On the Convergence and Straightness of Rectified Flow

Vansh Bansal* **Saptarshi Roy*** **Alessandro Rinaldo** **Purnamrita Sarkar**
Department of Statistics and Data Sciences, UT Austin
*Equal contribution.

Abstract

Flow Matching has become a cornerstone of modern generative models like Stable Diffusion 3, largely due to the efficiency of its Rectified Flow (RF) variant. The success of RF hinges on iteratively learning straight trajectories, pushing generation towards fewer sampling steps. However, the theoretical link between path geometry and sampling efficiency has been under-explored. This paper fills this gap by introducing a novel *Piecewise Straightness* parameter, $\gamma_{2,T}$. We establish the first Wasserstein convergence bound that explicitly links the discretization error of *any* general flow-model to $\gamma_{2,T}$, proving that minimizing curvature is the key to achieving high-fidelity, one-step sampling.

Building on this theory, we establish the first theoretical framework to analyze the straightness of RF. We begin by offering intuitive geometric arguments for simple cases before identifying sufficient conditions under which a single rectification step (1-RF) yields a perfectly straight or even a Monge optimal coupling. While whether these sufficient conditions are met depends on the problem geometry, they enable the first concrete proofs in this area. Critically, fulfilling these conditions makes the subsequent flow (2-RF) perfectly straight ($\gamma_{2,T} = 0$). This eliminates the discretization error in our bound and makes flawless, single-step sampling possible.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

In recent years, diffusion models have become the mainstream approach for image generation tasks (Ho et al., 2022; Balaji et al., 2022; Rombach et al., 2022). They leverage the score-based generative model (SGM) framework (Sohl-Dickstein et al., 2015; Ho et al., 2020), where data is gradually perturbed according to a pre-defined diffusion process, and the process is reversed using Stochastic Differential Equations (SDEs) for sample generation. While powerful, the stochastic nature of the reverse SDEs makes sampling computationally expensive as it requires fine a discretization. Deterministic alternatives, such as probability-flow ordinary differential equations (ODEs) and DDIM (Song et al., 2020b, 2023, 2020a; Lu et al., 2022; Zheng et al., 2023) can be faster but often produce less faithful outputs with the coarse discretizations needed for rapid sampling. The primary reason for this inaccuracy is the discretization error introduced by numerical solvers when approximating highly curved, nonlinear trajectories.

The key to overcoming this limitation lies in learning generative paths that are inherently straight, as this would minimize the error from numerical solvers. This has motivated the development of Flow Matching (FM) (Lipman et al., 2022; Albergo et al., 2023; Albergo and Vanden-Eijnden, 2022), a powerful framework that allows adoption of general probability paths to supervise ODE-based flow models. A prominent application of FM is Rectified Flow (RF) (Liu et al., 2023), which is uniquely designed to learn perfectly straight trajectories from a simple noise distribution to the target data. Through an iterative "reflow" procedure, RF progressively straightens the flow, thereby reducing the transport cost (Liu, 2022; Shaul et al., 2023).

Recent empirical works (Liu et al., 2024, 2023) have demonstrated RF's ability to generate high-quality images with just one or two discretization steps after 2-rectification procedures (2-RF). Moreover, (Lee et al., 2024) offered a heuristic explanation for why 2-

RF should often produce straight flows and developed an improved training routine to achieve this directly, avoiding the potential performance degradation from excessive reflowing. Despite these empirical breakthroughs, the underlying principle remains a heuristic. The formal connection between a flow's geometric straightness, its convergence rate, and the conditions under which RF can provably achieve such straightness remains elusive. In this paper, we establish this missing theoretical foundation by connecting a generative flow's geometry to its sampling efficiency.

First, we introduce a novel and robust Piecewise Straightness (PWS) parameter, $\gamma_{2,T}(\mathcal{Z})$, to quantify a flow's curvature. Using this metric, we establish the first 2-Wasserstein convergence bound that explicitly connects flow geometry to sampling efficiency. Our bound proves that the discretization error scales with $\mathcal{O}(\gamma_{2,T}/T^2)$, providing a rigorous theoretical justification for why straighter flows enable high-fidelity generation with fewer sampling steps. We also give ways to estimate $\gamma_{2,T}$ and show that for Gaussian-mixture target distributions it crucially depends on the maximum separation between the component means.

Building on this theory, we develop the first theoretical framework to analyze RF and identify a sufficient condition on the flow's Jacobian that makes the 2-RF flow perfectly straight ($\gamma_{2,T} = 0$). Furthermore, we identify a stronger commutativity condition under which this flow is not only straight but also the *Monge optimal* transport map. We also provide the first concrete proofs that these conditions are met for key multi-dimensional problems, such as Gaussian-to-Gaussian and Gaussian-to-Gaussian-mixture flows, confirming they yield optimal and straight flows.

Notations. For a matrix A, we define the matrix exponential as $\exp(A) := \sum_{k=}^{\infty} A^k/k!$. For two matrices A and B, we define the Lie-bracket operation as $[A, B] := AB - BA$.

2 PRELIMINARIES

Flow-based generative models define a mapping between samples X_0 from the noise distribution ρ_0 (typically standard Gaussian) and the samples X_1 from the target distribution ρ_1 through an ODE:

$$dZ_t = v(Z_t, t)\, dt, \quad Z_0 = X_0 \sim \rho_0, \tag{1}$$

where $v : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is a time-varying drift (velocity) field that defines a probability path between ρ_0 and ρ_1. A typical recipe for constructing v is to first consider a stochastic interpolation path $X_t = a_t X_0 + b_t X_1$ such that $a_0 = b_1 = 1$ and $a_1 = b_0 = 0$, and $(X_0, X_1) \sim \rho_0 \otimes \rho_1$. Then one can construct a drift field by optimizing the following:

$$v := \underset{f:\mathbb{R}^d\times[0,1]\to\mathbb{R}^d}{\arg\min} \int_0^1 \mathbb{E}\left[\left\|\dot{X}_t - f(X_t, t)\right\|_2^2\right] dt, \tag{2}$$

which has the solution $v(x,t) := v_t(x) = \mathbb{E}[\dot{X}_t \mid X_t = x]$. However, in practice, we parameterize v by a neural network class v_θ to solve (2) as the aforementioned conditional expectation is typically intractable. Let $\widehat{v}$ be an approximate solution of (2), and consider the ODE

$$d\tilde{Y}_t = \widehat{v}_t(\tilde{Y}_t)\, dt, \quad \tilde{Y}_0 \sim \rho_0. \tag{3}$$

As proposed in Liu et al. (2023), we apply the Euler discretization of the ODE to obtain our final sample estimates:

$$\widehat{Y}_{t_i} = \widehat{Y}_{t_{i-1}} + \widehat{v}_{t_{i-1}}(\widehat{Y}_{t_{i-1}})(t_i - t_{i-1}), \quad i \in [T], \tag{4}$$

where the ODE is discretized into T uniformly spaced steps, with $t_i = i/T$. The final sample estimate $\widehat{Y}_1$ follows the distribution $\widehat{\rho}_1 := \mathrm{Law}(\widehat{Y}_1)$.

The Optimal Transport (OT) problem : The OT problem, first formulated by Monge (1781), seeks to find a deterministic map $\mathcal{T}$ that transports mass from an initial distribution ρ_0 to a target distribution ρ_1 with minimal cost. This is expressed as

$$\inf_{\mathcal{T}} \mathbb{E}_{\rho_0}\left[c(\mathcal{T}(X_0) - X_0)\right] \quad \text{s.t.} \quad \mathrm{Law}(\mathcal{T}(X_0)) = \rho_1 \tag{5}$$

An equivalent dynamic formulation recasts this as finding an optimal continuous-time path $\{X_t\}_{t\in[0,1]}$ between the distributions. For convex cost functions c, the optimal path is the straight-line *displacement interpolant*, $X_t = tX_1 + (1-t)X_0$, which forms a geodesic in the Wasserstein space. This specific interpolant minimizes the kinetic energy of the flow, resulting in straight trajectories. RF, as we will discuss next, is uniquely designed to learn the drift function v_t corresponding to this displacement interpolant, simplifying the complex OT problem into a series of tractable least-squares optimization tasks. For the remainder of this paper, we assume $c = \|\cdot\|^2$ when referring to the OT problem (5) unless stated otherwise.

Rectified flow (RF) : RF (Liu et al., 2023) interpolates between the two distributions in straight line paths as $X_t = tX_1 + (1-t)X_0$ for $t \in [0,1]$, giving $v_t(x) = \mathbb{E}[X_1 - X_0 \mid X_t = x]$. Therefore, the resulting sampling ODE (1) also approximates a geodesic path between ρ_0 and ρ_1 that allows faster sampling.

Straight couplings and flows : In context of RFs, the optimization step (2) is known as a rectification step, and the solution path $\mathcal{Z} := \{Z_t\}_{t\in[0,1]}$ of ODE (1) is known as the rectified flow, denoted

by $\mathcal{Z} = \mathsf{Rectflow}(X_0, X_1)$. It is known that RF is marginal preserving Liu et al. (2023), i.e., $\text{Law}(X_t) = \text{Law}(Z_t)$. Hence, RF yields a new *dependent* coupling $(Z_0, Z_1) := \mathsf{Rectify}(X_0, X_1)$ between ρ_0 and ρ_1. Moreover, a coupling (X_0, X_1) is called a *straight coupling* if $\mathbb{E}[X_1 - X_0 \mid tX_1 + (1-t)X_0] = X_1 - X_0$, a.s. with respect to the distribution of (X_0, X_1) and for $t \sim \text{Unif}(0,1)$, i.e., the drift v along its linear interpolation is a constant function of time almost surely. Therefore, the flow generated by rectification of a straight coupling (X_0, X_1) is defined as a *straight flow*. Liu et al. (2023) showed that successive rectifications eventually lead to a near straight flow that allows faster sampling in the generation stage. However, quantitative bounds on the effects of straightness of a flow on its convergence rate still remain unknown.

3 STRAIGHTNESS OF A FLOW

This section introduces our novel parameters for quantifying the straightness of the ODE flow in (1). We will subsequently show that our more precise notions of straightness are critical for efficient numerical integration of flow-based models with fewer discretization steps.

Intuitively, a perfectly straight path has no curvature. For a parametric curve, $\alpha(t) := (t, Z_t)$ for $t \in [0,1]$, this corresponds to zero magnitude of its acceleration, i.e.,

$$\|\ddot{\alpha}(t)\|_2 = \|(0, \dot{v}_t(Z_t))\|_2 = \|\dot{v}_t(Z_t)\|_2 = 0 \qquad (6)$$

for all t. For the definition of a straight flow mentioned in Section 2, assuming $\mathcal{Z} = \{Z_t\}_{t\in[0,1]}$ is a twice differentiable curve defined through ODE (1) with $Z_0 \sim \rho_0 = N(0, I_d)$, we require (6) to hold for almost every $t \sim \text{Unif}(0,1)$. This motivates us to quantify the straightness of the entire flow $\mathcal{Z}$ by measuring its magnitude of acceleration along the path by the following two quantities:

Definition 1. *Let $\mathcal{Z} = \{Z_t\}_{t\in[0,1]}$ be twice-differentiable flow following the ODE (1).*

1. *The average straightness (AS) parameter of $\mathcal{Z}$ is defined as*

$$\gamma_1(\mathcal{Z}) := \int_0^1 \mathbb{E}\left[\|\dot{v}_t(Z_t)\|_2^2\right] dt.$$

2. *Let $0 = t_0 < t_1 < \ldots < t_T = 1$ be a partition of $[0,1]$ into T intervals of equal length. The piecewise straightness (PWS) parameter of the flow $\mathcal{Z}$ is defined as*

$$\gamma_{2,T}(\mathcal{Z}) := \max_{i\in[T]} \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \mathbb{E}\left[\|\dot{v}_t(Z_t)\|_2^2\right] dt.$$

The quantity $\gamma_1(\mathcal{Z})$ essentially captures the average curvature of the flow over time $t \in [0,1]$. On the other hand, $\gamma_{2,T}(\mathcal{Z})$ captures the maximum average curvature of $\mathcal{Z}$ over the sub-intervals $[t_{i-1}, t_i]$ for all $i \in [T]$. Therefore, $\gamma_{2,T}(\mathcal{Z})$ captures a more stringent notion of straightness, and it's easy to show that $\gamma_{2,T}(\mathcal{Z}) \geq \gamma_1(\mathcal{Z})$. See Lemma 3 in the appendix.

A small value for $\gamma_1(\mathcal{Z})$ or $\gamma_{2,T}(\mathcal{Z})$ indicates the flow is nearly straight. Below, we argue that these measures provide a more robust notion of straightness than the criterion

$$S(\mathcal{Z}) := \int_0^1 \mathbb{E}\left[\|Z_1 - Z_0 - v_t(Z_t)\|_2^2\right] dt$$

proposed by Liu et al. (2023). While, it can be shown that $\gamma_1(\mathcal{Z}) = \gamma_{2,T}(\mathcal{Z}) = 0 \implies S(\mathcal{Z}) = 0$, the quantity $S(\mathcal{Z})$ can be misleadingly small even for highly curved paths. For instance, consider the wavy flow given by $Z_t = Z_0 + (t, 50N^{-2}\sin(2\pi N t))^\top$. A straightforward calculation shows that $S(\mathcal{Z}) = O(N^{-2}) \to 0$ as $N \to \infty$, which incorrectly implies the path is becoming straighter. In reality, the flow just oscillates more rapidly. Our metrics, on the other hand, remain bounded away from zero ($\gamma_{2,T}(\mathcal{Z}) \geq \gamma_1(\mathcal{Z}) = 2 \times 10^4 \pi^4$), and correctly identify the increasing oscillations as a departure from straightness.

Estimating the straightness parameters : Since the true drift field is not typically tractable in practice, we give the following estimators for our straightness parameters which use the learnt field and do not require any extra computation:

$$\widehat{\gamma}_{2,T} = \frac{1}{(\Delta t)^2} \max_{i=1,\ldots,T-1} \left(\frac{1}{N} \sum_{j=1}^{N} \left\| \widehat{v}_i^{(j)} - \widehat{v}_{i-1}^{(j)} \right\|_2^2 \right) \qquad (7)$$

$$\text{and } \widehat{\gamma}_1 = \frac{1}{\Delta t} \frac{1}{N} \sum_{i=0}^{T-2} \sum_{j=1}^{N} \left\| \widehat{v}_{i+1}^{(j)} - \widehat{v}_i^{(j)} \right\|_2^2,$$

where $\widehat{v}_i^{(j)} = \widehat{v}_{t_i}(\widehat{Y}_{t_i}^{(j)})$ denotes the estimated drift for each of the $j \in [N]$ samples flowing through the discretized ODE (4).

4 WASSERSTEIN CONVERGENCE

In this section, we analyze error rates for the final sampling distribution of a learnt flow model in terms of the 2-Wasserstein distance from the target distribution. To this end, we make the following assumptions on the drift function and its estimate that are necessary for establishing our error bounds:

Assumption 1. *Assume that*

(a) *There exists* $\varepsilon_{\text{vl}} \geq 0$ *such that* $\max_{0\leq i\leq T} \mathbb{E}_{X_{t_i}\sim\rho_{t_i}} \left[\|v_{t_i}(X_{t_i}) - \widehat{v}_{t_i}(X_{t_i})\|_2^2\right] \leq \varepsilon_{\text{vl}}^2$.

(b) *There exists* $\widehat{L} > 0$ *such that* $\|\widehat{v}_t(x) - \widehat{v}_t(y)\|_2 \leq \widehat{L}\|x-y\|_2$ *almost surely.*

Assumption 1(a) requires $\widehat{v}_t$ to closely estimate the true drift v_t across all $t \in \{t_i\}_{i\in[T]}$, a standard condition in the diffusion model literature (Gupta et al., 2024; Li et al., 2024b,a; Chen et al., 2023) essential for controlling error rates. We emphasize that it is notably weaker than assuming a uniform bound on the estimation error for all $t \in [0,1]$. Assumption 1(b) imposes a Lipschitz condition (similar to *one-sided* Lipschitzness) on $\widehat{v}_t$, also common for score functions in both score-based and flow-based generative models (Chen et al., 2023; Kwon et al., 2022; Li et al., 2024b; Pedrotti et al., 2024; Boffi et al., 2024). Since $\widehat{v}$ is typically parameterized by neural networks with Lipschitz activations, this condition is both natural and practical. Moreover, in flow-based models, Lipschitz continuity of v_t is crucial for the well-posedness of the ODE (1) (Liu et al., 2023; Boffi et al., 2024), further justifying the use of Lipschitz-constrained networks in training.

4.1 Convergence under exact integration

We begin by analyzing the continuous-time flow (3) with estimated drift under exact integration. Our first main result bounds its Wasserstein distance to the target distribution in terms of the drift estimation error:

Theorem 1. *Let* ρ_1 *be absolutely continuous with respect to the Lebesgue measure in* $\mathbb{R}^d$. *Define* $\varepsilon^2(t) := \mathbb{E}_{X_t\sim\rho_t}\left[\|v_t(X_t) - \widehat{v}_t(X_t)\|^2\right]$ *for* $t \in [0,1]$, *and* $\tilde{\rho}_1 := Law(\tilde{Y}_1)$ *for* $\tilde{Y}_1$ *obtained through the flow in* (3). *Then, under Assumption 1(b), we have that*

$$W_2^2(\tilde{\rho}_1, \rho_1) \leq e^{1+2\widehat{L}} \int_0^1 \varepsilon^2(t)\, dt \quad \textit{almost surely.}$$

The bound presented in Theorem 1 closely resembles those established in prior works (Kwon et al., 2022; Pedrotti et al., 2024; Boffi et al., 2024), as it primarily depends on the estimation error $\varepsilon(t)$ for all $t \in [0,1]$. Specifically, if there exists an $\varepsilon > 0$ such that $\sup_{t\in[0,1]} \varepsilon^2(t) \leq \varepsilon^2$, then the squared 2-Wasserstein distance between the two distributions is of order $\mathcal{O}(\varepsilon^2)$. The full proof is deferred to Appendix C.1.

Remark 1. *The absolute continuity requirement in Theorem 1 can be relaxed. If the density of* ρ_1 *does not exist, then one can convolve* X_1 *with an independent noise* $W_\eta \sim N(0, \eta I_d)$ *for a very small* $\eta > 0$, *and consider the mollified distribution* $\rho_1^\eta := Law(X + W_\eta)$ *as the target distribution. Note that* ρ_1^η *is absolutely continuous and satisfies* $W_2^2(\rho_1^\eta, \rho_1) \leq \eta^2 d$. *Therefore, under the condition of Theorem 1, and using triangle inequality we have* $W_2^2(\tilde{\rho}_1, \rho_1) \lesssim \eta^2 d + e^{1+2\widehat{L}} \int_0^1 \varepsilon^2(t)\, dt$.

4.2 Convergence of the discretized flow

In this section we show that the more accurate perception of the straightness of a flow captured by our AS and PWS parameters is crucial for analyzing discretization error. We present our main result below:

Theorem 2. *Let Assumption 1 hold, and suppose the continuous-time flow* $\mathcal{Z} := \{Z_t\}_{t\in[0,1]}$ *is defined by the ODE* (1) *with a differentiable drift field* $v : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$. *Then the sampling distribution* $\widehat{\rho}_1$ *obtained through the discretized ODE* (4) *satisfies the following almost sure inequality:*

$$W_2^2(\widehat{\rho}_1, \rho_1) \leq \frac{27e^{4\widehat{L}}}{\max\{\widehat{L}^2, 1\}} \left(\frac{\gamma_{2,T}(\mathcal{Z})}{T^2} + \varepsilon_{\text{vl}}^2\right),$$

The term involving the PWS parameters could be referred to as an error term due to discretization. More importantly, the above Wasserstein error bound shows that $T = \Omega\left(\sqrt{\gamma_{2,T}(\mathcal{Z})/\epsilon}\right)$ is sufficient to achieve a discretization error of the order $O(\epsilon)$. Therefore, Theorem 2 indicates that if the flow is near-straight (i.e., $\gamma_{2,T}(\mathcal{Z}) \approx 0$), then accurate estimation of the data distribution can be achieved with a very few discretization steps. This finding indeed aligns with the prior empirical findings (Liu et al., 2023; Lee et al., 2024; Liu et al., 2024) related to the rectified flow. It is also consistent with the empirical behavior of Perflow (Yan et al., 2024), a methodology that has achieved improved performance by further straightening the rectified flow in each interval $[t_{i-1}, t_i]$ for all $i \in [T]$. The proof of the theorem can be found in Appendix C.3. Finally, the elementary inequality $\gamma_{2,T}(\mathcal{Z}) \leq T\gamma_1(\mathcal{Z})$ (see Appendix C.2) also yields the following corollary.

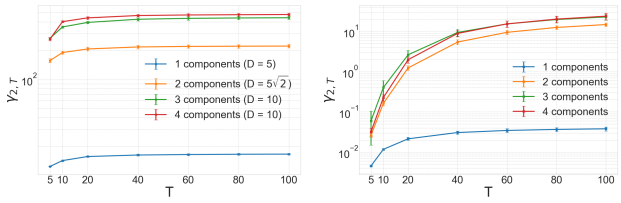
Corollary 1. *Under the same conditions of Theorem 2, we have the following almost sure inequality:*

$$W_2^2(\widehat{\rho}_1, \rho_1) \leq \frac{27e^{4\widehat{L}}}{\max\{\widehat{L}^2, 1\}} \left(\frac{\gamma_1(\mathcal{Z})}{T} + \varepsilon_{\text{vl}}^2\right).$$

4.3 Convergence rates for Rectified Flow

In this section, we focus our analysis on RF for an independent coupling between noise samples $X_0 \sim \rho_0 = N(0, I_d)$ and the target samples $X_1 \sim \rho_1$. We first obtain an expression for the acceleration $\dot{v}_t(x)$ and then obtains bounds on the Wasserstein error through $\gamma_{2,T}$ under certain assumptions.

By definition, we have $X_t = tX_1 + (1-t)X_0$ and $\rho_t =$



(a) (b)

Figure 1: For mixtures-of-Gaussians target distributions with varying components K and maximum mean separation D (see details in Appendix A.1), the figure shows (a) $\hat{\gamma}_{2,T}(\mathcal{Z})$ vs T for the 1-RF flow (b) $\hat{\gamma}_{2,T}(\mathcal{Z})$ vs T for the 2-RF flow

Law(X_t). Using Tweedie's formula, we obtain:

$$v_t(x) := \mathbb{E}\left[X_1 - X_0 \mid X_t = x\right] = \frac{x}{t} + \left(\frac{1-t}{t}\right) s_t(x), \tag{8}$$

where $s_t(x)$ is the score, i.e. the gradient of perturbed log density w.r.t x for $t \in (0,1]$ and $v_0(x) = \mathbb{E}[X_1] - x$; see Lemma 9.

Lemma 1. *Let $(X_0, X_1) \sim \rho_0 \otimes \rho_1$ and v be the RF field defined in (8). Then we have that:*

$$\dot{v}_t(x) = -\frac{s_t(x)}{t^2} - \frac{(1-t)^2}{t^2}\left(H_t(x)s_t(x) + \nabla_x \mathsf{Tr}(H_t(x))\right) \tag{9}$$

where $H_t(x) := \nabla_x s_t(x)$ is the gradient of the score function $s_t(x)$.

The result of Lemma 1 is critical, as it shows that the acceleration $\dot{v}_{t^-}$ and therefore the parameters $\gamma_{2,T}$ and γ_{1^-} is determined by the score s_t, Hessian H_t and the gradient of the Hessian's trace. While s_t is known to be a sub-Gaussian for $0 \le t < 1$ (Gupta et al., 2024, Lemma F.2), bounding the other two higher order terms requires the knowledge of the geometry of the target distribution ρ_1. As an example, the following lemma considers the expected squared acceleration norm for a Gaussian-mixture target with bounded mean separation.

Lemma 2. *Let $(X_0, X_1) \sim N(0, I_d) \otimes \rho_1$ where $\rho_1 = \sum_{i=1}^K \pi_k N(\mu_k, \sigma^2 I_d)$, such that $\mu_i \in \mathbb{R}^d$ and $D^2 := \max_{i,j} \|\mu_i - \mu_j\|_2^2$. For $\mathcal{Z} = \mathsf{Rectflow}(X_0, X_1)$ obtained by integrating (1) with drift field v defined in (8), we have*

$$\gamma_{2,T}(\mathcal{Z}) = \mathcal{O}\left(D^2 d^2 + D^4 d\right)$$

The lemma reveals a surprising result: for a fixed maximum-mean separation D, the straightness metric $\gamma_{2,T}$ is largely independent of the number of mixture components K. This theoretical finding is supported by our empirical results in Figure 1(a) which demonstrates that the value of $\hat{\gamma}_{2,T}$ estimated as in 7 is a function of D, increasing as D grows, while showing little variation with changing K when D is held constant. Moreover, it also immediately yields an $\mathcal{O}(d^2/T^2)$ bound on the W_2^2-error for a given D; see Lemma 6, suggesting that the discretization steps should scale linearly with d. We defer the proof to Appendix D.3.

Another attractive property of RF is its *reflow* procedure, where one iteratively rectifies the coupling generated by the preceding flow. Liu et al. (2023) show that their straightness parameter $S(\mathcal{Z})$ decreases with successive applications of the reflow procedure. In the same vein, Figure 1(b) shows that even our straightness parameter $\gamma_{2,T}$ decreases with successive rectifications.

Empirical studies argue that not only a single reflow (2-RF) is sufficient to generate a near-straight flow, but too many reflows also hurt the performance due to model collapse caused by error accumulation (Lee et al., 2024). However, most prior arguments are based on heuristics and a theoretical framework to argue the straightness of 2-RF remains elusive.

5 (WHEN) DOES 2-RF YIELD A STRAIGHT FLOW?

In this section, we pose the following question, which has mostly been addressed empirically for general target distributions previously.

When does 2-RF produce a provably straight flow, or equivalently, 1-RF yield a straight coupling?

Liu et al. (2023) prove that for one-dimensional target distributions, 2-RF does yield a straight flow since the 1-RF coupling is deterministic and monotonic, and therefore straight (Appendix F.2). We first illustrate that for some target distributions, such as Gaussians and two-Gaussians mixtures, 1-RF indeed yields a straight coupling even in d (≥ 2) dimensions. This provides the first concrete theoretical evidence for straightness of 2-RF beyond one-dimensional targets. Motivated by these examples, we develop a general framework for proving straightness and Monge optimality in Section 5.2 and 5.3 respectively.

5.1 Concrete illustrative examples

We start present a few examples of d-dimensional target distributions with $d \ge 2$, where 1-RF yeilds a straight coupling using the true RF drift field in (1).

Example 1 (Gaussian to Gaussian). *Let $\rho_0 = N(0, I_d)$ and $\rho_1 = N(\mu, \Sigma)$, where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ is a symmetric positive-definite matrix.*

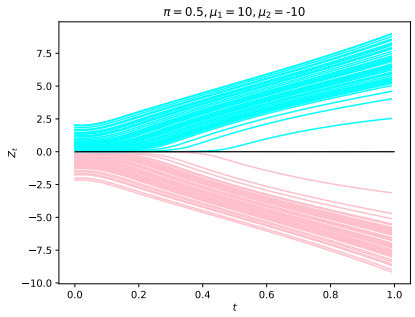


Figure 2: shows the flow of points from a standard Gaussian $\rho_0 = N(0, 1)$ to a symmetric mixture of two Gaussians $\rho_1 = .5N(10, 1) + .5N(-10, 1)$, where the black line represents $y = 0$.

Theorem 3. *Let $(X_0, X_1) \sim \rho_0 \times \rho_1$ be an independent coupling where ρ_0 and ρ_1 are specified in Example 1. The coupling $(Z_0, Z_1) = \mathsf{Rectify}(X_0, X_1)$ is straight and is given by $Z_1 = \Sigma^{1/2} Z_0 + \mu$.*

Proof sketch. In this simple case, one can check that the RF velocity is $v_t(Z_t) = \frac{x}{t} + \frac{1-t}{t}\Sigma_t^{-1}(t\mu - x)$, where $\Sigma_t = t^2\Sigma + (1-t)^2 I_d$. Therefore, ODE (1) can be solved exactly, yielding the solution $Z_1 = \Sigma^{1/2} Z_0 + \mu$. Consequently, this is also the Monge coupling between $N(0, I_d)$ and $N(\mu, \Sigma)$. Additionally, it is worth mentioning that the optimality of the 1-RF coupling is also a byproduct of a certain commutativity property of $\nabla_{Z_t} v_t(Z_t)$ (see Theorem 8). The details of the proof is deferred to Appendix E.1.

The rest of the section considers target distributions that are multimodal.

Example 2 (Gaussian to 2-mixture of Gaussian.). *Here, ρ_0 is $N(0, I_d)$ and ρ_1 is a mixture of two Gaussians with the isotropic covariance matrix, i.e., $\rho_1 := \pi N(\mu_1, \sigma^2 I_d) + (1-\pi) N(\mu_2, \sigma^2 I_d)$ with $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\pi \in [0, 1]$.*

Theorem 4. *Let $(X_0, X_1) \sim \rho_0 \otimes \rho_1$ be an independent coupling where ρ_0 and ρ_1 are specified in Example 2. Then 1-RF yields a straight coupling.*

Intuitive proof: First, note that $(Z_0, Z_1) := \mathsf{Rectify}(X_0, X_1)$ is an deterministic coupling. Therefore, the global invertibility of the map $H_t : Z_0 \mapsto (1-t)Z_0 + tZ_1$ is enough to ensure the straightness of (Z_0, Z_1). Now, consider a simple case of Example 2 with $\mu_1 = \mu = -\mu_2$. It turns out that in this case, the flow induced by v_t has an interesting geometric structure (see, for example, Figure 2). In particular, if z_0 is positive (negative), then z_t is also positive (negative) *for all* t. This holds coordinate-wise. Here z_0 and z_t are obtained from ODE (1). This follows from a very fundamental fact that flow ODE decouples into d one-dimensional ODEs after an orthonormal transformation. Since, in the transformed space, the linear interpolation paths of a straight coupling $(Z_{0,i}, Z_{1,i})$ can not intersect for scalar random variables, we must have that it is monotonically increasing, i.e., for $(z_{0,i}, z_{1,i})$ and $(z'_{0,i}, z'_{1,i})$ such that $z_{0,i} < z'_{0,i}$, we must have that $z_{1,i} < z'_{1,i}$ for each co-ordinate $i \in [d]$. This ensures that the RF transport map is co-ordinate wise increasing. Therefore, $H_t(\cdot)$ is invertible and 1-Rf yields a straight coupling. Furthermore, this also allows to conclude that the flow preserves the *quantiles* co-ordinate wise; see Lemma 7.

Finally, we come to the Gaussian mixture to Gaussian mixture setting.

Example 3 (2-mixture of Gaussians to 2-mixture of Gaussians). *Consider $\mu_{01} = (0, a)^\top, \mu_{02} = (0, -a)^\top$ and $\mu_{11} = (a, a)^\top, \mu_{12} = (a, -a)^\top$ for some $a > 0$. Let $X_0 \sim 0.5N(\mu_{01}, I_2) + 0.5N(\mu_{02}, I_2)$ and $X_1 \sim 0.5N(\mu_{11}, I_2) + 0.5N(\mu_{12}, I_2)$.*

Theorem 5. *Let $(X_0, X_1) \sim \rho_0 \otimes \rho_1$ be an independent coupling where ρ_0 and ρ_1 are specified in Example 3. Then 1-RF gives a straight coupling.*

The intuitive explanation is that even in this case, the flows along each coordinate decouples and leads to a straight coupling. We defer the proofs of Theorems 3, 4 and 5 to Appendix E. The above proof techniques, while intuitive, do not generalize to examples like a mixture of three or more Gaussians. They also do not provide a way to establish whether the 1-RF coupling gives a Monge map. Now we present an overarching theoretical framework for proving both.

5.2 Sufficient condition for straightness

In this section we provide a sufficient analytical condition that makes the 2-RF a straight flow. To this end, let us consider the ODE (1) with a fixed initial condition, i.e.,

$$dZ_t = v_t(Z_t)\, dt, \quad Z_0 = z_0, \tag{10}$$

where v_t is the solution of (2). For clarity, we denote the solution of the above ODE as $Z_t(z_0)$. It is well known that under a locally-Lipschitz field v_t, ODE (10) admits a unique solution if it's non-explosive (see Appendix B). These conditions are milder than those imposed in prior literature and are satisfied by a large class of target distributions such as Gaussian-mixtures. Therefore, in the remainder of this section, we always assume that (10) admits a unique solution.

Under the above conditions, Coddington et al. (1956, page 25-27) show that the Jacobian $J_t^{z_0} := \nabla Z_t(z) \,|_{z=z_0}$ obeys the following ODE:

$$\frac{dJ_t^{z_0}}{dt} = \nabla_{Z_t} v_t(Z_t(z_0)) J_t^{z_0}; \quad J_0^{z_0} = I_d. \tag{11}$$

Since the initial coupling is assumed to be independent, Equation (8) yields that $\nabla_z v_t(z) = (1/t)I_d + t^{-1}(1-t)\nabla_x^2 \log \rho_t(z)$, which is a symmetric matrix for all $t > 0$. Next, we state the sufficient condition on the Jacobian $J_1^{z_0}$ which makes 2-RF straight.

Assumption 2. *The minimum eigenvalue of* $J_1^{z_0} + J_1^{z_0\top}$ *is non-negative for all* $z_0 \in \mathbb{R}^d$.

Although, this assumption can be partially checked by simulating ODE (11) for a large set of initial values (see Appendix A.5), verifying it in practice is challenging, and it may not hold for general target distributions (see Figure 3(b)). However, in the later sections we theoretically show that Assumption 2 is satisfied for some motivating examples. We now state our main straightness result below:

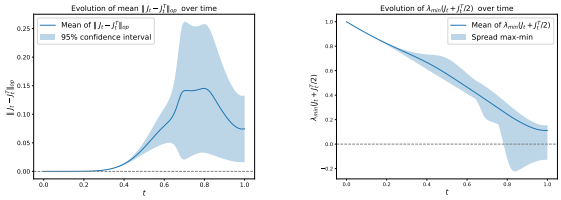
Theorem 6 (1-RF is straight)**.** *Under Assumption 2, 1-RF yields a straight coupling.*

As mentioned previous section, the main step of the proof relies on showing that the map $H_t(z_0) := (1-t)z_0 + tZ_1(z_0)$ is globally invertible almost surely in $t \in \text{Unif}([0,1])$. Plastock (1974) shows that a necessary and sufficient condition for a map $f\colon \mathbb{R}^d \to \mathbb{R}^d$ to be an homeomorphism is its local invertibility (namely, that the Jacobian of f is everywhere non-zero) and properness, i.e. $\|f(x)\|_2 \to \infty$ whenever $\|x\|_2 \to \infty$. When specialized to our problem, by (11) this amounts to showing that, for all $z_0 \in \mathbb{R}^d$, the determinant of $(1-t)I_d + tJ_1^{z_0}$ is non-zero, for almost all $t \in (0,1)$. While, in general, this condition is hard to verify, in Appendix E.4 we show that Assumption 2 guarantees both local invertibility and properness of the map under consideration. Thus, Theorem 6 provides an explicit condition for the straightness of the 1-RF coupling or, equivalently, the 2-RF flow.

While Assumption 2 seems stringent, we would show in Section 5.3 the surprising fact that both Examples 1 and 2 satisfy it. We further consider a more general family of target distributions, a mixture of more than two Gaussians, all with the same covariance matrix. We show that as long as the maximum pairwise mean separation is suitably small, 1-RF produces a straight coupling.

Theorem 7. *Let* $(X_0, X_1) \sim N(0, I_d) \otimes \rho_1$ *where* $\rho_1 := \sum_{j=1}^K \pi_j N(\mu_j, \sigma^2 I_d)$ *with mixture proportions* $\{\pi_j\}_{j\in[K]} \in (0,1)^K$. *If* $\max_{i\neq j} \|\mu_i - \mu_j\|_2^2 \leq 4\sigma^2$, *then 1-RF yields a straight coupling.*

We note that, for our framework to guarantee straightness, the bound on the maximum pairwise distance is important. See Figure 3(b) where a large pairwise mean separation leads to a violation of Assumption 2. The detailed proof is present in Appendix E.5. A similar straightness result is also true when



(a) (b)

Figure 3: The above plots show the evolution of mean $\|J_t^{z_0} - J_t^{z_0\top}\|_{op}$ and $0.5 * \lambda_{min}(J_t^{z_0} + J_t^{z_0\top})$ over 100 different random initial points $z_0 \sim N(0, I_2)$ for $\rho_1 = \sum_{k=1}^4 N(\mu_k, \sigma^2 I_2)$ with weights $\pi_k = k/10$ for $k \in \{1,2,3,4\}$, and $\mu_1 = (0,-2), \mu_2 = (-1,-2), \mu_3 = (1,3), \mu_4 = (2,0)$ and $\sigma = 0.1$.

$\rho_1 = \sum_{j=1}^K \pi_j N(\mu_j, \Sigma)$ for some positive definite matrix Σ. The details of the result and its proof is deferred to Appendix E.6.

We note that Theorem 7 assumes the ideal RF coupling, i.e., one generated by exactly integrating the true RF field. This is analytically unavailable even though the field is exactly known (see Lemma 12.) However, as we show in Lemma 2, even the discretizated ODE leads to small errors, resulting in a near straight 1-RF coupling, i.e., 2-RF flow is straight.

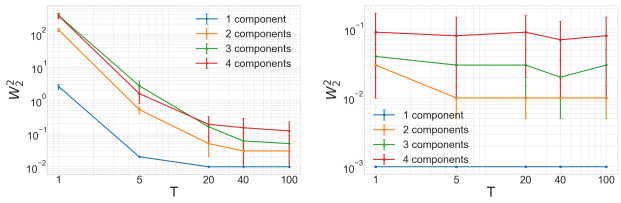
5.3 Connection to Monge map

It is important to clarify the distinction between straight and optimal couplings. While an OT (Monge) map for convex costs is always straight, the converse is not true; any straight coupling is not necessarily the optimal solution to the Monge problem. In this section, we show that 1-RF does lead to the Monge map between ρ_0 and ρ_1 under an additional sufficient condition on the velocity drift v_t. We emphasize that 1-RF may not yield a Monge map in general, and one needs to apply c-Rectified Flow (Liu, 2022) to obtain an approximate solution to Monge problem.

Theorem 8 (1-RF yields Monge map)**.** *Assume that* $[\nabla_{Z_t} v_t(Z_t(z_0)), \nabla_{Z_s} v_s(Z_s(z_0))] = 0$ *for all* $t \neq s$, *and initial points* z_0. *Then 1-RF yields the optimal transport plan for the OT problem* (5), *and its Jacobian satisfies*

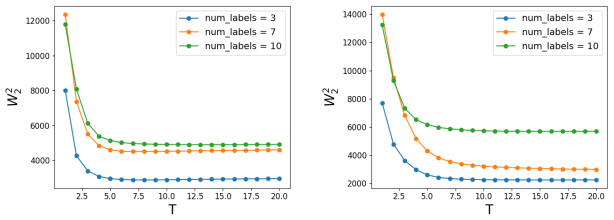
$$J_1^{z_0} = \exp\left(\int_0^1 \nabla_{Z_t} v_t(Z_t(z_0))\, dt\right).$$

The proof of the theorem is provided in Appendix F.1. Since $\nabla_{Z_t} v_t(Z_t(z_0))$ is symmetric for all $t \in [0,1]$, the matrix $J_1^{z_0}$ is symmetric and positive-definite. Brenier's theorem (Chewi et al., 2024) suggests that for



(a) (b)

Figure 4: For the same mixture-of-Gaussians target distributions and estimated drift model in Figure 1, we show (a) $W_2^2(\widehat{\rho}_1, \rho_1)$ vs T for the first RF (b) $W_2^2(\widehat{\rho}_1, \rho_1)$ vs T for the second RF.



(a) (b)

Figure 5: The figure shows $W_2^2(\widehat{\rho}_1, \rho_1)$ vs T for the first RF for (a) MNIST dataset (b) FashionMNIST dataset with varying number of labels.

$z_0 \mapsto Z_1(z_0)$ to be the Monge map, $Z_1 = \nabla\varphi$ for some *convex function* $\varphi : \mathbb{R}^d \to \mathbb{R}$, i.e., the Jacobian $J_1^{z_0} = \nabla^2\varphi(z_0)$ has to be a *symmetric and positive semi-definite* matrix. Thus, under commutativity condition above, *1-RF yields the ℓ_2-optimal (Monge) coupling, which is also straight.* However, we emphasize that the converse might not be true.

In one dimension, 1-RF always recovers the Monge map, as the commutativity condition holds trivially; see Appendix F.2. We show more examples below:

Proposition 1. *Let $(X_0, X_1) \sim \rho_0 \otimes \rho_1$ be an independent coupling. When ρ_0 and ρ_1 are specified in Example 1 or 2, 1-RF produces the optimal Monge map, and therefore a straight coupling.*

In these examples, one can easily verify that the conditions in Theorem 8 are satisfied. Therefore, it follows that 1-RF yields the straight and optimal coupling; see Appendix F.3.

In contrast, for a general mixture of Gaussians, the gradient of the velocity may not commute. So, even though under a certain condition on the means, the resulting 1-RF yields a straight coupling (Theorem 7), it may not be the Monge map (also see Figure 3).

6 EXPERIMENTS

In this section, we primarily explore the effect of the number of discretization steps T and the straightness parameter $\gamma_{2,T}(\mathcal{Z})$ on the W_2 distance between the target distribution and the sampling distribution of the first and second RFs. We present numerical experiments for some synthetic and real datasets. Additional experiments can be found in Appendix A. We provide the code at https://github.com/bansal-vansh/rectified-flow-straightness.

Synthetic data: We first start with the independent coupling $(X_0, X_1) \sim N(0, I_d) \otimes \rho_1$, where ρ_1 is a mixture of Gaussians with $K \in \{1, 2, 3, 4\}$ components and varying maximum mean separation D (see details in Appendix A.1). For all four cases, we estimate the RF drift using a feed-forward neural network and generate the 1-RF samples using the ODE 4 with T discretization steps. Figure 4(a)-(b) shows that $W_2(\widehat{\rho}_1, \rho_1)$ steeply decreases with increasing number of steps T, validating our Theorem 2. Moreover, we note that W_2 distance is consistently larger for the flow corresponding to a larger component mean separation D, owing to a larger value of the straightness parameter $\gamma_{2,T}(\mathcal{Z})$ as shown in Figure 1(a)-(b). Lastly, in accordance with the $\gamma_{2,T}$ plots in Figure 1(b), Figure 4(b), further empirically validates that the 2-RF for Gaussian mixtures produces a near-straight flow, since the Wasserstein error even with a single discretization step is close to 0.

Real data: For the real data experiments, we consider the MNIST and FashionMNIST datasets. In both examples, we train a UNet architecture-based network on training data to estimate the drift function and then evaluate the Wasserstein distance of the generated samples from the test split of the dataset. To emulate the behavior of having different number of modes, we consider three subsets of both the datasets consisting of the first 3 labels, the first 7 labels, and all 10 labels. We observe in Figure 5 that similar to the Gaussian mixture example, the presence of a higher number of components, possibly indicating towards increasing mode separation, increases the Wasserstein error, again indicating that the flow becomes less straight with the increasing number of modes.

7 CONCLUSION

This paper establishes the first formal link between flow geometry and sampling efficiency by introducing the piecewise straightness parameter, $\gamma_{2,T}$. We derive a novel Wasserstein convergence bound showing that the W_2^2-error scales with $\mathcal{O}(\gamma_{2,T}/T^2)$, providing

a rigorous foundation for why straighter flows enable high-fidelity, few-step generation.

Building on this, we propose the first theoretical framework to analyze straightness in Rectified Flow (RF). While the sufficient conditions we identify for achieving a perfectly straight flow ($\gamma_{2,T} = 0$) can be conservative in some cases, they eliminate the discretization error in our bound and allow for the first concrete proofs of perfect one-step generation in key multi-dimensional settings. Ultimately, our work provides a methodology to advance the study of flow straightness from an empirical heuristic to a provable principle.

Open problem: Recall that Assumption 2 is only sufficient to ensure global invertibility of the map $H_t(\cdot)$ almost surely for all $t \sim \text{Unif}([0,1])$. Even in our simulations, it appears to not be necessary in some cases, thus suggesting that $H_t(\cdot)$ remains globally invertible (almost surely in t) under milder conditions. In particular, we conjecture that $H_t(\cdot)$ is globally invertible almost surely in $t \sim \text{Unif}([0,1])$ as long as ρ_0 is the standard Gaussian and ρ_1 is a general mixture of Gaussians. Proving this conjecture remains a challenging open problem.

Acknowledgments. We thank Dr. Dheeraj Nagaraj at Google DeepMind India and Dr. Qiang Liu at UT Austin for engaging in helpful discussions which helped improve the manuscript significantly. We also thank Dr. Adam Klivans and the Institute for Foundations of Machine Learning (IFML) at UT Austin for providing the compute resources. The work was also supported by the NSF grants CCF-2019844, CCF-2505865, 2217069, NIH Award RF1NS121913, and DMS grant 2109155.

References

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching, 2024. URL https://arxiv.org/abs/2406.07507.

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

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport, 2024. URL https://arxiv.org/abs/2407.18163.

Earl A Coddington, Norman Levinson, and T Teichmann. *Theory of ordinary differential equations*. American Institute of Physics, 1956.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

Pablo Groisman and Julio D Rossi. Explosion time in stochastic differential equations with small diffusion. *Electronic Journal of Differential Equations*, 2007: 1–9, 2007.

Shivam Gupta, Aditya Parulekar, Eric Price, and Zhiyang Xun. Improved sample complexity bounds for diffusion model training, 2024. URL https://arxiv.org/abs/2311.13745.

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, 2022.

H. Kunita. Stochastic differential equations and stochastic flows of diffeomorphisms. In P. L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XII - 1982*, pages 143–303, Berlin, Heidelberg, 1984. Springer Berlin Heidelberg. ISBN 978-3-540-39109-8.

Dohyun Kwon, Ying Fan, and Kangwook Lee. Score-based generative modeling secretly minimizes the wasserstein distance. *Advances in Neural Information Processing Systems*, 35:20205–20217, 2022.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows, 2024. URL https://arxiv.org/abs/2405.20320.

Gen Li, Zhihan Huang, and Yuting Wei. Towards a mathematical theory for consistency training in diffusion models. *arXiv preprint arXiv:2402.07802*, 2024a.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*, 2024b.

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

Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport, 2022. URL https://arxiv.org/abs/2209.14577.

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

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2024.

C Lu, Y Zhou, F Bao, J Chen, and C Li. A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Proc. Adv. Neural Inf. Process. Syst., New Orleans, United States*, pages 1–31, 2022.

Wen-Xiu Ma, Xiang Gu, and Liang Gao. A note on exact solutions to linear differential equations by the matrix exponential. *Adv. Appl. Math. Mech*, 1(4): 573–580, 2009.

Wilhelm Magnus. On the exponential solution of differential equations for a linear operator. *Communications on pure and applied mathematics*, 7(4):649–673, 1954.

William Fogg Osgood. Beweis der existenz einer lösung der differentialgleichung dy dx= f\left (x, y\right) ohne hinzunahme der cauchy-lipschitz'schen bedingung. *Monatshefte für Mathematik und Physik*, 9: 331–345, 1898.

Francesco Pedrotti, Jan Maas, and Marco Mondelli. Improved convergence of score-based diffusion models via prediction-correction, 2024. URL https://arxiv.org/abs/2305.14164.

Roy Plastock. Homeomorphisms between banach spaces. *Transactions of the American mathematical society*, 200:169–183, 1974.

Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pages 388–394. Springer, 1992.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Filippo Santambrogio. *Optimal transport for applied mathematicians*. Springer, 2015.

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

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. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. *arXiv preprint arXiv:2405.07510*, 2024.

Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. *Advances in Neural Information Processing Systems*, 36:55502–55542, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable] Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]
2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] Yes

(b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] Yes

(c) Clear explanations of any assumptions. [Yes/No/Not Applicable] Yes

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable] No

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] Yes

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable] Yes

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] No

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable] Not applicable

(b) The license information of the assets, if applicable. [Yes/No/Not Applicable] Not applicable

(c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] Not applicable

(d) Information about consent from data providers/curators. [Yes/No/Not Applicable] Not Applicable

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable] Not Applicable

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] Not Applicable

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] Not Applicable

Appendix

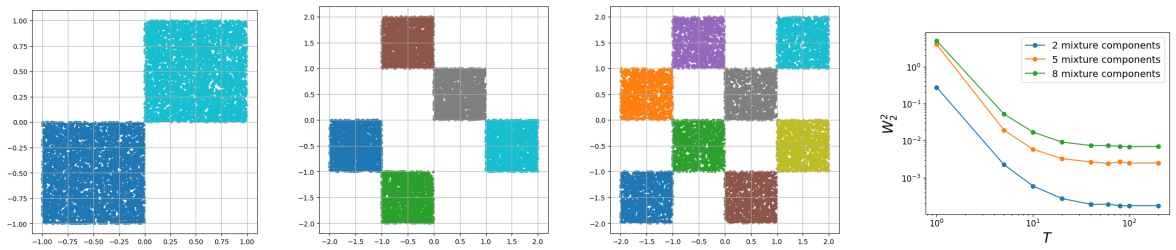
A EXPERIMENTAL DETAILS

A.1 Details of synthetic data experiments with mixture of Gaussian distributions

We choose the target distribution to be mixture of Gaussians with equal cluster probability and unit variance, where the number of components K varies within $\{1, 2, 3, 4\}$. In all the cases, we show that the actual Wasserstein error is closely characterized by $\gamma_{2,T}$. For $K = 1$, we set mean of the target distribution to be $\mu_1 = (5, 0)^\top$; for $K = 2$: $\mu_1 = (5, 0)^\top, \mu_2 = (0, 5)^\top$; for $K = 3$: $\mu_1 = (5, 0)^\top, \mu_2 = (0, 5)^\top, \mu_3 = (-5, 0)^\top$; for $K = 4$: $\mu_1 = (5, 0)^\top, \mu_2 = (0, 5)^\top, \mu_3 = (-5, 0)^\top, \mu_4 = (0, -5)^\top$.

A.2 Checker board example

We consider the checker-board distribution with 2, 5 and 8 components. We use training datasets of size 10,000 to train a feed-forward neural network in order to learn the velocity drift function and evaluate $W_2^2(\widehat{\rho}_1, \rho_1)$ using POT (Feydy et al., 2019) for different levels of discretization T over test data of size 5000. Figure 6(d) also shows that larger component size has a negative effect on the Wasserstein distance, which stems from the fact that a larger number of components typically pushes the flow away from straightness.



(a) (b) (c) (d)

Figure 6: (a) Checker-board distribution with 2 components. (b) Checker-board distribution with 5 components. (c) Checker-board distribution with 8 components. (d) shows the $W_2^2(\widehat{\rho}_1, \rho_1)$ vs T (on log-log scale) for the Checker-board distribution with varying components.

A.3 MNIST Dataset experiment

For MNIST data, we construct 3-different datasets. The first one only contains the digits $\{0, 1, 2\}$, the second one only contains $\{0, 1, 2, \ldots, 6\}$ and the final one contains $\{0, 1, 2, \ldots, 9\}$. Essentially, these datasets contain multiple modes which resembles the nature of the synthetic dataset examples discussed in the previous section. Figure **??**(a) shows that the Wasserstein distance is larger when there is more number of components in the dataset. Essentially, more components make the flow more non-straight, and hence convergence in Wasserstein is affected.

A.4 Additional Large-Scale Real Data Experiment on CelebA

To further examine whether our theoretical findings extend beyond relatively small benchmark datasets such as MNIST and FashionMNIST, we conducted an additional experiment on the CelebA dataset, which contains roughly 200,000 images and is substantially larger and more diverse. This experiment was designed to probe the effect of increasing modal complexity on the discretization error of Rectified Flow, and to test whether the trends predicted by our theory continue to hold at a larger scale.

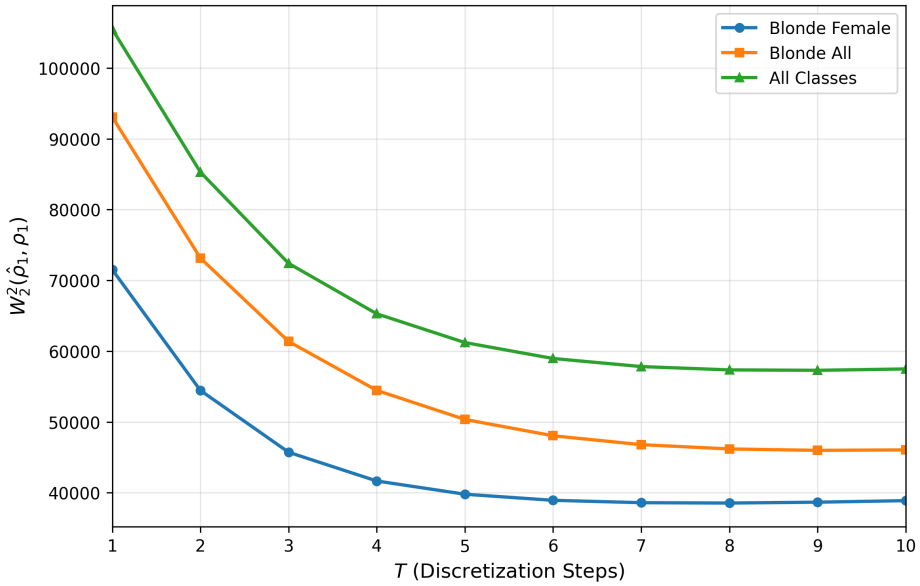


Figure 7: Squared Wasserstein error $W_2^2(\hat{\rho}_1, \rho_1)$ versus the number of discretization steps T on three CelebA subsets with increasing modal complexity. The discretization error decreases rapidly with T and stabilizes around $T \approx 8$. Moreover, for every fixed T, the error increases monotonically as the dataset becomes more multimodal, from *Blonde Female* to *Blonde All* to *All Classes*.

Experimental setup. We trained Rectified Flow models on three CelebA subsets with increasing diversity:

1. **Blonde Female**: a relatively homogeneous subset with the fewest modes,
2. **Blonde All**: a broader subset containing all blonde subjects,
3. **All Classes**: the full dataset subset with the highest diversity and the largest effective number of modes.

This ordering was chosen to approximately vary the underlying mode complexity from low to high while keeping the evaluation protocol fixed across all settings. For each trained model, we evaluated the squared 2-Wasserstein error between the generated distribution and the target data distribution under Euler discretization with varying numbers of sampling steps $T \in \{1, \ldots, 10\}$. The resulting values are reported in Figure 7

Observations. The CelebA experiment reveals two clear trends.

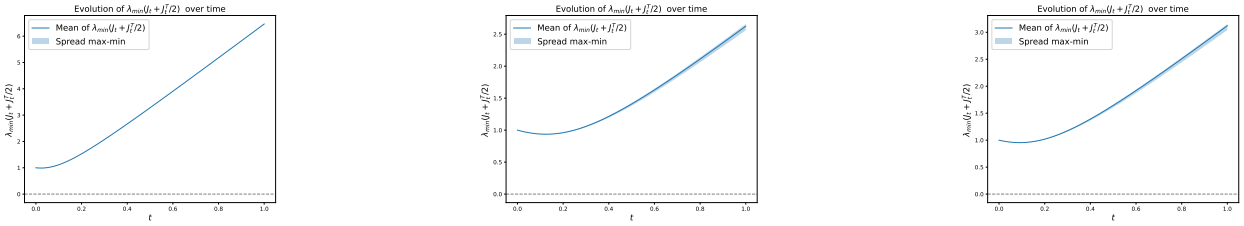
First, for all three subsets, the squared Wasserstein error decreases sharply as the number of discretization steps increases, and then stabilizes around $T \approx 8$. This is consistent with our theoretical analysis in Theorem 2, which predicts that discretization error should decrease as the flow is integrated more finely.

Second, for every fixed value of T, the error is smallest for *Blonde Female*, larger for *Blonde All*, and largest for *All Classes*. Since these three settings are ordered by increasing diversity and effective mode complexity, this provides further evidence that more complex multimodal datasets induce less straight flows and therefore require more discretization steps to achieve the same sampling fidelity.

Implication. These results support the broader relevance of our theory beyond small benchmark datasets. Although training on full ImageNet is outside the scope of the present work, the CelebA results already demonstrate the same qualitative scaling behavior predicted by our straightness-based analysis: increasing mode complexity worsens discretization error, while increasing T compensates for this by better resolving the flow trajectory.

Overall, this experiment strengthens the empirical evidence that the geometry of the learned flow—as captured by our straightness perspective—continues to govern sampling accuracy even in substantially larger and more realistic image datasets.

A.5 Verifying Assumption 2 empricially



(a) (b) (c)

Figure 8: Evolution of $0.5 \times \lambda_{\min}(J_t^{z_0} + J_t^{z_0\top})$ over time for 100 samples $z_0 \sim \mathcal{N}(0, I_2)$, under target distributions $\rho_1 = \sum_{i=1}^{K} \pi_i \mathcal{N}(\mu_i, \sigma^2 I)$ with $K \in \{2, 3, 4\}$, corresponding to subplots (a), (b), and (c), respectively.

Figure 8 shows an empirical verification of Assumption 2. The parameter settings given as follows:

(a) $K = 2$: $\mu_1 = (5, 1)$, $\mu_2 = (-7, -2)$, $\sigma = 6.5$, $\pi_1 = 0.6$, $\pi_2 = 0.4$.

(b) $K = 3$: $\mu_1 = (1, 2)$, $\mu_2 = (2, 0)$, $\mu_3 = (-1, -2)$, $\sigma = 2.5$, $\pi_1 = \pi_2 = 0.4$, $\pi_3 = 0.2$.

(c) $K = 4$: $\mu_1 = (1, 3)$, $\mu_2 = (2, 0)$, $\mu_3 = (-1, -2)$, $\mu_4 = (0, -2)$, $\sigma = 3$, $\pi_1 = 0.3$, $\pi_2 = 0.4$, $\pi_3 = 0.2$, $\pi_4 = 0.1$.

B EXISTENCE AND UNIQUENESS OF RECTIFIED FLOW

Prior works Liu et al. (2023); Liu (2022) assume (also Section 4) that the velocity is Lipschitz smooth, which is a sufficient condition for the existence of a unique solution to ODE (1). However, such conditions could be restrictive in practice. In this section, we will work with somewhat more pragmatic conditions on the true velocity functions v_t. Recall ODE (10), i.e.,

$$dZ_t = v_t(Z_t)\, dt, \quad Z_0 = z_0.$$

Definition 2. *For a positive integer k, a function $f : \mathbb{R}^d \to \mathbb{R}^d$ is said to be $\mathcal{C}^k$ if it is k-times continuously differentiable. Additionally, f is called a $\mathcal{C}^{1,1}$ function if f is a $\mathcal{C}^1$ function and its Jacobian is locally Lipschitz, i.e., for every $x \in \mathbb{R}^d$, there exists $\delta > 0$ and $L_{loc} > 0$ (which may depend on x) such that*

$$\max\{\|x - x_1\|_2, \|x - x_2\|_2\} \le \delta \Rightarrow \|\nabla_x f(x_1) - \nabla_x f(x_2)\|_{op} \le L_{loc} \|x_1 - x_2\|_2 .$$

Assumption 3. *We assume that the velocity function $v_t(\cdot)$ is a $\mathcal{C}^{1,1}$ function for all $t \in [0, 1]$.*

Note that, if $v_t(\cdot)$ is a $\mathcal{C}^2$ function, then it automatically satisfies Assumption 3. This is satisfied as long as the target distribution has the second moment (See Theorem 9). Therefore, we argue that Assumption 3 is less restrictive compared to the global Lipschitzness condition. Now we present a general existence and uniqueness result for Rectified Flow.

Proposition 2 (Existence and Uniqueness)**.** *Let $\mathbb{E}\|X_1\|_2 < \infty$ and the Assumption 3 hold. Also, assume that the solution to the ODE (10) satisfies the non-explosive condition*

$$\sup_{t \in [0,1]} \|Z_t(z_0)\|_2 < \infty \quad \text{for all initial values } z_0 \in \mathbb{R}^d. \tag{12}$$

Then there exists a unique solution to ODE (10).

The above proposition is a consequence of Theorem 5.2 of Kunita (1984). The non-explosivity condition is particularly important for the existence of the ODE as it avoids any singularities in the solution path. However, this condition is hard to verify apriori. However, we provide a sufficient condition for non-explosivity that is easier to check so that Proposition 2 can be of practical use.

Assumption 4 (Osgood type criterion (Osgood, 1898; Groisman and Rossi, 2007)). *Let $Z_t(z_0) \in \mathbb{R}^d$ be the solution of the ODE (10), where $(z_0, t) \in \mathbb{R}^d \times [0,1]$. There exists a non-negative locally-Lipschitz (or strictly increasing) function $h : \mathbb{R}_+ \to \mathbb{R}_+$ such that*

$$\int_{u_0}^{\infty} \frac{1}{h(u)}\, du > 1, \quad \text{for all } u_0 > 0, \tag{13}$$

and $\langle Z_t(z_0), v_t(Z_t(z_0))\rangle \le h(\|Z_t(z_0)\|_2^2)$, for all $(z_0, t) \in \mathbb{R}^d \times [0,1]$.

One sufficient condition is that $\sup_{t\in[0,1]} \langle x, v_t(x)\rangle \le h(\|x\|_2^2)$ for all $x \in \mathbb{R}^d$ and for a positive locally-Lipschitz (or strictly increasing) function h satisfying (13). The above criterion ensures that $\|Z_t(z_0)\|_2$ is always finite for all $t \in [0,1]$ (Groisman and Rossi, 2007), i.e., the solutions does not explode. To be precise, the integral in (13) quantizes the explosion time of $\|Z_t(z_0)\|_2$, and it ensures that the explosion time falls outside $[0,1]$. Moreover, as opposed to condition (12), this can be easily checked for a large class of target distributions, e.g., a general mixture of Gaussians. For example, for $(X_0, X_1) \sim N(0, I_d) \times \rho_1$ with $\rho_1 = \sum_{j=1}^{J} \pi_j N(\mu_j, \Sigma_j)$, it follows that $\sup_{t\in[0,1]} \langle x, v_t(x)\rangle \le A\|x\|_2^2 + B\|x\|_2$ for some $A, B > 0$ (see Appendix G.2). Therefore, $h(u) = Au + B\sqrt{u}$ is a valid choice and it also satisfies Assumption 4, as $\int_{u_0}^{\infty}(Au + B\sqrt{u})^{-1}\, du = \infty$ for all $u_0 > 0$. We now state the main result below.

Theorem 9. *Let $(X_0, X_1) \sim N(0, I_d) \times \rho_1$ such that $\mathbb{E}\|X_1\|_2 < \infty$. Then, the velocity $v_t(\cdot)$ satisfies Assumption 3. Moreover, under Assumption 4, there exists a unique solution to ODE (10).*

The above theorem gives a fairly general existence and uniqueness (without Lipschitz smoothness) result for Rectified Flow starting from an independent coupling that covers a large class of target distributions. Essentially, the first moment ensures that Assumption 3 is satisfied. Therefore, coupled with Assumption 4, the conditions of Proposition 2 are satisfied, and hence, the result follows. The complete proof is deferred to Appendix G.1.

C PROOFS FOR WASSERSTEIN CONVERGENCE BOUNDS

C.1 Proof of Theorem 1

Let $\{\rho_t\}_{t\in[0,1]}$ and $\{\tilde{\rho}_t\}_{t\in[0,1]}$ be distribution of the solution of (1) and (3) respectively. Let π_t be the optimal coupling between ρ_t and $\tilde{\rho}_t$. Therefore, using Corollary 5.25 of Santambrogio (2015), we have

$$\begin{aligned}
\frac{1}{2}\frac{dW_2^2(\rho_t, \tilde{\rho}_t)}{dt} &= \int \langle x - y, v_t(x) - \widehat{v}_t(y)\rangle \; d\pi_t(x,y) \\
&= \int \langle x - y, v_t(x) - \widehat{v}_t(x)\rangle \; d\pi_t(x,y) + \int \langle x - y, \widehat{v}_t(x) - \widehat{v}_t(y)\rangle \; d\pi_t(x,y) \\
&\le \frac{1}{2}\int \|x - y\|_2^2 \; d\pi_t(x,y) + \frac{1}{2}\int \|v_t(x) - \widehat{v}_t(x)\|_2^2 \; d\pi_t(x,y) + \widehat{L}\int \|x - y\|_2^2 \; d\pi_t(x,y) \\
&= (1/2 + \widehat{L})W_2^2(\rho_t, \tilde{\rho}_t) + \frac{b(t)}{2}.
\end{aligned}$$

Solving the above differential inequality leads to the following inequality

$$W_2^2(\rho_\tau, \tilde{\rho}_\tau) \le W_2^2(\rho_0, \tilde{\rho}_0) + e^{1+2\widehat{L}} \int_0^\tau b(t)\, dt.$$

The result follows by noting that $W_2^2(\rho_0, \tilde{\rho}_0) = 0$ and setting $\tau = 1$.

C.2 Comparison of straightness parameters

Lemma 3. *The AS and PWS parameters satisfy $\gamma_{2,T}(\mathcal{Z}) \ge \gamma_1(\mathcal{Z}) \ge S(\mathcal{Z})$. Moreover, $S(\mathcal{Z}) = 0$ if and only if $\gamma_1(\mathcal{Z}) = \gamma_{2,T}(\mathcal{Z}) = 0$.*

Recall that $S(\mathcal{Z}) = \int_0^1 \mathbb{E}\,\|Z_1 - Z_0 - v_t(Z_t)\|_2^2\ dt$. Also, note that $Z_1 - Z_0 = \int_0^1 v_u(Z_u)\ du$. Therefore, we have

$$\begin{aligned}
S(\mathcal{Z}) &= \int_0^1 \mathbb{E}\left\|\int_0^1 [v_u(Z_u) - v_t(Z_t)]\ du\right\|_2^2\ dt \\
&= \int_0^1 \mathbb{E}\left\|\int_0^1 \int_t^u \dot{v}_\tau(Z_\tau)\ d\tau\ du\right\|_2^2\ dt \\
&\le \int_0^1 \mathbb{E}\left[\int_0^1 |t-u| \int_{t\wedge u}^{t\vee u} \|\dot{v}_\tau(Z_\tau)\|_2^2\ d\tau\ du\right]\ dt \\
&\le \int_0^1 \mathbb{E}\int_0^1 \int_0^1 \|\dot{v}_\tau(Z_\tau)\|_2^2\ d\tau\ du \\
&\le \int_0^1 \mathbb{E}\,\|\dot{v}_\tau(Z_\tau)\|_2^2\ d\tau = \gamma_1(\mathcal{Z}).
\end{aligned}$$

Moreover, note that

$$\gamma_1(\mathcal{Z}) = \sum_{i=1}^T (t_i - t_{i-1}).\frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \mathbb{E}\,\|\dot{v}_\tau(Z_\tau)\|_2^2\ d\tau \le \gamma_{2,T}(\mathcal{Z}). \tag{14}$$

This shows that $S(\mathcal{Z}) \le \gamma_1(\mathcal{Z}) \le \gamma_{2,T}(\mathcal{Z})$.

For the second part, first note that the $t_i - t_{i-1} = 1/T$. Therefore,

$$\gamma_1(\mathcal{Z}) = \frac{1}{T}\sum_{i=1}^T \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \mathbb{E}\,\|\dot{v}_\tau(Z_\tau)\|_2^2\ d\tau \ge \frac{\gamma_{2,T}(\mathcal{Z})}{T}.$$

The above inequality along with (14) tells that $\gamma_1(\mathcal{Z}) = 0$ iff $\gamma_{2,T}(\mathcal{Z}) = 0$.

Now, due to the inequality $S(\mathcal{Z}) \le \gamma_1(\mathcal{Z})$, we have $S(\mathcal{Z}) = 0$ if $\gamma_1(\mathcal{Z}) = 0$. For the other direction, let us assume $S(\mathcal{Z}) = 0$. This shows that $v_t(Z_t) = Z_1 - Z_0$ almost surely in t and (Z_0, Z_1). This shows that $\dot{v}_t(Z_t) = 0$ almost surely. Hence the result follows.

C.3 Proof of Theorem 2

Recall that for a given partition $0 = t_0 < t_1 < \ldots < t_T = 1$ of the interval $[0,1]$ of equidistant points $\{t_i\}_{0\le i\le T}$ with $h := T^{-1}$, we follow the Euler discretized version of the of ODE (3) to obtain the sample estimates:

$$\widehat{Y}_{t_i} = \widehat{Y}_{t_{i-1}} + h\widehat{v}_{t_i}(\widehat{Y}_{t_i}), \quad \widehat{Y}_0 = Z_0.$$

Before analyzing the discretization error, we introduce the following interpolation process for $t \in [t_i, t_{i+1}]$ and each $i \in \{0, \ldots, T\}$:

$$\frac{\mathrm{d}}{\mathrm{dt}}\bar{Y}_t = \widehat{v}_{t_i}(\bar{Y}_{t_i}), \quad \bar{Y}_{t_i} = \widehat{Y}_{t_i}. \tag{15}$$

The above ODE flow gives us a continuous interpolation between $\widehat{Y}_{t_i}$ and $\widehat{Y}_{t_{i+1}}$. Coupled with the above flow equation and the ODE flow (4), we have the following *almost sure* differential inequality for $t \in [t_i, t_{i+1}]$:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{dt}}\|Z_t - \bar{Y}_t\|_2^2 &= 2\left\langle Z_t - \bar{Y}_t, \frac{\mathrm{d}}{\mathrm{dt}}Z_t - \frac{\mathrm{d}}{\mathrm{dt}}\bar{Y}_t\right\rangle \\
&= 2\left\langle Z_t - \bar{Y}_t, v_t(Z_t) - \widehat{v}_{t_i}(\bar{Y}_{t_i})\right\rangle \\
&\le \widehat{L}\|Z_t - \bar{Y}_t\|_2^2 + \|v_t(Z_t) - \widehat{v}_{t_i}(\bar{Y}_{t_i})\|_2^2/\widehat{L}
\end{aligned} \tag{16}$$

Multiplying $e^{-\widehat{L}(t-t_i)}$ on both sides of the above inequality and rearranging the terms leads to

$$e^{-\widehat{L}(t-t_i)}\frac{\mathrm{d}}{\mathrm{dt}}\|Z_t-\bar{Y}_t\|_2^2 - e^{-\widehat{L}(t-t_i)}\widehat{L}\|Z_t-\bar{Y}_t\|_2^2 \le e^{-\widehat{L}(t-t_i)}\|v_t(Z_t)-\widehat{v}_{t_i}(\bar{Y}_{t_i})\|_2^2/\widehat{L}.$$
$$\Leftrightarrow \frac{\mathrm{d}}{\mathrm{dt}}\{e^{-\widehat{L}(t-t_i)}\|Z_t-\bar{Y}_t\|_2^2\} \le e^{-\widehat{L}(t-t_i)}\|v_t(Z_t)-\widehat{v}_{t_i}(\bar{Y}_{t_i})\|_2^2/\widehat{L} \le \|v_t(Z_t)-\widehat{v}_{t_i}(\bar{Y}_{t_i})\|_2^2/\widehat{L}.$$
$$\Leftrightarrow \|Z_{t_{i+1}}-\widehat{Y}_{t_{i+1}}\|_2^2 \le e^{\widehat{L}(t_{i+1}-t_i)}\|Z_{t_i}-\widehat{Y}_{t_i}\|_2^2 + \frac{e^{\widehat{L}(t_{i+1}-t_i)}}{\widehat{L}}\int_{t_i}^{t_{i+1}}\|v_t(Z_t)-\widehat{v}_{t_i}(\bar{Y}_{t_i})\|_2^2\,dt.$$

Define $\Delta_i := \mathbb{E}\|Z_{t_i}-\widehat{Y}_{t_i}\|_2^2$. Using the above inequality we have

$$\begin{aligned}
&\Delta_{i+1}\\
&\le e^{\widehat{L}h}\Delta_i + \frac{e^{\widehat{L}h}}{\widehat{L}}\int_{t_i}^{t_{i+1}}\mathbb{E}\|v_t(Z_t)-\widehat{v}_{t_i}(\widehat{Y}_{t_i})\|_2^2\,dt\\
&\le e^{\widehat{L}h}\Delta_i\\
&+\frac{3e^{\widehat{L}h}}{\widehat{L}}\left\{\underbrace{\int_{t_i}^{t_{i+1}}\mathbb{E}\|v_t(Z_t)-v_{t_i}(Z_{t_i})\|_2^2\,dt}_{T_1}+\underbrace{\int_{t_i}^{t_{i+1}}\mathbb{E}\|v_{t_i}(Z_{t_i})-\widehat{v}_{t_i}(Z_{t_i})\|_2^2\,dt}_{T_2}+\underbrace{\int_{t_i}^{t_{i+1}}\mathbb{E}\|\widehat{v}_{t_i}(Z_{t_i})-\widehat{v}_{t_i}(\widehat{Y}_{t_i})\|_2^2\,dt}_{T_3}\right\}.
\end{aligned} \tag{17}$$

Now we will bound each of the last three terms on the right-hand side of the above inequality.

Bounding T_1. For the first term, we have

$$\begin{aligned}
\mathbb{E}\|v_t(Z_t)-v_{t_i}(Z_{t_i})\|_2^2 &= \mathbb{E}\left\|\int_{t_i}^{t}\frac{\mathrm{d}}{\mathrm{d}\tau}v_\tau(Z_\tau)\,d\tau\right\|_2^2\\
&\le (t-t_i)\int_{t_i}^{t}\mathbb{E}\left\|\frac{\mathrm{d}}{\mathrm{d}\tau}v_\tau(Z_\tau)\right\|_2^2 d\tau\\
&\le h^2\gamma_i,
\end{aligned} \tag{18}$$

where $\gamma_i = \frac{1}{t_{i+1}-t_i}\int_{t_i}^{t_{i+1}}\mathbb{E}\|\frac{\mathrm{d}}{\mathrm{d}\tau}v_\tau(Z_\tau)\|_2^2\,d\tau$. This shows that $T_1 \le h^3\gamma_i$.

Bounding T_2. The term T_2 is bounded by $h\varepsilon_{\mathrm{vl}}^2$ as $\mathbb{E}\|v_{t_i}(Z_{t_i})-\widehat{v}_{t_i}(Z_{t_i})\|_2^2 \le \varepsilon_{\mathrm{vl}}^2$ (Assumption 1(a)).

Bounding T_3. For the final term we will use that $\widehat{v}_{t_i}$ is $\widehat{L}$-Lipschitz. This entails that $T_3 \le \widehat{L}^2 h\Delta_i$. Plugging these bounds in the recursion formula (17), we get

$$\Delta_{i+1} \le e^{\widehat{L}h}(1+3\widehat{L}h)\Delta_i + 3e^{\widehat{L}h}(h^3\gamma_i + h\varepsilon_{\mathrm{vl}}^2)/\widehat{L}.$$

Solving the recursion yields

$$\Delta_T \le e^{T\widehat{L}h}(1+3\widehat{L}h)^T\Delta_0 + \frac{3h^3}{\widehat{L}}\left\{\sum_{k=1}^{T}e^{k\widehat{L}h}(1+3\widehat{L}h)^{k-1}\gamma_{T-k}\right\} + \frac{3h}{\widehat{L}}\left\{\sum_{k=1}^{T}e^{k\widehat{L}h}(1+3\widehat{L}h)^{k-1}\right\}\varepsilon_{\mathrm{vl}}^2.$$

Recall that $\gamma_{2,T}(\mathcal{Z}) := \max_k \gamma_k$. Note that $\Delta_0 = 0$ as $Z_0 = \widehat{Y}_0$. Therefore, we have

$$\Delta_T \le \frac{e^{4\widehat{L}}}{\widehat{L}^2}\left(\frac{\gamma_{2,T}(\mathcal{Z})}{T^2} + \varepsilon_{\mathrm{vl}}^2\right).$$

Here we used the fact that

$$\sum_{k=1}^{T}e^{k\widehat{L}h}(1+3\widehat{L}h)^{k-1} \le \frac{e^{4\widehat{L}}-1}{1+3\widehat{L}h-e^{-\widehat{L}h}} \le \frac{e^{4\widehat{L}}}{3\widehat{L}h}.$$

Therefore, we have

$$W_2^2(\widehat{\rho}_1,\rho_1) \le \Delta_T \le \frac{e^{4\widehat{L}}}{\widehat{L}^2}\left(\frac{\gamma_{2,T}(\mathcal{Z})}{T^2} + \varepsilon_{\mathrm{vl}}^2\right).$$

However, the above upper bound explodes for $\widehat{L} \to 0$. Therefore, we handle the case $\widehat{L} < 1$ in a slightly different manner.

Separately handling $\widehat{L} < 1$ case: We recall the decomposition (16). We will only change the last inequality in that decomposition, i.e., for $\alpha > 0$ we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}\|Z_t - \bar{Y}_t\|_2^2 &= 2\left\langle Z_t - \bar{Y}_t, \frac{\mathrm{d}}{\mathrm{dt}}Z_t - \frac{\mathrm{d}}{\mathrm{dt}}\bar{Y}_t \right\rangle \\ &= 2\left\langle Z_t - \bar{Y}_t, v_t(Z_t) - \widehat{v}_{t_i}(\bar{Y}_{t_i})\right\rangle \\ &\leq \alpha\|Z_t - \bar{Y}_t\|_2^2 + \|v_t(Z_t) - \widehat{v}_{t_i}(\bar{Y}_{t_i})\|_2^2/\alpha\end{aligned} \tag{19}$$

Therefore, following exactly similar steps as before, we arrive at the following recursion:

$$\Delta_{i+1} \leq e^{\alpha h}\left(1 + \frac{3\widehat{L}^2 h}{\alpha}\right)\Delta_i + \frac{3e^{\alpha h}}{\alpha}(h^3\gamma_i + h\varepsilon_{\mathrm{vl}}^2).$$

Solving this yields

$$\Delta_T \leq \frac{e^{\alpha + 3\widehat{L}^2/\alpha} - 1}{1 + 3\widehat{L}^2 h/\alpha - e^{-\alpha h}}\left(\frac{3h^3}{\alpha}.\gamma_{2,T}(\mathcal{Z}) + \frac{3h}{\alpha}.\varepsilon_{\mathrm{vl}}^2\right)$$

Note that $e^{\alpha+3\widehat{L}^2/\alpha} - 1 \leq e^{\alpha+3\widehat{L}/\alpha} - 1$ as $\widehat{L} < 1$. Additionally,

$$1 + 3\widehat{L}^2 h/\alpha - e^{-\alpha h} \geq 1 - e^{-\alpha h} \geq \alpha h e^{-\alpha h}.$$

Setting $\alpha = 1$, and using the above inequalities along with the fact that $h \leq 1$, we get

$$\frac{e^{\alpha+3\widehat{L}^2/\alpha} - 1}{1 + 3\widehat{L}^2 h/\alpha - e^{-\alpha h}} \leq \frac{e^{2+4\widehat{L}}}{h}.$$

Finally, using the above inequality we have

$$W_2^2(\widehat{\rho}_1, \rho_1) \leq \Delta_T \leq 27e^{4\widehat{L}}\left(\frac{\gamma_{2,T}(\mathcal{Z})}{T^2} + \varepsilon_{\mathrm{vl}}^2\right).$$

Combining this with previous upper bound we finally get the result.

D PROOFS FOR BOUNDING $\gamma_{2,T}$

Definitions : Let us define the key terms used in this derivation:

- z_t: The state (e.g., data point) at time t.
- $p_t(z_t)$: The probability density function of $X_t = (1-t)X_0 + tX_1$.
- $v_t(z_t)$: The velocity field that transports the probability density.
- $s_t(z_t) = \nabla_{z_t} \log p_t(z_t)$: The score function, which is the gradient of the log-probability density.
- $H_t(z_t) = \nabla_{z_t}^2 \log p_t(z_t)$: The Hessian of the log-probability density.
- $Y = \nabla_{z_t}\mathrm{div}(v_t(z_t))$: The gradient of the divergence of the velocity field.

The velocity field is defined as:

$$v_t(z_t) = \frac{z_t}{t} + \left(\frac{1-t}{t}\right)s_t(z_t) \tag{20}$$

D.1 General expression for $\dot{v}_t(z_t)$(Lemma 1)

Derivation of $\dot{v}_t(z_t)$:

We begin with the material derivative (or total derivative) of the velocity field $v_t(z_t)$, denoted by $\dot{v}_t(z_t)$.

$$\frac{d}{dt}v_t(z_t) = \nabla_{z_t} v_t(z_t) \cdot v_t(z_t) + \frac{\partial v_t(z_t)}{\partial t} \tag{21}$$

Expanding the partial derivative term using the product rule:

$$\begin{aligned}
\frac{\partial v_t}{\partial t} &= \frac{\partial}{\partial t}\left(\frac{z_t}{t} + \left(\frac{1-t}{t}\right) s_t(z_t)\right) \\
&= -\frac{z_t}{t^2} + \frac{\partial}{\partial t}\left(\left(\frac{1}{t} - 1\right) s_t(z_t)\right) \\
&= -\frac{z_t}{t^2} - \frac{1}{t^2} s_t(z_t) + \left(\frac{1-t}{t}\right) \frac{\partial s_t(z_t)}{\partial t}
\end{aligned}$$

Substituting this back gives the full expression for the material derivative:

$$\dot{v}_t(z_t) = \nabla_{z_t} v_t(z_t) \cdot v_t(z_t) - \frac{z_t}{t^2} - \frac{1}{t^2} s_t(z_t) + \left(\frac{1-t}{t}\right) \frac{\partial s_t(z_t)}{\partial t} \tag{22}$$

Next, we find an expression for $\frac{\partial s_t}{\partial t}$. We use the continuity equation, $\frac{\partial p_t}{\partial t} = -\text{div}(p_t v_t)$.

$$\begin{aligned}
\frac{\partial s_t}{\partial t} &= \frac{\partial}{\partial t} \nabla_{z_t} \log p_t = \nabla_{z_t}\left(\frac{\partial}{\partial t} \log p_t\right) = \nabla_{z_t}\left(\frac{1}{p_t}\frac{\partial p_t}{\partial t}\right) \\
&= \nabla_{z_t}\left(-\frac{1}{p_t}\text{div}(p_t v_t)\right) \\
&= \nabla_{z_t}\left(-\frac{1}{p_t}(\nabla_{z_t} p_t \cdot v_t + p_t \text{div}(v_t))\right) \\
&= \nabla_{z_t}\left(-(\frac{\nabla_{z_t} p_t}{p_t} \cdot v_t + \text{div}(v_t))\right) \\
&= \nabla_{z_t}[-s_t \cdot v_t - \text{div}(v_t)] \\
&= -\{H_t \cdot v_t + \nabla_{z_t} v_t \cdot s_t + Y_t\},
\end{aligned}$$

where $Y_t(z_t) = \nabla_{z_t} \text{div}(v_t(z_t))$. Now, we substitute the expression for $\frac{\partial s_t}{\partial t}$ into the equation for $\dot{v}_t$. The derivation follows the specific intermediate steps and cancellations from your notes.

$$\dot{v}_t = \nabla_{z_t} v_t \cdot v_t - \frac{z_t}{t^2} - \frac{s_t}{t^2} - \left(\frac{1-t}{t}\right)\{H_t \cdot v_t + \nabla_{z_t} v_t \cdot s_t + Y_t\}$$

Now substitute $s_t(z_t) = \frac{t v_t(z_t) - z_t}{1-t}$

$$= \cancel{\nabla_{z_t} v_t(z_t) \cdot v_t(z_t)} - \frac{z_t}{t^2} - \left(\frac{1-t}{t}\right)\left\{H_t(z_t) \cdot v_t(z_t) + \nabla_{z_t} v_t(z_t)\left(\frac{\cancel{t v_t(z_t)} - z_t}{1-t}\right) + Y_t(z_t)\right\} - \frac{1}{t^2} s_t(z_t)$$

Now substitute $\nabla_{z_t} v_t(z_t) = \frac{I + (1-t) H_t(z_t)}{t}$.

$$\begin{aligned}
\dot{v}_t(z_t) &= \cancel{-\frac{z_t}{t^2}} - \left(\frac{1-t}{t}\right)\left\{H_t(z_t)\left(\cancel{\frac{z_t}{t}} + \left(\frac{1-t}{t}\right) s_t(z_t)\right) - \left(\frac{\cancel{I + (1-t) H_t(z_t)}}{t}\right)\frac{z_t}{1-t} + Y_t(z_t)\right\} - \frac{s_t(z_t)}{t^2} \\
&= -\left(\frac{1-t}{t}\right)^2 H_t(z_t) s_t(z_t) - \left(\frac{1-t}{t}\right) Y_t(z_t) - \frac{1}{t^2} s_t(z_t)
\end{aligned}$$

Now, we find an expression for $Y_t(z_t)$ and substitute it.

$$\begin{aligned} Y_t(z_t) &= \nabla_{z_t} \mathrm{div}(v(z_t,t)) = \nabla_{z_t} \left(\mathrm{Tr}(\nabla_{z_t} v(z_t,t))\right) \\ &= \nabla_{z_t} \mathrm{Tr}\left(\frac{I}{t} + \left(\frac{1-t}{t}\right) H_t(z_t)\right) \\ &= \left(\frac{1-t}{t}\right) \nabla_{z_t} \mathrm{Tr}(H_t(z_t)) \quad (\text{since } \nabla \mathrm{Tr}(I) = 0) \end{aligned}$$

Substituting this into the simplified expression for $\dot{v}_t$ and rearranging gives the final result:

$$\dot{v}_t(z_t) = -\frac{1}{t^2}\left\{(1-t)^2[H_t(z_t)s_t(z_t) + \nabla_{z_t}\mathrm{Tr}(H_t(z_t))] + s_t(z_t)\right\} \tag{23}$$

D.2 Expression for $\dot{v}_t(z_t)$ for a target mixture of Gaussian distributions

Let $p_1(x) = \sum_{i=1}^K \pi_i \; p_{i,1}(x)$, where $p_{i,1}(x) = N(x \mid \mu_i, \sigma^2 I_d)$ be the target mixture of K Gaussians. Note that the density of $X_t = tX_1 + (1-t)X_0$ is another K-Gaussian mixture given by $p_1(x) = \sum_{i=1}^K \pi_i \; p_{i,t}(x)$, where $p_{i,t}(x) = N(x \mid t\mu_i, \sigma_t^2 I_d)$ and $\sigma_t^2 = t^2\sigma^2 + (1-t)^2$.

Derivation of the Score $s_t(z)$: The score of the mixture is the gradient of its log-density.

$$\begin{aligned} s_t(z) &= \nabla_z \log p_t(z) \\ &= \frac{\nabla_z p_t(z)}{p_t(z)} && \text{(Chain rule for log)} \\ &= \frac{\nabla_z \left(\sum_{i=1}^K \pi_i p_{i,t}(z)\right)}{p_t(z)} && \text{(Substitute definition of } p_t\text{)} \\ &= \frac{\sum_{i=1}^K \pi_i \nabla_z p_{i,t}(z)}{p_t(z)} && \text{(Linearity of gradient)} \end{aligned}$$

We use the identity $\nabla_z p_{i,t}(z) = p_{i,t}(z)\nabla_z \log p_{i,t}(z) = p_{i,t}(z)s_{i,t}(z)$.

$$\begin{aligned} s_t(z) &= \frac{\sum_{i=1}^K \pi_i p_{i,t}(z) s_{i,t}(z)}{p_t(z)} \\ &= \sum_{i=1}^K \frac{\pi_i p_{i,t}(z)}{\cdot p_t(z)} s_{i,t}(z) && \text{(Rearrange terms)} \end{aligned}$$

Recognizing the definition of the weight, $w_{i,t}(z) = \frac{\pi_i p_{i,t}(z)}{\cdot p_t(z)}$, we arrive at the final expression:

$$s_t(z) = \sum_{i=1}^K w_{i,t}(z) s_{i,t}(z) \tag{24}$$

This shows the score of the mixture is the weighted average of the scores of its components.

Derivation of the Hessian $H_t(z)$: The Hessian is the Jacobian of the score vector, $H_t(z) = \nabla_z s_t(z)^T$. We differentiate the expression for $s_t(z)$ using the product rule.

$$\begin{aligned} H_t(z) &= \nabla_z \left(\sum_{i=1}^K w_{i,t}(z) s_{i,t}(z)\right)^T \\ &= \sum_{i=1}^K \nabla_z (w_{i,t}(z) s_{i,t}(z))^T \\ &= \sum_{i=1}^K \left((\nabla_z w_{i,t}(z)) s_{i,t}(z)^T + w_{i,t}(z)(\nabla_z s_{i,t}(z)^T)\right) \end{aligned}$$

First, we find the gradient of the weight $w_{i,t}$.

$$\nabla_z w_{i,t} = \nabla_z \left(\frac{p_{i,t}}{\sum_j p_{j,t}} \right) = \frac{(\nabla_z p_{i,t})(\sum_j p_{j,t}) - p_{i,t}(\sum_j \nabla_z p_{j,t})}{(\sum_j p_{j,t})^2}$$

Substituting $\nabla_z p_{i,t} = p_{i,t} s_{i,t}$ and dividing by $(\sum_j p_{j,t})^2$, we get:

$$\nabla_z w_{i,t} = \frac{p_{i,t} s_{i,t}}{\sum_j p_{j,t}} - \frac{p_{i,t}}{(\sum_j p_{j,t})^2} \sum_j p_{j,t} s_{j,t} = w_{i,t} s_{i,t} - w_{i,t} s_t = w_{i,t}(s_{i,t} - s_t)$$

Now, substitute this back into the expression for the Hessian. Note that $\nabla_z s_{i,t}(z)^T = H_{i,t}(z) = -I_d/\sigma_t^2$.

$$\begin{aligned}
H_t(z) &= \sum_{i=1}^{K} \left(w_{i,t}(s_{i,t} - s_t) s_{i,t}^T + w_{i,t} H_{i,t} \right) \\
&= \left(\sum_i w_{i,t} H_{i,t} \right) + \left(\sum_i w_{i,t} s_{i,t} s_{i,t}^T \right) - s_t \left(\sum_i w_{i,t} s_{i,t}^T \right) \\
&= \left(\sum_i w_{i,t} (-\frac{I_d}{\sigma_t^2}) \right) + \mathbb{E}_{w_t}[s_{\cdot,t} s_{\cdot,t}^T] - s_t s_t^T \\
&= -\frac{I_d}{\sigma_t^2} + \mathrm{Cov}_{w_t}(s_{\cdot,t})(z)
\end{aligned}$$

where $\mathrm{Cov}_{w_t}(s_{\cdot,t})(z) := \mathbb{E}_{w_t}[s_{\cdot,t} s_{\cdot,t}^T] - s_t s_t^T$.

Derivation of $\nabla_z \mathbf{Tr}(H_t(z))$: We start with the trace of the Hessian.

$$\mathrm{Tr}(H_t(z)) = \mathrm{Tr}\left(-\frac{I_d}{\sigma_t^2} \right) + \mathrm{Tr}(\mathrm{Cov}_{w_t}(s_{\cdot,t})) = -\frac{d}{\sigma_t^2} + \mathbb{E}_{w_t}[\|s_{\cdot,t}\|^2] - \|s_t\|^2$$

Now we take the gradient of this expression with respect to z.

$$\nabla_z \mathrm{Tr}(H_t(z)) = \nabla_z \left(-\frac{d}{\sigma_t^2} + \sum_{i=1}^{K} w_{i,t}(z) \|s_{i,t}(z)\|^2 - \|s_t(z)\|^2 \right)$$

The first term is constant and its gradient is zero. We apply the product rule to the second term and the chain rule to the third.

$$\begin{aligned}
\nabla_z \mathrm{Tr}(H_t(z)) &= \sum_{i=1}^{K} \left((\nabla_z w_{i,t}) \|s_{i,t}\|^2 + w_{i,t} (\nabla_z \|s_{i,t}\|^2) \right) - \nabla_z \|s_t\|^2 \\
&= \sum_{i=1}^{K} \left(w_{i,t}(s_{i,t} - s_t) \|s_{i,t}\|^2 + w_{i,t} \left(\frac{-2 s_{i,t}}{\sigma_t^2} \right) \right) - 2 H_t s_t \\
&= \underbrace{\sum_i w_{i,t} s_{i,t} \|s_{i,t}\|^2}_{\mathbb{E}_{w_t}[s_{\cdot,t} \|s_{\cdot,t}\|^2]} - \underbrace{s_t \left(\sum_i w_{i,t} \|s_{i,t}\|^2 \right)}_{\text{This is } s_t \mathbb{E}_{w_t}[\|s_{\cdot,t}\|^2]} - \underbrace{\frac{2}{\sigma_t^2} \sum_i w_{i,t} s_{i,t}}_{\frac{2}{\sigma_t^2} s_t} - 2 H_t s_t
\end{aligned}$$

Grouping the terms gives the final expression:

$$\nabla_z \mathrm{Tr}(H_t(z)) = \mathbb{E}_{w_t}[s_{\cdot,t} \|s_{\cdot,t}\|^2] - s_t \mathbb{E}_{w_t}[\|s_{\cdot,t}\|^2] - \frac{2 s_t}{\sigma_t^2} - 2 H_t s_t \tag{25}$$

Final Simplified Expression for $\dot{v}_t(z_t)$: We now substitute the derived expression for $Y(z)$ into the general formula for $\dot{v}_t(z_t)$. The general formula is:

$$\dot{v}_t(z_t) = -\frac{1}{t^2}\left\{(1-t)^2[H_t(z)s_t(z) + Y(z)] + s_t(z)\right\}$$

Let's first simplify the term inside the square brackets, $H_t s_t + Y$.

$$\begin{aligned} H_t s_t + Y &= H_t s_t + \left(\mathbb{E}_{w_t}[s_{\cdot,t}\|s_{\cdot,t}\|^2] - s_t\mathbb{E}_{w_t}[\|s_{\cdot,t}\|^2] - \frac{2s_t}{\sigma_t^2} - 2H_t s_t\right) \\ &= \mathbb{E}_{w_t}[s_{\cdot,t}\|s_{\cdot,t}\|^2] - s_t\mathbb{E}_{w_t}[\|s_{\cdot,t}\|^2] - \frac{2s_t}{\sigma_t^2} - H_t s_t \end{aligned}$$

This expression can be made more compact by defining the covariance between the random vector $s_{\cdot,t}$ and the random scalar $\|s_{\cdot,t}\|^2$ under the discrete distribution w_t:

$$\mathrm{Cov}_{w_t}(s_{\cdot,t}, \|s_{\cdot,t}\|^2) = \mathbb{E}_{w_t}[s_{\cdot,t}\|s_{\cdot,t}\|^2] - \mathbb{E}_{w_t}[s_{\cdot,t}]\mathbb{E}_{w_t}[\|s_{\cdot,t}\|^2] = \mathbb{E}_{w_t}[s_{\cdot,t}\|s_{\cdot,t}\|^2] - s_t\mathbb{E}_{w_t}[\|s_{\cdot,t}\|^2]$$

This term is a vector that captures the asymmetry (or skewness) in the distribution of component scores. So, the simplified term is:

$$H_t s_t + Y = \mathrm{Cov}_{w_t}(s_{\cdot,t}, \|s_{\cdot,t}\|^2) - \frac{2s_t}{\sigma_t^2} - H_t s_t$$

Substituting this back into the formula for $\dot{v}_t$:

$$\dot{v}_t(z_t) = -\frac{\sigma^2}{\sigma_t^2}s_t - \frac{(1-t)^2}{t^2}\left[\mathrm{Cov}_{w_t}(s_{\cdot,t}, \|s_{\cdot,t}\|^2) - \mathrm{Cov}_{w_t}(s_{\cdot,t})s_t\right] \tag{26}$$

This is the final, simplified analytical expression for the time evolution of the velocity field for a Gaussian Mixture Model. It reveals that the dynamics are driven by the score (mean field), the Hessian-score product (curvature), and a third-order term related to the asymmetry of the mixture components.

D.3 Proof of Lemma 2

For ease of notation, we begin by deriving the expected square norm of the two covariance terms in Equation (26) for general Gaussian mixtures and then adapt it to the interpolating structure of RF.

Lemma 4 (Bound on the Expected Squared Norm of the Hessian)**.** *Let $p(x)$ be a mixture of K Gaussians, given by $p(x) = \sum_{k=1}^K \pi_k\ p_k(x)$, with $p_k(x) = \mathcal{N}(x \mid \mu_k, \Sigma)$, $\Sigma = \sigma^2 I_d$ and $D := \max_{i,j}\|\mu_i - \mu_j\|_2$. Let $s_k(x) = \Sigma^{-1}(\mu_k - x)$ be the score and $\gamma_k(x) = \pi_k p_k(x)/p(x)$ be the posterior weight of the k^{th} mixture component. Then:*

$$\mathbb{E}_x\left[\left\|Cov_{\gamma(x)}(\{s_k(x)\})\right\|_2^2\right] \le \frac{D^4}{\sigma^8}.$$

where $\mathrm{Cov}_{\gamma(x)}(\{s_k(x)\}) := \sum_{k=1}^K \gamma_k(x)s_k(x)s_k(x)^\top - s(x)s(x)^\top$.

Proof. By the property of covariance, we have $\mathrm{Cov}_{\gamma(x)}(\{s_k\}) = \Sigma^{-1}\mathrm{Cov}_{\gamma(x)}(\{\mu_k\})\Sigma^{-1}$. We recall that $\max_{i\neq j}\|\mu_1 - \mu_j\|_2 \le D$. Therefore, we have

$$\begin{aligned} \left\|\mathrm{Cov}_{\gamma(x)}(\{\mu_k\})\right\|_2 &\le \sum_k \gamma_k(x)\left\|\mu_k - m_{\gamma(x)}(\{\mu_k\})\right\|_2^2 \quad \text{where } m_{\gamma(x)}(\{\mu_k\}) := \sum_k \gamma_k(x)\mu_k \\ &\le D^2. \end{aligned}$$

The above bound yields $\left\|\mathrm{Cov}_{\gamma(x)}(\{s_k\})\right\|_2 \le D^2\left\|\Sigma^{-1}\right\|_2^2 = \frac{D^2}{\sigma^4}$, and which gives:

$$\mathbb{E}\left[\|\mathrm{Cov}(\{s_k\})\|_2^2\right] \le \frac{D^4}{\sigma^8}.$$

□

Lemma 5 (D-based Bound on a Third-Order Covariance Vector)**.** *Under the same GMM assumptions as Lemma 1, let* $\mathbf{C}(x) = Cov_{\gamma(x)}(\|s_\cdot(x)\|^2, s_\cdot(x))$*, where* $s_k(x) = \Sigma^{-1}(x - \mu_k)$*. Let* $D = \max_{i,j} \|\mu_i - \mu_j\|$*. The expected squared norm of* $\mathbf{C}(x)$ *is bounded by:*

$$\mathbb{E}\left[\|\mathbf{C}(x)\|_2^2\right] \leq \frac{D^2 d(d+2)}{\sigma^8}$$

Proof. From its definition, the norm of $\mathbf{C}(x)$ can be bounded as:

$$\|\mathbf{C}(x)\|_2 = \left\| \sum_{i,j} \gamma_i \gamma_j \|s_i\|^2 (s_i - s_j) \right\|_2 \leq \sum_{i,j} \gamma_i \gamma_j \|s_i\|^2 \|s_i - s_j\|_2.$$

We introduce a uniform bound on the pairwise distance term:

$$\|s_i - s_j\|_2 = \|\Sigma^{-1}(\mu_j - \mu_i)\|_2 \leq \|\Sigma^{-1}\|_2 \|\mu_j - \mu_i\|_2 \leq \|\Sigma^{-1}\|_2 D$$

Substituting this into the sum and simplifying, using $\sum_j \gamma_j = 1$:

$$\|\mathbf{C}(x)\|_2 \leq \sum_{i,j} \gamma_i \gamma_j \|s_i\|^2 (\|\Sigma^{-1}\|_2 D) = (\|\Sigma^{-1}\|_2 D) \left(\sum_i \gamma_i \|s_i\|^2 \right)$$

Squaring both sides and taking the expectation gives:

$$\mathbb{E}[\|\mathbf{C}(x)\|_2^2] \leq (\|\Sigma^{-1}\|_2 D)^2 \cdot \mathbb{E}\left[\left(\sum_i \gamma_i(x) \|s_i(x)\|^2 \right)^2 \right]$$

We bound the expectation term using Jensen's inequality:

$$\mathbb{E}\left[\left(\sum_i \gamma_i \|s_i\|^2 \right)^2 \right] \leq \mathbb{E}\left[\sum_i \gamma_i \|s_i\|^4 \right] = \sum_i \pi_i \mathbb{E}_{x \sim \mathcal{N}_i}[\|s_i(x)\|^4]$$

For the isotropic case, this sum evaluates to $\frac{d(d+2)}{\sigma^4}$. We also have $\|\Sigma^{-1}\|_2 = 1/\sigma^2$. Substituting these into the main inequality yields the final result:

$$\mathbb{E}[\|\mathbf{C}(x)\|_2^2] \leq \left(\frac{1}{\sigma^2} D \right)^2 \cdot \left(\frac{d(d+2)}{\sigma^4} \right) = \frac{D^2 d(d+2)}{\sigma^8}$$

□

Final bound on $\gamma_{2,T}$: By Gupta et al. (2024, Lemma F.3), we know that $\mathbb{E}\|s_t\|_2^2 = O(d)$, as p_t is mixture of Gaussian distribution (sub-Gaussian). We note that the maximum inter-mean distance at time t is tD, and the variance of each component at time t is $\sigma_t^2 = t^2\sigma^2 + (1-t)^2$. Recall that from the proof of Lemma 4, we have $\left\|\mathrm{Cov}_{\gamma(x)}(\{s_k\})\right\|_2 \leq \frac{D^2}{\sigma^4}$. If we apply Lemma 4 and Lemma 5 on (26), then we get

$$\mathbb{E}\|\dot{v}_t(z_t)\|_2^2 \leq C_\sigma \left(d + D^2 d^2 + D^4 d \right),$$

where $C_\sigma > 0$ is positive constant depending on σ.

Explicit bound on Wasserstein convergence rate

Lemma 6. *Let the target distribution* $\rho_1 = \sum_{k \in [K]} \pi_k N(\mu_k, \sigma^2 I_d)$*, and let* $\max_{i \neq j} \|\mu_i - \mu_j\|_2 \leq D$*. Consider the ODE flow (4) with* T *discretization steps and the true RF drift field given in Lemma 11. Then, the Lipschitz constant of the RF drift is* $\frac{(1+\sigma^2)^2}{\sigma^2}\left(1 + \frac{D^2}{2\sigma^2}\right)$ *and hence, the squared-Wasserstein convergence error scales as* $W_2^2(\widehat{\rho}_1, \rho_1) = \mathcal{O}(d^2/T^2)$.

Proof. To provide bound on the Lipschitz constant, we will simply obtain upper bound on the operator norm of $A_t := \nabla v_{z_t}(z_t, t)$. Following the notations in Lemma 12, we have each of the component variances to be $\Sigma_i = \sigma^2 I$ for all $i \in [K]$. This means that $\Sigma_{i,t} = \sigma_t^2 I$, where $\sigma_t^2 = (1-t)^2 + t^2\sigma^2$. Using (37), we have

$$A_t = \frac{1}{t}\{I - \frac{(1-t)}{\sigma_t^2} I\} + \frac{t(1-t)}{2\sigma_t^4} \underbrace{\sum_{i \neq j} w_{i,t} w_{j,t} (\mu_i - \mu_j)(\mu_i - \mu_j)^\top}_{B_t},$$

Now, we have

$$\|B_t\|_{op} \leq \frac{1}{\sigma_t^4} \times \max_{i \neq j} \|\mu_i - \mu_j\|_2^2 \times (\sum_{i \neq j} w_{i,t} w_{j,t}) \leq \frac{1}{\sigma_t^4} \times \max_{i \neq j} \|\mu_i - \mu_j\|_2^2.$$

The last inequality follows from the fact that $\sum_{i \neq j} w_{i,t} w_{j,t} = 1 - (\sum_{k \in [K]} w_{k,t}^2) \leq 1$. Recall that $\Sigma = \sigma^2 I_d$, and this entails that $\Sigma_t = \sigma_t^2 I_d$, where $\sigma_t^2 := (1-t)^2 + t^2\sigma^2$. Therefore, using the above expressions we have

$$\|A_t\|_{op} \leq \frac{|(1+\sigma^2)t - 1|}{\sigma_t^2} + \frac{D^2}{2\sigma_t^4}.$$

Next, we note that $\max_{t \in [0,1]} |(1+\sigma^2)t - 1| = \max\{\sigma^2, 1\} \leq 1 + \sigma^2$. This is due to the fact that $|(1+\sigma^2)t - 1|$ is strictly decreasing between $t \in [0, \frac{1}{1+\sigma^2}]$ and strictly increasing when $t \in [\frac{1}{1+\sigma^2}, 1]$.

Also, we define $\sigma_*^2 := \min_{t \in [0,1]} \sigma_t^2$. An elementary calculus shows that $\sigma_*^2 = \frac{\sigma^2}{1+\sigma^2}$. Therefore, we have

$$\|A_t\|_{op} \leq \frac{\max\{\sigma^2, 1\}}{\sigma_*^2} + \frac{D^2}{2\sigma_*^4} = \frac{(1+\sigma^2)^2}{\sigma^2} + \frac{D^2(1+\sigma^2)^2}{2\sigma^4} = \frac{(1+\sigma^2)^2}{\sigma^2}\left(1 + \frac{D^2}{2\sigma^2}\right)$$

Hence the Lipschitz constant is $\frac{(1+\sigma^2)^2}{\sigma^2}\left(1 + \frac{D^2}{2\sigma^2}\right)$. Now, the Wasserstein result is a direct application of Theorem 2 and Lemma 2 in conjunction, since the estimation error is zero. □

E PROOFS FOR STRAIGHTNESS OF 2-RF

E.1 Proof of Theorem 3

Let $X_0 \sim \mathcal{N}(0, I)$ and $X_1 \sim \mathcal{N}(\mu, \Sigma)$. Let $\Sigma_t = t^2\Sigma + (1-t)^2 I$. Then we have that $X_t \sim \mathcal{N}(t\mu, \Sigma_t)$,. Let the density of X_t be ξ_t and the score $s_t(x) = \nabla_x \log \xi_t(x) = \Sigma_t^{-1}(t\mu - x)$. Therefore, by using (8), the drift is given by:

$$\begin{aligned} v(x,t) &= \frac{x}{t} + \frac{1-t}{t}\Sigma_t^{-1}(t\mu - x) \\ &= (1-t)\Sigma_t^{-1}\mu + \frac{1}{t}\left(I - (1-t)\Sigma_t^{-1}\right)x \end{aligned}$$

Therefore, we have $\nabla_x v(x,t) = \frac{1}{t}\left(I - (1-t)\Sigma_t^{-1}\right)$. This shows that commutativity condition in Theorem 8 is satisfied and the $z_0 \mapsto Z_1(z_0)$ is the Monge map. To obtain the exact form, we want to solve the following ODE:

$$\frac{dZ_t}{dt} - \frac{1}{t}\left(I - (1-t)\Sigma_t^{-1}\right) Z_t = (1-t)\Sigma_t^{-1}\mu; \quad Z_0 = z_0 \tag{27}$$

Now we look at the structure of $I - (1-t)\Sigma_t^{-1}$. Let the eigendecomposition of $\Sigma = U\Lambda U^\top$. We will assume Σ is full rank. So,

$$I - (1-t)\Sigma_t^{-1} = U\Lambda_t U^\top$$

where $\frac{\Lambda_t}{t} = \frac{1}{t}\{I - (1-t)(t^2\Lambda + (1-t)^2 I)^{-1}\}$. This can also be written as:

$$\lambda_{t,i} = \frac{1}{t}\left\{1 - \frac{1-t}{t^2\lambda_i + (1-t)^2}\right\} = \frac{t(1+\lambda_i) - 1}{t^2\lambda_i + (1-t)^2}.$$

Substituting this into Equation (27), we have:

$$\frac{dZ_t}{dt} - U\text{diag}(\lambda_{t,1}, \ldots, \lambda_{t,d})U^\top Z_t = (1-t)\Sigma_t^{-1}\mu.$$

So, we first get the integrating factors of each eigenvalue.

$$I_i(t) = \frac{1}{\sqrt{(1+\lambda_i)t^2 - 2t + 1}}$$

Multiplying $U\text{diag}(I_1(t), \ldots, I_d(t))U^\top$ on both sides of the ODE and then solving we get:

$$U\Lambda_t' U^\top Z_t = U\Lambda_t'' U^\top \mu + \text{constant}$$

where $\lambda'_{t,i} = \frac{1}{\sqrt{(1+\lambda_i)t^2-2t+1}}$ and $\lambda''_{t,i} = \frac{t}{\sqrt{(1+\lambda_i)t^2-2t+1}}$

This yields,

$$\begin{aligned} &\Sigma^{-1/2} Z_1(z_0) - z_0 = \Sigma^{-1/2}\mu \\ &\Rightarrow Z_1(z_0) = \Sigma^{1/2} z_0 + \mu. \end{aligned} \tag{28}$$

This finishes the proof.

E.2 Proof of Theorem 4

We point out that straightness of 1-RF can be obtained via an intuitive argument in this case. First note that straightness of RF is invariant under rotation. Under a proper rotation, the d dimensional target distribution can be reduced to another where the means of the two components of the Gaussian mixture are sparse with two non-zero coefficients each, one of which is equal (lets say coordinate 1). Due to this it follows that ODE (10) gets decoupled and it can be analyzed coordinate-wise. So, essentially the d-dimensional problem gets reduced to one-dimensional case and the result follows from Proposition 3. We elaborate more on this below.

Intuitive proof of straightness:

Proof. Let $\tilde{X}_0, \tilde{X}_1 \in \mathbb{R}^d$ for $d \geq 2$, where $\tilde{X}_0 \sim \mathcal{N}(0, I)$ and $\tilde{X}_1 \sim \sum_{i=1}^2 \pi_i \mathcal{N}(\tilde{\mu}_i, \sigma^2 I)$ with $\sigma^2 = 1$ (for simplicity). We start with the matrix $\tilde{M} = \begin{bmatrix}\tilde{\mu}_1 & \tilde{\mu}_2\end{bmatrix}$ and perform a QR decomposition: $\tilde{M} = \tilde{Q}\tilde{R}$, where $\tilde{Q} \in \mathbb{R}^{d\times 2}$ is an orthonormal matrix that spans the subspace of $\tilde{\mu}_1$ and $\tilde{\mu}_2$.

Next, we extend $\tilde{Q}$ to a complete orthonormal basis for $\mathbb{R}^d$ using $\tilde{Q}' \in \mathbb{R}^{d\times(d-2)}$, which spans the orthogonal complement of the column space of $\tilde{Q}$. We define $Q = \begin{bmatrix}\tilde{Q} & \tilde{Q}'\end{bmatrix}^\top$. This projection guarantees that:

$$Q\tilde{\mu}_1 = (x_1, y_1, 0, \ldots, 0)^\top, \quad Q\tilde{\mu}_2 = (x_2, y_2, 0, \ldots, 0)^\top$$

i.e., only the first two components are non-zero.

To equalize one of the components, we apply a rotation matrix $R(\theta) \in \mathbb{R}^{d\times d}$, which rotates the first two components while leaving the others unchanged:

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & I_{d-2} \end{bmatrix}$$

We set θ as:

$$\theta = \tan^{-1}\left(\frac{y_2 - y_1}{x_1 - x_2}\right)$$

This ensures that the second components of $R(\theta)Q\tilde{\mu}_1$ and $R(\theta)Q\tilde{\mu}_2$ are identical.

Finally, we define the overall transformation as $P = R(\theta)Q$. This matrix $P \in \mathbb{R}^{d\times d}$ is orthonormal (and hence, invertible) since it is the product of two orthonormal matrices. The transformation P, not only makes the last $d-1$ coordinates of the means identical but also reduces the effective dimension of the flow to two.

Now, we rotate our space using the linear transformation P and obtain the distributions $X_0 = P\tilde{X}_0 \sim \mathcal{N}(0, I)$ and $X_1 = P\tilde{X}_1 \sim \sum_{i=1}^{2} \pi_i \mathcal{N}(\mu_i, \Sigma)$, where $\mu_i = P\tilde{\mu}_i$, $\Sigma = P\tilde{\Sigma}P^\top = I$. Also note that by the above construction of the transformation P, $\mu_{1,k} = \mu_{2,k} := c_k$. for all $k \in [d] \setminus \{1\}$. We first show that $(Z_0, Z_1) = \text{Rectify}(X_0, X_1)$ is straight and then argue that an invertible transformation does not hamper straightness.
To proceed, we apply the Rectify procedure on (X_0, X_1) and obtain the following ODE:

$$v_t(Z_t) = \frac{dZ_t}{dt} = \frac{(2t-1)Z_t}{\sigma_t^2} + \frac{1-t}{\sigma_t^2} \sum_{i=1}^{2} w_i(Z_t)\mu_i$$

For $k \in [d] \setminus \{1\}$, we have that

$$\frac{dZ_{t,k}}{dt} = \frac{(2t-1)Z_{t,k}}{\sigma_t^2} + c_k$$

Hence, using (28) the final mapping is just a translation given by $Z_{1,k} = Z_{0,k} + c_k$. However, for the first co-ordinate, for $g_t(Z_{t,1}) = \log\left(\frac{\pi_2}{\pi_1}\right) - \frac{1}{2\sigma_t^2}\left((Z_{t,1} - t\mu_{2,1})^2 - (Z_{t,1} - t\mu_{1,1})^2\right)$, we have

$$\frac{dZ_{t,1}}{dt} = \underbrace{\frac{(2t-1)Z_{t,1}}{\sigma_t^2} + \frac{1-t}{\sigma_t^2}\left(\frac{\mu_{1,1} + \mu_{2,1}\exp\left(g_t(Z_{t,1})\right)}{1 + \exp\left(g_t(Z_{t,1})\right)}\right)}_{v_1(Z_{t,1},t)}$$

Now using (37), it is easily verifiable that $\nabla v_1(Z_{t,1}, t)$ is bounded, i.e., $v_1(Z_{t,1}, t)$ is Lipschitz. Therefore, $Z_{t,1}(\cdot)$ is an increasing function. As a result $z_0 \mapsto Z_1(z_0)$ is co-ordinate wise increasing function and $H_t(z_0) := (1-t)z_0 + tZ_1(z_0)$ is an invertible map. Therefore, straightness of the resulting coupling follows. □

E.3 Proof of Theorem 5

Proof. Consider $\boldsymbol{\mu}_{01} = (0, a)^\top, \boldsymbol{\mu}_{02} = (0, -a)^\top$ and $\boldsymbol{\mu}_{11} = (a, a)^\top, \boldsymbol{\mu}_{12} = (a, -a)^\top$ for some $a > 0$. Let

$$X_0 \sim 0.5\mathcal{N}(\boldsymbol{\mu}_{01}, I) + 0.5\mathcal{N}(\boldsymbol{\mu}_{02}, I), \quad X_1 \sim 0.5\mathcal{N}(\boldsymbol{\mu}_{11}, I) + 0.5\mathcal{N}(\boldsymbol{\mu}_{12}, I).$$

In this case, the velocity functions in x and y-direction for 1-rectification turns out to be

$$u_t(x) = \frac{(2t-1)\,x}{\sigma_t^2} + \frac{(1-t)a}{\sigma_t^2},$$

$$v_t(y) = \frac{(2t-1)\,y}{\sigma_t^2} + \frac{a}{\sigma_t^2} \cdot \frac{\exp\left(-\frac{(y-a)^2}{2\sigma_t^2}\right)(1-2t) - \exp\left(-\frac{(y+a)^2}{2\sigma_t^2}\right)(1-2t) + \exp\left(-\frac{(y\ -\ (2t-1)a)^2}{2\sigma_t^2}\right) - \exp\left(-\frac{(y\ +\ (2t-1)a)^2}{2\sigma_t^2}\right)}{\exp\left(-\frac{(y-a)^2}{2\sigma_t^2}\right) + \exp\left(-\frac{(y+a)^2}{2\sigma_t^2}\right) + \exp\left(-\frac{(y\ -\ (2t-1)a)^2}{2\sigma_t^2}\right) + \exp\left(-\frac{(y\ +\ (2t-1)a)^2}{2\sigma_t^2}\right)}.$$

Next, we will take the derivative of $v_t(y)$ with respect to y. For notational brevity, let us define

$$e_1(y) = \exp\left(-\frac{(y-a)^2}{2\sigma_t^2}\right)(1-2t),$$

$$e_2(y) = \exp\left(-\frac{(y+a)^2}{2\sigma_t^2}\right)(1-2t),$$

$$e_3(y) = \exp\left(-\frac{(y-a(2t-1))^2}{2\sigma_t^2}\right),$$

$$e_4(y) = \exp\left(-\frac{(y+a(2t-1))^2}{2\sigma_t^2}\right).$$

Then we have

$$\left|\frac{dv_t(y)}{dy}\right| \leq \frac{2t-1}{\sigma_t^2} + \frac{a^2}{\sigma_t^4} \cdot \frac{4\{e_1(y)e_2(y) + e_2(y)e_3(y) + e_3(y)e_4(y) + e_4(y)e_1(y)\}}{(\sum_{j=1}^4 e_j(y))^2} \leq 2 + 4a^2.$$

We used the basic inequalities $4(ab+bc+cd+da) \leq (a+b+c+d)^2$ and $\sigma_t^2 \geq 1/2$ in the last step of the above display.

This shows that $v_t(y)$ is uniformly Lipschitz. This entails that the map $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ that sends y_0 to a point $y_1 \in \mathbb{R}$, and defined through the ODE

$$\frac{\mathrm{d}}{\mathrm{dt}} Y_t = v_t(Y_t);\ Y_0 = y_0,$$

is an injective map due to the uniqueness of the solution of the above ODE. Also, we denote by $Y_t^{y_0}$ the solution of the above ODE.

To show the strict increasing property of $\mathcal{T}$, let us consider the same ODE with $Y_0 = \tilde{y}_0 < y_0$. We also consider the solution $Y_t^{\tilde{y}_0}$. Consider the function $L_t := Y_t^{y_0} - Y_t^{\tilde{y}_0}$, which is also continuous in $t \in [0,1]$. To prove increasing property, it is enough to show that $L_1 > 0$. Let us assume that $L_1 \leq 0$. We already know $L_0 > 0$, and hence by Intermediate Value Property, we have there exists a $\tau \in (0,1]$ such that $L_\tau = 0$. This entails that there exists $y_\tau \in \mathbb{R}$ such that $Y_\tau^{y_0} = Y_\tau^{\tilde{y}_0} = y_\tau$. This shows that we have two different solutions of the ODE passing through (τ, y_τ), which is a contradiction. This proves the coveted strict increasing property of $\mathcal{T}$. Hence, *we have a straight coupling* by similar argument as in previous section. □

E.4 Proof of Theorem 6

Recall that we need to show that $\mathbb{E}(Z_1 - Z_0 \mid tZ_1 + (1-t)Z_0) = Z_1 - Z_0$ almost surely in t, Z_0. Therefore, it suffices to show that the function $H_t(z_0) := (1-t)z_0 + tZ_1(z_0)$ is an invertible map. We will equivalently show that $H_t(\cdot)$ is locally inveritible and a proper function Plastock (1974).

H_t is locally invertible : Note that $\nabla_{z_0} H_t(z_0) = (1-t)I_d + tJ_1^{z_0}$. Therefore, due to Assumption 2 we can conclude that

$$u^\top \nabla_{z_0} H_t(z_0) u \geq (1-t) \quad \text{for all } t \in [0,1] \text{ and } u \in \{w \in \mathbb{R}^d \mid \|w\|_2 = 1\}.$$

Specifically, $\nabla_{z_0} H_t(z_0)$ is invertible if $t < 1$. This shows that H_t is locally invertible for all z_0, as long as $t < 1$.

H_t is proper : We will show that $\|H_t(z_0)\|_2 \to \infty$ as $\|z_0\|_2 \to \infty$ for $t < 1$. Let us fix $z_1 \in \mathbb{R}^d$ and define the function $h_t(\lambda) = \langle H_t(\lambda z_0 + \bar{\lambda} z_1), z_0 - z_1 \rangle$, where $\lambda \in [0,1]$ and $\bar{\lambda} = 1 - \lambda$. By the Taylor's formula we have the following for some $\tilde{\lambda} \in (0,1)$:

$$\begin{aligned}
&h_t(1) = h_t(0) + (z_0 - z_1)^\top \nabla H_t(\tilde{z}_\lambda)(z_0 - z_1) \qquad ; \tilde{z}_\lambda = \tilde{\lambda} z_0 + (1-\tilde{\lambda}) z_1 \\
&\Rightarrow \langle H_t(z_0) - H_t(z_1), z_0 - z_1 \rangle \geq (1-t)\|z_0 - z_1\|_2^2 \\
&\Rightarrow \frac{\|H_t(z_0) - H_t(z_1)\|_2^2}{2(1-t)} \geq \frac{(1-t)}{2}\|z_0 - z_1\|_2^2 \quad \left(\text{Young's inequality: } ab \leq \frac{a^2}{2\eta} + \frac{\eta b^2}{2}\right)
\end{aligned}$$

The final inequality shows that $\lim_{\|z_0\|_2 \to \infty} \|H_t(z_0)\|_2 = \infty$.

H_t is globally invertible : H_t is locally invertible and proper. Then H_t is globally invertible due to Corollary 2.1 of Plastock (1974).

Straightness : This shows that $H_t(Z_0)$ is invertible almost surely in $Z_0 \sim N(0, I_d)$ for all $t \in [0,1)$. Then, we have

$$\begin{aligned}
V(Z_0, Z_1) &:= \mathbb{E}_{Z_0,t} \|Z_1 - Z_0 - \mathbb{E}\{Z_1 - Z_0 \mid H_t(Z_0)\}\|_2 \\
&= \int_0^1 \mathbb{E}_{Z_0}\Big[\|Z_1 - Z_0 - \mathbb{E}\{Z_1 - Z_0 \mid H_t(Z_0)\}\|_2\Big]\, dt \qquad (29) \\
&= 0.
\end{aligned}$$

Therefore, (Z_0, Z_1) is a straight coupling.

E.5 Proof of Theorem 7

We will show that Assumption 2 is satisfied in this case. For notational convenience we drop the superscript z_0 in $J_t^{z_0}$ and denote it by J_t. We start with the ODE

$$\frac{\mathrm{d}}{\mathrm{dt}} J_t = \nabla v(Z_t(z_0), t) J_t, \quad J_0 = I_d$$

Let $U_t = J_t^{-1}$. Then, elementary calculation shows that

$$\frac{\mathrm{d}}{\mathrm{dt}} U_t = -J_t^{-1} \dot{J}_t J_t^{-1} = -U_t \nabla v(Z_t(z_0), t).$$

Using (37), we get

$$A_t := \nabla v(Z_t(z_0), z_0) = \frac{1}{t}\left[I_d - (1-t)\Sigma_t^{-1}\right] + \frac{t(1-t)}{2} \underbrace{\sum_{i \neq j} w_{i,t} w_{j,t} \Sigma_t^{-1}(\mu_i - \mu_j)(\mu_i - \mu_j)^\top \Sigma_t^{-1}}_{B_t},$$

Recall that $\Sigma = \sigma^2 I_d$, and this entails that $\Sigma_t = \sigma_t^2 I_d$, where $\sigma_t^2 := (1-t)^2 + t^2\sigma^2$. Therefore,

$$\|B_t\|_{op} \leq \frac{1}{\sigma_t^4} \times \underbrace{\max_{i \neq j} \|\mu_i - \mu_j\|_2^2}_{=:D}.$$

Bounding $\|U_t\|_{op}$: Take any unit vector u. Since A_t is symmetric,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{dt}} u^T U_t U_t^T u &= -2u^T U_t A_t U_t^T u \\
&= -\frac{2}{t} u^\top U_t \left[I_d - (1-t)\Sigma_t^{-1}\right] U_t^\top u - \underbrace{t(1-t) u^\top U_t B_t U_t^\top u}_{\geq 0} \\
&\leq -2\frac{(1+\sigma^2)t - 1}{\sigma_t^2} \left(u^\top U_t U_t^\top u\right)
\end{aligned}$$

Hence for any unit vector u,

$$u^T U_t U_t^\top u \leq \frac{1}{\sigma_t^2} \Rightarrow \|U_t U_t^\top\|_{op} \leq \frac{1}{\sigma_t^2}$$

Hence

$$\|U_t\|_{op}^2 \leq \|U_t U_t^T\|_{op} \leq \frac{1}{\sigma_t^2}.$$

Lower bounding on $u^T U_t u$: Assuming $\|u\|_2 = 1$, we consider the evolution:

$$\frac{\mathrm{d}}{\mathrm{dt}} \left(u^T U_t u\right) = -\frac{(1+\sigma^2)t - 1}{\sigma_t^2} u^T U_t u - \frac{t(1-t)}{2} u^T U_t B_t u.$$

Using the bound:

$$|u^T U_t B_t u| \leq \|U_t\|_{op} \|B_t\|_{op} \leq \frac{D}{\sigma_t^5},$$

we get:

$$\frac{\mathrm{d}}{\mathrm{dt}}\left(u^T U_t u\right) \geq -\frac{(1+\sigma^2)t-1}{\sigma_t^2} u^T U_t u - \frac{t(1-t)}{2\sigma_t^5} D.$$

Define:

$$I(t) = \int_0^t \frac{(1+\sigma^2)s-1}{\sigma_s^2} ds = \frac{1}{2}\log(\sigma_t^2). \Rightarrow e^{I(t)} = \sigma_t$$

Multiplying by the integrating factor, we obtain:

$$\frac{\mathrm{d}}{\mathrm{dt}}\left(e^{I(t)} u^T U_t u\right) \geq -e^{I(t)} \cdot \frac{t(1-t)}{2\sigma_t^5} D = -\frac{t(1-t)}{2\sigma_t^4} D.$$

Integrating both sides, we obtain:

$$\sigma u^T U_1 u \geq 1 - \frac{D}{2}\int_0^1 \frac{s(1-s)}{\sigma_s^4} ds = 1 - \frac{D}{4\sigma^2}.$$

Therefore, if $D \leq 4\sigma^2$ then $u^T U_1 u \geq 0$ for all unit vector u. This ensures that Assumption 2 is satisfied as Lemma 8 yields

$$\lambda_{min}(J_1 + J_1^T) = 2 \min_{u:\|u\|_2=1} u^T J_1 u \geq 2\left(\min_{u:\|u\|_2=1} u^\top U_1 u\right) \times \lambda_{min}(J_1^T J_1) \geq 0.$$

Moreover, Assumption 4 is automatically satisfied (see Section G.2). Therefore, the straightness of 1-RF follows from Theorem 6.

E.6 A general version of Theorem 7

In this section, we present a slightly general version of Theorem 7 as follows.

Theorem 10. *Let $(X_0, X_1) \sim N(0, I_d) \otimes \rho_1$ where $\rho_1 := \sum_{j=1}^K \pi_j N(\mu_j, \Sigma)$ with mixture proportions $\{\pi_j\}_{j\in[K]} \in (0,1)^K$. Let m, M be minimum and maximum eigenvalues of $\Sigma^{1/2}$ respectively, and $\kappa := M/m$ be its condition number. If $\max_{i\neq j} \|\mu_i - \mu_j\|_2^2 \leq 2m^2(3-\kappa^2)$, then 1-RF yields a straight coupling.*

Proof. Our target distribution is $\pi_1 = \sum_{j=1}^K \pi_j N(\mu_j, \Sigma)$. WLOG, we can assume that

$$\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2).$$

If Σ is not diagonal, the let $\Sigma = P\Lambda P^\top$ be the spectral decomposition of Σ, where Λ is diagonal matrix and $PP^\top = I_d$. Now, recall that straightness is invariant under rotation. Therefore, we can always restrict ourselves to $P_\#^\top \rho_0 = N(0, I_d)$ and $P_\#^\top \rho_1 = \sum_{j=1}^K \pi_j N(P^\top \mu_j, \Lambda)$. If RF leads to straight coupling in the rotated frame, the it also does the same in the un-rotated one.

Keeping this in mind, we start with the ODE

$$\frac{\mathrm{d}}{\mathrm{dt}} J_t = \nabla v(Z_t(z_0), t) J_t, \quad J_0 = I_d.$$

Let $U_t = J_t^{-1}$. Then, elementary calculation shows that

$$\frac{\mathrm{d}}{\mathrm{dt}} U_t = -J_t^{-1} \dot{J}_t J_t^{-1} = -U_t \nabla v(Z_t(z_0), t).$$

We know

$$A_t := \nabla v(Z_t(z_0), z_0) = \frac{1}{t}\left[I_d - (1-t)\Sigma_t^{-1}\right] + \frac{t(1-t)}{2}\underbrace{\sum_{i\neq j} w_{i,t}w_{j,t}\Sigma_t^{-1}(\mu_i-\mu_j)(\mu_i-\mu_j)^\top\Sigma_t^{-1}}_{B_t},$$

Recall that $M = \max_{i\in[d]}\sigma_i$ and $m = \min_{i\in[d]}\sigma_i$. We also define $M_t^2 := (1-t)^2 + t^2M^2$ and $m_t^2 := (1-t)^2 + t^2m^2$. Therefore,

$$\|B_t\|_{op} \leq \frac{1}{m_t^4} \times \underbrace{\max_{i\neq j}\|\mu_i-\mu_j\|_2^2}_{=:D}.$$

Bounding $\|U_t\|_{op}$: Take any unit vector u. Since A_t is symmetric,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{dt}}u^TU_tU_t^Tu &= -2u^TU_tA_tU_t^Tu \\
&= -\frac{2}{t}u^\top U_t\left[I_d-(1-t)\Sigma_t^{-1}\right]U_t^\top u - \underbrace{t(1-t)u^\top U_tB_tU_t^\top u}_{\geq 0} \\
&= u^\top U_t\,\mathrm{diag}\left(-2\frac{(1+\sigma_1^2)t-1}{(1-t)^2+t^2\sigma_1^2},\ldots,-2\frac{(1+\sigma_d^2)t-1}{(1-t)^2+t^2\sigma_d^2}\right)U_t^\top u - \underbrace{t(1-t)u^\top U_tB_tU_t^\top u}_{\geq 0} \\
&\leq -2\frac{(1+m^2)t-1}{m_t^2}\left(u^\top U_tU_t^\top u\right)
\end{aligned}$$

The last inequality is true because the function $f_t(a) := -\frac{(1+a)t-1}{(1-t)^2+t^2a}$ is non-increasing over $a>0$ as $\frac{\partial f_t(a)}{\partial a} = -\frac{t(1-t)}{((1-t)^2+t^2a)^2} < 0$. Hence, the following holds for any unit vector u:

$$u^TU_tU_t^\top u \leq \frac{1}{m_t^2} \Rightarrow \|U_tU_t^\top\|_{op} \leq \frac{1}{m_t^2}.$$

Hence,

$$\|U_t\|_{op}^2 \leq \|U_tU_t^T\|_{op} \leq \frac{1}{m_t^2}. \tag{30}$$

Lower bounding on u^TU_tu : Assuming $\|u\|_2 = 1$, we again consider the ODE:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{dt}}\left(u^TU_tu\right) &= -\frac{1}{t}u^\top U_t\left[I_d-(1-t)\Sigma_t^{-1}\right]u - \frac{t(1-t)}{2}u^TU_tB_tu \\
&= u^\top U_t[f_t(m^2)I_d + \mathrm{diag}\left(f_t(\sigma_1^2)-f_t(m^2),\ldots,f_t(\sigma_d^2)-f_t(m^2)\right)]u - \frac{t(1-t)}{2}u^TU_tB_tu \\
&= -\frac{(1+m^2)t-1}{m_t^2}u^\top U_tu + u^\top U_t\mathrm{diag}\left(f_t(\sigma_1^2)-f_t(m^2),\ldots,f_t(\sigma_d^2)-f_t(m^2)\right)u - \frac{t(1-t)}{2}u^TU_tB_tu \\
&= -\frac{(1+m^2)t-1}{m_t^2}u^\top U_tu - u^\top U_t\mathrm{diag}\left(f_t(m^2)-f_t(\sigma_1^2),\ldots,f_t(m^2)-f_t(\sigma_d^2)\right)u - \frac{t(1-t)}{2}u^TU_tB_tu
\end{aligned}$$

As $f_t(\cdot)$ is decreasing function, we have

$$\begin{aligned}
f_t(m^2)-f_t(\sigma_k^2) &= -\frac{t(1-t)}{((1-t)^2+t^2\xi_k^2)^2}.(m^2-\sigma_k^2) \quad (\text{where } m\leq\xi_k\leq\sigma_k) \\
&= \frac{t(1-t)}{((1-t)^2+t^2\xi_k^2)^2}.(\sigma_k^2-m^2) \\
&\leq \frac{t(1-t)}{m_t^4}.(M^2-m^2) = \frac{t(1-t)}{m_t^4}.(\kappa^2-1)m^2
\end{aligned}$$

Next, Using (30) we get:

$$|u^T U_t B_t u| \leq \|U_t\|_{op}\|B_t\|_{op} \leq \frac{D}{m_t^5},$$

$$|u^\top U_t \text{diag}\left(f_t(m^2) - f_t(\sigma_1^2), \ldots, f_t(m^2) - f_t(\sigma_d^2)\right) u| \leq \frac{t(1-t)}{m_t^5}(\kappa^2 - 1)m^2.$$

we get:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(u^T U_t u\right) \geq -\frac{(1+m^2)t - 1}{m_t^2} u^T U_t u - \frac{t(1-t)}{m_t^5}(\kappa^2 - 1)m^2 - \frac{t(1-t)}{2m_t^5} D.$$

Define:

$$I(t) = \int_0^t \frac{(1+m^2)s - 1}{m_s^2} ds = \frac{1}{2}\log(m_t^2). \Rightarrow e^{I(t)} = m_t$$

Multiplying by the integrating factor, we obtain:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{I(t)} u^T U_t u\right) \geq -e^{I(t)} . \frac{t(1-t)}{m_t^5}(\kappa^2 - 1)m^2 - e^{I(t)} \cdot \frac{t(1-t)}{m_t^5} D = -\frac{t(1-t)}{m_t^4}(\kappa^2 - 1)m^2 - \frac{t(1-t)}{2m_t^4} D.$$

Integrating both sides, we obtain:

$$\begin{aligned} m u^T U_1 u &\geq 1 - \left\{\frac{D}{2} + (\kappa^2 - 1)m^2\right\} \int_0^1 \frac{s(1-s)}{m_s^4} ds \\ &= 1 - \left\{\frac{D}{2} + (\kappa^2 - 1)m^2\right\} . \frac{1}{2m^2}. \end{aligned}$$

The last inequality follows from Lemma 13. Therefore, if we have

$$D \leq 2m^2(3 - \kappa^2),$$

then $u^T U_1 u \geq 0$ for all unit vector u. This ensures that Assumption 2 is satisfied as Lemma 8 yields

$$\lambda_{min}(J_1 + J_1^T) = 2 \min_{u:\|u\|_2=1} u^T J_1 u \geq 2\left(\min_{u:\|u\|_2=1} u^\top U_1 u\right) \times \lambda_{min}(J_1^T J_1) \geq 0.$$

Moreover, Assumption 4 is automatically satisfied (see Section G.2). Therefore, the straightness of 1-RF follows from Theorem 6. □

F PROOFS FOR MONGE OPTIMALITY OF 2-RF

F.1 Proof of Theorem 8

Recall the ODE (11)

$$\frac{dJ_t^{z_0}}{dt} = \nabla_{Z_t} v_t(Z_t(z_0)) J_t^{z_0}; \quad J_0^{z_0} = I_d. \tag{31}$$

Due to the commutativity assumption, the unique solution to the above the ODE can be written in the following form

$$J_t^{z_0} = \exp\left(\int_0^t \nabla_{Z_u} v_u(Z_u(z_0))\, du\right), \tag{32}$$

where $\exp(A) := \sum_{k=0}^{\infty} A^k/k!$ for a $d \times d$ matrix A. We point the readers to Ma et al. (2009) and Section 5 of Magnus (1954) for discussions related to (32). Also, note that $\nabla_{Z_u} v_u(Z_u(z_0))$ is a symmetric matrix which

ensures that $J_t^{z_0}$ is a symmetric positive definite matrix for all $t \in [0,1]$. In particular, $J_1^{z_0} = \nabla_{z_0} Z_1(z_0)$ is also a symmetric positive definite matrix. Now, we will show that there exists a convex function φ such that $Z_1 = \nabla\varphi$. To show this, we will essentially use the symmetry of $\nabla_{z_0} Z_1(z_0)$. Let us define

$$\varphi(z) = \int_0^1 \langle Z_1(tz), z\rangle \; dt,$$

and note the following algebraic identity

$$\begin{aligned}
\frac{\partial Z_{1j}(tz)}{\partial t} &= \lim_{h\to 0} \frac{Z_{1j}(tz+hz) - Z_{1j}(tz)}{h} \\
&= \sum_{k=1}^d z_k . \frac{\partial Z_{1j}(u)}{\partial u_k}\Big|_{u=tz} \\
&= \frac{1}{t}\sum_{k=1}^d z_k . \frac{\partial Z_{1j}(tz)}{\partial z_k}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\frac{\partial\varphi(z)}{\partial z_j} &= \int_0^1 \sum_{k=1}^d \frac{\partial(z_k Z_{1k}(tz))}{\partial z_j}\, dt \\
&= \int_0^1 Z_{1j}(tz)\, dt + \int_0^1 \sum_{k=1}^d z_k \frac{\partial Z_{1k}(tz)}{\partial z_j}\, dt \\
&= \int_0^1 Z_{1j}(tz)\, dt + \int_0^1 \sum_{k=1}^d z_k \frac{\partial Z_{1j}(tz)}{\partial z_k}\, dt \quad \text{(Due to Symmetry)} \\
&= \int_0^1 Z_{1j}(tz)\, dt + \int_0^1 t . \frac{\partial Z_{1j}(tz)}{\partial t}\, dt \\
&= \int_0^1 Z_{1j}(tz) + tZ_{1j}(tz)\Big|_{t=0}^{t=1} - \int_0^1 Z_{1j}(tz)\, dt \\
&= Z_{1j}(z).
\end{aligned}$$

This shows that $Z_1 = \nabla\varphi$, and in fact $\nabla^2\varphi = \nabla Z_1 \succ 0$. Therefore, φ is also a convex function. Then the optimality of the coupling (Z_0, Z_1) follows from Theorem 1.48 of Santambrogio (2015).

F.2 RF in one-dimension yields Monge coupling

1-RF always yields a straight coupling as long as the solution to ODE (10) exists and is unique. This is because, in the one-dimensional setting, the commutativity condition in Theorem 8 is trivially met. Thus, we derive the following conclusion.

Proposition 3. *Let $(X_0, X_1) \sim N(0,1) \times \rho_1$ be a bivariate random vector in $\mathbb{R}^2$, and let the conditions in Theorem 9 hold. Then 1-RF yields the Monge transport map between $N(0,1)$ and ρ_1, and hence it also produces a straight coupling.*

The above theorem shows that 1-RF yields the Monge map for any target distribution (with a second moment) in one dimension. For example, the target distribution ρ_1 can be any log-concave or a general K-mixture of Gaussian distribution, and 1-RF will yield the Monge map between $N(0,1)$ and ρ_1.

However, it is instructive to point out that the straightness of the 1-RF can be understood through a much more intuitive and fundamental argument (Liu et al., 2023, Theorem D.10). In the one-dimensional case, uniqueness of the solution of ODE (10) implies that the map $z_0 \mapsto Z_t(z_0)$ is a monotonically increasing function for all $t \in (0,1]$. Then the straightness follows immediately from Lemma D.9 of Liu et al. (2023). In addition, the monotonicity property also ensure that all the quantiles are preserved:

Lemma 7. *Let $z_0 \in \mathbb{R}$ and write $z_t := Z_t(z_0)$. If the solution of ODE (1) is unique, then $\mathbb{P}(Z_t \le z_t)$ is a constant depending on z_0 for all t.*

Proof. We recall the ODE $\dot{Z}_t = v_t(Z_t)$ with $Z_0 = z_0$. As $x \mapsto v_t(x)$ is uniformly Lipschitz, there exists a unique solution $\{Z_t\}_{t\in[0,1]}$ such that $Z_0 = z_0$. Moreover, the map $Z_t : z_0 \mapsto z_t$ is monotonically increasing. To see this, let us assume $z_0 > \tilde{z}_0$, but $z_t < \tilde{z}_t$. Note that $G(\tau) := Z_\tau(z_0) - Z_\tau(\tilde{z}_0)$ is continuous in τ. Also, $G(0) > 0$ and $G(t) < 0$. By the intermediate value property, there exists a $t_0 \in [0, 1]$ such that $G(t_0) = 0$, i.e., $z_{t_0} = \tilde{z}_{t_0}$. This violates the uniqueness condition of the ODE solution. Hence, Z_t is monotonically increasing. By monotonicity, it follows that

$$\mathbb{P}(Z_t \le z_t) = \mathbb{P}(Z_0 \le z_0).$$

In addition, this monotonicity property also ensures that $Z_1 = \nabla\varphi$ for some convex function φ. This immediately shows that $Z_1(\cdot)$ is the Monge map (Santambrogio, 2015, Theorem 1.48). However, such arguments can not be easily generalized in higher dimensions and require deeper theoretical treatments as in Theorem 8. In the next sections, we move to examples in higher dimensions.

F.3 Proof of Proposition 1

Gaussian to Gaussian case : As shown in Section E.1, we have

$$\nabla_x v(x,t) = \frac{1}{t}\left(I - (1-t)\Sigma_t^{-1}\right).$$

Therefore, It is clear that $\nabla v(Z_t(z_0), t)$ and $\nabla v(Z_s(z_0), s)$ are commutative. Hence, the result follows from Theorem 8.

Gaussian to 2-mixture of Gaussian case : First, note that Assumption 4 is satisfied by the discussion in Section G.2. Therefore, it suffices to prove the commutativity of $\nabla v(Z_t(z_0), t)$ and $\nabla v(Z_s(z_0), s)$ for some $t < s$. Using (37), we have

$$A_t := \nabla v(Z_t(z_0), z_0) = \frac{(1+\sigma^2)t - 1}{\sigma_t^2} I_d + \frac{t(1-t)}{\sigma_t^4} w_{1,t} w_{2,t} (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top,$$

where $\sigma_t^2 := t^2\sigma^2 + (1-t)^2$. Now, it is evident that $A_t A_s = A_s A_t$. Hence, the result follows from Theorem 8.

G PROOFS FOR EXISTENCE OF RF

In this section, we collect the proofs of the main results of Appendix B.

G.1 Proof of Theorem 9

We start by analyzing the velocity function. Recall that

$$v_t(x) = \begin{cases} \frac{x}{t} + \left(\frac{1-t}{t}\right) s_t(x) & , 0 < t < 1 \\ \mathbb{E}(X_1) - x & , \quad t = 0 \\ x & , \quad t = 1. \end{cases}$$

where $s_t(x)$ is the (data) score function of $(1-t)X_0 + tX_1$. Let ϕ denote the standard gaussian density function in $\mathbb{R}^d$.

Verifying Assumption 3: For $t \in [0, 1)$ we have

$$\begin{aligned} s_t(x) &= \nabla_x \log\left(\int_{-\infty}^{\infty} (1-t)^{-d/2} \phi\left(\frac{x - ty}{1-t}\right) \rho_1(dy)\right) \\ &= \frac{\frac{1}{1-t}\int_{-\infty}^{\infty}\left(\frac{ty-x}{1-t}\right)\phi\left(\frac{x-ty}{1-t}\right)\rho_1(dy)}{\int_{-\infty}^{\infty}\phi\left(\frac{x-ty}{1-t}\right)\rho_1(dy)} \\ &= \frac{t}{(1-t)^2}\cdot\frac{\int_{-\infty}^{\infty} y\phi\left(\frac{x-ty}{1-t}\right)\rho_1(dy)}{\int_{-\infty}^{\infty}\phi\left(\frac{x-ty}{1-t}\right)\rho_1(dy)} - \frac{x}{(1-t)^2}. \end{aligned}$$

Therefore, $v_t(x) = \frac{\int_{-\infty}^{\infty} \left(\frac{y-x}{1-t}\right)\phi\left(\frac{x-ty}{1-t}\right)\rho_1(dy)}{\int_{-\infty}^{\infty}\phi\left(\frac{x-ty}{1-t}\right)\rho_1(dy)}$ for $t \in [0,1)$.

It is quite clear that $v_0(x)$ and $v_1(x)$ are $\mathcal{C}^2$ functions. Moreover, one can show that $v_t(x)$ is also $\mathcal{C}^2$ function for every $t \in (0,1)$ (∇_x and $\int$ are interchangeable due to moment condition). It suffices to show that $\Psi_1(x) := \int_{-\infty}^{\infty} y\phi\left(\frac{x-ty}{1-t}\right)\rho_1(dy)$ and $\Psi_2(x) := \int_{-\infty}^{\infty}\phi\left(\frac{x-ty}{1-t}\right)\rho_1(dy)$ are $\mathcal{C}^2$ functions and $\Psi_2 > 0$. Note that, $\Psi_2(x) = \mathbb{E}_{X_1\sim\rho_1}\phi\left(\frac{x-tX_1}{1-t}\right) > 0$. Now, we will show that $\Psi_1(x)$ is $\mathcal{C}^1$. One can similarly show that it is also $\mathcal{C}^2$ by following a similar argument.

We define

$$D(x,y) := \nabla_x\left[y\phi\left(\frac{x-ty}{1-t}\right)\right] = \frac{1}{(1-t)^2}y\left(ty-x\right)^\top \exp\left(-\frac{\|x-ty\|_2^2}{2(1-t)^2}\right).$$

Note that if $\|y\|_2^2 \geq 4\|x\|_2^2/t^2$, we have $\langle u, D(x,y)u\rangle \leq \frac{t\|y\|_2^2+\|y\|_2\|x\|_2}{(1-t)^2}\exp(-t\|y\|_2^2/4)$ for all $u \in \mathbb{S}^{d-1}$, as $\|ty-x\|_2^2 \geq (t^2/2)\|y\|_2^2 - \|x\|_2^2 \geq (t^2/4)\|y\|_2^2$. In addition, the upper bound is integrable w.r.t $\rho_1(dy)$. For $\|y\|_2^2 \leq 4\|x\|_2^2/t^2$, we have $\langle u, D(x,y)u\rangle \leq \frac{t\|y\|_2^2+\|y\|_2\|x\|_2}{(1-t)^2} \leq \frac{6\|x\|_2^2}{t(1-t)^2}$, and the upper bound is obviously integrabel w.r.t $\rho_1(dy)$. Therefore, we have

$$\nabla\Psi_1(x) = \int_{-\infty}^{\infty} D(x,y)\,\rho_1(dy).$$

The continuity also follows from generalized DCT. One can take a further derivative to show that Ψ_1 is $\mathcal{C}^2$ function, and follow the similar argument for $\Psi_2(x)$.

Non-explosive: For notational brevity, we write X_t instead of $X_t(z_0)$. Note that

$$\frac{\mathrm{d}}{\mathrm{dt}}\|X_t\|_2^2 = \langle X_t, v_t(X_t)\rangle \leq h(\|X_t\|_2^2).$$

Write $U_t := \|X_t\|_2^2$. Let V_t be a sequence of maps such that

$$\frac{\mathrm{d}}{\mathrm{dt}}V_t = h(V_t); \quad V_0 = U_0.$$

Due to Condition (13), we have $V_t < \infty$. Next, we claim that $U_t \leq V_t$ for all $t \in [0,1]$.

Under local-lipschitz property: If not, then there exist times t_0, t_1 such that

$$U_{t_0} = V_{t_0}, \quad \text{and} \quad U_t > V_t \quad \text{for all } t_0 < t \leq t_1.$$

Define $\Delta(t) := U_t - V_t$. Therefore, we have $\Delta(t_0) = 0$ and $\Delta(t) > 0$ for all $t \in (t_0, t_1]$. Let $w = U_{t_0} = V_{t_0}$. Due to local-Lipschitz property of h, there exists $\delta_w > 0$ and $L_w > 0$ such that

$$|w_1 - w| \vee |w_2 - w| < \delta_w \Rightarrow |h(w_1) - h(w_2)| \leq L_w|w_1 - w_2|.$$

Due to continuity of U_t and V_t at $t = t_0$, there exists $\eta > 0$ such that $t + \eta < t_1$ and for all $\eta' \leq \eta$ we have $|U_{t_0+\eta'} - w| \vee |V_{t_0+\eta'} - w| < \delta_w$. For , $t \in [t_0, t_0+\eta]$, we consider the ODE

$$\begin{aligned}
\dot{\Delta}(t) &= \dot{U}_t - \dot{V}_t \\
&= h(U_t) - h(V_t) \\
&\leq L_w|U_t - V_t| \quad \text{(local-Lipshcitzness)} \\
&= L_w\Delta(t) \qquad \text{(as } \Delta(t) > 0\text{)}.
\end{aligned}$$

Therefore, by Gronwall's lemma we have $\Delta(t) \leq \Delta(t_0)\exp(L_w t)$. This implies that $\Delta(t) \leq 0$ for $t \in (t_0, t_0+\eta]$, which is a contradiction to the fact that $\Delta(t) > 0$ for all $t \in (t_0, t_1]$. Hence, we have $U_t \leq V_t < \infty$ for all $t \in [0,1]$. This establishes the non-explosive property (Condition (12)) of the ODE.

Under strictly increasing property: In this case, we will show a stronger result, i.e., $U_t < V_t$ for all $t \in (0,1]$. If not, let $\tau := \inf\{t > 0 : U_t \geq V_t\}$. By definition, we have $\tau > 0$ and $U_\tau \geq V_\tau$. This implies that

$$\int_0^\tau h(U_t) - h(V_t)\,dt \geq 0 \Rightarrow \exists s \in (0,\tau) \text{ such that } h(U_s) \geq h(V_s).$$

Therefore, we have $U_s \geq V_s$, which contradicts the definition of τ. Hence, we have $U_t < V_t$ for all $t \in (0, 1]$. Now the result follows by applying Proposition 2.

G.2 Non-explosivity: Gaussian to a general mixture of Gaussian

First, for notational brevity, we write $\|u\|_\Sigma = \sqrt{u^\top \Sigma^{-1} u}$ for a positive-definite matrix Σ. Let $X_0 \sim N(0, I_d)$ and $X_1 \sim \sum_{i=1}^K \pi_i N(\mu_i, \Sigma_i)$. Let $X_t = tX_1 + (1-t)X_0$, then we have

$$v_t(x) = \frac{x}{t} + \frac{1-t}{t} s_t(x) \tag{33}$$

where, $s_t(x) = \nabla_x \log p_t(x)$ is given by

$$s_t(x) = \sum_i w_{i,t}(x) \Sigma_{i,t}^{-1} (t\mu_i - x),$$

$\Sigma_{i,t} = (1-t)^2 I_d + t^2 \Sigma_i$ and

$$w_{i,t}(x) = \frac{\pi_i \exp\left(\frac{-\|x - t\mu_i\|^2_{\Sigma_i}}{2}\right)}{\sum_j \pi_j \exp\left(\frac{-\|x - t\mu_j\|^2_{\Sigma_i}}{2}\right)}.$$

Therefore, we have

$$v_t(x) = \sum_i w_{i,t}(x) \left(I_d - (1-t)\Sigma_{i,t}^{-1}\right) \frac{x}{t} + (1-t) \sum_i w_{i,t}(x) \Sigma_{i,t}^{-1} \mu_i$$

Note that, if λ is an eigenvalue of Σ_i, then the corresponding eigenvalue of $\frac{1}{t}(I_d - (1-t)\Sigma_{i,t}^{-1})$ is $\frac{t^2(1+\lambda)-1}{(1-t)^2 + t\lambda^2} \leq (1 + \lambda^{-1})$. Therefore, $\left\|\frac{1}{t}(I_d - (1-t)\Sigma_{i,t}^{-1})\right\|_{op} \leq 1 + \left\|\Sigma_i^{-1}\right\|_{op} =: A_i$. Similar argument shows that $\left\|\Sigma_{i,t}^{-1}\right\|_{op} \leq A_i$. Therefore, we have

$$\langle x, v_t(x) \rangle \leq \underbrace{(\max_i A_i)}_{A} \|x\|_2^2 + \underbrace{(\max_i A_i \|\mu_i\|_2)}_{B} \|x\|_2 .$$

Therefore, Assumption 4 is satisfied with $h(u) = Au + B\sqrt{u}$ which is strictly monotonic function and $\int_{u_0}^\infty (Au + B\sqrt{u})^{-1}\, du = \infty$ for all $u_0 > 0$. Moreover, we have $\mathbb{E}\|X_1\|_2 < \infty$. Therefore, by Theorem 9 we conclude that the solution to the ODE (10) is unique.

H AUXILIARY RESULTS

Lemma 8. *Let $A \in \mathbb{R}^{d\times d}$ be an invertible matrix. Define $q(A) := \min_{u:\|u\|_2=1} u^\top A u$. Then the following inequality is true provided $q(A^{-1}) \geq 0$:*

$$q(A) \geq q(A^{-1}) q(A^\top A).$$

Proof. Let u be a unit vector such that $u^\top A u = q(A)$. As A is invertible, there exists a $v \in \mathbb{R}^d$ such that $u = A^{-1}v$. Note that we have $\|v\|_2^2 = u^\top A^\top A u \geq q(A^\top A) \geq 0$. Then, we have

$$\begin{aligned} q(A) &= u^\top A u \\ &= v^\top (A^{-1})^\top v \\ &= v^\top (A^{-1}) v \\ &\geq q(A^{-1}) \|v\|_2^2 \geq q(A^{-1}) q(A^\top A). \end{aligned}$$

□

Lemma 9. *Let $(X_0, X_1) \sim N(0, I_d) \otimes \rho_1$. Let the density of $X_t = tX + (1-t)Z$ be p_t, and the score to be $s_t(x) = \nabla \log p_t(x)$. Then, we have*

$$v_t(x) = \frac{x}{t} + \left(\frac{1-t}{t}\right) s_t(x).$$

Proof. First, note that due to Tweedie's formula (Robbins, 1992) we have $\mathbb{E}tX \mid X_t = x = x + (1-t)^2 s_t(x)$. Using this, we have

$$\begin{aligned} v_t(x) &= \mathbb{E}[X - Z \mid X_t = x] \\ &= \mathbb{E}[\frac{X - X_t}{1-t} \mid X_t = x] \\ &= \frac{x + (1-t)^2 s_t(x)}{t(1-t)} - \frac{x}{(1-t)} \quad \text{(applying Tweedie's formula)} \\ &= \frac{x}{t} + \left(\frac{1-t}{t}\right) s_t(x). \end{aligned} \tag{34}$$

□

Lemma 10. *Let $X_0 \sim \frac{1}{K_0}\sum_{i=1}^{K_0} N(\mu_{0i}, \sigma^2 I)$, and $X_1 \sim \frac{1}{K_1}\sum_{i=1}^{K_1} N(\mu_{1i}, \sigma^2 I)$ be independent, and define $X_t = tX_1 + (1-t)X_0$. Then, we have*

$$v_t(x) = \frac{x}{t} + \frac{(1-t)\sigma^2}{t}\left(\frac{1}{K_0}\sum_{i=1}^{K_0} \frac{p_t^{(i)}(x)}{p_t(x)}\left(s_t^{(i)}(x) - \frac{\mu_{0i}}{1-t}\right)\right),$$

where $p_t^{(i)}(x) = \frac{1}{K_1}\sum_{j=1}^{K_1} N(\underbrace{t\mu_{1j} + (1-t)\mu_{0i}}_{\mu_{tj}^{(i)}}, \sigma_t^2)$, $\sigma_t^2 = (t^2 + (1-t)^2)\sigma^2$.

$$s_t^{(i)}(x) = \nabla_x \log p_t^{(i)}(x) = \frac{1}{\sigma_t^2}\left(\sum_{j=1}^{K_1} w_j^{(i)}(x)\mu_{tj}^{(i)} - x\right),$$

and

$$w_j^{(i)}(x) = \frac{\exp\left(\frac{-\left\|x - \mu_{tj}^{(i)}\right\|^2}{2\sigma_t^2}\right)}{\sum_j \exp\left(\frac{-\left\|x - \mu_{tj}^{(i)}\right\|^2}{2\sigma_t^2}\right)}$$

Proof.

$$\begin{aligned} v_t(x) &= \mathbb{E}\left[X_1 - X_0 \mid X_t = x\right] \\ &= \mathbb{E}\left[\frac{X_1 - X_t}{1-t} \mid X_t = x\right] \\ &= \frac{1}{t(1-t)}\left(\mathbb{E}\left[tX_1 \mid X_t = x\right] - tx\right) \\ &= \frac{1}{t(1-t)}\left(\frac{1}{K_0}\sum_{i=1}^{K_0} \frac{p_t^{(i)}(x)}{p_t(x)}\mathbb{E}\left[tX_1 \mid X_t^{(i)} = x\right] - tx\right) \\ &= \frac{1}{t(1-t)}\left(\frac{1}{K_0}\sum_{i=1}^{K_0} \frac{p_t^{(i)}(x)}{p_t(x)}\left(x - (1-t)\mu_{0i} + \tilde{\sigma}_t^2 s_t^{(i)}(x)\right) - tx\right), \quad \text{where } \tilde{\sigma}_t^2 = (1-t)^2\sigma^2 \\ &= \frac{x}{t} + \frac{(1-t)\sigma^2}{t}\left(\frac{1}{K_0}\sum_{i=1}^{K_0} \frac{p_t^{(i)}(x)}{p_t(x)}\left(s_t^{(i)}(x) - \frac{\mu_{0i}}{1-t}\right)\right) \end{aligned}$$

□

Lemma 11. *Let $X_0 \sim \mathcal{N}(0, I_d)$ and $X_1 \sim \sum_i \pi_i N\left(\mu_i, \sigma_i^2 I\right)$ be independent. Let $X_t = tX_1 + (1-t)X_0$, with density p_t. Then, using Lemma 9, we have*

$$v_t(x) = \frac{x}{t} + \frac{1-t}{t} s_t(x) \tag{35}$$

where, $s_t(x) = \nabla_x \log p_t(x)$ is given by

$$s_t(x) = \sum_i w_{i,t}(x) \left(\frac{t\mu_i - x}{\sigma_{i,t}^2} \right),$$

$\sigma_{i,t}^2 = (1-t)^2 + t^2\sigma_i^2$ and

$$w_{i,t}(x) = \frac{\pi_i \exp\left(\frac{-\|x - t\mu_i\|^2}{2\sigma_{i,t}^2}\right)}{\sum_j \pi_j \exp\left(\frac{-\|x - t\mu_j\|^2}{2\sigma_{i,t}^2}\right)}$$

Proof. The result directly follows from Lemma 10 with $K_0 = 1$. □

Lemma 12. *Let $X_0 \sim N(0, I)$ and $X_1 \sim \sum_i \pi_i N(\mu_i, \Sigma_i)$ be independent random variables. Let $X_t = tX_1 + (1-t)X_0$ with density p_t. Then, we have*

$$v_t(x) = \frac{x}{t} + \frac{1-t}{t} s_t(x), \tag{36}$$

where, $s_t(x) = \nabla_x \log p_t(x)$ is given by

$$s_t(x) = \sum_i w_{i,t}(x) \Sigma_{i,t}^{-1} \left(t\mu_i - x\right),$$

$\Sigma_{i,t} = (1-t)^2 I_d + t^2 \Sigma_i$ and

$$w_{i,t}(x) = \frac{\frac{\pi_i}{\sqrt{\det(\Sigma_{i,t})}} \exp\left(\frac{-(x-t\mu_i)^\top \Sigma_{i,t}^{-1}(x-t\mu_i)}{2}\right)}{\sum_j \frac{\pi_j}{\sqrt{\det(\Sigma_{j,t})}} \exp\left(\frac{-(x-t\mu_j)^\top \Sigma_{j,t}^{-1}(x-t\mu_j)}{2}\right)}.$$

Note: One can also evaluate the exact derivative the drift v_t in the above case. For notational brevity, we define $\delta_{i,t} := \Sigma_{i,t}^{-1}(t\mu_i - z_t)$. Then, we have

$$\begin{aligned} \nabla_{z_t} v(z_t, t) = &\left\{ \sum_i \frac{1}{t} \left\{ I_d - (1-t)\Sigma_{i,t}^{-1} \right\} . w_{i,t}(z_t) \right\} \\ &+ \left(\frac{1-t}{t} \right) . \left\{ \sum_{i<j} w_{i,t}(z_t) w_{j,t}(z_t) (\delta_{i,t} - \delta_{j,t})(\delta_{i,t} - \delta_{j,t})^\top \right\} \end{aligned} \tag{37}$$

Lemma 13. *For $a, b > 0$, define*

$$I(a,b) = \int_0^1 s(1-s) \frac{\sqrt{(1-s)^2 + a^2 s^2}}{\left((1-s)^2 + b^2 s^2\right)^{5/2}} \, ds.$$

Then

$$I(a,b) = \frac{a^2 + ab + b^2}{3\, b^3 (a+b)}.$$

Proof. Begin with the substitution

$$u = \frac{s}{1-s}, \qquad s = \frac{u}{1+u}, \qquad ds = \frac{du}{(1+u)^2}.$$

Direct algebra shows that the integral becomes

$$I(a,b) = \int_0^\infty u\,\frac{\sqrt{1+a^2u^2}}{(1+b^2u^2)^{5/2}}\,du.$$

Next let $v = u^2$, so $u\,du = \frac{1}{2}dv$. Then

$$I(a,b) = \frac{1}{2}\int_0^\infty \frac{\sqrt{1+a^2v}}{(1+b^2v)^{5/2}}\,dv.$$

Now scale by $y = b^2v$, so $dv = \frac{1}{b^2}dy$ and

$$I(a,b) = \frac{1}{2b^2}\int_0^\infty \frac{\sqrt{1+\frac{a^2}{b^2}y}}{(1+y)^{5/2}}\,dy.$$

Set

$$r = \frac{a^2}{b^2}.$$

To convert the improper integral to a bounded interval, use

$$y = \frac{t}{1-t}, \qquad dy = \frac{dt}{(1-t)^2}, \qquad t \in [0,1].$$

A straightforward simplification yields

$$\int_0^\infty \frac{\sqrt{1+ry}}{(1+y)^{5/2}}\,dy = \int_0^1 \sqrt{1+(r-1)t}\,dt.$$

If $r \neq 1$, the elementary antiderivative gives

$$\int_0^1 \sqrt{1+(r-1)t}\,dt = \frac{2}{3}\,\frac{r^{3/2}-1}{r-1}.$$

Therefore,

$$I(a,b) = \frac{1}{2b^2}\cdot\frac{2}{3}\,\frac{r^{3/2}-1}{r-1} = \frac{1}{3b^2}\,\frac{r^{3/2}-1}{r-1}.$$

Substitute back $r = \frac{a^2}{b^2}$:

$$r^{3/2} = \frac{a^3}{b^3}, \qquad r-1 = \frac{a^2-b^2}{b^2},$$

so

$$I(a,b) = \frac{1}{3b^2}\cdot\frac{\frac{a^3}{b^3}-1}{\frac{a^2-b^2}{b^2}} = \frac{a^3-b^3}{3b^3\,(a^2-b^2)}.$$

Factor numerator and denominator:

$$a^3-b^3 = (a-b)(a^2+ab+b^2), \qquad a^2-b^2 = (a-b)(a+b).$$

Thus

$$I(a,b) = \frac{a^2+ab+b^2}{3\,b^3(a+b)}.$$

□