# FLOWS DON'T CROSS IN HIGH DIMENSION

**Teodora Reu**
Department of Computer Science
University of Oxford
teodora.reu@balliol.ox.ac.uk

**Sixtine Dromigny**
Department of Earth Science
University of Oxford

**Michael Bronstein**
Department of Computer Science
University of Oxford

**Francisco Vargas**
Department of Computer Science
University of Cambridge, Xaira Therapeutics

## ABSTRACT

Conditional Flow Matching (CFM) has emerged as a competitive framework for generative modeling, yet persistent concerns about trajectory crossings and their impact on gradient variance have influenced the development, of a new framework Rectify Flows. In this work, we rigorously analyze these assumptions through theoretical and empirical lenses. First, we prove that in high-dimensional spaces ($d > 2$), interpolating trajectories between source-target pairs almost surely never cross—a zero-measure phenomenon contradicting low-dimensional intuition. Second, we derive closed-form expressions for gradient variance under Gaussian distributions, revealing that suboptimal deterministic couplings (e.g., rotation-based pairings) incur dimension-dependent variance scaling. Empirically, we demonstrate that while 2D rotations inducing crossings amplify gradient noise, this effect diminishes linearly with dimension rather than abruptly vanishing. We also identify time-dependent variance patterns ($t \to 1$) uncorrelated with crossings, suggesting additional variance sources in CFM optimization.

## 1 INTRODUCTION

Diffusion models Sohl-Dickstein et al. (2015); Song et al. (2020); Ho et al. (2020); Song et al. (2023) have shown great promise in generating images (Ramesh et al., 2022), videos (Ho et al., 2022), and molecules (Hoogeboom et al., 2022). Data generation typically involves simulating a stochastic denoising process. To improve efficiency, this stochastic process is often transformed into an equivalent (marginal preserving) Ordinary Differential Equation (ODE) (Song et al., 2020), known as the probability flow ODE (PF-ODE).

Simultaneously, a new framework called Conditional Flow Matching (CFM) (Tong et al., 2023; Lipman et al., 2022) has emerged as a powerful approach to generative modeling, achieving performance comparable to diffusion models for various generation tasks (Davtyan et al., 2023; Zhao et al., 2024; Kapuśniak et al., 2024; Davis et al., 2024). However, both approaches train a neural network to approximate the small steps of a transformation, whether stochastic or deterministic. During training, the network learns to estimate the derivative of a function governing the evolution of the data, requiring integration during inference to reconstruct the final sample.

As a result, the major challenge in generative modeling is improving the efficiency of these frameworks by reducing the number of integration steps required during inference. In diffusion models, this has led to the development of knowledge distillation (e.g. consistency distillation in Song et al. (2023), progressive distillation in Salimans & Ho (2022)), and for CFM to Rectified Flows in Liu et al. (2022), to attain faster sampling.

In knowledge distillation-based methods, which currently represent the state-of-the-art in requiring a low number of steps at inference time (Luhman & Luhman, 2021; Salimans & Ho, 2022; Song et al., 2023; Zheng et al., 2023), a student model is trained to directly predict the solution of the PF-ODE.

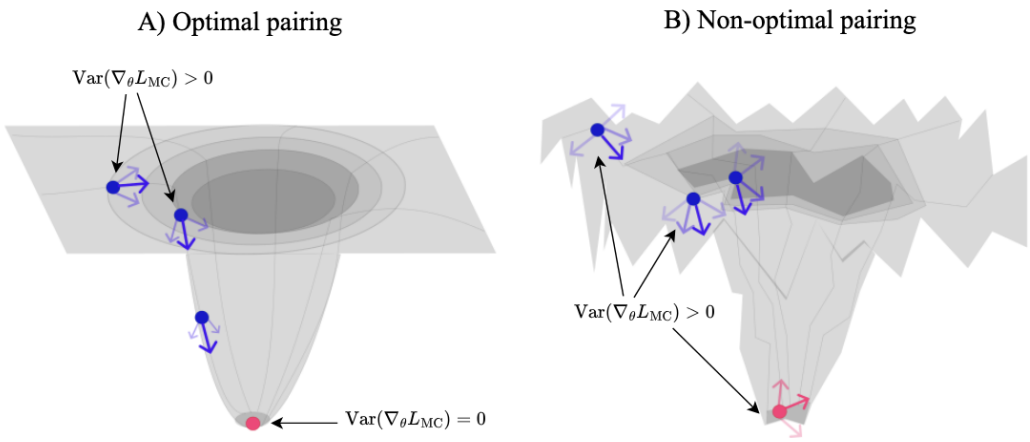



Figure 1: Schematic of a three-dimensional loss function $L_{\mathrm{MC}}(v_\theta)$ in grey shading. A) The loss function is represented for an optimal pairing between the samples $X_0$ and $X_1$. The blue points representing its positive variance ($\mathrm{Var}(\nabla_\theta L_{\mathrm{MC}}) > 0$) and the arrows represent the noise associated with the variance. The red point represents the optimal vector fields.

Similarly, in the context of CFM, methods like Rectified Flows as introduced by Liu et al. (2022) and Liu (2022) aim to enhance the integration process for more efficient sampling by converging to straighter vector fields, which are easier to integrate over. This is achieved by constructing deterministic couplings between the source and target distributions through the simulation of forward and backward dynamics of the learned vector fields. Although repeated applications of this process may introduce accumulated error (Liu et al., 2022), Roy et al. (2024) suggest that these couplings can be optimal for Gaussian-like distributions or exhibit non-crossing behavior, often referred to as *straight* trajectories in the literature.

It is assumed in the field that crossing paths (of the interpolant lines between the source and target distribution) add variance to the Monte Carlo approximated loss function, used for optimization are closely correlated (Fjelde et al., 2024). In this work, we challenge several prevailing assumptions in CFM and Rectified Flows:

- We theoretically prove that in high dimensions ($d > 2$), interpolating trajectories have zero-measure intersections, refuting concerns about trajectory crossings affecting training stability in Section 3.
- We derive bounds for the variance of the gradients of the CFM loss function under Gaussian source and target distributions in Section 4.
- We empirically analyze the correlation between trajectory crossings and gradient variance in Section 5. In two dimensions, we observe that trajectory crossings indeed affect the variance; however, as the dimensionality increases, we observe a linear decrease in variance rather than an abrupt drop, which is supported by our theoretical findings, though not entirely. We also observe that the variance increases as $t$ approaches 1, a phenomenon noted in Lee et al. (2024). This behavior cannot be explained solely by trajectory crossings, as the crossing from source to target should be symmetric around $t = 1/2$.

These findings concluded that while crossing might be correlated with the variance of gradients there seem to be other sources for its provenience.

**Research Question** How is the variance of the gradients correlated with, or bounded by, the incidence of crossing segments with endpoints in *pairs of samples* drawn from the source and target distributions?

## 2 THEORETICAL BACKGROUND

This section elaborates on the theoretical framework of Conditional Flow Matching (CFM) as introduced by Tong et al. (2023) and Lipman et al. (2022), adhering to their established notation. We start with a short description of Optimal Transport, followed by CFM, and then by Rectified Flows.

### 2.1 OPTIMAL TRANSPORT

Optimal Transport (OT) was initially posed as the Monge problem (Monge, 1781), seeking a map $T : \mathbb{R}^d \to \mathbb{R}^d$ that transports samples $X_0$ (with law $\pi_0$) into $X_1 = T(X_0)$ (with law $\pi_1$), minimizing an expected cost $\mathbb{E}[c(T(X_0) - X_0)]$ (Villani et al., 2009). Kantorovitch (1958) relaxed it by allowing stochastic couplings while preserving the same marginals, commonly referred to as the Kantorovich problem. If $\pi_0$ is absolutely continuous, both the Monge and Kantorovich problems share a deterministic optimal coupling. A time-continuous formulation connects $X_0 \sim \rho_0$ to $X_1 \sim \rho_1$ via $\{X_t\}_{t\in[0,1]}$; for a convex cost, the infimum is attained by a straight-line interpolant $X_t = (1-t)X_0 + tX_1$ (McCann, 1997).

### 2.2 CONDITIONAL FLOW MATCHING AND RECTIFIED FLOWS

Conditional Flow Matching (CFM) constructs a time-dependent ODE:

$$\mathrm{d}Z_t = v(Z_t, t)\,\mathrm{d}t, \quad Z_0 \sim \pi_0, \quad Z_1 \sim \pi_1, \tag{1}$$

with $v(\cdot, t)$ learned to align with linear trajectories connecting samples $X_0$ and $X_1$. Minimizing

$$\int_0^1 \mathbb{E}\big[\|X_1 - X_0 - v(X_t, t)\|^2\big]\,\mathrm{d}t, \quad X_t = (1-t)X_0 + tX_1, \tag{2}$$

yields $v(x, t) = \mathbb{E}[X_1 - X_0 \mid X_t = x]$. In practice, $v$ is parameterized by a neural network and trained via stochastic gradient descent on $(X_0, X_1)$ (Tong et al., 2023; Lipman et al., 2022).

Once trained, the ODE $\mathrm{d}Z_t = v(Z_t, t)\,\mathrm{d}t$ induces the same marginals as $X_t$ for all $t$. This flow can be "rectified" by simulating $\{Z_t\}_{t\in[0,1]}$; we denote this process as $\mathrm{RectFlow}(X_0, X_1)$, leading to rectified couplings $(Z_0, Z_1)$ (Liu et al., 2022).

**Are Rectified Couplings Better?** It has been shown that a single RectFlow step can transform random couplings of two Gaussians into an optimal transport (OT) coupling (Roy et al., 2024), and applying infinitely many RectFlows yields straight (non-crossing) trajectories (Liu et al., 2022). Empirically, this procedure also exhibits robust performance when combined with additional refinements (Lee et al., 2024). The key insight is that simulating the forward dynamics generates couplings $(Z_0, Z_1)$ in a continuous vector field whose integrated paths cannot intersect at any time $t$, returning more stable couplings. infinitely many applications of this algorithm can lead sometime to obtaining optimal pairings, that is when the parameterization of the network is such that it approximates the gradient of a vector field Liu (2022), noise is added to the interpolants Albergo et al. (2023), and the network is parametrized to approximate a convex function Kornilov et al. (2024).

### 2.3 LOSS AND VARIANCE OF GRADIENTS

Rewriting the CFM objective, from Equation 2:

$$L(v) \;=\; \mathbb{E}_{X_0\sim\pi_0}\big[\|T(X_0) - X_0 \;-\; v(X_t, t)\|^2\big], \quad X_t = (1-t)X_0 + t\,T(X_0), \tag{3}$$

where $T$ is the (deterministic or random) map satisfying $T_{\#}\pi_0 = \pi_1$. A Monte Carlo approximation uses samples $\{X_0^{(s)}\}_{s=1}^S$:

$$L_{\mathrm{MC}}(v) = \frac{1}{S}\sum_{s=1}^{S}\big\|\,T\big(X_0^{(s)}\big) - X_0^{(s)} - v\big(X_t^{(s)}, t\big)\big\|^2, \quad X_t^{(s)} = (1-t)X_0^{(s)} + t\,T(X_0^{(s)}). \tag{4}$$

This work studies the variance of $L_{\mathrm{MC}}(v_\theta)$, the gradient variance $\mathrm{Var}[\nabla_\theta L_{\mathrm{MC}}(v_\theta)]$, and the gradient variance $\mathrm{Var}[\nabla_\theta L_{\mathrm{MC}}(v_{\hat{\theta})}]$ where $v(\hat{\theta})$ is the optimal vector field, to assess training stability when $T$ pairs $(X_0, X_1)$ randomly versus deterministically.

## 3 INTERPOLATING LINES DON'T CROSS IN HIGHER DIMENSION

The success of the Rectified Flows framework relies on the fact that the couplings $(Z_0, Z_1)$, obtained by simulating the dynamics, exhibit non-crossing paths. To critically examine this framework, we investigate whether interpolating paths theoretically express crossings in higher dimensions. As our proofs show, the interpolating lines corresponding to the pairings $(X_0, X_1)$, never cross in dimensions higher than 2. Additionally, we extend our analysis by exploring the variance of gradients in a simplified setting, specifically in the context of linear transformations and Gaussian distributions.

As mentioned before crossing of interpolant lines is a big part of RectFlow framework. We first highlight the empirical evidence in Table 1 to motivate our theoretical results. This table shows the number of intersections for several datasets, starting with a shifted one-dimensional (1D) Gaussian embedded in two dimensions (leftmost column) and extending to higher-dimensional Gaussians and real-world data. As can be seen, the number of intersections is equal to the number of lines. This occurs because the probability of sampling parallel interpolating lines is zero—a fact we will discuss later. Additionally, the number of intersections occurring at distinct time points $t_i, t_j \in (0, 1)$ is equal to those occurring at identical time points $t_i = t_j \in (0, 1)$, due to the distributions being equally spaced from each other. For a Gaussian distribution embedded in a two-dimensional space, we observe that while intersections still occur at different time points, simultaneous intersections (i.e., occurring at the same time) drop to zero. This phenomenon will also be explained theoretically. Finally, moving to the last three columns, we observe that all intersections completely vanish.

Table 1: **For the first three columns:** Intersection results for 1000 randomly sampled pairs across various source/target distributions. The analysis spans from the common 1D Gaussian embedded in 2D space to $n$-dimensional Gaussian distributions in $n$-dimensional space. **For the last two columns:** Intersections between n-dimensional gaussians and 10000 random samples from CIFAR and MNIST. Notably, instead of a linear decrease in intersection occurrences over $t \in (0, 1)$, we observed an instantaneous drop to 0.

| **Source**<br>**Target** | $[0, \mathcal{N}(0,1)]$<br>$[5, \mathcal{N}(0,1)]$ | $\mathcal{N}(0, I_2)$<br>$\mathcal{N}(5, I_2)$ | $\mathcal{N}(0, I_d)$<br>$\mathcal{N}(5, I_d)$ for $d > 2$ | $\mathcal{N}(0, I_{728})$<br>MNIST | $\mathcal{N}(0, I_{1024})$<br>CIFAR10 |
|---|---|---|---|---|---|
| **Intersections for** $t_i, t_j \in (-\infty, \infty)$ | 499,500 | 499,500 | 0 | 0 | 0 |
| **Intersections for** $t_i, t_j \in (0, 1)$ | 247,967 | 214,091 | 0 | 0 | 0 |
| **Intersections for** $t_i = t_j = \hat{t}$ | 247,967 | 0 | 0 | 0 | 0 |

The following result explains why the number of interpolating lines is as high as the possible pairings in the first and second columns. The proofs for the following proposition and lemmas are in Appendix A.1.

**Lemma 1.** *Let $x_0, x_1 \sim \pi_0$, and $y_0, y_1 \sim \pi_1$, where $\pi_0, \pi_1$ are probability distributions on $\mathbb{R}^d$ admitting a density. Define the linear interpolants $l_i(t_i, x_i, y_i) = (1 - t_i)x_i + t_i y_i$ for $i \in \{0, 1\}$. Then for $d = 2$, the probability that $l_0$ and $l_1$ intersect is 1.*

For the remaining cells in the table, which are not zero, in the first column, it is expected that the intersections for $t_i, t_j \in (0, 1)$ and $t_i = t_j$ are equal. This is because, for this particular setting of source and target distributions, all intersections occur when $t_i = t_j$. By examining the first axis, we observe that $0(1 - t_i) + 5(t_i) = 0(1 - t_j) + 5(t_j)$, which results in the two intersection times being equal. The number of these intersections is roughly half of the total number of intersections. This is because, depending on the order in which we sample, the two samples from the target will result in an intersection about half of the time.

Continuing we explain the drop in zero-measure for the increase in dimensionality for any absolutely continuous probability measures $\pi_0, \pi_1$. In the following proposition, we establish that for source and target distributions supported in $\mathbb{R}^d$, for $d > 1$ the probability of two interpolating segments $l_i(x_0, x_1, t_i) = (1 - t_i)x_0 + t_i x_1$ for $t \in [0, 1]$ and for $i \in [0, 1]$ at any time crossing at $t_i = t_j = t \in [0, 1]$ is zero. Furthermore, that in higher-dimensional spaces, specifically for $\pi_0, \pi_1 \in \mathbb{R}^d$ with $d > 2$, such intersections do not occur for any $t_i, t_j \in [0, 1]$. These results can be intuitively understood through a well-known theorem which states that the probability of an event confined to a subspace of lower dimension than the ambient probability space is zero.

**Proposition 1.** *Let $x_0, x_1 \sim \pi_0(x)$ and $y_0, y_1 \sim \pi_1(y)$, where $\pi_0$ and $\pi_1$ are probability distributions on $\mathbb{R}^d$ admitting a density. Let the lines $l_i(x_i, y_i, t_i) = (1 - t_i)x_i + t_i y_i$. For $d \geq 2$, the lines*

*$l_0$ and $l_1$ cross at $t = t_i = t_j \in (0, 1)$ with probability $0$. Similarly, for $d > 2$ the two lines intersect for $t_i, t_j \in (0, 1)$ with probability 0.*

Since the probability of sampling pairs of crossing lines in high dimension has zero measure, the expected number of crossing lines when sampling more pairs, will of course be zero. In the following section, we want to explore how is the variance of the gradients affected in general by crossings.

## 4 Variance of Gradients and Crossings

In the following subsections, we will study the impact that the choice of pairing has on the variance of the gradients. We will first discuss the impact that deterministic pairings have on the gradients. A common misconception is that deterministic pairings are generally beneficial. We demonstrate, in this context, that they can also be detrimental by introducing a transformation (a rotation matrix) that has the potential to maximize the number of crossings at time $t = 1/2$. Simultaneously, we will study the variance expressed by Optimal Transport (OT) couplings, both around the optimal vector field and around any arbitrary vector field.

In this section, we study two deterministic couplings between two Gaussians: $\mathcal{N}(0, I_d)$ and $\mathcal{N}(\mu, M_d)$, where $M_d$ is a symmetric and positive definite matrix. Since we know that the optimal coupling is given by $T_{OT}(X_0) = M_d^{1/2} X_0 + \mu$, we can analyze the variance of the vector fields as they learn from these pairings. We then introduce a second transformation, which is still deterministic but far from optimal, namely $T_{rOT}(X_0) = M_d^{1/2} R_\alpha X_0 + \mu$, where $R_\alpha$ is a rotation matrix that rotates $X_0$ around the origin by an angle of $\alpha^\circ$ degrees. Since the rotation matrix will cause the interpolating lines to intersect, we aim to observe the effect of these intersections on the variance of the gradients. As it can be seen in Figure 2 at $180^o$ we have the most variance, fact which we also observe in our theoretical results. The proofs for the following proposition and lemmas can be found in Appendix A.2.

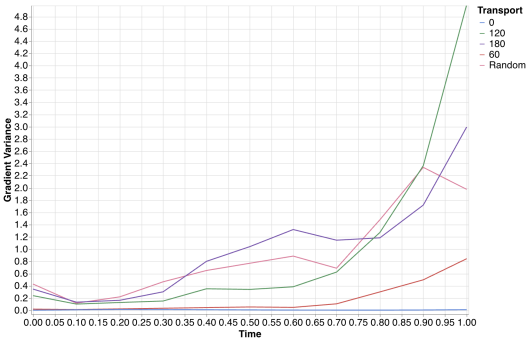



Figure 2: Gradient variance for various rotation matrices and random pairing. For a $180^\circ$ rotation, we observe a peak around $t = \frac{1}{2}$. Notably, for angles greater than $120^\circ$, higher variance is experienced than with random pairing.

**Proposition 2.** *Let $\pi_0 \sim \mathcal{N}(0, \boldsymbol{I}_d)$, and $\pi_1 \sim \mathcal{N}(\mu, \boldsymbol{M}_d)$, where $M_d$ is positive definite and symmetric matrix. Let the following deterministic maps:*

$$T_{OT}(x_0) = M_d^{1/2} x_0 + \mu = x_1, \quad \textit{and} \quad T_{rOT}(x_0) = M_d^{1/2} R x_0 + \mu = x_1. \tag{5}$$

*Let the linear transformation $v_{\theta,\Theta}$ with be $v_{\theta,\Theta}(x_t, t) = \boldsymbol{\Theta}[x_t, t] + \boldsymbol{\theta}$, where $[,]$ is concatenation. Let:*

$$L_{MC}^{OT}(\boldsymbol{\Theta}, \boldsymbol{\theta}) = \frac{1}{N} \sum \|T_{OT}(X_0) - X_0 - v_{\boldsymbol{\Theta},\boldsymbol{\theta}}(X_t, t)\|^2, \tag{6}$$

$$L_{MC}^{rOT}(\boldsymbol{\Theta}, \boldsymbol{\theta}) = \frac{1}{N} \|T_{rOT}(X_0) - X_0 - v_{\boldsymbol{\Theta},\boldsymbol{\theta}}(X_t, t)\|^2, . \tag{7}$$

*Then for optimal $(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}})$ we have:*

$$\textit{Var}[\nabla_{\boldsymbol{\theta}} L_{MC}^{OT}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}})] = 0, \quad \textit{Var}[\nabla_{\boldsymbol{\theta}} L_{MC}^{rOT}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}})] = \frac{4}{N} (\mathbf{1}^\top (M_d^{1/2} R - R)(M_d^{1/2} R - R)^\top 1^\top), \tag{8}$$

$$\textit{Var}[\nabla_{\boldsymbol{\Theta}} L_{MC}^{OT}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}})] = 0, \quad \textit{and} \quad \textit{Var}[\nabla_{\boldsymbol{\Theta}} L_{MC}^{rOT}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}})] = \frac{8}{N} tr((M^{1/2}(R - \boldsymbol{I}))^2). \tag{9}$$

**Remark 1.** *First of all, a variance of* $0$ *at the optimal parameters for optimal pairings is expected. It is also worth noting that the higher the values in* $M_d$*, and the higher the trace of the values in* $M_d$*, the higher the variance of the gradients. However, in practice, one usually standardizes their target distribution. For* $tr(M_d) = d$*, we observe that* $R$ *plays a significant role in the bounding. For a rotation transformation, when* $R = -I_d$ *(i.e., a* $180°$ *rotation, with the maximal intersection number at* $t = 1/2$*), and* $M_d = I_d$*, the gradient variance is given by* $Var(\nabla_\theta L_{MC}^{rOT}(\theta)) = \frac{16d}{N}$*, and* $Var(\nabla_\Theta L_{MC}^{rOT}(\theta)) = \frac{24d}{N}$.

A key takeaway from this subsection is that, for this simple deterministic pairing, the variance of the gradients scales with the dimensionality, the complexity of the rotation (i.e., how many dimensions it collapses and the extent of line intersections), and the variance of the target distribution.

## 5 Neural Network Simulations

In this section, we extend our previous analysis by parameterizing the vector field with a neural network. In particular, we employ a three-layer multi-layer perceptron (MLP) with a hidden dimension of 64 and SELU activations. Our experimental investigation is designed to address the following questions:

- **Deterministic and Optimal** $T$**:** How does high variance in the target distribution affect the gradient variance for optimal pairings? Additionally, how do batch size and dimensionality influence the gradient variance around the optimum when the target distribution is shifted?
- **Random** $T$**:** To what extent does the complexity of the target distribution influence the results, and how does the gradient variance scale with respect to dimensionality and batch size?
- **Deterministic but Non-optimal** $T$**:** How is the network affected by trajectory crossings induced by a rotation matrix?

### 5.1 Deterministic and Optimal $T$

In this subsection, we investigate how the variance in the target distribution affects the variance of the gradient norm around the optimally learned neural network. Since we are restricted to optimal pairings, we study this behavior exclusively between Gaussian distributions, which are the only ones for which optimal pairings are known. Accordingly, we consider a source distribution $\mathcal{N}(0, I_d)$ and a target distribution $\mathcal{N}(5, \text{var} \times I_d)$ across increasing dimensions.

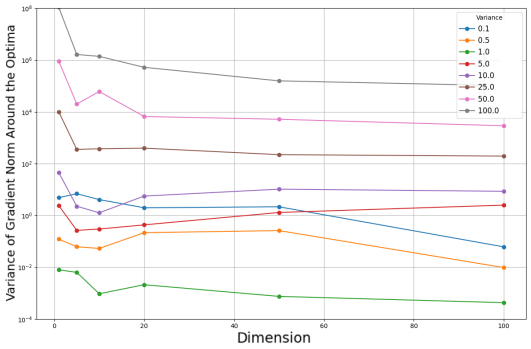
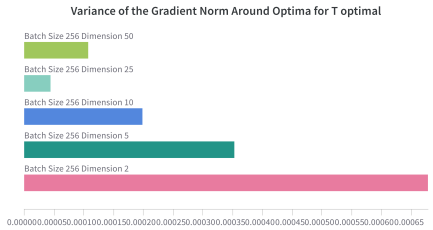


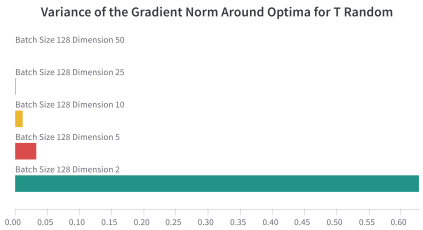

Figure 3: Variance of the gradient norm for optimal transport between two normal distributions, $\mathcal{N}(0, I_d)$ and $\mathcal{N}(5, \text{var} \times I_d)$. The decrease of the gradient variance is linear in dimension and exponential in the increase in the variance of the target distribution.

As seen in Figure 3, the variance of the gradients increases significantly with the variance of the target distribution. Another notable observation is that for optimal pairings, the variance of the gradients around the optimum appears to decrease with increasing dimension. This is unusual; for instance, in Proposition 2, the variance of the gradients is expected to increase with dimension. Therefore, this must be a neural network-specific behavior.

From Figure 4a, we observe that the variance of the gradient norm decreases linearly in dimension.




(a) Comparison of the effects of dimension increase variance of gradients for optimal $T$.




(b) Comparison of the effects of dimension increase variance of gradients for random $T$.

Figure 4: As dimension increases the variance decreases, linearly for optimal $T$, and exponentially for random $T$. This could be explained by our theoretical findings.

### 5.2 Random $T$

The source distribution is $\mathcal{N}(0, I_d)$, and the target distribution is a standardized Gaussian mixture. We investigate how the number of modes (representing the complexity of the dataset), batch sizes, and the dimensionality of the dataset interplay with the values of interest.

We believe that the exponential decrease in the variance of gradient norms around the optimum with increasing dimensionality is a phenomenon that is difficult to explain, as illustrated in Figures 4 and 5. In contrast, the observed linear decrease with batch size is expected. Another notable observation is that the complexity of the standardized target—measured by the number of medians present—does not significantly increase the difficulty for the network to learn the target.

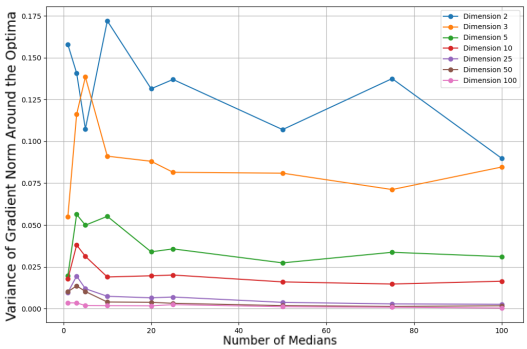



Figure 5: Variance of the gradient norm as the complexity of the distribution increases for random pairings. As the dimensionality of the data increases, the gradient norm decreases. Interestingly, the complexity of the data (measured in the number of medians the gaussian mixture has) does not appear to significantly influence the final result.

### 5.3 Deterministic but Non-Optimal $T$

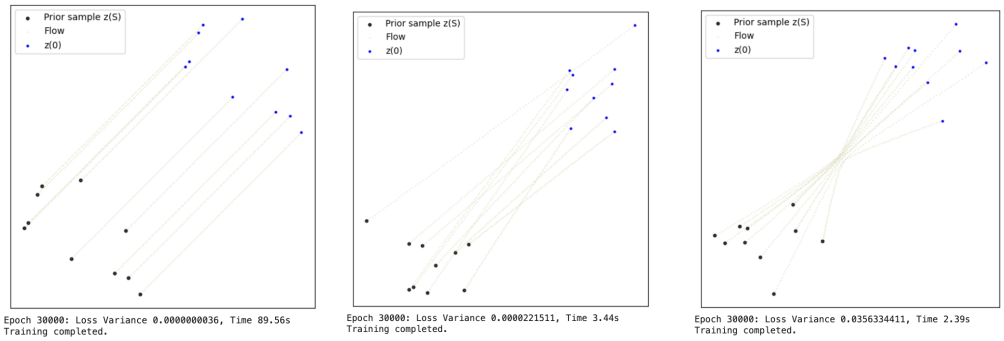
This experiment explores a simplified setting to study the variance of the loss function. We define the transformation as $T(X_0) = (X_0 + 5) \times \text{Rotation Matrix}(\alpha)$,

where $\theta$ denotes the rotation angle. As shown in Figure 6, the loss variance peaks when the rotation angle is maximal. Notably, the learned pairing appears to be closer to the optimal pairing than to the rotated pairing, as exemplified by the $90°$ rotation.

**Neural Network Experiment Conclusion:** A key takeaway from this section, and a potential area for further research, is the decreasing behavior of gradient variance with the dimensionality of the space, which occurs for both optimal and random pairings. For random pairing, this could be explained by the lack of intersections between interpolant lines in high dimensions, but for now, this remains a hypothesis.

## 6 Conclusion

In this work, we investigated the theoretical and empirical properties of Conditional Flow Matching (CFM) with a focus on trajectory crossings and gradient variance. We demonstrated that in high-




(a) $\alpha = 0^\circ$, low variance (b) $\alpha = 90^\circ$, low variance (c) $\alpha = 180^\circ$, high variance

Figure 6: Variance analysis for different rotation angles. The variance in the $180^\circ$ rotation case (c) is very high, but the paings seem to be close to optimal. Interestingly, while the $90^\circ$ rotation (b) results in a rotated target distribution.

dimensional spaces ($d > 2$), interpolating trajectories have zero-measure intersections, addressing concerns about trajectory crossings affecting training stability. Furthermore, we provided theoretical bounds on the variance of gradients under Gaussian source and target distributions and analyzed how these variances evolve in different pairing scenarios.

Our empirical findings support the theoretical results, showing that interpolating lines rarely intersect in high dimensions and that variance in gradient norms behaves unexpectedly for different conditions. Specifically, we observed an exponential decrease in gradient variance with increasing dimensionality, a phenomenon that remains an open question for further investigation. Additionally, we studied how deterministic and random pairings impact gradient variance, revealing that suboptimal deterministic couplings, such as those induced by rotation matrices, significantly increase gradient variance. This serves as a response to recent speculations on the impact crossings have on variance gradient (Fjelde et al., 2024), and on the quality of generated couplings (Roy et al., 2024).

**Further Research** To extend this work, we plan to generalize the variance results from Section 4 to more sophisticated target distributions. Specifically, we aim to: (1) Analyze neural parameterizations where $v_\theta$ is implemented as an MLP rather than linear transformation, (2) Establish dimension-dependent variance bounds for arbitrary couplings $T$, and (3) Formalize the relationship between gradient variance decay and trajectory non-crossing in high dimensions

We hypothesize that MLP-based analyses could explain the exponential variance reduction observed empirically (Figure 5) through the lens of implicit regularization in overparameterized networks. Furthermore, we aim to connect these variance properties to our theoretical results on measure-theoretic trajectory non-crossing in $\mathbb{R}^d$ (Proposition 1).

## REFERENCES


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.

Oscar Davis, Samuel Kessler, Mircea Petrache, Ismail Ilkan Ceylan, Michael Bronstein, and Avishek Joey Bose. Fisher flow matching for generative modeling over discrete data. *arXiv preprint arXiv:2405.14664*, 2024.

Aram Davtyan, Sepehr Sameni, and Paolo Favaro. Efficient video prediction via sparsely conditioned flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23263–23274, 2023.

Tor Fjelde, Emile Mathieu, and Vincent Dutordoir. An introduction to flow matching, January 2024. URL `https://mlg.eng.cam.ac.uk/blog/2024/01/20/flow-matching.html`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:1–13, 2022.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Leonid Kantorovitch. On the translocation of masses. *Management science*, 5(1):1–4, 1958.

Kacper Kapuśniak, Peter Potaptchik, Teodora Reu, Leo Zhang, Alexander Tong, Michael Bronstein, Avishek Joey Bose, and Francesco Di Giovanni. Metric flow matching for smooth interpolations on the data manifold. *arXiv preprint arXiv:2405.14780*, 2024.

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Alexander Korotin. Optimal flow matching: Learning straight trajectories in just one step. *arXiv preprint arXiv:2403.13117*, 2024.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. *arXiv preprint arXiv:2405.20320*, 2024.

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. *arXiv preprint arXiv:2209.14577*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

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

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pp. 666–704, 1781.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Saptarshi Roy, Vansh Bansal, Purnamrita Sarkar, and Alessandro Rinaldo. 2-rectifications are enough for straight flows: A theoretical insight into wasserstein convergence. *arXiv e-prints*, pp. arXiv–2410, 2024.

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

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

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*, 2020.

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

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional flow matching: Simulation-free dynamic optimal transport. *arXiv preprint arXiv:2302.00482*, 2(3), 2023.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Wenliang Zhao, Minglei Shi, Xumin Yu, Jie Zhou, and Jiwen Lu. Flowturbo: Towards real-time flow-based image generation with velocity refiner. *arXiv preprint arXiv:2409.18128*, 2024.

Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In *International conference on machine learning*, pp. 42390–42402. PMLR, 2023.

# A APPENDIX

## A.1 INTERSECTION PROOFS

**Lemma 1.** *Let $x_0, x_1 \sim \mathcal{N}(0, I_d)$, and $y_0, y_1 \sim p(x)$, where $p(x)$ is a probability distribution on $\mathbb{R}^d$ admitting a density. Define the linear interpolants $l_i(t) = (1-t)x_i + ty_i$ for $i \in \{0, 1\}$. Then:*

- *For $d = 2$, the probability that $l_0$ and $l_1$ intersect is 1.*
- *For $d \geq 3$, the probability that $l_0$ and $l_1$ intersect is 0.*

*Proof.* Throughout this proof, we use the following theorem: Let $X$ be a random variable with a continuous probability distribution in $\mathbb{R}^n$, and let $A$ be a lower-dimensional subset of $\mathbb{R}^n$.

Since $\mathbb{P}$ admits a density, it follows that it is absolutely continuous wrt to the Lebesgue measure $\lambda$ then as $\lambda(A) = 0$ by absolute continuity, we have that $\mathbb{P}(X \in A) = 0$.

For $d = 2$, the lines $l_0(t)$ and $l_1(t)$ lie in the same plane. This implies three possibilities: the lines are either parallel, intersect for $t \in [0, 1]$, or intersect for $t \notin [0, 1]$. The probability that the two lines are parallel is zero because

$$\mathbb{P}(l_0(t) \parallel l_1(t)) = \mathbb{P}(y_1 \in A) = 0,$$

where $A$ is the line through $y_0$ that is parallel to $l_0$. Hence, the two lines intersect with probability 1. However, whether the intersection occurs within $t \in [0, 1]$ depends on the start and target distributions.

For $d \geq 3$, two lines intersect only if they are coplanar. However, in dimensions $d \geq 3$, the event that the two lines lie in the same plane requires that their spans are constrained to a lower-dimensional subspace (a plane) in $\mathbb{R}^d$. Since the set of coplanar configurations corresponds to a subset of $\mathbb{R}^d$ with dimension 2, the probability of this event is 0, as stated above.

Therefore, for $d \geq 3$, the probability of $l_0(t)$ and $l_1(t)$ intersecting is 0. □

**Lemma 2.** *Let $x_0, x_1 \sim \pi_0(x)$ and $y_0, y_1 \sim \pi_1(y)$, where $\pi_0$ and $\pi_1$ are probability distributions in $\mathbb{R}^d$ admitting a density, with $d \geq 2$. The lines $l_i(t) = (1-t)x_i + ty_i$, for $t \in (0, 1)$, intersect at $t = \hat{t}$ with probability zero.*

*Proof.* Suppose the lines $l_0(t)$ and $l_1(t)$ intersect at some $t = \hat{t} \in (0, 1)$. This implies

$$l_0(\hat{t}) = l_1(\hat{t}),$$

which expands to

$$(1-\hat{t})x_0 + \hat{t}y_0 = (1-\hat{t})x_1 + \hat{t}y_1.$$

Rearranging terms, we find

$$y_1 = \frac{1-\hat{t}}{\hat{t}}(x_0 - x_1) + y_0.$$

To compute the probability of such an intersection, note that $y_1$ must lie exactly on the affine subspace defined by the above equation, which is a one-dimensional line segment in $\mathbb{R}^d$.

The joint probability of the points $(x_0, x_1, y_0, y_1)$ can be written as

$$\mathbb{P}(l_0(t) \text{ intersects } l_1(t) \text{ for } t \in (0,1)) = \mathbb{P}(x_0, x_1, y_0) \cdot \mathbb{P}(y_1 = \frac{1-\hat{t}}{\hat{t}}(x_0 - x_1) + y_0 \mid x_0, x_1, y_0).$$

Since $\pi_1(y)$ is continuous and differentiable, the probability density of $y_1$ lying on any lower-dimensional subspace (e.g., a line segment) in $\mathbb{R}^d$, with $d > 2$, is zero. Therefore,

$$\mathbb{P}(y_1 = \frac{1-\hat{t}}{\hat{t}}(x_0 - x_1) + y_0 \mid x_0, x_1, y_0) = 0,$$

which implies

$$\mathbb{P}(l_0(t) \text{ intersects } l_1(t) \text{ at } t = \hat{t} \text{ for } t \in (0,1)) = 0.$$

Hence, the lines $l_0(t)$ and $l_1(t)$ intersect with probability zero for $t \in (0, 1)$. □

**Lemma 3.** *Let $x_0, x_1 \sim \pi_0(x)$ and $y_0, y_1 \sim \pi_1(y)$, where $\pi_0$ and $\pi_1$ are continuous and differentiable probability distributions in $\mathbb{R}^d$, with $d > 2$. The lines $l_i(t_i) = (1-t_i)x_i + t_i y_i$, for $t_i \in (0,1)$, intersect at with probability zero.*

*Proof.* The proof of this lemma is similar to the previous one, with the mention of now what will happen in the end is that $y_1$ will be conditioned to belong on a 2 dimensional probability space for $0 \leq t_0, t_1 \leq 1$, which never happens in a probability space with higher dimension. Suppose the lines $l_0(t_0)$ and $l_1(t_1)$ intersect at some $t_0 = \hat{t}_0, t_1 = \hat{t}_1 \text{for} t_0, t_1 \in (0,1)$. This implies

$$l_0(\hat{t}_0) = l_1(\hat{t}_1),$$

which expands to

$$(1-\hat{t}_0)x_0 + \hat{t}_0 y_0 = (1-\hat{t}_1)x_1 + \hat{t}_1 y_1.$$

Rearranging terms, we find

$$y_1 = \frac{(1-\hat{t}_0)x_0 - (1-\hat{t}_1)x_1 + \hat{t}_0 y_0}{\hat{t}_1}.$$

which for $\hat{t}_0, \hat{t}_1 \in (0,1)$ we have that $y_1$ would belong to a 2D surface. Just like before we have:

$\mathbb{P}(l_0(t_0) \text{ intersects } l_1(t_1) \text{ for } \hat{t}_i \in (0,1)) = \mathbb{P}(x_0, x_1, y_0) \cdot$

$$\mathbb{P}\left(y_1 = \frac{(1-\hat{t}_0)x_0 - (1-\hat{t}_1)x_1 + \hat{t}_0 y_0}{\hat{t}_1} \,\middle|\, x_0, x_1, y_0\right) = 0.$$

□

**Proposition 1.** *Let $x_0, x_1 \sim \pi_0(x)$ and $y_0, y_1 \sim \pi_1(y)$, where $\pi_0$ and $\pi_1$ is a probability distribution on $\mathbb{R}^d$ admitting a density. Let the lines $l_i(x_i, y_i, t_i) = (1-t_i)x_i + t_i y_i$. For $d \geq 2$, the lines $l_0$ and $l_1$ cross at $t = t_i = t_j \in (0,1)$ with probability 0. Similarly, for $d > 2$ the two lines intersect for $t_i, t_j \in (0,1)$ with probability 0.*

*Proof.* Follows by Lemmas 2, 3. □

### A.2 VARIANCE PROOFS

**Lemma 4.** *Let $\pi_0 \sim \mathcal{N}(0, \boldsymbol{I}_d)$, and $\pi_1 \sim \mathcal{N}(\mu, \boldsymbol{I}_d)$. Let $T_0$ be a deterministic optimal transport map with $T_0(x_0) = x_0 + \mu = x_1$, and $T_1(x_0) = Rx_0 + \mu = x_1$ be another deterministic, but not optimal transport map, with $R$ a rotation matrix that rotates the reference distribution. Let the $v$ be $v(x_t, \boldsymbol{\theta}, \boldsymbol{\Theta}) = \hat{v}(x_0, \boldsymbol{\theta}, \boldsymbol{\Theta}) = \boldsymbol{\Theta} x_0 + \boldsymbol{\theta}$, so just a linear transformation. Let:*

$$L_{MC}^{OT}(\boldsymbol{\Theta}, \boldsymbol{\theta})] = \frac{1}{N}\|T_0(X_0^{(n)}) - X_0^{(n)} - v_{\boldsymbol{\Theta},\boldsymbol{\theta}}(X_0^{(n)})\|^2, \tag{10}$$

$$L_{MC}^{rOT}(\boldsymbol{\Theta}, \boldsymbol{\theta})] = \frac{1}{N}\|T_1(X_0^{(n)}) - X_0^{(n)} - v_{\boldsymbol{\Theta},\boldsymbol{\theta}}(X_0^{(n)})\|^2, . \tag{11}$$

*Then for optimal $(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}}) = (\mathbf{0}, \mu)$ we have:*

$$\textit{Var}[\nabla_\theta L_{MC}^{OT}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}})] = 0, \quad \textit{and} \quad \textit{Var}[\nabla_\theta L_{MC}^{rOT}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}})] = \frac{4}{N}\mathbf{1}^\top (R - I_d)(R - I_d)^\top \mathbf{1}. \tag{12}$$

*Proof.* First since the transport map is linear (maybe we can sonider something a bit more interesting than same variance) we can write the loss

$$\text{Var}\left[\nabla_{\boldsymbol{\Theta}_i} \frac{1}{N} \sum [||\mu - \boldsymbol{\Theta} X_0^{(n)} - \boldsymbol{\theta}||^2]\right] \tag{13}$$

$$= \text{Var}\left[-\frac{2}{N} \sum X_0^\top (\mu - \boldsymbol{\Theta} X_0^{(n)} - \boldsymbol{\theta})\right] \tag{14}$$

Where we have used that :

$$\nabla_{\boldsymbol{\Theta}_i}[\boldsymbol{\Theta} X_0^{(n)}]_i = \nabla_{\boldsymbol{\Theta}_i} \boldsymbol{\Theta}_i^\top X_0^{(n)} = X_0^{(n)} \tag{15}$$

and for the bias we have

$$\text{Var}\left[\nabla_{\boldsymbol{\theta}} \frac{1}{N} \sum [||\mu - \boldsymbol{\Theta} X_0^{(n)} - \boldsymbol{\theta}||^2]\right] \tag{16}$$

$$= \text{Var}\left[-\frac{2}{N} \sum \mathbf{1}^\top (\mu - \boldsymbol{\Theta} X_0^{(n)} - \boldsymbol{\theta})\right] \tag{17}$$

$$= \frac{4\text{Var}[\mathbf{1}^\top \boldsymbol{\Theta} X_0^{(n)}]}{N} = \frac{4}{N} \mathbf{1}^\top \boldsymbol{\Theta} \boldsymbol{\Theta}^\top \mathbf{1} \tag{18}$$

For rotation

$$\text{Var}\left[\nabla_{\boldsymbol{\theta}} \frac{1}{N} \sum [||(R - \boldsymbol{I}) X_0 + \mu - \boldsymbol{\Theta} X_0^{(n)} - \boldsymbol{\theta}||^2]\right] \tag{19}$$

$$= \text{Var}\left[-\frac{2}{N} \sum \mathbf{1}^\top ((R - \boldsymbol{I}) X_0 + \mu - \boldsymbol{\Theta} X_0^{(n)} - \boldsymbol{\theta})\right] \tag{20}$$

$$= \text{Var}[\mathbf{1}^\top (R - \boldsymbol{I} - \boldsymbol{\Theta}) X_0^{(n)}] \tag{21}$$

$$= \frac{4}{N} (\mathbf{1}^\top \boldsymbol{\Theta} \boldsymbol{\Theta}^\top \mathbf{1} + \mathbf{1}^\top \boldsymbol{\Theta} (R - \boldsymbol{I})^\top \mathbf{1} + \mathbf{1}^\top (R - \boldsymbol{I}) \boldsymbol{\Theta}^\top \mathbf{1} + \mathbf{1}^\top (R - \boldsymbol{I})(R - \boldsymbol{I})^\top \mathbf{1}), \tag{22}$$

As we don't need to apply any linear transformation to $X$ for the optimal parameters, because $X$ is already a Gaussian distribution $\mathcal{N}(0, \boldsymbol{I}_d)$. Therefore, we only need to learn the shift parameter $\boldsymbol{\theta}$. This means that the optimal transport map's parameter matrix $\boldsymbol{\Theta}$ is simply the zero matrix, $0$.

For the optimal coupling, the variance of the gradients around the optimal parameters is $0$, indicating no variation. However, for the case involving the rotated transport map, the variance of the gradients is scales with the rotation, and it peaks at $180^o$, when the number of crossings at time $t = 1/2$ is maximal (all lines cross). □

**Proposition 2.** *Let $\pi_0 \sim \mathcal{N}(0, \boldsymbol{I}_d)$, and $\pi_1 \sim \mathcal{N}(\mu, \boldsymbol{M}_d)$, where $M_d$ is positive definite and symmetric matrix. Let the following deterministic optimal maps with:*

$$T_0(x_0) = Gx_0 + \mu = x_1, \quad \textit{and} \quad T_1(x_0) = GRx_0 + \mu = x_1, \tag{23}$$

*with $G = M_d^{1/2}$ which is positive definite and maybe symmetric. Let the $v$ from Equations **??** be $v(x_t, \boldsymbol{\theta}, \boldsymbol{\Theta}) = \hat{v}(x_0, \boldsymbol{\theta}, \boldsymbol{\Theta}) = \boldsymbol{\Theta} x_0 + \boldsymbol{\theta}$, so just a linear transformation. Let:*

$$L_{MC}^{OT}(\boldsymbol{\Theta}, \boldsymbol{\theta}) = \frac{1}{N} \sum \|T_0(X_0) - X_0 - v_{\boldsymbol{\Theta}, \boldsymbol{\theta}}(X_0)\|^2, \tag{24}$$

$$L_{MC}^{rOT}(\boldsymbol{\Theta}, \boldsymbol{\theta}) = \frac{1}{N} \|T_1(X_0) - X_0 - v_{\boldsymbol{\Theta}, \boldsymbol{\theta}}(X_0)\|^2, . \tag{25}$$

*Then for optimal $(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}}) = (G - \boldsymbol{I}, \mu)$ we have:*

$$\textit{Var}[\nabla_{\boldsymbol{\theta}} L_{MC}^{OT}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}})] = 0, \quad \textit{and} \quad \textit{Var}[\nabla_{\boldsymbol{\theta}} L_{MC}^{rOT}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}})] = \frac{4}{N} (\mathbf{1}^\top (GR - R)(GR - R)^\top 1^\top), \tag{26}$$

*and,*

$$\textit{Var}[\nabla_{\boldsymbol{\Theta}} L_{MC}^{OT}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}})] = 0, \quad \textit{and} \quad \textit{Var}[\nabla_{\boldsymbol{\Theta}} L_{MC}^{rOT}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}})] = \frac{8}{N} tr((G(R - \boldsymbol{I}))^2). \tag{27}$$

*Proof.* For $T_0$, we have almost just like before:

$$\text{Var}\left[\nabla_{\boldsymbol{\theta}} \frac{1}{N} \sum [||(GX_0^{(n)} - X_0^{(n)} + \mu - \boldsymbol{\Theta} X_0^{(n)} - \boldsymbol{\theta}||^2]\right] \tag{28}$$

$$= \text{Var}\left[-\frac{2}{N} \sum \mathbf{1}^\top (\mu + (G - \boldsymbol{\Theta} - \boldsymbol{I}) X_0^{(n)} - \boldsymbol{\theta})\right] \tag{29}$$

$$= \frac{4}{N} \text{Var}[\mathbf{1}^\top (G - \boldsymbol{\Theta} - \boldsymbol{I}) X_0^{(n)}] = \frac{4}{N} \mathbf{1}^\top (G - \boldsymbol{\Theta} - \boldsymbol{I})(G - \boldsymbol{\Theta} - \boldsymbol{I})^\top \mathbf{1} \tag{30}$$

$$\text{Var}\left[\nabla_{\boldsymbol{\Theta}_i} \frac{1}{N} \sum [||(GX_0^{(n)} - X_0^{(n)} - \mu - \boldsymbol{\Theta} X_0^{(n)} - \boldsymbol{\theta}||^2]\right] \tag{31}$$

$$= \text{Var}\left[-\frac{2}{N} \sum X_0^\top (\mu + (G - \boldsymbol{I} - \boldsymbol{\Theta}) X_0^{(n)} - \boldsymbol{\theta})\right] \tag{32}$$

Note that the optimal our vector field is optimal when $\hat{\boldsymbol{\Theta}} = G - \boldsymbol{I}$, and $\hat{\boldsymbol{\theta}} = \mu$. Then we have:

$$Var[\nabla_{\boldsymbol{\theta}} L_{MC}(v(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}}, X_0))] = 0, \tag{33}$$

and,

$$Var[\nabla_{\boldsymbol{\Theta}} L_{MC}^{OT}(v(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}}, X_0))] = \frac{4}{N} \text{Var}[X_0^\top (\mu - \hat{\boldsymbol{\theta}})] = \frac{4||\mu - \hat{\boldsymbol{\theta}}||^2}{N} = 0. \tag{34}$$

For rotation we have:

$$\text{Var}\left[\nabla_{\boldsymbol{\theta}} \frac{1}{N} \sum [||(GRX_0^{(n)} - X_0^{(n)} - \mu - \boldsymbol{\Theta} X_0^{(n)} - \boldsymbol{\theta}||^2]\right] \tag{35}$$

$$= \text{Var}\left[-\frac{2}{N} \sum \mathbf{1}^\top (\mu + (GR - \boldsymbol{\Theta} - I_2) X_0^{(n)} - \boldsymbol{\theta})\right] \tag{36}$$

$$= \frac{4}{N} \text{Var}[\mathbf{1}^\top (G - \boldsymbol{\Theta} - I_2) X_0^{(n)}] = \tag{37}$$

and at the optima,

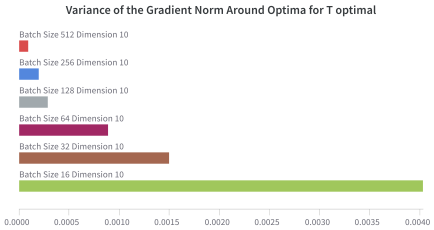
$$\frac{4}{N} \text{Var}(\mathbf{1}^\top (GR - R) X_0) = \frac{4}{N} (\mathbf{1}^\top (GR - R)(GR - R)^\top 1^\top) \tag{38}$$

For variance of the $\boldsymbol{\Theta}$ around the optima, we have:

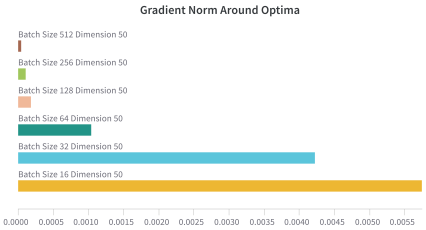
$$Var[\nabla_{\boldsymbol{\Theta}} L_{MC}^{rOT}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\theta}}))] = \frac{4}{N} \text{Var}[X_0^\top G(R - I) X_0] = \frac{8}{N} \text{tr}((M^{1/2}(R - I))^2). \tag{39}$$

The last equality comes from the variance of a quadratic form for Gaussian random vectors, that is if we have $X \sim \mathcal{N}(0, I_d)$. □

### A.3 Other experiments




(a) Comparison of the effects of batch size increase variance of gradients for optimal $T$.




(b) Comparison of the effects of batch size increase variance of gradients for random $T$.

Figure 7: As dimension increases the variance decreases, linearly for optimal $T$, and linearly for random $T$.