 0$ and $s = 1$,

$$\mathbb{E}\left[\|Z_1 - Z_0\|^2\right]^{1/2} - W_2(\pi_0, \pi_1) \leq 2\mathbb{E}\left[\|Z_1 - Z_0 - v_0(Z_0)\|^2\right]^{1/2}.$$

Hence, if one step Euler is exact, then the rectified coupling (Z_0, Z_1) is the L2 optimal coupling.

Proof. Let $\hat{\pi}_{s|t}$ the distribution of $\hat{Z}_{s|t} = \Phi_{s|t}^{\texttt{Euler}}(Z_t) = Z_t + (s-t)v_t(Z_t)$. By Theorem 9, the mapping is the gradient of a convex function. By Brenier Theorem, $\Phi_{s|t}^{\texttt{Euler}}$ forms the L2 optimal transport between π_t and $\hat{\pi}_{s|t}$, that is,

$$W_2(\pi_t, \hat{\pi}_{s|t}) = \mathbb{E}\left[\left\|\hat{Z}_{s|t} - Z_t\right\|^2\right]^{1/2}.$$

By triangle inequalities,

$$\begin{aligned}
\mathbb{E}\left[\|Z_s - Z_t\|^2\right]^{1/2} &\le \mathbb{E}\left[\left\|\hat{Z}_{s|t} - Z_t\right\|^2\right]^{1/2} + \mathbb{E}\left[\left\|\hat{Z}_{s|t} - Z_s\right\|^2\right]^{1/2} \\
&\le W_2(\pi_t, \hat{\pi}_{s|t}) + \mathbb{E}\left[\left\|\hat{Z}_{s|t} - Z_s\right\|^2\right]^{1/2} \\
&\le W_2(\pi_t, \pi_s) + W_2(\hat{\pi}_{s|t}, \pi_s) + \mathbb{E}\left[\left\|\hat{Z}_{s|t} - Z_s\right\|^2\right]^{1/2} \\
&\le W_2(\pi_t, \pi_s) + 2\mathbb{E}\left[\left\|\hat{Z}_{s|t} - Z_s\right\|^2\right]^{1/2},
\end{aligned}$$

where we used $W_2(\hat{\pi}_{s|t}, \pi_s) \le \mathbb{E}\left[\left\|\hat{Z}_{s|t} - Z_s\right\|^2\right]^{1/2}$.

□

CHAPTER FIVE

Stochastic Solvers

Given an ODE $\mathrm{d}Z_t = v_t(Z_t)\,\mathrm{d}t$ with ρ_t as the density function of Z_t, it is always possible to convert it into a stochastic differential equation (SDE) by compounding the ODE with the Langevin dynamics of ρ_t:

$$\mathrm{d}\tilde{Z}_t = \underbrace{v_t(\tilde{Z}_t)\,\mathrm{d}t}_{\text{Rectified flow}} + \underbrace{\sigma_t^2 \nabla \log \rho_t(\tilde{Z}_t)\,\mathrm{d}t + \sqrt{2}\,\sigma_t\,\mathrm{d}W_t}_{\text{Langevin dynamics}}, \qquad \tilde{Z}_0 = Z_0, \quad (5.1)$$

where we add a Langevin dynamics component, thereby transforming the deterministic dynamics into a stochastic one. However, since the target ρ_t of the Langevin dynamics is set to match the original marginal distribution of the ODE, the Langevin dynamics remains in equilibrium and does not alter the distribution of the process at each step. This ensures that the SDE and ODE share the same marginal distributions at each step, i.e., $\mathrm{Law}(\tilde{Z}_t) = \mathrm{Law}(Z_t)$, even though the trajectory-wise distributions of $\{Z_t\}$ and $\{\tilde{Z}_t\}$ are clearly different.

Implementing (5.1) of course requires knowing the score function $\nabla \log \rho_t$ in addition to v_t, which is not avaiable generally. However, if the ODE is the rectified flow induced from an affine interpolation of an independent coupling $(X_0 \perp\!\!\!\perp X_1)$ with Gaussian noise X_0, we can express $\nabla \log \rho_t$ using v_t in a closed form with Tweedie's formula as shown in Section 4.1:

$$\nabla \log \rho_t(x) = \frac{\alpha_t v_t(x) - \dot{\alpha}_t x}{\beta_t(\dot{\alpha}_t \beta_t - \alpha_t \dot{\beta}_t)}.$$

This is what allows us to freely switch between ODEs and SDEs during inference after the RF velocity field v_t is learned, without additional retraining.

However, given that ODEs are simpler and faster to solve, what motivates the introduction of diffusion noise? What are the pros and cons of using SDEs instead of ODEs? It turns out that introducing Langevin dynamics provides a *self-correcting mechanism* that helps reduce outliers when the trajectory deviates from ρ_t due to practical errors. On the other hand, since the estimation of $\nabla \log \rho_t$ is itself imperfect, the SDE introduces errors of its own. In particular, a large diffusion coefficient tends to make samples more concentrated — that is, *larger diffusion yields more concentrated samples*. We draw understandings on why this is the case.

5.1 Langevin Dynamics as a Guardrail

One of the problems with rectified flow is that errors may accumulate over time as we solve the ODE $\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t$. These errors can arise from both the approximation error of the model v_t and the numerical discretization error. If the estimate $\hat{Z}_t$ at inference deviates from the true distribution $Z_t \sim \rho_t$, there is no built-in mechanism in the ODE to correct it. Instead, the error may amplify when $\hat{Z}_t$ deviates from the true distribution ρ_t, as the model v_t is estimated less accurately in the low-density regions of ρ_t, which are rarely sampled during training.

It would be ideal if we could introduce a mechanism that dynamically corrects errors or acts as a "guardrail" to keep the estimated distribution $\hat{Z}_t$ close to the true distribution ρ_t when significant deviations occur. One approach is to use the Langevin dynamics of ρ_t to steer $\hat{Z}_t$ towards ρ_t during such deviations. To introduce this concept, some quick background is in order.

Background (Stochastic Differential Equations). A stochastic differential equation (SDE) introduces randomness into an ordinary differential equation (ODE) and is given by

$$\mathrm{d}Y_t = v_t(Y_t)\,\mathrm{d}t + \sigma_t\,\mathrm{d}W_t, \tag{5.2}$$

where v_t represents the deterministic drift, and σ_t denotes the diffusion coefficient, and W_t is a random process that drives the stochasticity. We assume that W_t is a Brownian motion (or a Wiener process), which satisfies $W_s \sim \texttt{Normal}(W_t, s-t)$ for any $s \geq t$ and $W_0 = 0$.

The Euler–Maruyama (or simply Euler) discretization of this SDE with time step ε approximates the solution as

$$Y_{t+\varepsilon} = Y_t + \varepsilon v_t(Y_t) + \sigma_t(W_{t+\varepsilon} - W_t). \tag{5.3}$$

For Brownian motion, the noise increment is $W_{t+\varepsilon} - W_t = \sqrt{\varepsilon}\,\xi_t$, where $\xi_t \sim \mathcal{N}(0,1)$.

Although a full treatment of SDEs is involved, for the purpose of understanding, it is sufficient to know that (5.2) is the limit of (5.3) in a suitable sense as $\varepsilon \to 0^+$. All properties of the SDE can be derived from the discretized form (5.3) by taking the limit as $\varepsilon \to 0^+$.

Background (Fokker–Planck Equation). Let ρ_t be the density function of Y_t in the SDE (5.2) at time t. A major result to know is the Fokker–Planck equation, which governs the evolution of ρ_t:

$$\frac{\partial \rho_t(x)}{\partial t} = -\nabla\cdot\left(\bar{v}_t(x)\rho_t(x)\right), \quad \text{with} \quad \bar{v}_t(x) = v_t(x) - \frac{\sigma_t^2}{2}\nabla \log \rho_t(x).$$

It resembles the continuity equation for ODEs, but introduces a negative gradient term $-\frac{\sigma_t^2}{2}\nabla \log \rho_t(x)$ in the drift due to randomness. This negative gradient term decreases the probability density ρ_t, resulting in a "diffusion" effect that drives particles *away* from regions of high probability in ρ_t.

Background (Langevin Dynamics). For a distribution with a density function ρ^*, its (overdamped) Langevin dynamics is

$$\mathrm{d}Y_t = \sigma_t^2 \nabla \log \rho^*(Y_t) + \sqrt{2}\sigma_t \mathrm{d}W_t,$$

where $\sigma_t > 0$ is the diffusion coefficient. It is expected that the distribution of Y_t converges to that of ρ^* at convergence when $t \to \infty$. To see this quickly, note that the density ρ_t of Y_t at time t satisfies the Fokker-Planck equation

$$\partial_t \rho_t = -\sigma_t^2 \nabla \cdot (\bar{v}_t \rho_t), \quad \text{with} \quad v_t = \nabla \log \rho_t^* - \nabla \log \rho_t.$$

Obviously, ρ^* is an invariant measure of the process, because $\bar{v}_t = 0$ when $\rho_t = \rho^*$.

It is known that Langevin dynamics can be viewed as the gradient flow of the KL divergence $\mathrm{KL}(\rho_t \,||\, \rho^*)$ under the 2-Wasserstein metric: We can interpret $\mathrm{grad}_{\rho_t} \mathrm{KL}(\rho_t \,||\, \rho^*) := \nabla \cdot ((\nabla \log \rho_t^* - \nabla \log \rho_t)\rho_t)$ as the gradient of the KL divergence functional $\mathrm{KL}(\rho_t \,||\, \rho^*)$ w.r.t. ρ_t under the 2-Wasserstein metric, and hence the Langevin dynamics can be viewed as a gradient flow in the space of distributions:

$$\partial_t \rho_t = -\sigma_t^2 \mathrm{grad}_{\rho_t} \mathrm{KL}(\rho_t \,||\, \rho^*).$$

To use Langevin dynamics as a correction mechanism, at each time t, before advancing to the next time point, we can simulate a short segment of Langevin dynamics to correct the distribution of $\hat{Z}_t$ towards ρ_t. Specifically, starting from $Z_{t,0} = \hat{Z}_t$, we simulate the Langevin dynamics associated with ρ_t for a certain amount of time τ:

$$\mathrm{d}Z_{t,\tau} = \sigma_\tau^2 \nabla \log \rho_t(Z_{t,\tau})\, \mathrm{d}\tau + \sqrt{2}\, \sigma_\tau\, \mathrm{d}W_\tau, \qquad \tau \geq 0, \tag{5.4}$$

where τ represents the simulation time of the Langevin dynamics.

At each time t, the SDE starts from an initial point $Z_{t,0}$, and it is expected that $Z_{t,\infty}$ will follow the distribution ρ_t. However, if we have been following the rectified flow, the distribution is already close to ρ_t, so it is not necessary to simulate the Langevin dynamics for too long. Hence, it can be sufficient to perform only one step of Langevin dynamics per rectified flow update. Furthermore, the Langevin and flow updates can be merged into a single step, resulting in the combined SDE system:

$$\mathrm{d}\tilde{Z}_t = \underbrace{v_t(\tilde{Z}_t)\mathrm{d}t}_{\text{Rectified flow}} + \underbrace{\sigma_t^2 \nabla \log \rho_t(\tilde{Z}_t)\mathrm{d}t + \sqrt{2}\sigma_t \mathrm{d}W_t}_{\text{Langevin dynamics}}, \qquad \tilde{Z}_0 = Z_0, \tag{5.5}$$

Here, the rectified flow component is responsible for driving the density function ρ_t forward according to the original plan, while the Langevin component acts as a "negative feedback loop" that corrects distributional drift when it appears, but without introducing any bias when the distributions are already well aligned. This is because, if the simulation has been accurate and $\tilde{Z}_t$ follows the correct distribution ρ_t, the Langevin dynamics for ρ_t remain in equilibrium at each time t, and therefore do not contribute to the evolution of the density.

5.2 The SDEs Preserve Marginals

In the following, we provide a rigorous proof that $\tilde{Z}_t$, which follows the SDE in (5.5), preserves the marginal distribution of the ODE $\mathrm{d}Z_t = v_t(Z_t)\,\mathrm{d}t$ when ρ_t is the density of Z_t. As we see below, the proof requires careful treatment to inductively establish the preservation of the marginal distribution over time, starting from the initialization $\tilde{Z}_0 = Z_0$. This is achieved by using the KL divergence formula in Lemma 2.

Theorem 11. Let ρ_t be the density function of Z_t following ODE $\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t$, whose solution is unique on $t \in [0,1]$. Assume $\nabla \log \rho_t(x)$ and $v_t(x)$ are continuously differentiable.

Then the marginal distributions of $\tilde{Z}_t$ of (5.5) matches that of Z_t from rectified flow $\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t$:

$$\mathrm{Law}(\tilde{Z}_t) = \mathrm{Law}(Z_t), \quad \forall t \in [0,1].$$

Proof. Let $\tilde{\rho}_t$ be the density function of $\tilde{Z}_t$. By Fokker Planck equation,

$$\begin{aligned}
\partial_t \tilde{\rho}_t &= -\nabla\cdot((v_t + \sigma_t^2 \nabla \log \rho_t(x))\tilde{\rho}_t) + \sigma_t^2\, \nabla\cdot(\nabla\tilde{\rho}_t) \\
&= -\nabla\cdot((v_t + \sigma_t^2(\nabla \log \rho_t(x) - \nabla \log \tilde{\rho}_t(x)))\tilde{\rho}_t) \\
&= -\nabla\cdot(\bar{v}_t\tilde{\rho}_t),
\end{aligned}$$

where we define $\bar{v}_t = v_t + \sigma_t^2(\nabla \log \rho_t - \nabla \log \tilde{\rho}_t)$. Therefore, $\tilde{Z}$ shares the same marginal distributions with $\bar{Z}_t$ following the ODE below:

$$\mathrm{d}\bar{Z}_t = \bar{v}_t(\bar{Z}_t)\mathrm{d}t = (v_t(\bar{Z}_t) + \sigma_t^2(\nabla \log \rho_t(\bar{Z}_t) - \nabla \log \tilde{\rho}_t(\bar{Z}_t)))\mathrm{d}t, \quad \bar{Z}_0 = Z_0.$$

Applying Lemma 2 for KL divergence between marginal distributions of $\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t$ and $\mathrm{d}\bar{Z}_t = \bar{v}_t(\bar{Z}_t)\mathrm{d}t$, we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\tilde{\rho}_t \mid\mid \rho_t) &= \mathbb{E}\left[(\nabla \log \tilde{\rho}_t(\bar{Z}_t) - \nabla \log \rho_t(\bar{Z}_t))^\top(\bar{v}_t(\bar{Z}_t) - v_t(\bar{Z}_t))\right] \\
&= -\sigma_t^2 \mathbb{E}\left[\left\|\nabla \log \tilde{\rho}_t(\bar{Z}_t) - \nabla \log \rho_t(\bar{Z}_t)\right\|^2\right] \le 0.
\end{aligned}$$

This suggests that

$$\mathrm{KL}(\tilde{\rho}_t \mid\mid \rho_t) \le \mathrm{KL}(\tilde{\rho}_0 \mid\mid \rho_0) = 0,$$

where we used $\tilde{\rho}_0 = \rho_t$ by initialization. Hence, $\tilde{\rho}_t = \rho_t$ for $t \ge 0$. □

5.3 SDEs with Independent Gaussian X_0

In the case of affine interpolation $X_t = \alpha_t X_1 + \beta_t X_0$ with an independent coupling $(X_0 \perp\!\!\!\perp X_1)$, one can express $\nabla \log \rho_t$ as

$$\nabla \log \rho_t(x) = \beta_t^{-1}\mathbb{E}\left[\nabla \log \rho_0(X_0) \mid X_t = x\right], \tag{5.6}$$

where ρ_0 is the density function of the noise X_0.

Further, if X_0 is a Gaussian distribution, i.e., $X_0 \sim \texttt{Normal}(\mu_0, \Sigma_0)$, then the score function is $\nabla \log \rho_0(x) = -\Sigma_0^{-1}(x - \mu_0)$, which is a linear

function of x. Hence, we can push the conditional expectation in (5.6) into the input side of $\nabla \log \rho_0(\cdot)$, and hence express $\nabla \log \rho_t$ using with expected noise $\hat{x}_{0|t}(x) = \mathbb{E}[X_0|X_t = x]$:

$$\begin{aligned}\nabla \log \rho_t(x) &= \beta_t^{-1}\mathbb{E}[\nabla \log \rho_0(X_0)|X_t = x] \\ &= \beta_t^{-1}\nabla \log \rho_0(\mathbb{E}[X_0|X_t = x]) \qquad //\nabla \log \rho_0 \text{ is linear} \\ &= \beta_t^{-1}\nabla \log \rho_0(\hat{x}_{0|t}(x)).\end{aligned}$$

With this, we can rewrite the SDE as

$$\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t + \frac{\sigma_t^2}{\beta_t}\nabla \log \rho_0(\hat{x}_{0|t}(Z_t))\mathrm{d}t + \sqrt{2}\sigma_t \mathrm{d}W_t.$$

Because β_t converges to zero when $t \to 1$, it is more stable to rewrite it as

$$\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t + \gamma_t \nabla \log \rho_0(\hat{x}_{0|t}(Z_t))\mathrm{d}t + \sqrt{2\beta_t\gamma_t}\mathrm{d}W_t. \tag{5.7}$$

where we take the diffusion variance to be $\sigma_t^2 = \beta_t\gamma_t$. In this case, if γ_t is bounded, then the diffusion variance $\sigma_t^2 = \beta_t\gamma_t$ decays to zero with rate β_t as $t \to 1$.

Another suggestive form is obtained by setting $\sigma_t^2 = \beta_t^2 e_t$, which decays to zero faster with β_t^2:

$$\mathrm{d}Z_t = \underbrace{v_t(Z_t)\mathrm{d}t}_{\text{Rectified flow}} + \beta_t \underbrace{\left(e_t \nabla \log \rho_0(\hat{x}_{0|t}(Z_t))\mathrm{d}t + \sqrt{2e_t}\mathrm{d}W_t\right)}_{\text{Langevin on noise distribution } \rho_0}. \tag{5.8}$$

This form is particularly intuitive, as the term in the brackets can be viewed as Langevin dynamics on the noise distribution ρ_0. Since $Z_t \overset{\text{law}}{=} X_t = \alpha_t X_1 + \beta_t X_0$, the Langevin term on the noise space is scaled by the coefficient β_t to obtain the update for $\mathrm{d}Z_t$.

Algorithm 1 SDE Sampler of Rectified Flow

Inputs: The rectified flow $\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t$ induced by interpolation path $X_t = \alpha_t X_1 + \beta_t X_0$ with an independent coupling $X_0 \perp\!\!\!\perp X_1$ and a Gaussian $X_0 \sim \texttt{Normal}(\mu_0, \Sigma_0)$. A non-negative sequence γ_t controlling the noise magnitude.

Algorithm: Numerically solve the SDE initialized from $Z_0 \sim \pi_0$:

$$\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t - \gamma_t(\hat{x}_{0|t}(Z_t) - \mu_0)\mathrm{d}t + \sqrt{2\beta_t\gamma_t}\Sigma_0^{1/2}\mathrm{d}W_t,$$

where $\hat{x}_{0|t}(x) = (\alpha_t v_t(x) - \dot{\alpha}_t x)/(\dot{\alpha}_t\beta_t - \alpha_t\dot{\beta}_t)$.
In particular, one discretization scheme is

$$\hat{Z}_{t+\varepsilon_t} = \hat{Z}_t + \varepsilon_t(v_t(Z_t) - \gamma_t(\hat{x}_{0|t}(Z_t) - \mu_0)) + \sigma_{\texttt{diff}}(\xi_t - \mu_0),$$

where $\xi_t \sim \pi_0$, and $\sigma_{\texttt{diff}} = \sqrt{2\beta_t\gamma_t}$ (for Euler-Maruyama method), or $\sigma_{\texttt{diff}} = \sqrt{\beta_t^2 - (\beta_t - \gamma_t)^2}$ for the stable variant in Remark 29.
Return the estimated X_1.

Remark 26. To expand on the intuition further, note that each Z_t can be decomposed into

$$Z_t = \alpha_t \hat{X}_{1|t} + \beta_t \hat{X}_{0|t}, \tag{5.9}$$

where $\hat{X}_{1|t} = \hat{x}_{1|t}(Z_t)$ and $\hat{X}_{0|t} = \hat{x}_{0|t}(Z_t)$ are the estimated end points given $X_t = Z_t$ and its slope $\dot{X}_t = v_t(Z_t)$.

In particular, the $\hat{X}_{0|t}$ term is the estimated noise. We can view Langevin dynamics as injecting some fresh noise to $\hat{X}_{0|t}$ without changing its distribution. This is achieved by applying Langevin update w.r.t. the noise distribution ρ_0:

$$\tilde{X}_{0|t} \simeq \hat{X}_{0|t} + e_t \nabla \log \rho_0(\hat{X}_{0|t})\mathrm{d}t + \sqrt{2e_t}\mathrm{d}W_t.$$

We then combine the updated $\hat{X}'_{0|t}$ with $\tilde{X}_{0|t}$ to obtain an updated Z_t:

$$\tilde{Z}_t = \alpha_t \hat{X}_{1|t} + \beta_t \tilde{X}_{0|t}. \tag{5.10}$$

Combining (5.9) and (5.10):

$$\begin{aligned}\tilde{Z}_t &= Z_t + \beta_t(\tilde{X}_{0|t} - \hat{X}_{1|t}) \\ &= Z_t + \beta_t(e_t \nabla \log \rho_0(\hat{X}_{0|t})\mathrm{d}t + \sqrt{2e_t}\mathrm{d}W_t).\end{aligned}$$

This yields (5.8). Note, however, this is purely an intuition explanation, because the posterior expectation $\hat{X}_{0|t} = \mathbb{E}[X_0|X_t]$ does not actually follow ρ_0, except at $t = 0$ when $\hat{X}_{0|t} = X_0$.

Example 12 (SDEs with Standard Gaussian X_0). If $X_0 \sim \texttt{Normal}(0, I)$ is the standard Gaussian noise, we have $\nabla \log \rho_0(x) = -x$, and the SDE (5.7) reduces to

$$\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t - \gamma_t \hat{x}_{0|t}(Z_t)\mathrm{d}t + \sqrt{2\beta_t\gamma_t}\mathrm{d}W_t.$$

We can further convert $\hat{x}_{0|t}(x)$ to $v_t(x)$ via Lemma 9. Note that

$$\hat{x}_{0|t}(x) = \frac{\dot{\alpha}_t x - \alpha_t v_t(x)}{\dot{\alpha}_t \beta_t - \alpha_t \dot{\beta}_t}.$$

We have

$$\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t + \frac{\gamma_t}{\lambda_t}(\alpha_t v_t(Z_t) - \dot{\alpha}_t Z_t)\mathrm{d}t + \sqrt{2\beta_t\gamma_t}\mathrm{d}W_t, \tag{5.11}$$

where $\lambda_t = \dot{\alpha}_t\beta_t - \alpha_t\dot{\beta}_t$. Note that λ_t is bounded away from typical interpolations. We have $\lambda_t = 1$ for straight interpolation.

Example 13 (The SDE of DDPM). The noise schedule γ_t (or equivalently σ_t and e_t) is a parameter that we can choose freely at inference time. A particular choice, which recovers the continuous limit of

DDPM, and the SDEs in Song et al. [2020b], is

$$\gamma_t^{\texttt{DDPM}} = \frac{\lambda_t}{\alpha_t},$$

with which the SDE in (5.11) becomes

$$\mathrm{d}Z_t = 2v_t(Z_t)\mathrm{d}t - \frac{\dot{\alpha}_t}{\alpha_t} Z_t \mathrm{d}t + \sqrt{2\frac{\beta_t \lambda_t}{\alpha_t}} \mathrm{d}W_t.$$

This choice is singular at $t = 0$ because $\alpha_t = 0$ as required by the interpolation process. Hence, an approximation or modification of γ_t is needed in practice.

In fact, the default DDPM uses $\alpha_t = \exp(-\frac{1}{4}a(1-t)^2 - \frac{b}{2}(1-t))$ with $a = 19.9$ and $b = 0.1$, which does not satisfy $\alpha_0 = 0$. Alternatively. it might be better to use an α that validates $\alpha_0 = 0$, but modify the coefficient to $\gamma_t^{\texttt{DDPM}} = \frac{\lambda_t}{\alpha_t + \varepsilon}$ with $\varepsilon > 0$, or something similar.

Example 14 (SDE for Straight Interpolation). When $\beta_t = 1 - \alpha_t$, we have $\lambda_t = \dot{\alpha}_t(1-\alpha_t) + \alpha_t\dot{\alpha}_t = \dot{\alpha}_t$. Equation (5.11) reduces to

$$\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t + \frac{\gamma_t}{\dot{\alpha}_t}\left(\alpha_t v_t(Z_t) - \dot{\alpha}_t Z_t\right)\mathrm{d}t + \sqrt{2(1-\alpha_t)\gamma_t}\mathrm{d}W_t.$$

In particular, when $\alpha_t = t$ and $\beta_t = 1 - t$, we have

$$\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t + \gamma_t\left(tv_t(Z_t) - Z_t\right)\mathrm{d}t + \sqrt{2(1-t)\gamma_t}\mathrm{d}W_t. \tag{5.12}$$

If $\gamma_t = 1$ is constant, this leads to a linearly decaying coefficient $\sigma_t^2 = (1-t)$.

With $\gamma^{\texttt{DDPM}} = 1/t$, we have

$$\mathrm{d}Z_t = 2v_t(Z_t)\mathrm{d}t - \frac{1}{t}Z_t\mathrm{d}t + \sqrt{2\frac{1-t}{t}}\mathrm{d}W_t. \tag{5.13}$$

Example 15 (SDE for Spherical Interpolation). When $\alpha_t^2 + \beta_t^2 = 1$, we have $\dot{\alpha}_t\alpha_t = \dot{\beta}_t\beta_t = 0$, and hence

$$\lambda_t = \dot{\alpha}_t\beta_t - \alpha_t\dot{\beta}_t = \dot{\alpha}_t\beta_t + \frac{\dot{\alpha}_t\alpha_t^2}{\beta_t} = \frac{\dot{\alpha}_t}{\beta_t}.$$

The SDE becomes

$$\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t + \frac{\gamma_t\beta_t}{\dot{\alpha}_t}\left(\alpha_t v_t(Z_t) - \dot{\alpha}_t Z_t\right)\mathrm{d}t + \sqrt{2\beta_t\gamma_t}\mathrm{d}W_t.$$

With $\gamma_t^{\texttt{DDPM}} = \frac{\dot{\alpha}_t}{\alpha_t\beta_t}$, it gives

$$\mathrm{d}Z_t = 2v_t(Z_t)\mathrm{d}t - \frac{\dot{\alpha}_t}{\alpha_t}Z_t\mathrm{d}t + \sqrt{2\frac{\dot{\alpha}_t}{\alpha_t}}\mathrm{d}W_t.$$

If $\alpha_t = \sin(\frac{\pi}{2}t)$ and $\beta_t = \cos(\frac{\pi}{2}t)$, we have $\lambda_t = \frac{\pi}{2}$, and

$$\mathrm{d}Z_t = \left((1 + \tfrac{2\gamma_t}{\pi}\sin(\tfrac{\pi}{2}t))v_t(Z_t) - \gamma_t \cos(\tfrac{\pi}{2}t)Z_t\right)\mathrm{d}t + \sqrt{2\cos(\tfrac{\pi}{2}t)\gamma_t}\mathrm{d}W_t.$$

5.4 Diffusion May Cause Over-Concentration

Although things work out nicely in theory, we need to be careful that the introduced score function $\nabla \log \rho_t(x)$ itself has errors, and it may introduce undesirable effects if we rely on it too much (by using a large σ_t). This is indeed the case in practice. As shown in the figure below, when we increase the noise magnitude σ_t, the generated samples tend to cluster closer to the centers of the Gaussian modes.

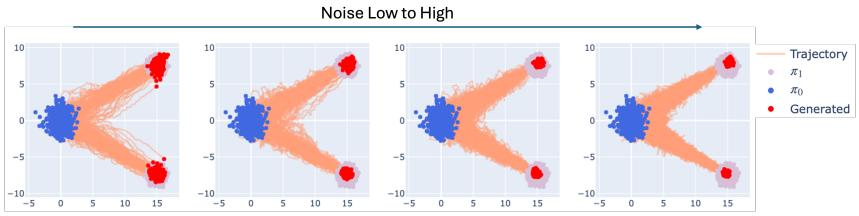


Figure 5.1: Generated samples with varying noise magnitude σ_t.

So, larger diffusion yields more concentrated results? This appears counterintuitive at first glance. Why does this happen?

To see this, assume the estimated velocity field is $\hat{v}_t \approx v_t$. The corresponding estimated score function from Tweedie's formula becomes:

$$\nabla \log \hat{\rho}_t(x) = \frac{1}{\lambda_t \beta_t}\left(\alpha_t \hat{v}_t(x) - \dot{\alpha}_t x\right).$$

Because β_t must converge to 0 as $t \to 1$, the estimated score function $\nabla \log \hat{\rho}_t(x)$ would diverge to infinity in this limit. On the other hand, the true magnitude of $\nabla \log \rho_t(x)$ may be finite, thus being significantly overestimated when t is close to 1. Since $\nabla \log \rho_t(x)$ points toward the centers of mass of clusters, its overestimation leads to an overly concentrated distribution around these centers.

Remark 27 (Role of Noise). In summary, the Langevin guardrail can become too *excessive*, causing over-concentration. It is the score function $\nabla \log \rho_t(x)$ that drives this concentration, rather than the noise itself, as one might initially assume from the ODE vs. SDE dichotomy. The noise component in Langevin dynamics compensates for the concentration induced by the score function, but it does not necessarily prevent it when the score is overestimated.

In the context of text-to-image generation, this over-concentration effect often produces overly smoothed images, which sometimes appear *cartoonish*. Such over-smoothing eliminates fine details and high-frequency variations, resulting in outputs with a blurred appearance.

5.5 Natural Euler Discretization of SDEs

Similar to the natural Euler discretization of the ODE system, it is possible to consider natural variants of Euler(-Maruyama) discretization for the SDE system, when the interpolation is not straight. As expected, the natural Euler discretization is equivalent to the vanilla Euler discretization of the SDE under the straight interpolation with a reparametrization of time and noise schedule.

In the following, we first discuss the idea of natural Euler discretization for SDEs. and then study the equivariance property analogous to natural Euler ODE samplers.

Natural Euler Discretization of SDEs

Recall that the update from $\hat{Z}_t$ to $\hat{Z}_{t+\varepsilon}$ of the natural Euler sampler of the rectified flow $\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t$ based on the affine interpolation $X_t = \alpha_t X_1 + \beta_t X_0$ is

$$\begin{aligned} &\hat{Z}_{t+\varepsilon} = \alpha_{t+\varepsilon}\hat{X}_{1|t} + \beta_{t+\varepsilon}\hat{X}_{0|t}, \\ &\text{with} \quad \hat{X}_{0|t} = \hat{x}_{0|t}(\hat{Z}_t), \quad \hat{X}_{1|t} = \hat{x}_{1|t}(\hat{Z}_t), \end{aligned} \tag{5.14}$$

where $\hat{Z}_{t+\varepsilon}$ is extracted from the interpolation curve that passes through the point $\hat{Z}_t$ with a slope equal to the RF velocity field $v_t(\hat{Z}_t)$. Here ε is the step size, and $\hat{x}_{0|t}$ and $\hat{x}_{1|t}$ are obtained from v_t via formula $\hat{x}_{0|t}(x) = (-\alpha_t v_t(x) + \dot{\alpha}_t x)/\lambda_t$ and $\hat{x}_{1|t}(x) = (\beta_t v_t(x) - \dot{\beta}_t x)/\lambda_t$ with $\lambda_t = \dot{\alpha}_t\beta_t - \alpha_t\dot{\beta}_t$.

To generalize this to the SDE, we need to randomize the update in a certain way. This can be achieved by introducing randomness into the noise prediction $\hat{X}_{0|t}$. Since $\hat{X}_{0|t}$ is the posterior prediction of $X_0 \sim \rho_0$, we may want to continuously "inject" refreshed noises $\xi_t \sim \rho_0$ into $\hat{X}_{0|t}$ to keep it close to the prior distribution ρ_0. We consider the following update

$$\hat{X}_{0|t}^{\texttt{Randomized}} = (1-\vartheta_t)\hat{X}_{0|t} + \sqrt{1-(1-\vartheta_t)^2}\,\xi_t, \qquad \xi_t \sim \rho_0,$$

where $\vartheta_t \in [0,1]$ specifies the proportion of fresh noise we want to inject. This update is justified in the following sense: If ρ_0 is a zero-mean Gaussian, and $\hat{X}_{0|t}$ and ξ_t are independent draws from ρ_0, then $\hat{X}_{0|t}^{\texttt{Randomized}}$ also follows ρ_0.

Now, replacing $\hat{X}_{0|t}$ with $\hat{X}_{0|t}^{\texttt{Randomized}}$ in (5.14) yields the stochastic neural Euler update:

$$\begin{aligned} \hat{Z}_{t+\varepsilon} &= \alpha_{t+\varepsilon}\hat{X}_{1|t} + \beta_{t+\varepsilon}\hat{X}_{0|t}^{\texttt{Randomized}} \\ &= \alpha_{t+\varepsilon}\hat{X}_{1|t} + \beta_{t+\varepsilon}(1-\vartheta_t)\hat{X}_{0|t} + \beta_{t+\varepsilon}\sqrt{1-(1-\vartheta_t)^2}\xi_t. \end{aligned} \tag{5.15}$$

As we show in the sequel, if we take $\vartheta_t = \varepsilon e_t$ to be proportional to the step size ε, then (5.15) can serve as an approximation scheme for the SDE in (5.8) as the step size ε tends to zero.

The SDE Limit

To verify the SDE limit, we need to exam the update (5.15) in the limit of infinitesimal step size ($\varepsilon \to 0$) and show that it approaches to standard Euler-Maruyama discretization of SDE (5.8).

We first rewrite (5.15) into

$$\hat{Z}_{t+\varepsilon} = \hat{Z}^{\texttt{ODE}}_{t+\varepsilon} - \beta_{t+\varepsilon}\vartheta_t \hat{X}_{0|t} + \beta_{t+\varepsilon}\sqrt{1-(1-\vartheta_t)^2}\xi_t.$$

where $\hat{Z}^{\texttt{ODE}}_{t+\varepsilon} = \alpha_{t+\varepsilon}\hat{X}_{1|t} + \beta_t \hat{X}_{0|t}$. Because $\hat{Z}^{\texttt{ODE}}_{t+\varepsilon} = \hat{Z}_t + \varepsilon v_t(\hat{Z}_t) + o(\varepsilon)$, and $1-(1-\vartheta_t)^2 = 2\vartheta_t + o(\varepsilon)$, we have

$$\hat{Z}_{t+\varepsilon} \simeq \hat{Z}_t + \varepsilon v_t(\hat{Z}_t) - \beta_t \vartheta_t \hat{x}_{0|t}(\hat{Z}_t) + \beta_t\sqrt{2\vartheta_t}\xi_t.$$

Taking $\vartheta_t = \varepsilon e_t$, this is the Euler-Maruyama update of the SDE in (5.8):

$$\mathrm{d}Z_t = v_t(Z_t)\mathrm{d}t - \beta_t(e_t \hat{x}_{0|t}(\hat{Z}_t) + \sqrt{2e_t}\mathrm{d}W_t).$$

To be clear, the derivation above is purely heuristic. Since $\hat{X}_{0|t}$ is the posterior mean prediction, it does not follow the prior ρ_0, and hence $\hat{X}^{\texttt{Randomized}}_{0|t}$ does not have the same distribution as $\hat{X}_{0|t}$. The correctness of the scheme only holds when ϑ_t (and the step size) is sufficiently small to ensure convergence to the SDE limit, as we show below.

Example 16 (Connection to DDPM). Using $\alpha_t \hat{X}_{1|t} + \beta_t \hat{X}_{0|t} = \hat{Z}_t$, we can rewrite the update in (5.15) in terms of $\hat{x}_{0|t}(x) = \mathbb{E}[X_0 \mid X_t = x]$:

$$\begin{aligned}\hat{Z}_{t+\varepsilon} &= \alpha_{t+\varepsilon}\frac{\hat{Z}_t - \beta_t \hat{X}_{0|t}}{\alpha_t} + \beta_{t+\varepsilon}(1-\vartheta_t)\hat{X}_{0|t} + \beta_{t+\varepsilon}\sqrt{1-(1-\vartheta_t)^2}\xi_t \\ &= \frac{\alpha_{t+\varepsilon}}{\alpha_t}\hat{Z}_t + \left(\beta_{t+\varepsilon}(1-\vartheta_t) - \frac{\alpha_{t+\varepsilon}\beta_t}{\alpha_t}\right)\hat{x}_{0|t}(\hat{Z}_t) + \beta_{t+\varepsilon}\sqrt{1-(1-\vartheta_t)^2}\xi_t.\end{aligned}$$

Rearranging the terms gives

$$\frac{\hat{Z}_{t+\varepsilon}}{\alpha_{t+\varepsilon}} = \frac{\hat{Z}_t}{\alpha_t} + \left(\frac{\beta_{t+\varepsilon}}{\alpha_{t+\varepsilon}}(1-\vartheta_t) - \frac{\beta_t}{\alpha_t}\right)\hat{x}_{0|t}(\hat{Z}_t) + \frac{\beta_{t+\varepsilon}}{\alpha_{t+\varepsilon}}\sqrt{1-(1-\vartheta_t)^2}\xi_t. \tag{5.16}$$

Taking $\varsigma_t = \beta_{t+\varepsilon}\sqrt{1-(1-\vartheta_t)^2}$, we have $\beta_{t+\varepsilon}(1-\vartheta_t) = \sqrt{\beta_{t+\varepsilon}-\varsigma_t^2}$. Hence,

$$\frac{\hat{Z}_{t+\varepsilon}}{\alpha_{t+\varepsilon}} = \frac{\hat{Z}_t}{\alpha_t} + \left(\frac{\sqrt{\beta^2_{t+\varepsilon}-\varsigma_t^2}}{\alpha_{t+\varepsilon}} - \frac{\beta_t}{\alpha_t}\right)\hat{x}_{0|t}(\hat{Z}_t) + \frac{\varsigma_t}{\alpha_{t+\varepsilon}}\xi_t.$$

In the case of $\alpha_t^2 + \beta_t^2 = 1$, this coincides with the generalized DDPM in Equation 12 of Song et al. [2020a].

Remark 28 (Invariance). Similar to the case of deterministic natural Euler samplers, stochastic natural Euler update in Equation (5.16) with ϑ_t on time grid $\{t_i\}$, is equivalent to the stochastic Euler update on $X'_t = \alpha'_t X_1 + \beta'_t X_0$ on time grid $\{t'_i\}$ with $t_i = \tau(t'_i)$ and noise coefficient $\vartheta'_{t'} = \vartheta_t$, and noise $\xi'_{t'} = \xi_t$.

To see this, note that the stochastic natural Euler method of $X'_t =$

$\alpha'_t X_1 + \beta'_t X_0$ is

$$\frac{\hat{Z}'_{t'+\varepsilon'}}{\alpha'_{t'+\varepsilon'}} = \frac{\hat{Z}'_t}{\alpha'_{t'}} + \left(\frac{\beta'_{t'+\varepsilon'}}{\alpha'_{t'+\varepsilon'}}(1-\vartheta'_{t'}) - \frac{\beta'_{t'}}{\alpha'_{t'}} \right) \hat{x}'_{0|t'}(\hat{Z}'_{t'}) + \frac{\beta'_{t'+\varepsilon'}}{\alpha'_{t'+\varepsilon'}} \sqrt{1-(1-\vartheta'_{t'})^2}\xi'_{t'}.$$

Assume $\hat{Z}'_t = \frac{1}{\omega_{t'}}\hat{Z}_t$ with $t = \tau(t')$. By the invariance properties of Proposition 2, note that all the terms in the update are invariant under the transform, that is, $\frac{\hat{Z}'_{t'}}{\alpha_{t'}} = \frac{\hat{Z}_t}{\alpha_t}$, $\frac{\beta'_{t'}}{\beta_{t'}} = \frac{\beta_t}{\alpha_t}$, $\vartheta'_{t'} = \vartheta_t$, $\hat{x}'_{0|t'}(\hat{Z}'_{t'}) = \hat{x}_{0|t}(\hat{Z}_t)$, $\xi'_{t'} = \xi_t$. Hence the update above is equivalent to Equation (5.16).

Example 17 (Noise Schedule of DDPM). The noise schedule ϑ_t is a parameter that we need to decide at inference time. The following choice, which, in the case of $\alpha_t^2 + \beta_t^2 = 1$, recovers the default configuration of the original DDPM algorithm:

$$\vartheta_t^{\texttt{DDPM}} = 1 - \frac{\alpha_t}{\alpha_{t+\varepsilon}}\frac{\beta_{t+\varepsilon}}{\beta_t}. \tag{5.17}$$

To see this, note that DDPM takes $\varsigma_t^{\texttt{DDPM}} = \frac{\beta_{t+\varepsilon}}{\beta_t}\sqrt{1-\frac{\alpha_t^2}{\alpha_{t+\varepsilon}^2}}$ in the case of $\alpha_t^2 + \beta_t^2 = 1$ (see Section 4.1 of Song et al. [2020a]).

Using $\alpha_t^2 + \beta_t^2 = 1$, we have

$$\begin{aligned}
\frac{1}{\beta_t^2} - \frac{\alpha_t^2}{\alpha_{t+\varepsilon}^2\beta_t^2} &= \frac{\alpha_{t+\varepsilon}^2 - \alpha_t^2}{\alpha_{t+\varepsilon}^2\beta_t^2} \\
&= \frac{\alpha_{t+\varepsilon}^2(\alpha_t^2+\beta_t^2) - \alpha_t^2(\alpha_{t+\varepsilon}^2+\beta_{t+\varepsilon}^2)}{\alpha_{t+\varepsilon}^2\beta_t^2} \\
&= 1 - \frac{\alpha_t^2\beta_{t+\varepsilon}^2}{\alpha_{t+\varepsilon}^2\beta_t^2}.
\end{aligned}$$

Hence, solving $\varsigma_t^{\texttt{DDPM}} = \beta_{t+\varepsilon}\sqrt{1-(1-\vartheta_t^{\texttt{DDPM}})^2}$ yields (5.17).

Remark 29 (Stable Discretization of Langevin Dynamics). Let us consider a generic Lagevin dynamics $\mathrm{d}Z_t = e_t \nabla \log \rho(Z_t)\mathrm{d}t + \sqrt{2e_t}\mathrm{d}W_t$. Time discretization with the standard Euler-Maruyama method yields

$$\hat{Z}_{t+\varepsilon} = \hat{Z}_t + \varepsilon e_t \nabla \log \rho(\hat{Z}_t) + \sqrt{2\varepsilon e_t}\xi_t, \quad \xi_t \sim \texttt{Normal}(0, I), \tag{5.18}$$

where $\varepsilon > 0$ is the step size.

One problem with this scheme is that it causes bias towards increasing variance. To see this, consider the simplest case when $\rho \sim \texttt{Normal}(0, I)$, and hence the update becomes

$$\hat{Z}_{t+\varepsilon} = (1-\varepsilon e_t)\hat{Z}_t + \sqrt{2\varepsilon e_t}\xi_t. \tag{5.19}$$

Calculating the variance yields

$$\mathrm{Var}(\hat{Z}_{t+1}) = (1-\varepsilon e_t)^2 \mathrm{Var}(\hat{Z}_t) + 2\varepsilon e_t.$$

If $\mathrm{Var}(\hat{Z}_t) = 1$, then $\mathrm{Var}(\hat{Z}_{t+1}) = (1-\varepsilon e_t)^2 + 2\varepsilon e_t = 1 + \varepsilon^2 e_t^2$, which is larger than 1. In fact, the invariance distribution of (5.19) is

$$\texttt{Normal}(0, \sigma_{\texttt{ex}}^2), \quad \text{with} \quad \sigma_{\texttt{ex}}^2 = 2\varepsilon e_t/(2\varepsilon e_t - \varepsilon^2 e_t^2) > 1, \quad \text{if } \varepsilon < 1/(2e_t).$$

Further, if $\varepsilon e_t > 1/2$, the update divergence and exists no invariant distribution.

To correct this bias and instability, we can modify the discretization in (5.18) into

$$\hat{Z}_{t+\varepsilon} = \hat{Z}_t + \varepsilon e_t \nabla \log \rho(\hat{Z}_t) + \sqrt{1-(1-\varepsilon e_t)^2}\xi_t. \tag{5.20}$$

If $\rho \sim \texttt{Normal}(0, I)$, it reduces to

$$\hat{Z}_{t+\varepsilon} = (1-\varepsilon e_t)\hat{Z}_t + \sqrt{1-(1-\varepsilon e_t)^2}\xi_t.$$

This can be viewed as "taking out" εe_t fraction of $\hat{Z}_t$ and replace it with a "refresh noise" ξ_t with a proper magnitude to ensure the correct variance. One can easily verify that it converges to the correct invariant distribution $\rho \sim \texttt{Normal}(0, I)$ for any $\varepsilon > 0$. In addition, (5.20) approaches to (5.18) as $\varepsilon \to 0$, because by Taylor approximation:

$$1-(1-\varepsilon e_t)^2 = 2\varepsilon e_t + \mathcal{O}(\varepsilon^2).$$

CHAPTER SIX

Reward Tilting

Assume we have trained a rectified flow $\mathrm{d}Z_t = v_t(X_t)\mathrm{d}t$ from an interpolation process $\{X_t\}$ induced from a data distribution $X_1 \sim \pi_1$. Assume that we are given a non-negative reward function $r(x)$ at the inference time, which defines a tilted data distribution π_1^r:

$$\pi_1^r(x) = \frac{\pi_1(x)r(x)}{A_r}, \qquad A_r = \int \pi_1(x)r(x)\mathrm{d}x.$$

Here $\pi_1^r(x)$ is obtained by weighting the density $\pi_1(x)$ by $r(x)$ and then normalize. The question is how to modify the original rectified flow to sample from π_1^r, preferably without retraining the model.

Ideally, we would like to obtain the rectified flow trained on data drawn from π^r, following the same rectified flow training procedure. One approach is to first induce an tilted interpolation process and study its induced rectified flow. Specifically, let $\{X_t^r\}$ be the reward-tilted variant of the original interpolation process $\{X_t\}$, obtained by reweighing the probability of each trajectory of $\{X_t\}$ with the reward $r(X_1)$ evaluated on the terminal state X_1. Specifically, let $\mathbb{P}$ be the probability measure of the original process $\{X_t\}$, then $\{X_t^r\}$ is the process with the law $\mathbb{P}^r$ satisfying

$$\frac{\mathrm{d}\mathbb{P}^r(\{x_t\})}{\mathrm{d}\mathbb{P}(\{x_t\})} = \frac{r(x_1)}{A_r}.$$

The goal is to study the properties of the rectified flow $\mathrm{d}Z_t^r = v_t^r(Z_t^r)\,\mathrm{d}t$ induced by the tilted process $\{X_t^r\}$ and to understand how it is related to and constructed from the rectified flow of the original process $\{X_t\}$.

6.1 General Case

Theorem 12. Let $\mathrm{d}Z^r = v_t^r(Z_t^r)\mathrm{d}t$ be the rectified flow induced by the tilted process $\{X_t^r\}$. We have

Marginal Distribution

The shared marginal distribution ρ_t^r of X_t^r and Z_t^r satisfies

$$\rho_t^r(x) = \rho_t(x)\frac{\mathbb{E}\left[r(X_1) \mid X_t = x\right]}{\mathbb{E}\left[r(X_1)\right]}, \tag{6.1}$$

and hence

$$\nabla \log \rho_t^r(x) = \nabla \log \rho_t(x) + \nabla \log \mathbb{E}\left[r(X_1) \mid X_t = x\right]. \tag{6.2}$$

Initial Distribution

In particular, the initial distribution of Z_t^r is

$$\rho_0^r(x) = \rho_0(x) \frac{\mathbb{E}\left[r(X_1) \mid X_0\right]}{\mathbb{E}\left[r(X_1)\right]}.$$

If (X_0, X_1) is the independent coupling of ρ_0 and ρ_1, we have $\rho_0^r = \rho_0$, that is, the tilting does not modify the initial distribution.

Transition Probability

For any $s, t \in [0,1]$, the density of X_t^r given X_t^r satisfies

$$\rho_{X_s^r | X_t^r}(x_s \mid x_t) = \rho_{X_s | X_t}(x_s \mid x_t) \frac{\mathbb{E}\left[r(X_1) \mid X_s = x_s,\ X_t = x_t\right]}{\mathbb{E}\left[r(X_1) \mid X_t = x_t\right]}.$$

In particular, if X_t is a Markov process and $s \geq t$, we have

$$\rho_{X_s^r | X_t^r}(x_s \mid x_t) = \rho_{X_s | X_t}(x_s \mid x_t) \frac{\mathbb{E}\left[r(X_1) \mid X_s = x_s\right]}{\mathbb{E}\left[r(X_1) \mid X_t = x_t\right]}.$$

Velocity Field

The velocity field $v_t^r(x) := \mathbb{E}[\dot{X}_t^r \mid X_t = x]$ of X_t^r is

$$v_t^r(x) = \frac{\mathbb{E}\left[r(X_1)\dot{X}_t \mid X_t = x\right]}{\mathbb{E}\left[r(X_1) \mid X_t = x\right]}. \tag{6.3}$$

Proof. 1) Let $\rho_{X_s | X_t}(x_s \mid x_t)$ be the density function of X_s given X_t, and $\rho_{X_s^r | X_t^r}$ that of X_s^r given X_t^r. By the definition of tilting, we have

$$\begin{aligned}
&\rho_{X_1^r}(x_1) = \frac{\rho_{X_1}(x_1) r(x_1)}{\mathbb{E}\left[r(X_1)\right]}, \\
&\rho_{X_t^r | X_1^r}(x_t \mid x_1) = \rho_{X_t | X_1}(x_t \mid x_1), \quad \forall t \in [0,1], \\
&\rho_{X_s^r | X_t^r, X_1^r}(x_s, \mid x_t, x_1) = \rho_{X_s | X_t, X_1}(x_s, \mid x_t, x_1) \quad \forall t, s \in [0,1],
\end{aligned} \tag{6.4}$$

where the last two equation holds because all probabilities condi-

tioned on $X_t = X_t^r = x_t$ are the same for $\{X_t\}$ and $\{X_t\}^r$. Hence

$$\begin{aligned}
\rho_t^r(x_t) &= \int \rho^r_{X_t^r \mid X_1^r}(x_t \mid x_1)\rho_{X_1^r}(x_1)\mathrm{d}x_1 \\
&= \frac{1}{\mathbb{E}\left[r(X_1)\right]} \int \rho_{X_t|X_1}(x_t \mid x_1)\rho_{X_1}(x_1)r(x_1)\mathrm{d}x_1 \\
&= \frac{1}{\mathbb{E}\left[r(X_1)\right]} \int \rho_{X_1|X_t}(x_1 \mid x_t)\rho_{X_t}(x_t)r(x_1)\mathrm{d}x_1 \qquad \text{//Bayes rule} \\
&= \rho_{X_t}(x_t)\frac{\int \rho_{X_1|X_t}(x_1 \mid x_t)r(x_1)\mathrm{d}x_1}{\mathbb{E}\left[r(X_1)\right]} \\
&= \rho_{X_t}(x_t)\frac{\mathbb{E}\left[r(X_1) \mid X_t = x_t\right]}{\mathbb{E}\left[r(X_1)\right]}.
\end{aligned}$$

2) For the transition probabilities, let us start with the case when $s = 1$:

$$\begin{aligned}
\rho_{X_1^r|X_t^r}(x_1 \mid x_t) &= \frac{\rho_1^r(x_1)\rho_{X_t^r|X_1^r}(x_t \mid x_1)}{\rho_t^r(x_t)} \qquad \text{//Bayes rule } p_{X|Y}p_Y = p_X p_{Y|X} \\
&= \frac{r(x_1)}{\mathbb{E}\left[X_1|X_t = x_t\right]}\frac{\rho_1(x_1)\rho_{X_t|X_1}(x_t \mid x_1)}{\rho_t(x_t)} \qquad \text{//By (6.4) and (6.1)} \\
&= \frac{r(x_1)}{\mathbb{E}\left[r(X_1) \mid X_t = x_t\right]}\rho_{X_1|X_t}(x_1 \mid x_t).
\end{aligned}$$

For the more general case,

$$\begin{aligned}
\rho_{X_s^r|X_t^r}(x_s \mid x_t) &= \int \rho_{X_s^r|X_t^r,X_1^r}(x_s \mid x_t, x_1)\rho_{X_1^r|X_t^r}(x_1 \mid x_t)\mathrm{d}x_1 \\
&= \int \rho_{X_s|X_t,X_1}(x_s \mid x_t, x_1)\rho_{X_1|X_t}(x_1 \mid x_t)\frac{r(x_1)}{\mathbb{E}\left[r(X_1) \mid X_t = x_t\right]}\mathrm{d}x_1 \\
&= \int \rho_{X_1|X_s,X_t}(x_1 \mid x_s,\ x_t)\rho_{X_s|X_t}(x_s \mid x_t)\frac{r(x_1)}{\mathbb{E}\left[r(X_1) \mid X_t = x_t\right]}\mathrm{d}x_1 \\
&= \frac{\rho_{X_s|X_t}(x_s \mid x_t)}{\mathbb{E}\left[r(X_1) \mid X_t = x_t\right]} \int \rho_{X_1|X_s,X_t}(x_1 \mid x_s,\ x_t)r(x_1)\mathrm{d}x_1 \\
&= \rho_{X_s|X_t}(x_s \mid x_t)\frac{\mathbb{E}\left[r(X_1) \mid X_s = x_s, X_t = x_t\right]}{\mathbb{E}\left[r(X_1) \mid X_t = x_t\right]}.
\end{aligned}$$

2) The derivation is similar by noting that $\rho_{\dot{X}_t^r|X_t^r,X_1^r} = \rho_{\dot{X}_t|X_t,X_1}$:

$$\begin{aligned}
v_t^r(x_t) &= \mathbb{E}\left[\dot{X}_t^r \mid X_t^r\right] \\
&= \int v_t \rho_{\dot{X}_t^r|X_t^r,X_1^r}(v_t \mid x_t, x_1)\rho_{X_1^r|X_t^r}(x_1 \mid x_t)\mathrm{d}v_t\mathrm{d}x_1 \\
&= \int v_t \rho_{\dot{X}_t|X_t,X_1}(v_t \mid x_t, x_1)\frac{\rho_{X_1|X_t}(x_1 \mid x_t)r(x_1)}{\mathbb{E}\left[r(X_1) \mid X_t = x_t\right]}\mathrm{d}v_t\mathrm{d}x_1 \\
&= \frac{\int v_t r(x_1)\rho_{\dot{X}_t \mid X_t,X_1}(v_t \mid x_t, x_1)\mathrm{d}v_t\mathrm{d}x_1}{\mathbb{E}\left[r(X_1) \mid X_t = x_t\right]} \\
&= \frac{\mathbb{E}\left[r(X_1)\dot{X}_t \mid X_t = x_t\right]}{\mathbb{E}\left[r(X_1) \mid X_t = x_t\right]}.
\end{aligned}$$

□

From $v_t^r(x) = \frac{\mathbb{E}\left[r(X_1)\dot{X}_t \mid X_t = x_t\right]}{\mathbb{E}[r(X_1) \mid X_t = x]}$ in Eq 6.3, the velocity field of the tilted process is determined by the r-weighted expectation of $\dot{X}_t$. This does not seem to provide much insight for post-training tilting, because it is not available unless it is explicitly trained to do so. Assume we have we have a set of rewards $\{r_\vartheta \colon \vartheta \in \Theta\}$ parameterized by ϑ during training, we can certainly learn a meta network

$$v_t(x;\ \vartheta) = \frac{\mathbb{E}\left[r_\vartheta(X_1)\dot{X}_t \mid X_t = x_t\right]}{\mathbb{E}\left[r_\vartheta(X_1) \mid X_t = x\right]}.$$

In the canonical case of $X_t = \alpha_t X_1 + \beta_t X_0$ with $X_0 \perp\!\!\!\perp X_1$ and $X_0 \sim \texttt{Normal}(0, I)$. We have

$$v_t^r(x) = v_t(x) + \beta_t^2 \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right) \nabla \log \mathbb{E}\left[r(X_1) \mid X_t = x\right].$$

Assume X_1 is close to deterministic conditioned on X_t, we have $\mathbb{E}\left[r(X_1)|X_t = x\right] \approx r(\mathbb{E}\left[X_1 \mid X_t = x\right]) = r\left(\frac{\dot{\beta}_t x - \beta_t v_t(x)}{\dot{\beta}_t \alpha_t - \beta_t \dot{\alpha}_t}\right)$. Hence,

$$\nabla_x \log \mathbb{E}\left[r(X_1)|X_t = x\right] \approx \frac{\dot{\beta}_t I - \beta_t \nabla v_t(x)}{\dot{\beta}_t \alpha_t - \beta_t \dot{\alpha}_t} \nabla \log r\left(\frac{\dot{\beta}_t x - \beta_t v_t(x)}{\dot{\beta}_t \alpha_t - \beta_t \dot{\alpha}_t}\right).$$

This yields

$$v_t^r(x) \approx v_t(x) + \frac{\beta_t}{\alpha_t}(\dot{\beta}_t I - \beta_t \nabla v_t(x)) \nabla \log r\left(\frac{\dot{\beta}_t x - \beta_t v_t(x)}{\dot{\beta}_t \alpha_t - \beta_t \dot{\alpha}_t}\right).$$

If we use this in practice, we need to handle the singularity at $t = 0$ due to the $1/\alpha_t$ term.

6.2 Training-Free Gaussian Tilting

Let v_t, ρ_t the velocity and density of the rectified flow induced from $X_t = \alpha_t X_1 + \beta_t X_0$ with $(X_0, X_1) \sim \pi_0 \times \pi_1$. Let v_t^r, ρ_t^r corresponding to $(X_0, X_1) \sim \pi_0 \times \pi_1^r$, where

$$\pi_1^r(x) = \frac{\pi_1(x) r(x)}{Z_r}, \qquad Z_r = \int \pi_1(x) r(x) \mathrm{d}x.$$

In this case, we can express the rectified flow of the titled process can be explicitly expressed using the original rectified flow. It turns out it is easy to state the relation in terms of the expected target $\hat{x}_{1|t}(x) = \mathbb{E}\left[X_1|X_t = x\right]$, which can be then converted to that of the velocity field v_T.

Lemma 20. Consider the Gaussian reward function $r(x) = \exp(-\eta \frac{\|x - x^*\|^2}{2})$ for a reference point $x^* \in \mathbb{R}^d$ and magnitude $\eta \in \mathbb{R}$. Define $\hat{x}_{1|t}(x) = \mathbb{E}\left[X_1 \mid X_t = x\right]$ and $\hat{x}_{1|t}^r(x) = \mathbb{E}\left[X_1 \mid X_t = x\right]$. Then,

$$\hat{x}_{1|t}^r(x) = \hat{x}_{1|\tilde{t}}(\tilde{x}),$$

where $\tilde{t}$ and $\tilde{x} \in \mathbb{R}^d$ are defined by (x, t) via

$$\frac{\alpha_s^2}{\beta_s^2} = \frac{\alpha_t^2}{\beta_t^2} + \eta, \qquad x = \frac{\beta_s^2}{\alpha_s}(\frac{\alpha_t}{\beta_t^2}x + \eta x^*).$$

If α_t/β_t is monotonically increasing w.r.t. t, then the solution of $\tilde{t}$ is unique and lies on $[t, 1]$ if $\eta \geq 0$, and lies in $[0, t]$ if $0 \geq \eta \geq -\frac{\alpha_t^2}{\beta_t^2}$.

Proof.

$$\begin{aligned}
\hat{x}_{1|\tilde{t}}(\tilde{x}) &= \mathbb{E}\left[\dot{X}_{\tilde{t}} \mid X_{\tilde{t}} = \tilde{x}\right] \\
&= \frac{\int \pi_1(x_1) \exp(-\frac{\|y-\alpha_s x_1\|^2}{2\beta_s^2}) x_1 \mathrm{d}x_1}{\int \pi_1(x_1) \exp(-\frac{\|y-\alpha_s x_1\|^2}{2\beta_s^2}) \mathrm{d}x_1} \\
&= \frac{\int \pi_1(x_1) r(x_1) \exp(-\frac{\|x-\alpha_s x_1\|^2}{2\beta_t^2}) x_1 \mathrm{d}x_1}{\int \pi_1(x_1) r(x_1) \exp(-\frac{\|x-\alpha_t x_1\|^2}{2\beta_t^2}) \mathrm{d}x_1} \qquad \text{//Lemma 22.} \\
&= \hat{x}^r_{1|t}(x).
\end{aligned}$$

□

By Lemma 21, we can convert the formula of $\hat{x}^r_{1|t}(x)$ to get the tilted velocity field $v^r_t(x)$.

Proposition 5. Under the condition above, we have

$$\begin{aligned}
v^r_t(x) &= \frac{\dot{\beta}_t}{\beta_t}x + \alpha_t(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})\hat{x}_{1|s}(y) \\
&= \frac{\dot{\beta}_t}{\beta_t}x + \frac{\alpha_t \kappa_t}{\alpha_s \kappa_s}(v_s(y) - \frac{\dot{\beta}_s}{\beta_s}y),
\end{aligned}$$

where $\kappa_t = \frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}$.

Proof. By Lemma 21,

$$\begin{aligned}
v^r_t(x) &= \frac{\dot{\beta}_t}{\beta_t}x + \alpha_t(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})\hat{x}^r_{1|t}(x) \\
&= \frac{\dot{\beta}_t}{\beta_t}x + \alpha_t \kappa_t \hat{x}_{1|s}(y)
\end{aligned}$$

Also note that $v_s(y) = \frac{\dot{\beta}_s}{\beta_s}y + \alpha_s \kappa_s \hat{x}_{1|s}(y)$. Hence,

$$v^r_t(x) = \frac{\dot{\beta}_t}{\beta_t}x + \frac{\alpha_t \kappa_t}{\alpha_s \kappa_s}(v_s(y) - \frac{\dot{\beta}_s}{\beta_s}y).$$

□

Example 18. With the straight interpolation $\alpha_t = t$, $\beta_t = 1 - t$, the

Algorithm 2 Gaussian Tilting for Rectified Flow

Input: 1) A pretrained rectified flow v_t induced by $X_t = \alpha_t X_1 + \beta_t X_0$ with $X_0 \perp\!\!\!\perp X_1$ and $X_0 \sim \texttt{Normal}(0, I)$, and $X_1 \sim \pi_1$.
2) A Gaussian reward $r(x) = \exp(-\eta \|x - x^*\|^2)$ for a target $x^* \in \mathbb{R}^d$ and magnitude $\eta \in \mathbb{R}$.
Output: A sample from the tilted distribution $\pi_1^r(x) \propto \pi_1(x) r(x)$.
Algorithm: Get the tilted velocity field v_t^r using the procedure below, and solve the rectified flow with v_t^r initialized from π_0 using any RF samplers.

//The case of straight interpolation ($\alpha_t = t, \beta_t = 1 - t$):
Define $v_t^r(x) =$`get_tilted_velocity_straight`(x, t):

0) If $\eta < -t^2/(1-t)^2$: Return $v_t^r(x) = v_t(x)$.

1) Get the tilted time s and position y:

$$s = \frac{\sqrt{t^2 + \eta(1-t)^2}}{1 - t + \sqrt{t^2 + \eta(1-t)^2}}, \qquad y = \frac{t(1-s)^2}{s(1-t)^2} x + \frac{(1-s)^2}{s} \eta x^*.$$

2) Predict the target position $\hat{x}_{1|t}$ from (s, y):

$$\hat{x}_{1|t} = y + (1 - s) v_s(y),$$

3) Get the tilted velocity field at (x, t):

$$v_t^r(x) = \frac{\hat{x}_{1|t} - x}{1 - t}.$$

Return $v_t^r(x)$.

//For general α_t, β_t:
Define $v_t^r(x) =$`get_tilted_velocity_affine`(x, t):

0) If $\eta < -\alpha_t^2/\beta_t^2$: Return $v_t^r(x) = v_t(x)$.

1) Get the tilted time s and position y by solving

$$\frac{\alpha_s^2}{\beta_s^2} = \frac{\alpha_t^2}{\beta_t^2} + \eta, \qquad y = \frac{\beta_s^2}{\alpha_s}\left(\frac{\alpha_t}{\beta_t^2} x + \eta x^*\right).$$

2) Predict the target position $\hat{x}_{1|t}$ from (s, y):

$$\hat{x}_{1|t} = \frac{1}{\alpha_s} \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t} \right)^{-1} \left(v_s(y) - \frac{\dot{\beta}_s}{\beta_s} y \right)$$

3) Get the tilted velocity field at (x, t):

$$v_t^r(x) = \frac{\dot{\beta}_t}{\beta_t} x + \alpha_t \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t} \right) \hat{x}_{1|t}.$$

Return $v_t^r(x)$.

tilted time and positions are

$$s = \frac{\sqrt{t^2 + \eta(1-t)^2}}{1 - t + \sqrt{t^2 + \eta(1-t)^2}},$$

and

$$\begin{aligned} y &= \frac{t(1-s)^2}{s(1-t)^2}x + \frac{(1-s)^2}{s}\eta x^* \\ &= (1 - t + \sqrt{t^2 + \eta(1-t)^2}) \left(\frac{t}{\sqrt{t^2 + \eta(1-t)^2}}x + \eta \frac{(1-t)^2}{\sqrt{t^2 + \eta(1-t)^2}}x^* \right) \\ &= \frac{t}{s}x + \frac{\eta(1-t)^2}{s}x^*. \end{aligned}$$

The resulting tilted velocity field is

$$v_t^r(x) = \frac{1-s}{1-t}v_s(y) + \frac{1}{1-t}(y - x).$$

In particular, at $t = 0$, $s = \frac{\sqrt{\eta}}{1+\sqrt{\eta}}$, $y = \frac{1}{(1+\sqrt{\eta})\sqrt{\eta}}x^*$, which advances in time, and move towards x^*. When $t \approx 1$, we have $s \approx 1$ and $y \approx x$. It is because no tilting is needed in the final stage, if proper tilting is made during the process.

Remark 30. In practice, we can use a negative η, in which case the tilting corresponds to introducing a repulsive force against the target x^*. Although using $\eta < 0$ is not entirely theoretically valid, as it leads to a singularity near $t = 0$, we can mitigate this issue by applying a threshold within the singularity regime. This approach still yields a practical procedure that introduces repulsiveness against x^*. See Algorithm 2.

Lemma 21. For any X_t satisfying $X_t = \alpha_t X_1 + \beta_t X_0$, let

$$v_t(x) = \mathbb{E}\left[\dot{X}_t \mid X_t = x\right], \qquad \hat{x}_{1|t}(x) = \mathbb{E}\left[X_1 \mid X_t = x\right].$$

Then

$$v_t(x) = \frac{\dot{\beta}_t}{\beta_t}x + \alpha_t(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})\hat{x}_{1|t}(x).$$

Proof.

$$\begin{aligned} v_t(x) &= \mathbb{E}\left[\dot{X}_t \mid X_t = x\right] \\ &= \mathbb{E}\left[\dot{\alpha}_t X_1 + \dot{\beta}_t X_0 \mid X_t = x\right] \\ &= \mathbb{E}\left[\dot{\alpha}_t X_1 + \dot{\beta}_t \frac{(X_t - \alpha_t X_1)}{\beta_t} \mid X_t = x\right] \\ &= \frac{\dot{\beta}_t}{\beta_t}x + \alpha_t(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t})\hat{x}_{1|t}(x). \end{aligned}$$

□

Lemma 22. For $t \in [0,1]$, $\eta \in \mathbb{R}$, and $x, x^* \in \mathbb{R}$, let s and y solve

$$\frac{\alpha_s^2}{\beta_s^2} = \frac{\alpha_t^2}{\beta_t^2} + \eta, \qquad y = \frac{\beta_s^2}{\alpha_s}\left(\frac{\alpha_t}{\beta_t^2}x + \eta x^*\right).$$

Then for any $x_1 \in \mathbb{R}^d$,

$$-\frac{\|x - \alpha_t x_1\|^2}{2\beta_t^2} - \eta\frac{\|x_1 - x^*\|^2}{2} = -\frac{\|y - \alpha_s x_1\|^2}{2\beta_s^2} + const,$$

where $const = \frac{\|y\|^2}{2\beta_s^2} - \frac{\|x\|^2}{2\beta_t^2} - \frac{\eta}{2}\|x^*\|^2$ is a constant w.r.t. x_1.

Proof. We have

$$\frac{\|x - \alpha_t x_1\|^2}{\beta_t^2} + \eta\|x_1 - x^*\|^2 = \left(\frac{\alpha_t^2}{\beta_t^2} + \eta\right)\|x_1\|^2 - 2x_1^\top\left(\frac{\alpha_t}{\beta_t^2}x + \eta x^*\right) + \frac{\|x\|^2}{\beta_t^2} + \eta\|x^*\|^2.$$

Set s and y such that

$$\frac{\alpha_t^2}{\beta_t^2} + \eta = \frac{\alpha_s^2}{\beta_s^2}, \qquad \frac{\alpha_t}{\beta_t^2}x + \eta x^* = \frac{\alpha_s}{\beta_s^2}y.$$

Then we have

$$\begin{aligned}\frac{\|x - \alpha_t x_1\|^2}{\beta_t^2} + \eta\|x_1 - x^*\|^2 &= \frac{\alpha_s^2}{\beta_s^2}\|x_1\|^2 - 2\frac{\alpha_s}{\beta_s^2}y^\top x_1 + \frac{\|y\|^2}{\beta_s^2} - \frac{\|y\|^2}{\beta_s^2} + \frac{\|x\|^2}{\beta_t^2} + \eta\|x^*\|^2 \\ &= \frac{\|y - \alpha_s x_1\|^2}{\beta_s^2} - \frac{\|y\|^2}{\beta_s^2} + \frac{\|x\|^2}{\beta_t^2} + \eta\|x^*\|^2.\end{aligned}$$

□

Bibliography

Scipy rk45 function. `https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.RK45.html`. Accessed: 2022-08-19.

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.

Luigi Ambrosio and Gianluca Crippa. Existence, uniqueness, stability and differentiability properties of the flow associated to weakly differentiable vector fields. In *Transport equations and multi-D hyperbolic conservation laws*, pages 3–57. Springer, 2008.

Luigi Ambrosio, Elia Brué, and Daniele Semola. *Lectures on optimal transport*. Springer, 2021.

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

Abdul Fatir Ansari, Ming Liang Ang, and Harold Soh. Refining deep generative models via discriminator gradient flow. In *International Conference on Learning Representations*, 2020.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, Joydeep Ghosh, and John Lafferty. Clustering with bregman divergences. *Journal of machine learning research*, 6(10), 2005.

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022.

Fausto Bernardini, Joshua Mittleman, Holly Rushmeier, Cláudio Silva, and Gabriel Taubin. The ball-pivoting algorithm for surface reconstruction. *IEEE transactions on visualization and computer graphics*, 5(4):349–359, 1999.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Charlotte Bunne, Ya-Ping Hsieh, Marco Cuturi, and Andreas Krause. Recovering stochastic dynamics via gaussian Schrödinger bridges. *arXiv preprint arXiv:2202.05722*, 2022.

Yuri Burda, Roger B Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *ICLR (Poster)*, 2016.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2020.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Tianrong Chen, Guan-Horng Liu, and Evangelos A Theodorou. Likelihood training of Schrödinger bridge using forward-backward sdes theory. *arXiv preprint arXiv:2110.11291*, 2021.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Max Daniels, Tyler Maunu, and Paul Hand. Score-based generative neural networks for large-scale optimal transport. *Advances in neural information processing systems*, 34:12955–12965, 2021.

Valentin De Bortoli, Arnaud Doucet, Jeremy Heng, and James Thornton. Simulating diffusion bridges with score matching. *arXiv preprint arXiv:2111.07243*, 2021a.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34, 2021b.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Joseph L Doob and JI Doob. *Classical potential theory and its probabilistic counterpart*, volume 549. Springer, 1984.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

Conor Durkan and Yang Song. On maximum likelihood training of score-based generative models. *arXiv e-prints*, pages arXiv–2101, 2021.

Alessio Figalli and Federico Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*. 2021.

Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam M Oberman. How to train your neural ode. *arXiv preprint arXiv:2002.02798*, 2020.

R Flamary, N Courty, D Tuia, and A Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 1, 2016.

Hans Föllmer. An entropy approach to the time reversal of diffusion processes. In *Stochastic Differential Systems Filtering and Control*, pages 156–163. Springer, 1985.

Giulio Franzese, Simone Rossi, Lixuan Yang, Alessandro Finamore, Dario Rossi, Maurizio Filippone, and Pietro Michiardi. How much is enough? a study on diffusion times in score-based generative models. *arXiv preprint arXiv:2206.05173*, 2022.

Ruiqi Gao, Emiel Hoogeboom, Jonathan Heek, Valentin De Bortoli, Kevin P. Murphy, and Tim Salimans. Diffusion meets flow matching: Two sides of the same coin. 2024. URL `https://diffusionflow.github.io/`.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

C. G. Gray and E. F. Taylor. When action is not least. *American Journal of Physics*, 75:434–458, 2007.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *arXiv preprint arXiv:2205.11495*, 2022.

Ulrich G Haussmann and Etienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205, 1986.

Eric Heitz, Laurent Belcour, and Thomas Chambon. Iterative α-(de) blending: A minimalist deterministic diffusion model. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–8, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*, pages 2722–2730. PMLR, 2019.

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

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022b.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34, 2021.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

Yifan Jiang, Shiyu Chang, and Zhangyang Wang. TransGAN: Two pure transformers can make one strong GAN, and that can scale up. *Advances in Neural Information Processing Systems*, 34, 2021.

Bowen Jing, Gabriele Corso, Renato Berlinghieri, and Tommi Jaakkola. Subspace diffusion generative models. *arXiv preprint arXiv:2205.01490*, 2022.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022a.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022b.

Valentin Khrulkov and Ivan Oseledets. Understanding DDPM latent codes through optimal transport. *arXiv preprint arXiv:2202.07477*, 2022.

Young-Heon Kim and Emanuel Milman. A generalization of caffarelli's contraction theorem via (reverse) heat flow. *Mathematische Annalen*, 354(3):827–862, 2012.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34: 21696–21707, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Peter E Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.

Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2020.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *Advances in Neural Information Processing Systems*, 34:14593–14605, 2021.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. *arXiv preprint arXiv:2201.12220*, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Thomas G Kurtz. Equivalence of stochastic equations and martingale problems. In *Stochastic analysis 2010*, pages 113–130. Springer, 2011.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

Hugo Lavenant and Filippo Santambrogio. The flow map of the fokker–planck equation does not provide optimal transport. *Applied Mathematics Letters*, page 108225, 2022.

Junhyeok Lee and Seungu Han. Nu-wave: A diffusion probabilistic model for neural audio upsampling. *arXiv preprint arXiv:2104.02321*, 2021.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. *arXiv preprint arXiv:2405.20320*, 2024.

Antoine Lejay. The girsanov theorem without (so much) stochastic analysis. In *Séminaire de Probabilités XLIX*, pages 329–361. Springer, 2018.

Christian Léonard, Sylvie Rœlly, and Jean-Claude Zambrini. Reciprocal processes. a measure-theoretical point of view. *Probability Surveys*, 11: 237–269, 2014.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022.

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

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024. URL `https://doi.org/10.48550/arXiv.2412.06264`.

Qiang Liu. On rectified flow and optimal coupling. *preprint*, 2022.

Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. *arXiv preprint arXiv:2112.01573*, 2021.

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*, 2022a.

Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022b.

Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022c.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.

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

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021a.

Shitong Luo and Wei Hu. Score-based point cloud denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4583–4592, 2021b.

Zhaoyang Lyu, Xudong Xu, Ceyuan Yang, Dahua Lin, and Bo Dai. Accelerating diffusion models via early stop of the diffusion process. *arXiv preprint arXiv:2205.12524*, 2022.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *arXiv preprint arXiv:2103.16091*, 2021.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Marco Moriconi. Condition for minimal harmonic oscillator action. *American Journal of Physics*, 85(8):633–634, 2017.

Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In *International conference on machine learning*, pages 25858–25889. PMLR, 2023.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9223–9232, 2021.

George Papamakarios, Eric T Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(57):1–64, 2021.

Stefano Peluchetti. Non-denoising forward-time diffusions. 2021.

Stefano Peluchetti. Diffusion bridge mixture transports, schrödinger bridge problems and generative modeling. *Journal of Machine Learning Research*, 24(374):1–51, 2023.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022a.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022b.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. *arXiv preprint arXiv:2110.02999*, 2021.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. *Advances in neural information processing systems*, 29, 2016.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Simo Särkkä and Arno Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling StyleGAN to large diverse datasets. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, pages 1–10, 2022.

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

Neta Shaul, Juan Perez, Ricky TQ Chen, Ali Thabet, Albert Pumarola, and Yaron Lipman. Bespoke solvers for generative flow models. *arXiv preprint arXiv:2310.19075*, 2023.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2C: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34:12533–12548, 2021.

Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020b.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34, 2021.

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Anastasiya Tanana. Comparison of transport map generated by heat flow interpolation and the optimal transport brenier map. *Communications in Contemporary Mathematics*, 23(06):2050025, 2021.

Giulio Trigila and Esteban G Tabak. Data-driven optimal transport. *Communications on Pure and Applied Mathematics*, 69(4):613–648, 2016.

Belinda Tzen and Maxim Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Conference on Learning Theory*, pages 3084–3114. PMLR, 2019.

Benigno Uria, Iain Murray, and Hugo Larochelle. Rnade: The real-valued neural autoregressive density-estimator. *Advances in Neural Information Processing Systems*, 26, 2013.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

David A Van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

Francisco Vargas, Pierre Thodoroff, Austen Lamacraft, and Neil Lawrence. Solving Schrödinger bridges via maximum likelihood. *Entropy*, 23(9): 1134, 2021.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

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

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

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

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Gefei Wang, Yuling Jiao, Qian Xu, Yang Wang, and Can Yang. Deep generative learning via Schrödinger bridge. In *International Conference on Machine Learning*, pages 10794–10804. PMLR, 2021.

Rose E Wang, Esin Durmus, Noah Goodman, and Tatsunori Hashimoto. Language modeling via stochastic processes. *arXiv preprint arXiv:2203.11370*, 2022a.

Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-GAN: Training gans with diffusion. *arXiv preprint arXiv:2206.02262*, 2022b.

Antoine Wehenkel and Gilles Louppe. Diffusion priors in variational autoencoders. *arXiv preprint arXiv:2106.15671*, 2021.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

Lemeng Wu, Chengyue Gong, Xingchao Liu, Mao Ye, and Qiang Liu. Diffusion-based molecule generation with informative prior bridges. *arXiv preprint*, 2022.

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. *arXiv preprint arXiv:2112.07804*, 2021.

Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022.

Mao Ye, Lemeng Wu, and Qiang Liu. First hitting diffusion models for generating manifold, graph and categorical data. *Advances in Neural Information Processing Systems*, 35:27280–27292, 2022.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

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

Qinsheng Zhang, Molei Tao, and Yongxin Chen. gDDIM: Generalized denoising diffusion implicit models. *arXiv preprint arXiv:2206.05564*, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. EGSDE: Unpaired image-to-image translation via energy-guided stochastic differential equations. *arXiv preprint arXiv:2207.06635*, 2022.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020.

Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Truncated diffusion probabilistic models. *arXiv preprint arXiv:2202.09671*, 2022.

Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.