Conditional Wasserstein Distances with Applications in Bayesian OT Flow Matching

Jannis Chemseddine, Paul Hagemann, Christian Wald, Gabriele Steidl

March 28, 2024

Abstract

In inverse problems, many conditional generative models approximate the posterior measure by minimizing a distance between the joint measure and its learned approximation. While this approach also controls the distance between the posterior measures in the case of the Kullback–Leibler divergence, this is in general not hold true for the Wasserstein distance. In this paper, we introduce a conditional Wasserstein distance via a set of restricted couplings that equals the expected Wasserstein distance of the posteriors. Interestingly, the dual formulation of the conditional Wasserstein-1 flow resembles losses in the conditional Wasserstein GAN literature in a quite natural way. We derive theoretical properties of the conditional Wasserstein distance, characterize the corresponding geodesics and velocity fields as well as the flow ODEs. Subsequently, we propose to approximate the velocity fields by relaxing the conditional Wasserstein distance. Based on this, we propose an extension of OT Flow Matching for solving Bayesian inverse problems and demonstrate its numerical advantages on an inverse problem and class-conditional image generation.

1 Introduction

Many sampling algorithms for the posterior $P_{X|Y=y}$ in Bayesian inverse problems

$$Y = f(X) + \Xi \tag{1}$$

with a forward operator $f : \mathbb{X} \to \mathbb{Y}$, and a noise model Ξ, perform learning on some joint measures. This means, given some observations y of Y, a probability measure P_{Y,G_θ} is learned, where G_θ also depends on y. Most approaches minimize (or upper bound) some loss of the form

$$L(\theta) = D(P_{Y,X}, P_{Y,G_\theta}),$$

where D denotes a suitable distance on the space of probability measures. For instance, this is done in the framework of conditional (stochastic) normalizing flows [7, 26, 25, 54], conditional GANs [40] or conditional gradient flows for the Wasserstein metric [18, 24]. Note

that the robustness of such conditional generative models was shown under the assumption that the expected error to the posterior $E_Y\left[W_1(P_{X|Y=y}, P_{Z|Y=y})\right]$ is small in [3], which shows that it is important to relate this quantity to the approximation of the joint measures.

In [31], the authors investigated the relation between the joint measures $D(P_{Y,Z}, P_{Y,X})$ and its relation to the expected error between the posteriors $E_Y\left[D(P_{Z|Y=y}, P_{X|Y=y})\right]$. For the Kullback–Leibler divergence $D = \mathrm{KL}$, it follows by the chain rule [16, Theorem 2.5.3] that

$$E_Y\left[\mathrm{KL}(P_{X|Y=y}, P_{Z|Y=y})\right] = \mathrm{KL}(P_{Y,X}, P_{Y,Z}).$$

Such results are important as they show that it is possible to approximate the posterior via the joint distribution. Unfortunately, we have for the Wasserstein-1 distance that in general only

$$W_1(P_{Y,X}, P_{Y,Z}) \leq \mathbb{E}_{y\sim P_Y}\left[W_1(P_{X|Y=y}, P_{Z|Y=y})\right] \tag{2}$$

holds true, in contrast to the equality claim in [31, Theorem 2]. A simple counterexample is given in Appendix A. Intuitively, strict inequality can arise when the optimal transport (OT) plan needs to transport mass in the Y-component. This is the motivation for considering only plans that do not have mass transport in the Y-component. This leads us to the definition of conditional Wasserstein distances $W_{p,Y}$, where admissible transport plans are restricted to the set $\Gamma^4_Y = \Gamma^4_Y(P_{Y,X}, P_{Y,Z})$ of 4-plans α fulfilling in addition $(\pi^{1,3})_\sharp\alpha = \Delta_\sharp P_Y$, where $\Delta(y) = (y, y)$ is the diagonal map:

$$W^p_{p,Y} \coloneqq \inf_{\alpha\in\Gamma^4_Y} \int \|(y_1, x_1) - (y_2, x_2)\|^p \,\mathrm{d}\alpha$$

Inspired by [31], we show that this conditional Wasserstein distance indeed fulfills

$$W^p_{p,Y}(P_{Y,X}, P_{Y,Z}) = \mathbb{E}_{y\sim P_Y}\left[W^p_p(P_{X|Y=y}, P_{Z|Y=y})\right].$$

Further, we prove results on gradient flows [6] with respect to our conditional Wasserstein distance: we show the connections to the continuity equation, verify that there exists a velocity field with no mass transport in Y-direction and recover a corresponding ODE formulation. Indeed, this conditional Wasserstein distance can be used to explain a numerical observation made by [18, 24], namely that rescaling the Y-component leads to velocity fields with no mass transport in Y-direction in the limit. Using these ideas, we propose to relax the conditional Wasserstein distance to allow "small amounts" of mass transport in Y-direction.

Then, we use our insights to design efficient posterior sampling algorithms. By leveraging recent ideas of flow matching [35, 36], we design Bayesian OT flow matching. Note that the recent approaches of [55, 53] do not respect the OT in X-direction as they always choose the diagonal coupling. This leads to awkward situations, where the optimal Y-diagonal coupling is not recovered even between Gaussians, see Example 9. We use our proposed Bayesian OT flow matching and verify its advantages on a Gaussian mixture toy problem and on class conditional image generation on the CIFAR10 dataset.

Contributions

- We introduce conditional Wasserstein distances and highlight their relevance to conditional Wasserstein GANs in inverse problems.
- We derive theoretical properties of the conditional Wasserstein distance and establish geodesics in this conditional Wasserstein space, with velocity plans having no transport in Y-direction.
- We show that the conditional Wasserstein distance can be used in conditional generative approaches and demonstrate the advantages on MNIST particle flows [24, 4]. We propose a version of OT flow matching [50, 44] for inverse problems which uses a relaxed version of our conditional Wasserstein distance, and show that it overcomes obstacles from previous flow matching versions for inverse problems [53].

Related work Our work is in the intersection of conditional generative modelling [1, 7, 40] and (computational) OT [43, 51]. The recent work [22, Theorem 2] derives an inequality based on restricting the admissible couplings in the their OT formulation to so-called conditional sub-couplings. Note that their reformulation is only a restatement of the expected value, but does not relate it to the joint distributions. Those authors also look for geodesics in the Wasserstein space, but pursue a different approach. While we relate it to the velocity fields in gradient flows, they pursue an autoencoder/GAN idea.

The closest work, which appeared after our first version of this paper, is [29]. Unfortunately, we became aware of those paper when our paper was close to its finish. Here the authors define the conditional optimal transport problem and calculate its dual. Their work is more focused on the infinite dimensional setting, whereas we consider the velocity fields needed for the flow matching application.

In the OT literature, there has been a collection of class conditional OT distances used in domain adaption [41, 45]. In particular, conditional OT as in [48] is relevant as they consider OT plans for each condition y minimizing $\mathbb{E}_y[W_1(P_{X|Y=y}, G(\cdot, y)_\# P_Z)]$. However they relax their problem using a KL divergence. The papers on Wasserstein gradient flows [5, 23] investigate conditional Wasserstein distances from a different point of view for defining the so-called geometric tangent space of the 2-Wasserstein space. Geometric tangent spaces play a crucial role in Wasserstein flows of maximum mean discrepancies with Riesz kernels in [27] and their neural variants in [4]. In [24, Remark 7], an inequality between the joint Wasserstein and the expected value over the conditionals is derived, where the result requires compactly supported measures and certain regularity of the associated posterior densities. In [12]. the supervised training of conditional Monge maps is proposed, for which the authors solved the dual problem using convex neural networks. The authors of [37] also considers an amortized objective between the conditional distributions and propose a relaxation, which only needs samples from the joint distribution involving maximum mean discrepancies. Numerically, we first verify our theoretical statements based on

particle flows, which were also used in [4, 24]. Further, we apply our framework to solve inverse problems using Bayesian flow matching [53, 55] and OT flow matching [35, 36, 50, 44].

This paper is an extension of our first ArXiv version [13] on conditional Wasserstein distances with more focus on gradient flows and flow matching.

Outline of the paper In Section 2, we recall preliminaries from OT. Then, in Section 3, we introduce conditional Wasserstein distances of joint probability measures, and show their relation to the expectation over the Wasserstein distance of the conditional probabilities. Moreover, we highlight the connection to work on geometric tangent spaces. In Section 4, we calculate the dual of our conditional Wasserstein-1 distance and show how a loss function used in the conditional Wasserstein GAN literature arises in a natural way. In Section 5, we deal with geodesics with respect to the conditional Wasserstein distance, prove properties of the corresponding velocity fields, showing that they vanish in the Y-component, and show existence for flow ODEs. We propose a relaxation of the conditional Wasserstein distance which appears to be useful for numerical computations in Section 6. We combine our findings with OT flow matching to get Bayesian flow matching in Section 6. Finally, in Section 8, we present numerical results: we verify a convergence result using particle flows to MNIST, and demonstrate the advantages of our Bayesian OT flow matching procedure on a Gausian mixture model toy example and on CIFAR10 class-conditional image generation. All proofs are postponed to the appendix.

2 Preliminaries

Throughout this paper, we will use the following notation. These are basics from from optimal transport theory and can be found in [51]. By $\mathcal{P}(\mathbb{X})$, we denote the set of probability measures on $\mathbb{X} \subseteq \mathbb{R}^n$ and by $\mathcal{P}_p(\mathbb{X})$, $p \in [1, \infty)$ the subset of measures with finite p-th moments. For $\mu \in \mathcal{P}(\mathbb{X})$ and a measurable function $F : \mathbb{X} \to \mathbb{Y}$, we define the the *push forward measure* by $F_\sharp \mu = F \circ \mu^{-1}$. For a product space $\prod_{i=1}^K \mathbb{X}_i$, we denote the projection onto the $i_1, \dots, i_k$-th component by $\pi^{i_1,\dots,i_k}$. The *Wasserstein-p metric* [51] on $\mathcal{P}_p(\mathbb{X})$ is given by

$$\begin{aligned} W_p(\mu, \nu) &\coloneqq \Big(\min_{\gamma \in \Gamma} \int_{\mathbb{X}^2} \|x - y\|^p \, \mathrm{d}\gamma(x, y) \Big)^{\frac{1}{p}} \\ &= \Big(\min_{\gamma \in \Gamma} \mathbb{E}_{(x,y)\sim\gamma} \big[\|x - y\|^p \big] \Big)^{\frac{1}{p}}. \end{aligned} \tag{3}$$

where $\Gamma = \Gamma(\mu, \nu)$ denotes the set of all probability measures $\gamma \in \mathcal{P}(\mathbb{X} \times \mathbb{X})$ with marginals $\pi^1_\sharp \gamma = \mu$ and $\pi^2_\sharp \gamma = \nu$ and $\| \cdot \|$ is the Euclidean distance on $\mathbb{R}^n$, see [51]. If $\mu \in \mathcal{P}_p(\mathbb{X})$ is absolutely continuous, then, for $p \in (1, \infty)$, there exits a unique optimal transport map

$T \in L^p_\mu(\mathbb{X}, \mathbb{X})$, also known as Monge map, which solves

$$\min_{T \,\text{measurable}} \Big\{ \int_{\mathbb{X}} \|x - T(x)\|^p \,\mathrm{d}\mu(x) \quad \text{such that} \quad T_\sharp \mu = \nu \Big\}.$$

Further, this optimal map is related to the optimal transport plan γ in (3) by $\gamma = (\mathrm{Id}, T)_\sharp \mu$, see [51]. The same holds true for empirical measures with the same number of points, see [43, Proposition 2.1]. In this paper, we ask for relations between joint and posterior probabilities: for random variables $X, Z \in B \subseteq \mathbb{R}^m$ and $Y \in A \subseteq \mathbb{R}^d$, we are interested in Wasserstein distances between $P_{Y,X}, P_{Y.Z} \in \mathcal{P}_p(A \times B)$ and $P_{X|Y=y}, P_{Z|Y=y} \in \mathcal{P}_p(B)$. Since $\pi^1_\sharp P_{Y,X} = P_Y$ as well as $\pi^1_\sharp P_{Y,Z} = P_Y$, we see that the joint probabilities belong indeed to the subset

$$\mathcal{P}_{p,Y}(A \times B) \coloneqq \{\gamma \in \mathcal{P}_p(A \times B) : \pi^1_\sharp \gamma = P_Y\}.$$

For $p = 2$, this set was considered as set of velocity plans at P_Y in [6, Sect. 12.4] and [23, Sect. 4]. It is the basis for defining the so-called geometric tangent space of $\mathcal{P}_2(\mathbb{R}^d)$ which was used, e.g. in [27, 4] for handling (neural) Wasserstein gradient flows of maximum mean discrepancies.

We will frequently apply the disintegration formula [6, Theorem 5.3.1] which says that for a measure $\gamma \in \mathcal{P}(A \times B)$ with $\pi^1_\sharp \gamma = \mu \in \mathcal{P}(A)$, there exists a μ-a.e. uniquely determined Borel family of probability measures $(\gamma_y)_{y \in A}$ such that

$$\int_{A \times B} f(y, x) \,\mathrm{d}\gamma(y, x) = \int_A \int_B f(y, x) \,\mathrm{d}\gamma_y(x) \mathrm{d}\mu(y)$$

for any Borel measurable map $f : A \times B \to [0, +\infty]$. In particular, for $\gamma = P_{Y,X} \in \mathcal{P}(A \times B)$, the disintegration formula reads as

$$\int_{A \times B} f(y, x) \,\mathrm{d}P_{Y,X}(y, x) = \int_A \int_B f(y, x) \,\mathrm{d}P_{X|Y}(x) \mathrm{d}P_Y(y). \tag{4}$$

3 Conditional Wasserstein distance

We have already seen that in general we can only expect inequality in (2). Towards equality, we introduce a conditional Wasserstein distance which allows only couplings which leave the Y-component invariant. To this end, we introduce the set of special 4-plans

$$\Gamma^4_Y = \Gamma^4_Y(P_{Y,X}, P_{Y,Z}) \coloneqq \Big\{ \alpha \in \Gamma(P_{Y,X}, P_{Y,Z}) : \pi^{1,3}_\sharp \alpha = \Delta_\sharp P_Y \Big\},$$

where $\Delta : A \to A^2$, $y \mapsto (y, y)$ is the diagonal map. Note that $\Delta^{-1}(y_1, y_2) = \emptyset$ if $y_1 \neq y_2$ and $\Delta^{-1}(y_1, y_2) = y$ if $y_1 = y_2 = y$. Then, we define the *conditional Wasserstein-p distance*, $p \in [1, \infty)$ by

$$W_{p,Y}(P_{Y,X}, P_{Y,Z}) \coloneqq \Big(\inf_{\alpha \in \Gamma^4_Y} \int_{(A \times B)^2} \|(y_1, x_1) - (y_2, x_2)\|^p \,\mathrm{d}\alpha \Big)^{\frac{1}{p}}. \tag{5}$$

Indeed we will see in Corollary 2 that this is a metric on $\mathcal{P}_{p,Y}(A \times B)$.

In terms of Monge maps, this means that we are considering functions $(\mathrm{Id}, T(y,\cdot)) : (y,x) \mapsto (y, T(y,x))$, where $T : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^m$ and $(\mathrm{Id}, T(y,\cdot))_{\#} P_{Y,X} = P_{Y,Z}$. The following proposition gives the desired equivalence of the form (2). The proof is given in Appendix B.

Proposition 1. *The following relations holds true.*

i) The conditional Wasserstein-p distance (5) *fulfills*

$$W_{p,Y}^p(P_{Y,X}, P_{Y,Z}) = \mathbb{E}_Y\big[W_p^p(P_{X|Y=y}, P_{Z|Y=y})\big]. \tag{6}$$

ii) Let $\alpha \in \Gamma_{p,Y}^4$ is an optimal plan in (5) *with disintegration α_{y_1,y_2} with respect to $\pi_\sharp^{1,3}\alpha$. Then $\alpha_{y,y} \in \mathcal{P}(\mathbb{R}^{2m})$ is an optimal plan for $W_p(P_{X|Y=y}, P_{Z|Y=y})$ for P_Y-a.e. $y \in A$.*

iii) There exists a collection of optimal plans $\alpha_y \in \Gamma(P_{X|Y=y}, P_{Z|Y=y})$, $y \in A$ for $W_p(P_{X|Y=y}, P_{Z|Y=y})$ such that

$$\alpha \coloneqq \int_A \mathrm{d}\delta_{y_1}(y_2)\,\mathrm{d}\alpha_{y_1}(x_1,x_2)\mathrm{d}P_Y(y_1) \tag{7}$$

is a well-defined coupling in Γ_Y^4 which is optimal in (5).

For $p = 2$, it was shown in [6, Sect. 12.4] and [23, Sect. 4] that the square root of the right-hand side in (4) is a metric on $\mathcal{P}_{2,Y}(A \times B)$. The proof can be generalized in a straightforward way for $p \in [1,\infty)$. Thus, by Proposition 1, we have the following corollary.

Corollary 2. *The conditional Wasserstein distance $W_{p,Y}$ is a metric on $\mathcal{P}_{p,Y}(A \times B)$.*

Interestingly, for $p = 2$, there was also given an equivalent definition by [23] of $\mathcal{W}_{p,Y}$, namely

$$\mathcal{W}_{p,Y}(P_{Y,X}, P_{Y,Z}) \coloneqq \inf_{\beta \in \Gamma_Y^3(P_{Y,X}, P_{Y,Z})} \Big(\int_{A \times B^2} \|x_1 - x_2\|^p \,\mathrm{d}\beta(y, x_1, x_3) \Big)^{\frac{1}{p}}$$

with the set of 3-plans

$$\Gamma_Y^3(P_{Y,X}, P_{Y,Z}) \coloneqq \{\beta \in \mathcal{P}_p(A \times B^2) : \pi_\sharp^{1,2}\beta = P_{Y,X}, \pi_\sharp^{1,3}\beta = P_{Y,Z}\}.$$

The relation between the admissible 3-plans and 4-plans is given in the following proposition which proof can be found in the appendix.

Proposition 3. *The map $\pi_\sharp^{1,2,4} : \Gamma_Y^4(P_{Y,X}, P_{Y,Z}) \to \Gamma_Y^3(P_{Y,X}, P_{Y,Z})$ is a bijection and for every $\alpha \in \Gamma_Y^4(P_{Y,X}, P_{Y,Z})$ it holds*

$$\int_{(A\times B)^2} \|(y_1,x_1) - (y_2,x_2)\|^p \,\mathrm{d}\alpha = \int_{A \times B^2} \|x_1 - x_2\|^p \,\mathrm{d}\pi_\sharp^{1,2,4}\alpha.$$

4 Dual formulation of $W_{1,Y}$ and relation to GAN loss

Interestingly, the conditional Wasserstein-1 distance recovers loss functions in the conditional Wasserstein GAN literature [1, 31, 38]. Wasserstein GANs [8] aim to sample from a target distribution P_X based on a simpler distribution P_Z, a generator $G = G_\theta$ is learned such that the Wasserstein-1 distance in its dual formulation

$$W_1(P_X, G_\# P_Z) = \max_{f \in \mathrm{Lip}_1} \left\{ \mathbb{E}_X[f] - \mathbb{E}_Z\big[f \circ G\big] \right\}$$

is minimized, where Lip_1 denotes the set of 1-Lipschitz continuous functions. At the same time, a discriminator $f = f_\omega$ is learned such that the final Wasserstein GAN learning problem becomes

$$\min_\theta \max_\omega \{\mathbb{E}_X[f] - \mathbb{E}_Z[f \circ G]\} \quad \text{subject to} \quad f \in \mathrm{Lip}_1.$$

Usually, the the 1-Lipschitz condition is enforced via so-called weight-clipping [8].

In [1], this approach was generalized to inverse problems. Assume that $A \subset \mathbb{R}^d$ and $B \subset \mathbb{R}^m$ are compact sets throughout this section. For given $y \in A$, an optimal $h(y, \cdot) \in \mathrm{Lip}_1$ is found in

$$W_1(P_{X|Y=y}, G(y,\cdot)_\# P_Z) = \max_{h(y,\cdot) \in \mathrm{Lip}_1} \left\{ \mathbb{E}_{X|Y=y}[h(y,x)] - \mathbb{E}_Z\big[h(y, G(y,\cdot))\big] \right\}.$$

Now the authors take the expectation value on both sides and exchange expectation and maximum to get, together with (4), the relation

$$\mathbb{E}_Y[W_1(P_{X|Y=y}, G(y,\cdot)_\# P_Z] = \max_h \{\mathbb{E}_{Y,X}[h] - \mathbb{E}_{Y,Z}[h(y, G(y,\cdot)]\}, \tag{8}$$

where the maximum is taken over functions h which are Lipschitz-1 continuous in the second variable. However, exchanging maximum and expectation value requires that $(y,x) \mapsto h(y,x)$ is measurable which is not always the case. This „gap" was fixed under stronger assumptions, e.g. on the posterior, in [38].

In this section, we show that the dual formulation of the conditional Wasserstein distance $W_{1,Y}$ leads naturally to the desired loss on the right-hand side of (8) for an appropriate regular function set for h. More precisely, we have the following theorem which is proved in the appendix.

Theorem 4. *Let $A \subset \mathbb{R}^d$ and $B \subset \mathbb{R}^m$ be compact sets. Then it holds*

$$W_{1,Y}(P_{Y,X}, P_{Y,Z}) = \sup_{h \in \mathcal{F}} \{\mathbb{E}_{Y,X}[h] - \mathbb{E}_{Y,Z}[h(y, G(y,\cdot)]\},$$

where $\mathcal{F}$ denotes the set of bounded, upper semi-continuous functions $h : A \times B \to \mathbb{R}$ satisfying $|h(y,x_1) - h(y,x_2)| \le \|x_1 - x_2\|$ for all $y \in A$ and all $x_1, x_2 \in B$.

5 Geodesics and velocity fields

In this section, we deal with geodesics and velocity fields in $\big(\mathcal{P}_{Y,2}(\mathbb{R}^d \times \mathbb{R}^m), W_{2,Y}\big)$. We restrict our attention to $p = 2$ and $A = \mathbb{R}^d$, $B = \mathbb{R}^m$. Coming from inverse problems, we have considered probability measures $P_{X,Y}$ related to random variables $(Y, X) \in \mathbb{R}^d \times \mathbb{R}^m$. When switching to flows, it is more convenient to address equivalently just probability measures on $\mathbb{R}^d \times \mathbb{R}^m$.

Let us recall some results which can be found, e.g. in [6] for our setting. A curve $\mu\colon [0,1] \to \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^m)$ is called a *geodesic* if

$$W_2(\mu_s, \mu_t) = |s - t| W_2(\mu_0, \mu_1) \quad \text{for all} \quad s, t \in [0,1].$$

The Wasserstein space is geodesic, i.e. any two measures $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^m)$ can be connected by a geodesic. Let $e_t : (\mathbb{R}^d \times \mathbb{R}^m)^2 \to \mathbb{R}^n \times \mathbb{R}^m$, $t \in [0,1]$ by defined by

$$e_t(y_1, x_1, y_2, x_2) \coloneqq \big((1-t)\pi^{1,2} + t\pi^{3,4}\big)\,(y_1, x_1, y_2, x_2) = (1-t)(y_1, x_1) + t(y_2, x_2).$$

Any geodesic $\mu\colon [0,1] \to \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^m)$ connecting $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^m)$ is determined by an optimal plan $\alpha \in \Gamma(\mu_0, \mu_1)$ in (3) via

$$\mu_t \coloneqq (e_t)_\sharp \alpha, \quad t \in [0,1]. \tag{9}$$

Conversely, any optimal plan $\alpha \in \Gamma(\mu_0, \mu_1)$ gives by (9) rise to a geodesic connecting μ_0 and μ_1. The following lemma considers curves defined by (9) which connect measures $\mu_0, \mu_1 \in \mathcal{P}_{2,Y}(\mathbb{R}^d \times \mathbb{R}^m)$. Its proof is given in the appendix and is similar to [6, Theorem 7.2.2].

Lemma 5. *Let $\mu_0, \mu_1 \in \mathcal{P}_{2,Y}(\mathbb{R}^d \times \mathbb{R}^m)$ and let $\alpha \in \Gamma^4_Y(\mu_0, \mu_1)$ be an optimal plan in (5). Then the following holds true.*

i) The curve $\mu_t \coloneqq (e_t)_\sharp \alpha$ is a geodesic in $(\mathcal{P}_{2,Y}(\mathbb{R}^d \times \mathbb{R}^m), W_{2,Y})$.

ii) The curve $(\mu_t)_y \coloneqq (1-t)\pi^1 + t\pi^2)_\sharp \alpha_{y,y}$ is the disintegration of μ_t with respect to P_Y. Further, $(\mu_t)_y$ is a geodesic in $(\mathcal{P}_2(\mathbb{R}^m), W_2)$ for P_Y-a.e. $y \in \mathbb{R}^d$.

iii) μ_t is weakly continuous.

By the following proposition, the above geodesic μ_t has an associated vector field v_t which satisfy a continuity equation. Moreover, informally speaking, the associated vector field v_t does not transport any mass in the y-component. This is related to the observation in [18].

Proposition 6. *Let $\mu_0, \mu_1 \in \mathcal{P}_{2,Y}(\mathbb{R}^d \times \mathbb{R}^m)$. Let $\alpha \in \Gamma^4_Y(\mu_0, \mu_1)$ be an optimal plan in (5) and $\mu_t = (e_t)_\sharp \alpha$, $t \in [0,1]$. Then there exists a vector field $v_t \in L^2_{\mu_t}(\mathbb{R}^d \times \mathbb{R}^m, \mathbb{R}^d \times \mathbb{R}^m)$ such that the following relations are fulfilled:*

i) $(e_t)_\sharp\left((y_2, x_2) - (y_1, x_1))\,\alpha\right) = v_t\mu_t$,

ii) $\|v_t\|_{L^2_{\mu_t}} \leq W_{2,Y}(\mu_0, \mu_1)$,

iii) for $j \leq d$, *we have* $(v_t)_j = 0$ μ_t*-a.e.*,

iv) μ_t, v_t *fulfill the continuity equation*

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0$$

in a distribultional sense, i.e. we have for all $\varphi \in C_c^\infty(\mathbb{R}^d \times \mathbb{R}^m)$ *that*

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^d \times \mathbb{R}^m} \varphi \,\mathrm{d}\mu_t = \int_{\mathbb{R}^d \times \mathbb{R}^m} \langle \nabla \phi, v_t \rangle \,\mathrm{d}\mu_t.$$

The proof is given in Appendix D and parts $i), ii), iv)$ are adapted from the proofs of [5, Theorem 17.2, Lemma 17.3.]

Furthermore, since by Lemma 5 iii), a geodesic induced by an optimal $W_{2,Y}$ plan is weakly continuous, we obtain the following proposition from [6, Proposition 8.1.8] which is needed in the numerical section.

Proposition 7. *Let* $\mu_0, \mu_1 \in \mathcal{P}_{2,Y}(\mathbb{R}^d \times \mathbb{R}^m)$. *Let* $\alpha \in \Gamma_Y^4(\mu_0, \mu_1)$ *be an optimal plan in* (5) *and* $\mu_t = (e_t)_\sharp \alpha$, $t \in [0, 1]$. *Assume that the corresponding Borel vector field* v_t *from Proposition 6 fulfills*

$$\int_0^1 \left(\sup_B (\|v_t\|_{L^2_{\mu_t}}) + \mathrm{Lip}(v_t, B) \right) \mathrm{d}t < \infty \tag{10}$$

for all compact subsets $B \subset \mathbb{R}^d \times \mathbb{R}^m$, *where* $\mathrm{Lip}(v_t, B)$ *denotes the Lipschitz constant of* v_t *on* B. *Then, for* μ_0*-a.e.* $(y, x) \in \mathbb{R}^d \times \mathbb{R}^m$, *the ODE*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\phi_t &= v_t(\phi_t), \\ \phi_0(y, x) &= (y, x), \end{aligned}$$

admits a unique global solution and $\mu_t = (\phi_t)_\sharp \mu_0$, $t \in [0, 1]$.

For special cases we can drop the requirements (10) on the Borel vector field as the following proposition, which is proved in the appendix, shows.

Proposition 8. *For* $y_i \in \mathbb{R}^d$, $I = 1, \ldots, n$, *let* $P_Y \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. *Let* $\mu_0, \mu_1 \in \mathcal{P}_{2,Y}(\mathbb{R}^d \times \mathbb{R}^m)$ *fulfill one of the following conditions:*

i) $\mu_{0,y}, \mu_{1,y}$ *are empirical measures with the same number of particles* $n \in \mathbb{N}$ *for* P_Y *a.e.* $y \in \mathbb{R}^d$. *Let* T_{y_i} *be a choice of optimal transport maps between* μ_{0,y_i} *and* μ_{1,y_i} *and let* α *be the corresponding optimal plan* $\alpha \in \Gamma_Y^4(\mu_0, \mu_1)$, *or*

ii) $\mu_{0,y}, \mu_{1,y}$ *for* P_Y*-a.e.* $y \in \mathbb{R}^d$ *are absolutely continuous with densities* $\rho_{0,y}, \rho_{1,y}$ *which are supported on open, convex, bounded, connected, subsets* $\Omega_{0,y}, \Omega_{1,y}$ *on which they are bounded away from* 0 *and* ∞*. Assume further that* $\rho_{0,y} \in C^2(\Omega_{0,y}), \rho_{1,y} \in C^2(\Omega_{1,y})$ *and let* T_y *be the associated optimal transport maps and* $\alpha \in \Gamma^4_Y(\mu_0, \mu_1)$ *be the associated optimal transport plan.*

Let $\mu_t = (e_t)_\sharp \alpha$ *with associated vector field* $v_t \in L^2_{\mu_t}$*, where* $(v_t)_j = 0$ *for all* $j \leq d$*. Then there is a representative of* v_t *such that the flow equation*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\phi_t &= v_t(\phi_t) \\ \phi_0(y,x) &= (y,x) \end{aligned}$$

admits a global solution and $\mu_t = (\phi_t)_\sharp \mu_0$*. Furthermore, for* $T \in L^2_{\mu_0}$ *defined by* $T(y_i, x) \coloneqq (y_i, T_{y_i}(x))$*, we have*

$$v_t(\phi_t(y,x)) = T(y,x) - (y,x) = \big(0, \pi^2 \circ T(y,x) - x\big)$$

for μ_0*-a.e.* $(y,x) \in \mathbb{R}^d \times \mathbb{R}^m$*.*

A Benamou-Brenier type formula for $W_{p,Y}$ is given in Appendix F.

6 Relaxation of the conditional Wasserstein distance

When working with conditional Wasserstein distances, we are facing the following problems:

1. We cannot use standard optimal transport algorithms [20] out of the box.
2. Assume that P_Y is not empirical and let $\mu \in \mathcal{P}_{2,Y}(\mathbb{R}^d \times \mathbb{R}^m)$. Then it is impossible to approximate μ by empirical measures in the $W_{2,Y}$ topology, since $\mathcal{P}_{2,Y}(\mathbb{R}^d \times \mathbb{R}^m)$ does not contain any empirical measures.
3. Assume that we are interested in the optimal plan $\alpha \in \Gamma^4_Y(\mu_0, \mu_1)$, but we are only given empirical measures $\mu_{n,0}, \mu_{n,1}$, which are W_2 approximations of μ_0, μ_1. Let Y_n be a random variable distributed as $\pi^1_\sharp \mu_{n,0}$. Even if we assume that $\pi^1_\sharp \mu_{n,0} = \pi^1_\sharp \mu_{n,1}$, Example 9 shows that we cannot guarantee that there exists a sequence of the optimal plans $\alpha_n \in \Gamma^4_{Y_n}(\mu_{n,0}, \mu_{n,1})$ that converges in any sense to α.

Example 9. We consider independent, standard normal distributed random variables $Y, X, Z \in \mathbb{R}$. Let $\mu = \nu \coloneqq P_{Y,X}$. Now we sample $(y_i, x_i, z_i) \sim (Y, X, Z)$ and define

$$\mu_n \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{y_i, x_i}, \quad \nu_n \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{y_i, z_i},$$

i.e. $\mu_n \to \mu$ and $\nu_n \to \nu$ as $n \to \infty$ in the W_2-topology. Let Y_n be a random variable distributed like $\frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. Then $\Gamma_{2,Y_n}(\mu_n, \nu_n)$ contains exactly one plan $\alpha_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i,x_i,y_i,z_i}$ which is clearly optimal. Let $\Delta : \mathbb{R}^3 \to \mathbb{R}^4$ be defined by $\Delta(y,x,z) = (y,x,y,z)$. Then $\hat{\alpha} := \lim_n \alpha_n = \Delta_\sharp (P_Y \otimes P_X \otimes P_Z)$ in the W_2-topology. Moreover, $\hat{\alpha} \in \Gamma_{2,Y}(\mu,\nu)$ and

$$\begin{aligned}\int_{R^4} \|(y_1,x_1)-(y_2,x_2)\|^2 \mathrm{d}\hat{\alpha} &= \int_{\mathbb{R}^3} \|(y_1,x_1)-(y_1,x_2)\|^2 \,\mathrm{d}(P_Y \otimes P_X \otimes P_Z) \\ &= \int_{\mathbb{R}^2} \|x_1-x_2\|^2 \,\mathrm{d}(P_X \otimes P_Z) > 0 = W_{2,Y}(\mu,\nu).\end{aligned}$$

Hence $\hat{\alpha}$ is not an optimal coupling, although it is a limit of optimal couplings in the W_2 sense.

In order to overcome the above drawbacks, we define a cost function for which the OT plan $\alpha \in \Gamma(\mu_o, \mu_1)$ only approximately fulfills $\pi^{1,3}_\sharp \alpha = \Delta_\sharp P_Y$:

$$d^p_\beta((y_1,x_1),(y_2,x_2)) = \|x_1-x_2\|^p + \beta\|y_1-y_2\|^p, \quad p \in [1,\infty),\ \beta > 0.$$

For large values of β, it is very costly to move mass in y-direction. Then we denote by $W_{p,\beta}$ the OT distance with respect to the cost d^p_β, i.e. for $\mu_0, \mu_1 \in \mathcal{P}_p(A \times B)$ we set

$$W_{p,\beta}(\mu_0,\mu_1)^p \coloneqq \inf_{\alpha \in \Gamma(\mu_0,\mu_1)} \int_{(A\times B)^2} \mathrm{d}^p_\beta((y_1,x_1),(y_2,x_2))\,\mathrm{d}\alpha.$$

The same idea has been pursued in [2], where the authors rescaled the y-costs to obtain a blocktriangular map in the Knothe-Rosenblatt setting [33, 46] and similarly in [29]. Note that [2] was published on ArXiv after our first version of the present paper.

Proposition 10. *Let $\mu_0, \mu_1 \in \mathcal{P}_{p,Y}(\mathbb{R}^d \times \mathbb{R}^m)$ and let α^β be a sequence of optimal transport plans for μ_0, μ_1 with respect to $W_{p,\beta}$. Then, for $\beta \to \infty$, we have*

$$\int_{\mathbb{R}^{2d}} \|y_1-y_2\|^p \,\mathrm{d}\pi^{1,3}{}_\#(\alpha^\beta) \to 0.$$

Remark 11. *Alternatively, instead of rescaling the costs d_β we would also rescale the inputs, which was done for instance in [18, 24]. Take for instance (as we do numerically) the cost function $d^2_\beta = \|x_1-x_2\|^2 + \beta\|y_1-y_2\|^2$. Then this is equivalent to rescaling the Y-component by $\sqrt{\beta}$.*

The following proposition shows that the issue raised in Example 9 is addressed by W_{2,d_β}, at least on compact sets.

Proposition 12. *Assume that $\mu, \nu \in P_{2,Y}(A \times B)$ for compact subsets $A \subset \mathbb{R}^d, B \subset \mathbb{R}^m$ and let μ_n, ν_n be empirical measures that converge in W_2 to μ, ν. Then for a sequence $\beta_k \to \infty$ there exists an increasing subsequence n_k and optimal plans $\alpha_{n_k} \in \Gamma(\mu_{n_k}, \nu_{n_k})$ for $W_{2,d_{\beta_k}}(\mu_{n_k}, \nu_{n_k})$ such that α_{n_k} converges with respect to W_2 to an optimal plan $\alpha \in \Gamma^4_Y(\mu,\nu)$ for $W_{2,Y}(\mu,\nu)$.*

7 Bayesian flow matching

In this section, we combine the conditional Wasserstein distance with Bayesian flow matching. We start by briefly recalling flow matching and its OT variant. Then we turn to the conditional setting, where we describe a method from the literature, which we call "diagonal" Bayesian flow matching and introduce our new OT Bayesian flow matching.

Remark 13. *Usually, in conditional generative modelling, the word "conditional" appears the context of inverse problems (or solving class conditional problems). However, in the vanilla flow matching [35] the word "conditional" is used for the loss and the whole procedure is referred to as "conditional flow matching". Therefore, we will call the flow matching procedure for inverse problems "Bayesian flow matching".*

Flow matching and OT flow matching aim to sample from a target distribution P_X by learning the velocity field v_t of a flow ODE [14]

$$\begin{aligned} \frac{d}{dt}\phi_t(x) &= v_t(\phi_t(x)), \quad t \in [0,1], \\ \phi_0(x) &= x, \end{aligned} \tag{11}$$

which transports samples from an initial distribution P_Z to those from P_X. Once an approximate velocity field v_t^θ is learned, it can be replaced in (11) to get a neural ODE [14].

Flow Matching The flow matching framework [35, 36] learns v_t^θ based on *linear interpolation* between independent Z and X, i.e.

$$X_t = (1-t)\ Z + t\ X$$

and consequently

$$X - Z = \frac{d}{dt} X_t = v_t(X_t).$$

In a minibatch setting with $(\boldsymbol{z}, \boldsymbol{x}) = ((z_i, x_i))_{i=1}^I$ sampled from the product distribution $P_Z \otimes P_X$, this becomes

$$v_t(\boldsymbol{x}_t) \coloneqq \boldsymbol{x} - \boldsymbol{z}, \quad \boldsymbol{x}_t := (1-t)\boldsymbol{z} + t\boldsymbol{x}.$$

Consequently, an approximating velocity field v_t^θ can be learned by minimizing the loss function

$$L_{\mathrm{FM}}(\theta) \coloneqq \mathbb{E}_{(z,x)\sim P_Z\otimes P_X, t\sim U([0,1])}\left[\|v_t^\theta(x_t) - (x-z)\|^2\right].$$

The objective L_{FM} can be also derived differently, with ideas inspired by the score matching framework [30, 52]. Then instead of regressing to the true velocity field at x_t, they show that regressing to it has the same gradients when one conditions at x ("conditional" flow matching [35, Theorem 2]), which leads to the same loss formulation.

OT Flow Matching In contrast to the above linear interpolation, the authors in [50, 44] suggested to use the $W_2(P_Z, P_X)$ coupling γ, respectively its Monge map T and the corresponding *McCann interpolation* [39]

$$X_t \coloneqq T_t(Z) = (1-t)Z + tT(Z)$$

which leads to

$$T(Z) - Z = \frac{d}{dt} X_t = v_t(X_t).$$

By Proposition 16, the associated Borel vector field of the geodesic is given by $v_t = (T - \mathrm{Id}) \circ T_t^{-1}$. In a minibatch setting, this corresponds to sampling $(\boldsymbol{z}, \boldsymbol{x})$ from $P_Z \otimes P_X$ and calculating the optimal plan γ between $\frac{1}{I}\sum_{i=1}^I \delta_{z_i}$ and $\frac{1}{I}\sum_{i=1}^I \delta_{x_i}$, see (12), which is supported on $(z_i, T(z_i))_{i=1}^I$. Consequently, the loss function becomes

$$L_{\mathrm{OT}}(\theta) \coloneqq \mathbb{E}_{(z,x)\sim\gamma, t\sim U([0,1])}\big[\|v_t^\theta(x_t) - (x-z)\|^2\big],$$

where $x_t := T_t(z, x)$.

Let us turn to the conditional setting. In inverse problems, samples from the posterior measure are usually not available. In the conditional setting the corresponding flow ODE reads

$$\begin{aligned} \frac{d}{dt}\phi_t(y,x) &= v_t(\phi_t(y,x)), \quad t \in [0,1], \\ \phi_0(y,x) &= (y,x). \end{aligned}$$

To this end, we pick the target measure as the joint distribution $P_{Y,X}$ and start from $P_{Y,Z}$. We do not want mass movement in Y-direction, as this would mean the measurement would change and we would not sample the posterior, which amounts to the Y-component of v_t being (close to) zero, cf. Proposition 6.

Diagonal Bayesian Flow Matching In [53, 55] flow matching is extended to the conditional setting. Given independent Z and (Y, X) we again consider the linear interpolation between Z and X given by

$$X_t = (1-t)\ Z + t\ X.$$

Then (Y, X_t) interpolates between (Y, Z) and (Y, X). Consequently

$$(0, X - Z) = \frac{d}{dt}(Y, X_t) = v_t(Y, X_t).$$

Given a minibatch $(\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{x}) = ((z_i, y_i, x_i))_{i=1}^I$ sampled from the product distribution $P_Z \otimes P_{Y,X}$, this becomes

$$v_t(\boldsymbol{y}, \boldsymbol{x}_t) \coloneqq (0, \boldsymbol{x} - \boldsymbol{z}), \quad \boldsymbol{x}_t \coloneqq (1-t)\boldsymbol{z} + t\boldsymbol{x}.$$

This yields the diagonal Bayesian flow matching loss

$$L_{Y,\mathrm{FM}}(\theta) = \mathbb{E}_{(z,y,x)\sim P_Z\otimes P_{Y,X}, t\sim U([0,1])}[\|v_t^\theta(y,x_t) - (x-z)\|^2].$$

Under the assumption $y_i \neq y_j$ for $i \neq j$ this diagonal matching coincides with the only admissible plan in the conditional Wasserstein distance. In general however, according to Example 9, this approach does not approximate OT plans as X and Z are essentially independently.

OT Bayesian Flow Matching For $P_{Y,Z}, P_{Y,X}$ as in Proposition 8 there exists an optimal plan $\alpha \in \Gamma_Y^4(P_{Y,Z}, P_{Y,X})$ and corresponding map T. Furthermore there exists a vector field $v_t \in L^2_{\mu_t}$ such that there exists a solution ϕ_t to the flow equation which satisfies

$$v_t(\phi_t(y,x)) = T(y,x) - (y,x) = \big(0, (\pi^2 \circ T)(y,x) - x\big).$$

Setting

$$X_t \coloneqq T_t(Y,Z) = (1-t)Z + t(\pi^2 \circ T)(Y,Z)$$

we have that (Y, X_t) interpolates between (Y,Z) and (Y,X) and

$$(0, (\pi^2 \circ T)(Y,Z) - Z) = \frac{d}{dt}(Y, X_t) = v_t(Y, X_t).$$

Given a minibatch $(\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{x}) = ((z_i, y_i, x_i))_{i=1}^I$ sampled from the product distribution $P_Z \otimes P_{Y,X}$, we can calculate the optimal map T and corresponding conditional coupling α between $(\boldsymbol{y}, \boldsymbol{z})$ and $(\boldsymbol{y}, \boldsymbol{x})$. The plan α by construction is only supported on $(y_i, z_i, y_i, (\pi^2 \circ T)(y_i, x_i)))$. Now let $(y, z, y, x) \sim \alpha$, then we have

$$v_t(y, x_t) = T(y,z) - (y,z) = (y,x) - (y,z) = (0, x-z)$$

where $x_t \coloneqq T_t(y,z)$. This gives rise to the following loss

$$L_{Y,\mathrm{OT}}(\theta) = \mathbb{E}_{((y,z,y,x)\sim\alpha, t\sim U([0,1])}[\|v_t^\theta(y,x_t) - (x-z)\|^2].$$

In practice, we use Proposition 10 to approximate the optimal coupling α. Therefore we allow small errors in the Y-component, in order to move more optimally in the X-direction, which is more in the spirit of Proposition 12. Numerically, instead of taking the optimal transport plan with respect to the modified cost function, we rescale the Y-part, see remark 11.

8 Numerical experiments

In this section, we want to show cases in which it is beneficial to use the conditional Wasserstein distance. First, we verify that the convergence result for an increasing parameter β given in Proposition 10 for particle flows to MNIST [17]. Then we show the advantages of our Bayesian OT flow matching procedure on a Gausian mixture model (GMM) toy example and on CIFAR10 [34] class-conditional image generation.

8.1 Particle flow convergence

In this example, we minimize $W_{Y,\beta}(P_{Y,X}, P_{Y,Z})$ for the empirical measures We consider the particle flow, i.e., the flow from (Y, Z) to (Y, X) for empirical distributions, where we compute a particle flow [4] from (Y, Z) to (Y, X). We follow a particle flow path, i.e., we define a curve $z(0)_i \sim \mathcal{N}(0, I)$ which follows

$$\dot{z}(t) = -\eta \nabla_{z(t)} \mathcal{D}\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{y_i,x_i}, \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i,z(t)_i}\right),$$

for an appropriate "distance" $\mathcal{D}$, step size η and given joint samples $(y_i, x_i)_{i=1}^n$. We choose $\mathcal{D}$ as an approximation of $W_{2,\beta}$ by rescaling Y and using the Sinkorn divergence as the sample based distance measure [21, 19], where the "blur" parameter is chosen so small that it is close to the Wasserstein distance. This way we can numerically verify the convergence in Proposition 10. Note that there are no neural networks involved in this example.

We see in Fig. 1 that increasing β indeed yields plans which transport no mass in Y-direction anymore, which has the consequence that the generated images fit the corresponding label. It can be seen that for $\beta = 5$ each row only has one type of digit.

$\beta = 1$ $\beta = 3$ $\beta = 4$ $\beta = 5$

Figure 1: Class conditional MNIST particle flow for different choices of β. With increasing β the labels are better fitted.

8.2 GMM example

Here we use an experimental setup from [26]: in (1), we choose P_X to be a GMM in $\mathbb{R}^5$ with 10 mixture components, uniformly chosen means in $[-1, 1]$ and standard deviation 0.1. We apply linear diagonal forward model $f = (f_{i,j})_{i,j=1}^5$ with $f_{i,i} = \frac{0.1}{i+1}$ and zero components otherwise and Gaussian noise with standard deviation 0.1. This yields a posterior measure $P_{X|Y=y}$ which is a also a GMM [26, Lemma 11]. Therefore we can sample and evaluate the true posterior as groundtruth. We train a diagonal Bayesian flow matching model according to $L_{Y,FM}$ and our OT Bayesian flow matching according to $L_{Y,OT}$ with POT [20]

for the same amount of time on a fixed dataset of size 10000, where we choose the best model according to a validation set of size 2000. We parameterize both models with around 140k parameters. We evaluate them using the Sinkhorn distance [21, 19] with the package GeomLoss averaged with 100 posteriors and over 10 training and test runs with randomly sampled mixtures. The sampling is done via an explicit Euler discretization of 10 steps. Our proposed model trained according to $L_{Y,OT}$ with $\beta = 20$ obtains an average Sinkhorn distance of **0.0235**, whereas the naive model obtains a value of **0.0255**. In Figure 2 one can see that both models approximate the posterior very well.

8.3 Class conditional image generation

We apply our Bayesian OT flow matching for conditional image generation. We choose the condition Y to be the class labels in order to generate samples of CIFAR10 for a given class. We simulate the flows for different values of β, by which we mean that we rescale the Y-component by β as mentioned in Remark 11. We also simulate flows using the "diagonal" plans which coincide with the diagonal Bayesian flow matching objective [53]. For high quality inference we use an adaptive step size solver (Runge-Kutta of order 5). We also compare quantitatively when sampling with a Euler scheme with 20 steps. The samples in Fig. 3 are generated using the adaptive step size solver and sorted by class labels. For low values of β we see that the resulting samples do not match their class labels, increasing β leads to accurate class representations. The samples are generated given the labels of the training samples, therefore we see improved FID results as β increases. The diagonal flow matching objective has the correct class representations, however since the associated couplings are not optimal our experiments suggest that this leads to higher variance during training and therefore slightly lower image quality, see [50] for more details on the advantages of OT based flow matching. We evaluate the methods on a fixed number of epochs, for completeness we note that the diagonal method improves to an FID (AD) of 4.88 under additional training.

9 Conclusions

Inspired from applications in Bayesian inverse problems, we introduced conditional Wasserstein distances. We managed to rewrite these distances as expectations of the Wasserstein distances with respect to the observation. Therefore we are able to directly infer posterior guarantees when trained with the corresponding losses. Furthermore, we calculated the dual of the conditional Wasserstein-1 distance, when the probability measures are compactly supported and recovered well-known conditional Wasserstein GAN losses. We established corresponding velocity fields in the gradient flow theory and used our results to design a new Bayesian flow matching algorithm. Theoretically, improving Proposition 8 for non-discrete P_Y distributions would be the next step. In particular one would need to show measurability of the "conditional Monge maps", which does not seem to be easy. Further, from a practical

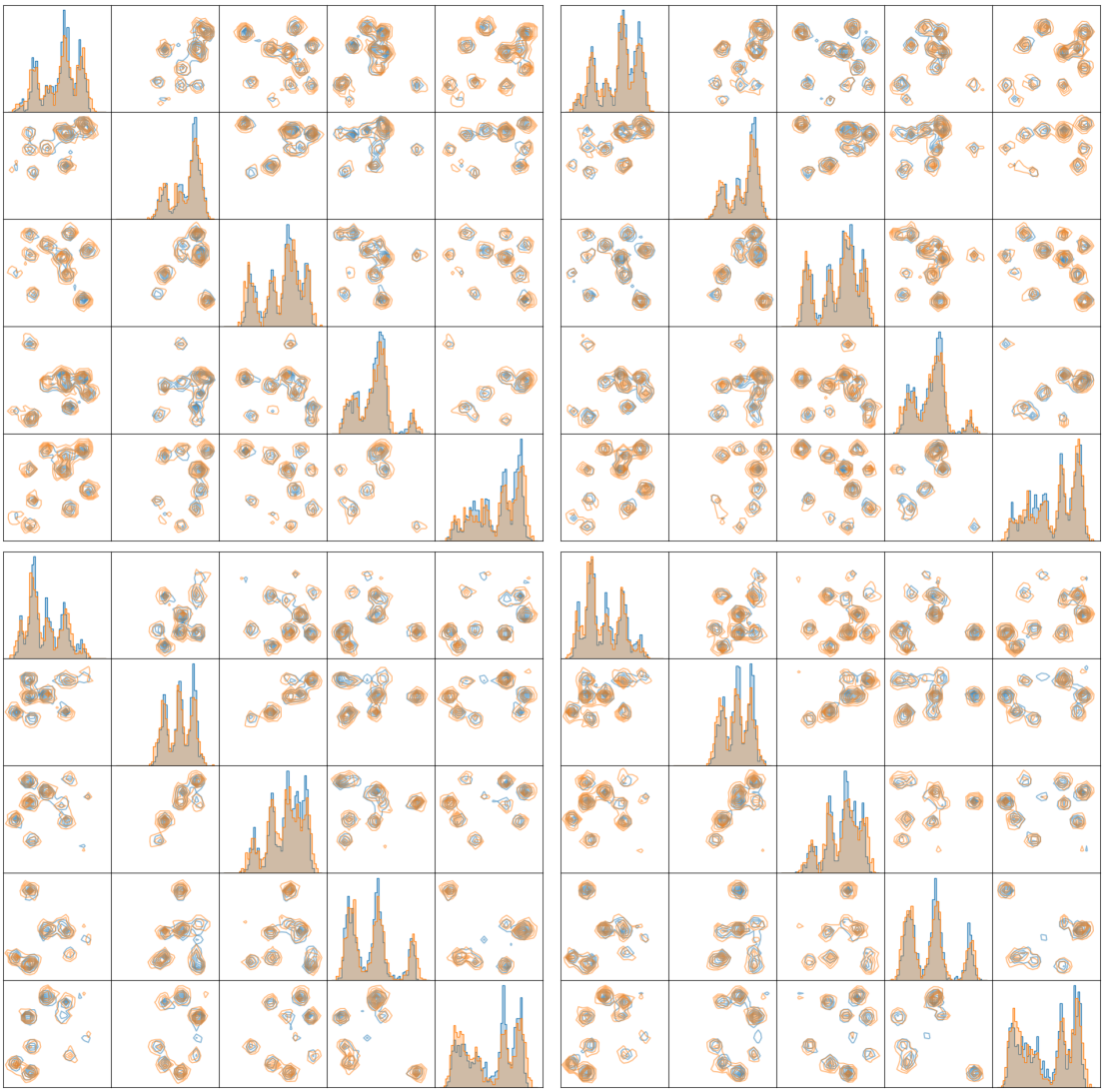

Figure 2: Posterior histograms for different methods with diagonal Bayesian flow matching on the left and our OT Bayesian flow matching on the right. Ground truth posterior is in orange and model prediction in blue.

viewpoint, a clear use case would be in conditional domain translation, i.e., when the latent distribution is not a standard Gaussian, but given by some data distribution. There, finding a good OT matching and making use of our proposed framework could improve existing algorithms.

	1	3	5	20	100	diagonal
FID (AD)	5.3	5.1	4.79	4.57	**4.38**	5.34
FID (Euler)	11.18	11.05	10.82	10.13	**9.98**	10.76

Figure 3: Class Conditional CIFAR results for different choices of β and for the diagonal couplings. Additionally FID results are reported using an adaptive step size solver (AD) and an Euler scheme with 20 steps (Euler).

Acknowledgement: P. Hagemann acknowledges funding from from the DFG within the project SPP 2298 "Theoretical Foundations of Deep Learning", C. Wald and G. Steidl gratefully acknowledge funding by the DFG within the SFB “Tomography Across the Scales” (STE 571/19-1, project number: 495365311).

References

[1] J. Adler and O. Öktem. Deep Bayesian inversion. *arXiv preprint arXiv:1811.05910*, 2018.

[2] J. Alfonso, R. Baptista, A. Bhakta, N. Gal, A. Hou, I. Lyubimova, D. Pocklington, J. Sajonz, G. Trigila, and R. Tsai. A generative flow for conditional sampling via optimal transport. *arXiv preprint arXiv:2307.04102*, 2023.

[3] F. Altekrüger, P. Hagemann, and G. Steidl. Conditional generative models are provably robust: Pointwise guarantees for Bayesian inverse problems. *Transactions on Machine Learning Research*, 2023.

[4] F. Altekrüger, J. Hertrich, and G. Steidl. Neural Wasserstein gradient flows for discrepancies with riesz kernels. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 664–690. PMLR, 23–29 Jul 2023.

[5] L. Ambrosio, E. Brué, and D. Semola. *Lectures on Optimal Transport*. UNITEXT. Springer International Publishing, 2021.

[6] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[7] L. Ardizzone, C. Lüth, J. Kruse, C. Rother, and U. Köthe. Guided image generation with conditional invertible neural networks. *arXiv preprint arXiv:1907.02392*, 2019.

[8] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

[9] J. Aubin and I. Ekeland. *Applied Nonlinear Analysis*. Dover Books on Mathematics Series. Dover Publications, 2006.

[10] G. Basso. A hitchhikers guide to Wasserstein distances. *Online manuscript available at* `https://api.semanticscholar.org/CorpusID:51801464`, 2015.

[11] V. Bogachev. *Weak convergence of measures*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[12] C. Bunne, A. Krause, and M. Cuturi. Supervised training of conditional Monge maps. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[13] J. Chemseddine, P. Hagemann, and C. Wald. Y-diagonal couplings: Approximating posteriors with conditional Wasserstein distances. *arXiv preprint arXiv:2310.13433v*, 2023.

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

[15] R. T. Q. Chen. torchdiffeq, 2018.

[16] T. Cover. *Elements of Information Theory*. John Wiley & Sons, Ltd, 2005.

[17] L. Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[18] C. Du, T. Li, T. Pang, S. Yan, and M. Lin. Nonparametric generative modeling with conditional sliced Wasserstein flows. *arxiv preprint arXiv:2305.02164*, 2023.

[19] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé, and G. Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

[20] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, et al. Pot: Python Optimal Transport. *The Journal of Machine Learning Research*, 22(1):3571–3578, 2021.

[21] A. Genevay, G. Peyre, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR, 09–11 Apr 2018.

[22] Y. geun Kim, K. Lee, Y. Choi, J.-H. Won, and M. C. Paik. Wasserstein geodesic generator for conditional distributions. *arXiv preprint arXiv:2308.10145*, 2023.

[23] N. Gigli. On the geometry of the space of probability measures endowed with the quadratic Optimal Transport distance. *PhD Thesis*, 2008. cvgmt preprint.

[24] P. Hagemann, J. Hertrich, F. Altekrüger, R. Beinert, J. Chemseddine, and G. Steidl. Posterior sampling based on gradient flows of the MMD with negative distance kernel. *Proceedings ICLR*, 2024.

[25] P. Hagemann, J. Hertrich, and G. Steidl. Generalized normalizing flows via Markov chains. In *Non-local Data Interactions: Foundations and Applications*. Cambridge University Press, 2022.

[26] P. Hagemann, J. Hertrich, and G. Steidl. Stochastic normalizing flows for inverse problems: A Markov chains viewpoint. *SIAM/ASA Journal on Uncertainty Quantification*, 10(3):1162–1190, 2022.

[27] J. Hertrich, M. Gräf, R. Beinert, and G. Steidl. Wasserstein steepest descent flows of discrepancies with Riesz kernels. *Journal of Mathematical Analysis and Applications*, 2023.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[29] B. Hosseini, A. W. Hsu, and A. Taghvaei. Conditional optimal transport on function spaces. *arXiv preprint arXiv:2311.05672*, 2024.

[30] A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[31] Y.-g. Kim, K. Lee, and M. C. Paik. Conditional Wasserstein generator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7208–7219, 2023.

[32] D. P. Kingma and J. Ba. Adam: a method for stochastic optimization. In *Proceedings of the ICLR '15*, 2015.

[33] H. Knothe. Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1):39 – 52, 1957.

[34] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[35] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[36] X. Liu, C. 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.

[37] P. Manupriya, R. K. Das, S. Biswas, S. Chandhok, and S. N. Jagarlapudi. Empirical Optimal Transport between conditional distributions. *arXiv preprint arXiv:2305.15901*, 2023.

[38] J. Martin. About exchanging expectation and supremum for conditional Wasserstein GANs. *arXiv preprint arXiv:2103.13906*, 2021.

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

[40] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*, 2014.

[41] T. Nguyen, V. Nguyen, T. Le, H. Zhao, Q. H. Tran, and D. Phung. Cycle class consistency with distributional Optimal Transport and knowledge distillation for unsupervised domain adaptation. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

[42] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[43] G. Peyré and M. Cuturi. Computational Optimal Transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[44] A.-A. Pooladian, H. Ben-Hamu, C. Domingo-Enrich, B. Amos, Y. Lipman, and R. Chen. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, 2023.

[45] A. Rakotomamonjy, R. Flamary, G. Gasso, M. E. Alaya, M. Berar, and N. Courty. Optimal Transport for conditional domain matching and label shift. *Mach. Learn.*, 111(5):1651–1670, May 2022.

[46] M. Rosenblatt. Remarks on a Multivariate Transformation. *The Annals of Mathematical Statistics*, 23(3):470 – 472, 1952.

[47] F. Santambrogio. Optimal Transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[48] E. G. Tabak, G. Trigila, and W. Zhao. Data driven conditional Optimal Transport. *Machine Learning*, 110:3135–3155, 2021.

[49] J. Thickstun. Kantorovich-Rubinstein duality. *Online manuscript available at* `https://courses.cs.washington.edu/courses/cse599i/20au/resources/L12_duality.pdf`.

[50] A. Tong, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, K. Fatras, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

[51] C. Villani. *Optimal Transport: Old And New*, volume 338. Springer, 2009.

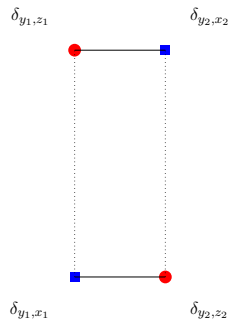

Figure 4: Visualization of the example. The blue dots belong to $P_{Y,X}$ and the red dots to $P_{Y,Z}$. The optimal coupling is the solid line, while the "diagonal" coupling is the dotted one.

[52] P. Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[53] J. B. Wildberger, M. Dax, S. Buchholz, S. R. Green, J. H. Macke, and B. Schölkopf. Flow matching for scalable simulation-based inference. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[54] C. Winkler, D. Worrall, E. Hoogeboom, and M. Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

[55] Q. Zheng, M. Le, N. Shaul, Y. Lipman, A. Grover, and R. T. Q. Chen. Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*, 2023.

A Counterexample for equality in equation 2

We provide a simple example showing that we cannot expect equality in (2). Recall that for two empirical measures $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{a_i}$ and $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{b_i}$, $a_i, b_i \in \mathbb{R}^d$, the Wasserstein-$p$ distance, $p \in [1, \infty)$ can be written as

$$W_p^p(\mu, \nu) = \inf_{\sigma \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^n \|a_i - b_{\sigma(i)}\|^p, \tag{12}$$

where $\mathcal{S}_n$ is the set of permutations on $\{1, \dots, n\}$, see [43, Proposition 2.1].

On the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with $\Omega = \{\omega_1, \omega_2\}$, $\mathcal{A} = 2^\Omega$ and $\mathbb{P}(\omega_1) = \mathbb{P}(\omega_2) = \frac{1}{2}$, we define the random variables $X, Y : \Omega \to \mathbb{R}$ by

	X	Y	Z
ω_1	0	0	n
ω_2	n	1	0

, $\quad n > 1.$

Then we have

$$P_{Y,X} = \frac{1}{2}\delta_{0,0} + \frac{1}{2}\delta_{1,n}, \quad P_{Y,Z} = \frac{1}{2}\delta_{1,0} + \frac{1}{2}\delta_{0,n}$$

which implies by (12) that

$$W_1(P_{Y,X}, P_{Y,Z}) = \frac{1}{2}\min\big\{\|(0,0)-(1,0)\| + \|(1,n)-(0,n)\|, \\ \|(0,0)-(0,n)\| + \|(1,n)-(1,0)\|\big\} = 1.$$

On the other hand, we get

$$P_{X|Y=0} = \delta_0, \quad P_{X|Y=1} = \delta_n, \quad P_{Z|Y=0} = \delta_n, \quad P_{Z|Y=1} = \delta_0, \quad P_Y = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1,$$

so that

$$\mathbb{E}_y[W_1(P_{X|Y=y}, P_{Z|Y=y})] = n = nW_1(P_{Y,X}, P_{Y,Z}).$$

Note that if we forbid the coupling to move mass across the y-direction, we actually would obtain equality, which motivates our definition of conditional Wasserstein distance, for an illustration see Fig. 4.

Note that in [31], the summation metric is considered, i.e. $\|(x_1,y_1)-(x_2,y_2)\|_{sum} = \|x_1-x_2\| + \|y_1-y_2\|$ for which our counterexample is still valid.

B Proofs of Section 3

Proof of Proposition 1. i) First we show $\geq$. Let α_{y_1,y_2} be the disintegration of some $\alpha \in \Gamma_Y^4(P_{Y,X}, P_{Y,Z})$ with respect to $\pi^{1,3}_\sharp\alpha$. Then we obtain

$$\begin{aligned} I(\alpha) &= \int_{(A\times B)^2} \|(y_1,x_1)-(y_2,x_2)\|^p \,\mathrm{d}\alpha(y_1,x_1,y_2,x_2) \\ &= \int_{A^2}\int_{B^2} \|(y_1,x_1)-(y_2,x_2)\|^p \,\mathrm{d}\alpha_{y_1,y_2}(x_1,x_2)\mathrm{d}\pi^{1,3}_\sharp\alpha(y_1,y_2) \\ &= \int_{A^2}\int_{B^2} \|(y_1,x_1)-(y_2,x_2)\|^p \,\mathrm{d}\alpha_{y_1,y_2}(x_1,x_2)\,\mathrm{d}\Delta_\sharp P_Y(y_1,y_2) \\ &= \int_{A}\int_{B^2} \|(y,x_1)-(y,x_2)\|^p \,\mathrm{d}\alpha_{y,y}(x_1,x_2)\mathrm{d}P_Y(y) \\ &= \int_{A}\int_{B^2} \|x_1-x_2\|^p \,\mathrm{d}\alpha_{y,y}(x_1,x_2)\mathrm{d}P_Y(y). \end{aligned} \tag{13}$$

Next, we show that $\alpha_{y,y} \in \Gamma(P_{X|Y=y}, P_{Z|Y=y})$ a.e., which means $\pi^1_\sharp \alpha_{y,y} = P_{X|Y=y}$ and $\pi^2_\sharp \alpha_{y,y} = P_{Z|Y=y}$ a.e.. Using (4), we obtain indeed for all Borel measurable functions $f : A \times B \to [0, \infty]$ that

$$\begin{aligned}
\int_A \int_B f(y, x_1) \, \mathrm{d}\pi^1_\sharp(\alpha_{y,y})(x_1) \mathrm{d}P_Y(y) &= \int_{A^2} \int_B f(y_1, x_1) \, \mathrm{d}\pi^1_\sharp \alpha_{y_1,y_2}(x_1) \, \mathrm{d}(\Delta)_\sharp P_Y(y_1, y_2) \\
&= \int_{A^2} \int_B f(y_1, x_1) \, \mathrm{d}\pi^1_\sharp(\alpha_{y_1,y_2})(x_1) \, \mathrm{d}\pi^{1,3}_\sharp \alpha(y_1, y_2) \\
&= \int_{A^2 \times B^2} f(y_1, x_1) \, \mathrm{d}\alpha_{y_1,y_2}(x_1, x_2) \, \mathrm{d}\pi^{1,3}_\sharp \alpha(y_1, y_2) \\
&= \int_{A^2 \times B^2} f(y_1, x_1) \, \mathrm{d}\alpha = \int_{A \times B} f(y_1, x_1) \, \mathrm{d}\pi^{1,2}_\sharp \alpha(y_1, x_1) \\
&= \int_{A \times B} f(y_1, x_1) \, \mathrm{d}P_{Y,X}(y_1, x_1).
\end{aligned}$$

Consequently, we have $I(\alpha) \geq \mathbb{E}_{y \sim P_Y} \mathbb{E}\big[W^p_p(P_{X|Y=y}, P_{Z|Y=y})\big]$ and since $W^p_{p,Y}(P_{Y,X}, P_{Y,Z}) = \inf_\alpha I(\alpha)$, this gives assertion.

Now we prove the opposite direction $\leq$. For any $y \in A$, let $\alpha_y \in \Gamma(P_{X|Y=y}, P_{Z|Y=y})$ be an optimal plan, i.e.

$$\mathcal{W}^p_p(P_{X|Y=y}, P_{Z|Y=y}) = \int_{B^2} \|x_1 - x_2\|^p \, \mathrm{d}\alpha_y(x_1, x_2).$$

This implies

$$\int_A \mathcal{W}^p_p(P_{X|Y=y}, P_{Z|Y=y}) \, \mathrm{d}P_Y(y) = \int_A \int_{B^2} \|x_1 - x_2\|^p \, \mathrm{d}\alpha_{y,y}(x_1, x_2) \mathrm{d}P_Y(y). \tag{14}$$

ii) For an optimal $\alpha \in \Gamma^4_Y(P_{Y,X}, P_{Y,Z})$, we have by Part i) and (13) that

$$\begin{aligned}
W^p_{p,Y}(P_{Y,X}, P_{Y,Z}) &= \int_A W^p_p(P_{X|Y=y}, P_{Z|Y=y}) \, \mathrm{d}P_y(y) \\
&= \int_A \int_{B^2} \|x_1 - x_2\|^p \, \mathrm{d}\alpha_{y,y}(x_1, x_2) \mathrm{d}P_Y(y).
\end{aligned}$$

Hence we get

$$0 = \int_A \Big(\int_{B^2} \|x_1 - x_2\|^p \, \mathrm{d}\alpha_{y,y}(x_1, x_2) - W^p_p(P_{X|Y=y}, (P_{Z|Y=y}) \Big) \mathrm{d}P_Y(y).$$

The inner integrand is nonnegative which finally implies that it is zero P_Y-a.e. and therefore $\alpha_{y,y}$ is an optimal plan in $W_p(P_{X|Y=y}, P_{Z|Y=y})$.

iii) Let α be defined by (7), i.e.,

$$\begin{aligned}&\int\limits_{(A\times B)^2} f(y_1,x_1,y_2,x_2)\,\mathrm{d}\alpha(y_1,x_1,y_2,x_2)\\&=\int\limits_A\int\limits_{A\times B^2} f(y_1,x_1,y_2,x_2)\,\mathrm{d}(\delta_{y_1}\times\alpha_{y_1})(y_2,x_1,x_2)\mathrm{d}P_Y(y_1)\end{aligned}$$

for all Borel measurable functions $f:(A\times B)^2\to[0,+\infty]$. Indeed α is a well defined probability measure on $(A\times B)^2$ by the following reasons: by [6, Lemma 12.4.7], we can choose a Borel family $(\alpha_y)_y$. For any Borel set $\mathcal{S}\subseteq A\times B\times B$, we have

$$(\delta_y\times\alpha_y)(\mathcal{S})=\int_{A\times B^2}1_{\mathcal{S}}(\tilde{y},x_1,x_2)\,\mathrm{d}(\delta_y\times\alpha_y)(\tilde{y},x_1,x_2)=\int_{B^2}1_{\mathcal{S}}(y,x_1,x_2)\,\mathrm{d}\alpha_y.$$

By [6, Equation 5.3.1] the function $y\mapsto\int_{B^2}1_{\mathcal{S}}(y,x_1,x_2)\mathrm{d}\alpha_y$ is Borel measurable. Consequently also $y\mapsto\delta_y\times\alpha_y(\mathcal{S})$ is Borel measurable and thus α is well defined. Then (14) can be rewritten as

$$\int_A\mathcal{W}_p^p(P_{X|Y=y},P_{Z|Y=y})\,\mathrm{d}P_Y(y)=\int_{(A\times B)^2}\|(y_1,x_1)-(y_2,x_2)\|^p\,\mathrm{d}\alpha(y_1,x_1,y_2,x_2).\quad(15)$$

It remains to show that $\alpha\in\Gamma_Y(P_{Y,X},P_{Y,Z})$ which means $\pi^{1,3}_\sharp\alpha=\Delta_\sharp P_Y$, $\pi^{1,2}_\sharp\alpha=P_{Y,X}$ and $\pi^{3,4}_\sharp\alpha=P_{Y,Z}$. The first equality follows from

$$\begin{aligned}\int_{A^2}f(y_1,y_2)\mathrm{d}\pi^{1,3}_\sharp\alpha&=\int_{(A\times B)^2}(f\circ\pi^{1,3})(y_1,x_1,y_2,x_2)\,\mathrm{d}\alpha(y_1,x_1,y_2,x_2)\\&=\int_{(A\times B)^2}f(y_1,y_2)\,\mathrm{d}\delta_{y_1}(y_2)\,\mathrm{d}\alpha_{y_1}(x_1,x_2)\mathrm{d}P_Y(y_1)\\&=\int_A f(y,y)\,\mathrm{d}P_Y(y)=\int_{A^2}f(y_1,y_2)\,\mathrm{d}(\Delta_\sharp P_Y)(y_1,y_2)\end{aligned}$$

for all Borel functions $f:A^2\to[0,+\infty]$, and the second one from

$$\begin{aligned}\int_{A\times B}f(y,x)\mathrm{d}\pi^{1,2}_\sharp\alpha(y,x)&=\int_{(A\times B)^2}f(y_1,x_1)\,\mathrm{d}\delta_{y_1}(y_2)\mathrm{d}\alpha_{y_1}(x_1,x_2)\mathrm{d}P_Y(y_1)\\&=\int_{A\times B}f(y,x)\mathrm{d}\pi^1_\sharp\alpha_y(x)\,\mathrm{d}P_Y(y)\\&=\int_{A\times B}f(y,x)\mathrm{d}P_{X|Y=y}(x)\,\mathrm{d}P_Y(y)\\&=\int_{A\times B}f(y,x)\,\mathrm{d}P_{Y,X}(y,x)\end{aligned}$$

for all Borel functions $f : A \times B \to [0, +\infty]$. The third equality follows analogously.

The final assertion follows from (15) and the equality relation (6). □

Proof of Proposition 3. Let $\kappa : A \times B^2 \to (A \times B)^2$ be defined by $(y, x_1, x_2) \mapsto (y, x_1, y, x_2)$. We show that $\kappa_\sharp : \Gamma_Y^3(P_{Y,X}, P_{Y,Z}) \to \Gamma_Y^4(P_{Y,X}, P_{Y,Z})$ is the inverse of $\pi_\sharp^{1,2,4}$. Since $\mathrm{Id}_{(A\times B^2)} = \pi^{1,2,4} \circ (\Delta \circ \pi^1, \pi^2, \pi^3)$, it remains to show that $\kappa_\sharp \circ \pi_\sharp^{1,2,4} = \mathrm{Id}_{\Gamma_Y^4(P_{Y,X},P_{Y,Z})}$. For $\alpha \in \Gamma^4(P_{Y,X}, P_{Y,Z})$ and Borel measurable function $f : (A \times B)^2 \to [0, +\infty]$, we have

$$\begin{aligned}
&\int_{(A\times B)^2} f(y_1, x_1, y_2, x_2)\, \mathrm{d}\kappa_\sharp \pi_\sharp^{1,2,4}\alpha = \int_{(A\times B)^2} f(y_2, x_1, y_2, x_2)\, \mathrm{d}\alpha(y_1, x_1, y_2, x_2) \\
&= \int_{A^2}\int_{B^2} f(y_2, x_1, y_2, x_2)\, \mathrm{d}\alpha_{y_1,y_2}(x_1, x_2)\mathrm{d}\pi_\sharp^{1,3}\alpha(y_1, y_2) \\
&= \int_{(A\times B)^2} f(y_1, x_1, y_2, x_2)\, \mathrm{d}\alpha_{y_1,y_2}(x_1, x_2)\mathrm{d}\Delta_\sharp P_Y(y_1, y_2) \\
&= \int_{(A\times B)^2} f(y_1, x_1, y_2, x_2)\, \mathrm{d}\alpha.
\end{aligned}$$

The second claim follows by

$$\begin{aligned}
&\int_{(A\times B)^2} \|(y_1, x_1) - (y_2, x_2)\|^p\, \mathrm{d}\alpha = \int_{A^2}\int_{B^2} \|(y_1, x_1) - (y_2, x_2)\|^p\, \mathrm{d}\alpha_{y_1,y_2}\mathrm{d}\pi_\sharp^{1,3}\alpha(y_1, y_2) \\
&= \int_{A^2}\int_{B^2} \|(y_1, x_1) - (y_2, x_2)\|^p\, \mathrm{d}\alpha_{y_1,y_2}\mathrm{d}\Delta_\sharp P_Y(y_1, y_2) \\
&= \int_{A}\int_{B^2} \|(y, x_1) - (y, x_2)\|^p \mathrm{d}\alpha_{y,y}\mathrm{d}P_Y(y) \\
&= \int_{A}\int_{B^2} \|x_1 - x_2\|^p\, \mathrm{d}\pi_\sharp^{2,3,4}\alpha. \qquad \square
\end{aligned}$$

C Proofs of Section 4

The proof uses similar arguments as the short notes [49] and [10], which are derivations for the dual for the "usual" Wasserstein distance. We adapt these ideas for our conditional Wasserstein distance.

Proof of Proposition 4. Let $C_b = C_b(A \times B)$ be the space of continuous bounded functions on $A \times B$ and $\mathcal{S}$ the set of nonnegative finite Borel measures α on $(A \times B)^2$ which are supported at most on the y-diagonal. By [47, Section 1.2], we know that

$$\sup_{f,g\in C_b(A\times B)} \mathbb{E}_{Y,X}[f] + \mathbb{E}_{Y,Z}[g] - \int_{(A\times B)^2} (f + g)\, \mathrm{d}\alpha = \begin{cases} 0 & \text{if } \alpha \in \Gamma(P_{Y,X}, P_{Y,Z}), \\ \infty & \text{otherwise.} \end{cases}$$

Using this relation, we obtain

$$\begin{aligned} W_{1,Y}(P_{Y,X}, P_{Y,Z}) &= \inf_{\alpha \in \Gamma_Y^4} \int \|(y_1, x_1) - (y_2, x_2)\| \, \mathrm{d}\alpha \\ &= \inf_{\alpha \in \mathcal{S}} \sup_{f,g \in C_b} L(\alpha, f, g) \end{aligned}$$

with the Lagrangian

$$\begin{aligned} L(\alpha, f, g) \coloneqq \mathbb{E}_{Y,X}[f] + \mathbb{E}_{Y,Z}[g] \\ + \int_{(A \times B)^2} \|(y_1, x_1) - (y_2, x_2)\| - f(y_1, x_1) - g(y_2, x_2) \, \mathrm{d}\alpha. \end{aligned} \tag{16}$$

By Corollary 15 below, strong duality holds true, so that we can exchange infimum and supremum to get

$$W_{1,Y}(P_{Y,X}, P_{Y,Z}) = \sup_{f,g \in C_b} \inf_{\alpha \in \mathcal{S}} L(\alpha, f, g).$$

From this, we see that the optimal f, g have to fulfill

$$f(y, x_1) + g(y, x_2) \leq \|x_1 - x_2\| \tag{17}$$

for all $y \in A$, since otherwise the attained infimum is $-\infty$. Therefore we have for the optimal f, g that $L(\alpha, f, g) \geq \mathbb{E}_{Y,X}[f] + \mathbb{E}_{Y,Z}[g]$ and choosing the plan $\alpha = 0 \in \mathcal{S}$, we obtain

$$\inf_{\alpha \in \mathcal{S}} L(\alpha, f, g) = \mathbb{E}_{P_{Y,X}}[f] + \mathbb{E}_{P_{Y,Z}}[g]$$

for all $(f, g) \in \tilde{\mathcal{F}}$, where

$$\tilde{\mathcal{F}} \coloneqq \{(f, g) \in (C_b(A \times B))^2 : f(y, x_1) + g(y, x_2) \leq \|x_2 - x_2\|\}.$$

Consequently, we get

$$W_{1,Y}(P_{Y,X}, P_{Y,Z}) = \sup_{(f,g) \in \tilde{\mathcal{F}}} \mathbb{E}_{Y,X}[f] + \mathbb{E}_{Y,Z}[g]. \tag{18}$$

For $(f, g) \in \tilde{\mathcal{F}}$, we define $\tilde{f}(y, x) \coloneqq \inf_{u \in B} \|x - u\| - g(y, u)$. Then

$$\begin{aligned} \tilde{f}(y, x) &= \inf_{u \in B} \{\|x - u\| - g(y, u)\} \\ &\leq \inf_{u \in B} \{\|x - z\| + \|z - u\| - g(y, u)\} \\ &= \tilde{f}(y, z) + \|x - z\| \end{aligned}$$

shows the 1-Lipschitz continuity of $\tilde{f}$ with respect to the second component. Using (17) we obtain that $\tilde{f}(y,x) \geq f(y,x)$. Since $\tilde{f}(y,x) \leq \|x-x\| - g(y,x)$, we conclude

$$f(y,x) \leq \tilde{f}(y,x) \leq -g(y,x). \tag{19}$$

Thus, $\tilde{f}$ is bounded. As pointwise infimum over continuous functions, $\tilde{f}$ is upper semicontinuous in (y,x). In summary, we have that $\tilde{f} \in \mathcal{F}$. By (18) and (19), we conclude

$$W_{1,Y}(P_{Y,X}, P_{Y,Z}) = \sup_{(f,g)\in\tilde{\mathcal{F}}} \{\mathbb{E}_{Y,X}[f] + \mathbb{E}_{Y,Z}[g]\} \leq \sup_{h\in\mathcal{F}} \{\mathbb{E}_{Y,X}[h] - \mathbb{E}_{Y,Z}[h]\}$$

and further for $\alpha \in \Gamma_Y^4(P_{Y,X}, P_{Y,Z}) \subset \Gamma(P_{Y,X}, P_{Y,Z})$ that

$$\begin{aligned}
\sup_{h\in\mathcal{F}} \{\mathbb{E}_{Y,X}[h] - \mathbb{E}_{Y,Z}[h]\} &\leq \sup_{h\in\mathcal{F}} \inf_{\alpha\in\Gamma_Y^4} \int_{(A\times B)^2} h(y_1,x_1) - h(y_2,x_2) \mathrm{d}\alpha \\
&= \sup_{h\in\mathcal{F}} \inf_{\alpha\in\Gamma_Y^4} \int_{(A\times B)^2} h(y_1,x_1) - h(y_1,x_2) \mathrm{d}\alpha \\
&\leq \inf_{\alpha\in\Gamma_Y^4} \int_{(A\times B)^2} \|x_1 - x_2\| \mathrm{d}\alpha \\
&= \inf_{\alpha\in\Gamma_Y^4} \int_{(A\times B)^2} \|(y_1,x_1) - (y_2,x_2)\| \mathrm{d}\alpha \\
&= W_{1,Y}(P_{Y,X}, P_{Y,Z}).
\end{aligned}$$

Thus, $W_{1,Y}(P_{Y,X}, P_{Y,Z}) = \sup_{h\in\mathcal{F}}\{\mathbb{E}_{Y,X}[h] - \mathbb{E}_{Y,Z}[h]\}$, which finishes the proof. □

The proof of strong duality relies on the following minimax principle from [9, Theorem 7 Chapter 6].

Theorem 14. *Let X be a convex subset of a topological vector space, and Y be a convex subset of a vector space. Assume $f : X \times Y \to \mathbb{R}$ satisfies the following conditions:*

i) For every $y \in Y$, the map $x \mapsto f(x,y)$ is lower semi continuous and convex.

ii) There exists y_0 such that $x \mapsto f(x,y_0)$ is inf-compact, i.e the set $\{x \in X : f(x,y_0) \leq a\}$ is relatively compact for each $a \in \mathbb{R}$.

iii) For every $x \in X$, the map $y \to f(x,y)$ is convex.

Then it holds

$$\inf_{x\in X} \sup_{y\in Y} f(x,y) = \sup_{y\in Y} \inf_{x\in X} f(x,y).$$

Based on the theorem we can prove the desired strong duality relation.

Corollary 15. *For the Lagrangian in* (16) *it holds*

$$\inf_{\alpha\in\mathcal{S}} \sup_{f,g\in C_b} L(\alpha,f,g) = \sup_{f,g\in C_b} \inf_{\alpha\in\mathcal{S}} L(\alpha,f,g).$$

Proof. We will verify the conditions in Theorem 14. Recall that $\mathcal{S}$ is the set of finite nonnegative Borel measures α on $(A\times B)^2$ such that there exists a finite nonegative finite measure β on B with $\pi^{1,3}_\sharp\alpha = \Delta_\sharp\beta$. Let $\mathcal{M}$ be the topological vector space of finite signed Borel measures on $(A\times B)^2$ with weak convergence topology. Thus, since the pushforward is linear on $\mathcal{S}$, we conclude that $\mathcal{S}$ is a convex subset. Now we use Theorem 14 with $X \coloneqq \mathcal{S}$, $Y \coloneqq C_b \times C_b$ and $f \coloneqq L$.

Verifying i) The map $\alpha \mapsto L(\alpha,f,g)$ is linear and continuous on $\mathcal{S}$ under the weak convergence of measures. This follows from the fact that the integrand of α in $L(\alpha,f,g)$ is in $C_b((A\times B)^2)$.

Verifying iii) Note that for any $\alpha\in\mathcal{S}$ the map $(f,g)\mapsto L(\alpha,f,g)$ is linear in (f,g) and therefore convex.

Verifying ii) Setting $f(y,x)\coloneqq -1, g(y,x)\coloneqq -1$ for all (y,x), we will show that for any fixed $a\in R$, the set

$$\mathcal{S}_a := \{\alpha\in\mathcal{S} : L(\alpha,-1,-1)\le a\}$$

is relatively compact. Since the integrand is bounded from below by 2 and $\mathcal{S}$ only contains nonnegative measures, the measures in $\mathcal{S}_a$ are uniformly bounded in the total variation norm, since otherwise

$$L(\alpha,-1,-1) = 2 + \int_{(A\times B)^2} \|(y_1,x_1)-(y_2,x_2)\| + 2\,\mathrm{d}\alpha$$

can become arbitrary large which contradicts $L(\alpha,-1,-1)\le a$. Therefore the compactness of A,B implies that $\mathcal{S}_a$ is a family of tight measures. By [11, Theorem 8.6.7], the set $\mathcal{S}_a$ is relatively compact in the weak topology. □

D Proofs of Section 5

Proof of Proposition 5. i) We have $\mu_t\in\mathcal{P}_{p,Y}(\mathbb{R}^d\times\mathbb{R}^m)$ for every $t\in[0,1]$ by

$$(\pi^1)_\sharp\mu_t = \pi^1_\sharp(e_t)_\sharp\alpha = ((1-t)\pi^1+t\pi^3)_\sharp(\pi^{1,3})_\sharp\alpha = ((1-t)\pi^1+t\pi^2)_\sharp\Delta_\sharp P_Y = P_Y.$$

For $s,t\in[0.1]$, let $\alpha_{s,t}\coloneqq(e_s,e_t)_\sharp\alpha$. By definition we see that $\alpha_{s,t}\in\Gamma(\mu_s,\mu_t)$. Further $\pi^{1,3}_\sharp\alpha_{s,t} = \Delta_\sharp P_Y$ follows from

$$\pi^{1,3}\circ(e_s,e_t) = \big((1-s)\pi^1+s\pi^2,(1-t)\pi^1+t\pi^2\big)\circ\pi^{1,3}$$

and consequently

$$\begin{aligned}\pi^{1,3}_\sharp \alpha_{s,t} &= \big((1-s)\pi^1 + s\pi^2, (1-t)\pi^1 + t\pi^2\big)_\sharp \pi^{1,3}_\sharp \alpha \\ &= \big(\big((1-s)\pi^1 + s\pi^2, (1-t)\pi^1 + t\pi^2\big) \circ \Delta\big)_\sharp P_Y = \Delta_\sharp P_Y.\end{aligned}$$

In summary, we see that $\alpha_{s,t} \in \Gamma_{2,Y}(\mu_s, \mu_t)$. Thus, we have

$$\begin{aligned}W^2_{2,Y}(\mu_s, \mu_t) &\le \int_{(\mathbb{R}^d \times \mathbb{R}^m)^2} \|(y_1, x_1) - (y_2, x_2)\|^2 \,\mathrm{d}\alpha_{s,t} \\ &= \int_{(\mathbb{R}^d \times \mathbb{R}^m)^2} \big\|(t-s)\left((x_1, y_1) - (x_2, y_2)\right)\big\|^2 \,\mathrm{d}\alpha \\ &= |t-s|^2 \, W^2_{2,Y}(\mu_0, \mu_1). \end{aligned} \tag{20}$$

Finally, the desired equality follows like in [6, Theorem 7.2.2]for $0 \le s \le t \le 1$ by

$$W_{2,Y}(\mu_0, \mu_1) \le W_{2,Y}(\mu_0, \mu_s) + W_{2,Y}(\mu_s, \mu_t) + W_{2,Y}(\mu_t, \mu_1) \le W_{2,Y}(\mu_0, \mu_1),$$

which implies equality in (20).
ii) First, we show $(e_t)_\sharp \alpha = (\pi^1, (1-t)\pi^2 + t\pi^4)_\sharp \alpha$. For any Borel measurable function $f : \mathbb{R}^d \times \mathbb{R}^m \to [0, \infty]$, we have indeed

$$\begin{aligned}\int_{\mathbb{R}^d \times \mathbb{R}^m} f \,\mathrm{d}(e_t)_\sharp \alpha &= \int_{(\mathbb{R}^d \times \mathbb{R}^m)^2} f((1-t)y_1 + ty_2, (1-t)x_1 + tx_2) \,\mathrm{d}\alpha \\ &= \int_{\mathbb{R}^{2d}} \int_{\mathbb{R}^{2m}} f((1-t)y_1 + ty_2, (1-t)x_1 + tx_2) \,\mathrm{d}\alpha_{y_1,y_2} \mathrm{d}\pi^{1,3}_\sharp \alpha \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^{2m}} f((1-t)y + ty, (1-t)x_1 + tx_2) \,\mathrm{d}\alpha_{y,y} \mathrm{d}P_Y \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^m} f \,\mathrm{d}(\pi^1, (1-t)\pi^2 + t\pi^4)_\sharp \alpha.\end{aligned}$$

Using the above relation, we obtain

$$\begin{aligned}&\int_{\mathbb{R}^d \times \mathbb{R}^m} f \,\mathrm{d}((\mu_t)_y \otimes P_Y) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^m} f(y, x) \,\mathrm{d}((1-t)\pi^1 + t\pi^2)_\sharp \alpha_{y,y}(x) \mathrm{d}P_Y(y) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^{m \times m}} f(y, (1-t)x_1 + tx_2) \,\mathrm{d}\alpha_{y,y}(x_1, x_2) \mathrm{d}P_Y(y) \\ &= \int_{\mathbb{R}^{2d}} \int_{\mathbb{R}^{2m}} f(y_1, (1-t)x_1 + tx_2) \,\mathrm{d}\alpha_{y_1,y_2}(x_1, x_2) \mathrm{d}\Delta_\sharp P_Y(y_1) \\ &= \int_{(\mathbb{R}^d \times \mathbb{R}^m)^2} f(y_1, (1-t)x_1 + tx_2) \mathrm{d}\alpha(y_1, x_1, y_2, x_2) \\ &= \int_{(\mathbb{R}^d \times \mathbb{R}^m)^2} f \,\mathrm{d}(e_t)_\sharp \alpha = \int_{\mathbb{R}^d \times \mathbb{R}^m} f \,\mathrm{d}\mu_t,\end{aligned}$$

which proves that $(\mu_t)_y$ is indeed the disintegration of μ_t with respect to P_Y. By Proposition 1 ii) we know that $\alpha_{y,y} \in \mathcal{P}(\mathbb{R}^{2m})$ is optimal in (3) for P_y-a.e. $y \in \mathbb{R}^d$. By (9) this implies that $(\mu_t)_y$ is a geodesic in $\mathcal{P}_2(\mathbb{R}^m)$.

iii) Recall (see [6, Section 5.1]) that a sequence $\mu_k \in \mathcal{P}(\mathbb{R}^n)$ is said to converge weakly to $\mu \in \mathcal{P}(\mathbb{R}^n)$ if $\lim_{k\to\infty} \int_{\mathbb{R}^n} f(x)\,\mathrm{d}\mu_k(x) = \int_{\mathbb{R}^n} f(x)\,\mathrm{d}\mu(x)$ for all $f \in C_b(\mathbb{R}^n)$. By the dominated convergence theorem, we have for $\mu_s = (e_s)_\sharp \alpha$ and every $f \in C_b(\mathbb{R}^d \times \mathbb{R}^m)$ that

$$\begin{aligned}\lim_{s\to t} \int_{\mathbb{R}^d\times\mathbb{R}^m} f\,\mathrm{d}\mu_s &= \lim_{s\to t} \int_{(\mathbb{R}^d\times\mathbb{R}^m)^2} f((1-s)(y_1,x_1) - s(y_2,x_2))\,\mathrm{d}\alpha \\ &= \int_{(\mathbb{R}^d\times\mathbb{R}^m)^2} f((1-t)(y_1,x_1) - t(y_2,x_2))\mathrm{d}\alpha = \int_{\mathbb{R}^d\times\mathbb{R}^m} f\,\mathrm{d}\mu_t,\end{aligned}$$

which finishes the proof. □

Proof of Proposition 6. The statements i),ii) can given in [5, Lemma 17.3] (with $e \coloneqq e_t$, $\mu = \alpha$, $v = (y_2, x_2) - (y_1, x_1)$, $w = v_t$).

The continuity equation in iv) follow as in the proof of [5, Theorem 17.2].

Towards iii), note that it holds for any Borel measurable set $U \subseteq (\mathbb{R}^d \times \mathbb{R}^m)^2$ and $j \le d$ that

$$\begin{aligned}\left|\int_U (y_2)_j - (y_1)_j\,\mathrm{d}\alpha\right| &\le \int_U |(y_2)_j - (y_1)_j|\,\mathrm{d}\alpha \le \int_{\pi^{1,3}(U)} |(y_2)_j - (y_1)_j|\,\mathrm{d}\pi^{1,3}_\sharp\alpha \\ &= \int_{\pi^{1,3}(U)} |(y_2)_j - (y_1)_j|\,\mathrm{d}\Delta_\sharp P_Y \\ &= \int_{\Delta^{-1}(\pi^{1,3}(U))} |y_j - y_j|\,\mathrm{d}P_Y = 0.\end{aligned}$$

Thus, for any Borel measurable set $V \subseteq \mathbb{R}^d \times \mathbb{R}^m$ and $j \le d$, we obtain by Part i) that

$$\int_V (v_t)_j\,\mathrm{d}\mu_t = \int_V \mathrm{d}(e_t)_\sharp((y_2)_j - (y_1)_j)\alpha) = \int_{e_t^{-1}(V)} ((y_2)_j - (y_1)_j)\,\mathrm{d}\alpha = 0.$$

This implies $(v_t(y,x))_j = 0$ for μ_t-a.e. $(y,x) \in \mathbb{R}^d \times \mathbb{R}^m$. □

For the proof of Proposition 8 we need the following proposition. Since we have not found a proof in the literature, we give it for convenience.

Proposition 16. *Let $\mu_0, \mu_1 \in (\mathcal{P}_2(\mathbb{R}^m), W_2)$ which fulfill one of the following conditions:*

i) μ_0, μ_1 are empirical measures with the same number of points and T is an optimal map with associated optimal plan $\alpha \in \Gamma_2(\mu_0, \mu_1)$, or

ii) μ_0, μ_1 *both admit densities* ρ_0, ρ_1 *which are supported on open, convex, bounded, connected subsets* $\Omega_0, \Omega_1 \subset \mathbb{R}^m$ *on which they are bounded away from* 0 *and* ∞*. Assume further that* $\rho_0 \in C^2(\Omega_0), \rho_1 \in C^2(\Omega_1)$*. Let* T *be the optimal Monge map with associated optimal plan* $\alpha \in \Gamma_2(\mu_0, \mu_1)$*.*

Let $\mu_t = (e_t)_\sharp \alpha$ *and* $v_t \in L^2_{\mu_t}(\mathbb{R}^m, \mathbb{R}^m)$ *with* $v_t \mu_t = (e_t)_\sharp (x_2 - x_1)\alpha$ *which then satisfy the continuity equation. Then there exists a solution of the flow equation*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\phi_t &= v_t(\phi_t), \\ \phi_0(x) &= x, \end{aligned}$$

such that $\mu_t = \phi_{t,\sharp}\mu_0$*. Furthermore, we have*

$$v_t(\phi_t(x)) = T(x) - x$$

for μ_0*-a.e.* $x \in \mathbb{R}^m$*.*

Proof. i): Let $\mu_0 = \frac{1}{n}\sum_{i=1}^n \delta_{a_i}, \mu_1 = \frac{1}{n}\sum_{i=1}^n \delta_{b_i}$. The optimal plan is then $\alpha = \frac{1}{n}\sum_{i=1}^n \delta_{a_i, T(a_i)}$. Using $e_{t,\sharp}(x_2 - x_1)\alpha = v_t\mu_t$ and $\mu_t = \frac{1}{n}\sum_{i=1}^n \delta_{T_t(a_i)}$ for $T_t(x) = (1-t)x + tT(x)$ we can conclude

$$v_t((1-t)a_i + tT(a_i)) = T(a_i) - a_i.$$

Furthermore, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}T_t(a_i) = T(a_i) - a_i = v_t(T_t(a_i)),$$

and thus $\phi_t := T_t$ fulfills the flow equation and $v_t(\phi_t(x)) = T(x) - x$ for μ_0-a.e. $x \in \mathbb{R}^m$.

ii): First, note that by [5, (16.12)] if there exists an invertible Monge map T then the geodesic between μ_0, μ_1 fulfills the continuity equation with vector field

$$v_t = (T - \mathrm{Id}) \circ T_t^{-1}$$

where $T_t = (1-t)\mathrm{Id} + tT$. By Caffarelli's regularity Theorem [51, Theorem 12.5, ii)], we get the existence of a unique Monge map $T \in C^1(\Omega_1)$ mapping μ_0 to μ_1 and $U \in C^1(\Omega_2)$ mapping μ_1 to μ_0. By [5, Theorem 5.2] we know that $T \circ U = \mathrm{Id}$ on Ω_2 and $U \circ T = \mathrm{Id}$ on Ω_2 and thus $T : \Omega_1 \to \Omega_2$ is a C^1 diffeomorphism and in particular $\det(\nabla T) \neq 0$ on Ω_1. Since we know by [6, Proposition 6.2.12] that ∇T is positive definite μ_1 a.e. on Ω_1 we can deduce from $\det(\nabla T) \neq 0$ that ∇T is positive definite on Ω_1. Consequently for $T_t = (1-t)\mathrm{Id} + tT$ we have that $\nabla T_t = (1-t)\mathrm{Id} + t\nabla T$ is positive definite on Ω_1 and thus the image of Ω_1 under T_t is open. Furthermore, we know by the proof of [6, Proposition 6.2.12] that T_t as a Monge map from μ_0 to μ_t is injective on all points where ∇T_t is positive definite, which is on the whole Ω_1, and thus T_t is a diffeomorphism onto its image. Consequently, it possesses

a C^1 inverse $T_t^{-1} : T_t(\Omega_1) \to \Omega_1$. Then $v_t := (T - \mathrm{Id}) \circ T_t^{-1} : T_t(\Omega_1) \to \mathbb{R}^m$ we have that v_t is measurable since it is continuous on $T_t(\Omega_1)$ and the same is true for $\phi_t = T_t : \Omega_1 \to \mathbb{R}^m$. Furthermore, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_t(x) = T(x) - x = (T - \mathrm{Id}) \circ T_t^{-1}(T_t(x)) = v_t(\phi_t(x)).$$

Since we can set $\phi_t(x) = x$ on $\mathbb{R}^m \setminus \Omega_1$ and $v_t(x) = 0$ for $x \in \mathbb{R}^m \setminus T_t(\Omega_1)$, we obtain the claim. □

Proof of Proposition 8. We will use the results from Proposition 16 and stack them with respect to y_i. The main obstruction is the measurability of the resulting objects which we address in the following.
For $e_t((y_1, x_1), (y_2, x_2)) = (1-t)(y_1, x_1) + t(y_2, x_2)$ and $\tilde{e}_t(x_1, x_2) = (1-t)x_1 + tx_2$, it holds

$$\begin{aligned}
\int_{\mathbb{R}^d \times \mathbb{R}^m} f(y, x)\, \mathrm{d}(e_t)_\sharp((y_2, x_2) - (y_1, x_1)\alpha) &= \int_{(\mathbb{R}^d \times \mathbb{R}^m)^2} f \circ e_t((y_2, x_2) - (y_1, x_1))\, \mathrm{d}\alpha \\
&= \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^{2m}} f \circ e_t((y_i, x_1), (y_i, x_2))\, \mathrm{d}\alpha_{y_i} \\
&= \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^{2m}} f((y_i, \tilde{e}_t(x_1, x_2)))\, (0, x_2 - x_1)\mathrm{d}\alpha_{y_i} \\
&= \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^{2m}} f((y_i, \tilde{e}_t(x_1, x_2)))\, (0, x_2 - x_1)\mathrm{d}\alpha_{y_i}
\end{aligned}$$

and thus $(e_t)_\sharp((y_2, x_2) - (y_1, x_1)\alpha) = \frac{1}{n}\sum_{i=1}^n \delta_{y_i} \otimes (0, (\tilde{e}_t)_\sharp((x_2 - x_1)\alpha_{y_i}))$. Combining with Proposition 6, we conclude

$$v_t\mu_t = \frac{1}{n}\sum_{i=1}^n \delta_{y_i} \otimes (0, (\tilde{e}_t)_\sharp((x_2 - x_1)\alpha_{y_i}))\,.$$

Furthermore, we have

$$v_t\mu_t = \int_{\mathbb{R}^d} v_t \mathrm{d}\mu_{t,y}\mathrm{d}P_Y = \frac{1}{n}\sum_{i=1}^n \delta_{y_i} \otimes v_t(y_i, \cdot)\mu_{t,y_i},$$

which implies $(\tilde{e}_t)_\sharp\left((x_2 - x_1)\alpha_{y_i}\right) = \pi^2 \circ (v_t(y_i, \cdot))\mu_{t,y_i}$ for all $i \in \{1, \ldots, n\}$. By Proposition 16 we know that there exists $\tilde{v}_{t,y_i} \in L^2(\mu_{t,y_i})$ with $(\tilde{e}_t)_\sharp\left(x_2 - x_1)\right)\alpha_{y_i}) = \tilde{v}_{t,y_i}(\cdot)\mu_{t,y_i}$ such

that there exists a μ_{0,y_i}-measurable solution ϕ_{t,y_i} of

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\phi_{t,y_i} &= \tilde{v}_{t,y_i}\left(\phi_{t,y_i}\right)\\ \phi_{0,y_i}(x) &= x\end{aligned}$$

for μ_{0,y_i} a.e. $x \in \mathbb{R}^m$ and $\mu_{t,y_i} = (\phi_{t,y_i})_\sharp \mu_{0,y_i}$. Since P_Y is a finite empirical measure also $\phi_t : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d \times \mathbb{R}^m$ defined on (y_i, x) as $(y_i, \phi_{t,y_i}(x))$ is μ_t measurable and $\tilde{v}_t : (y_i, x) \mapsto (0, \tilde{v}_{t,y_i}(x))$ is in $L^2_{\mu_t}$ and coincides with v_t as element of $L^2_{\mu_t}$. The latter is true since they coincide on $\{y_i\} \times \mathbb{R}^m$ up to a μ_{t,y_i} null set $\mathcal{N}_i$ because of

$$\pi^2 \circ (v_t(y_i, \cdot))\mu_{t,y_i} = (\tilde{e}_t)_\sharp \left((x_2 - x_1)\alpha_{y_i}\right) = \tilde{v}_{t,y_i}\mu_{t,y_i}.$$

Thus they coincide up to the set

$$\cup_{i=1}^n \{y_i\} \times \mathcal{N}_i \cup \{(y, x) \in \mathbb{R}^{d+m} : y \notin \{y_1, \ldots, y_n\}\}$$

which is a μ_t null set. Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_t = \tilde{v}_t(\phi_t)$$

for μ_0-a.e. $(y, x) \in \mathbb{R}^d \times \mathbb{R}^m$. Furthermore

$$\begin{aligned}(\phi_t)_\sharp \mu_0(a \times b) &= \int_{(y,\phi_{t,y}(x)) \in a \times b} \mathrm{d}\mu_0 = \int_{y \in a} \int_{\phi_{t,y}(x) \in b} \mathrm{d}\mu_{0,y}(x) \mathrm{d}P_Y(y)\\ &= \int_a \int_b \mathrm{d}\phi_{t,y,\sharp}\mu_{0,y} \mathrm{d}P_Y(y) = \int_a \int_b \mathrm{d}\mu_{t,y} \mathrm{d}P_Y(y)\\ &= \mu_t(a \times b)\end{aligned}$$

shows $\mu_t = (\phi_t)_\sharp \mu_0$. The last claim follows from

$$\tilde{v}_t((y_i, \phi_{t,y_i}(x)) = (0, T_{y_i}(x) - x)$$

for μ_{0,y_i}-a.e. $x \in \mathbb{R}^d$. □

E Proofs of Section 6

Proof of Proposition 10. Denote by α_{opt} the optimal transport plan associated to the conditional Wasserstein metric $W_{p,Y}$. Since it is only Y diagonally supported, we have that

$$\|(y_1, x_1) - (y_2, x_2)\|^p = d^p_\beta((y_1, x_1), (y_2, x_2))$$

for α_{opt} a.e. $(y_1, x_1, y_2, x_2) \in (A \times B)^2$. Thus, for an optimal plan α for $W_{p,\beta}$, we conclude

$$\begin{aligned} W_{p,Y}(\mu_0, \mu_1)^p &= \int_{(A\times B)^2} \|(y_1, x_1) - (y_2, x_2)\|^p \,\mathrm{d}\alpha = \int_{(A\times B)^2} d_\beta^p((y_1, x_1), (y_2, x_2)) \,\mathrm{d}\alpha \\ &\geq \int_{(A\times B)^2} d_\beta^p((y_1, x_1), (y_2, x_2)) \,\mathrm{d}\alpha \\ &\geq \int_{B^2} \|x_1 - x_2\|^p \,\mathrm{d}\pi^{2,4}_\sharp\alpha + \beta \int_{A^2} \|y_1 - y_2\|^p \,\mathrm{d}\pi^{1,3}_\sharp\alpha \\ &\geq \beta \int_{A^2} \|y_1 - y_2\|^p \, d\pi^{1,3}_\sharp\alpha \end{aligned}$$

and thus the claim. □

In order to proof Proposition 12 we need the following lemma which is a variant of [6, Proposition 7.1.3].

Lemma 17. *Let $\beta > 0$, let $A \subset \mathbb{R}^d, B \subset \mathbb{R}^m$ be compact sets and let $\mu_n \to \mu$, $\nu_n \to \nu$ in $(\mathcal{P}_2(A \times B), W_2)$. Then there exists a subsequence of optimal plans α_{n_k} for $W_{2,d_\beta}(\mu_{n_k}, \nu_{n_k})$ and an optimal plan $\alpha \in P_2\left((A \times B)^2\right)$ for $W_{2,d_\beta}(\mu, \nu)$ such that $\alpha_{n_k} \to \alpha$ with respect to $\left(\left(\mathcal{P}_2\left((A \times B)\right)^2\right), W_2\right)$.*

Proof. Let $f : A \times B \to \sqrt{\beta}A \times B$ be defined by $(y, x) \mapsto (\sqrt{\beta}y, x)$. Then for $\mu_1, \mu_2 \in P_2(A \times B)$ we have that $W_{2,d_\beta}(\mu_1, \mu_2) = W_2(f_\sharp\mu_1, f_\sharp\mu_2)$ since there is a bijection of couplings

$$\alpha \mapsto (f, f)_\sharp\alpha \tag{21}$$

and we can compute

$$\int_{\left(A\times\sqrt{(\beta)}B\right)^2} \|(y_2, x_2) - (y_1, x_2)\|^2 \mathrm{d}(f, f)_\sharp\alpha = \int_{(A\times B)^2} \beta\|y_2 - y_1\|^2 + \|x_2 - x_1\| \mathrm{d}\alpha$$

which implies that optimal couplings are mapped to optimal couplings. Furthermore

$$\begin{aligned} W_2^2(f_\sharp\mu_1, f_\sharp\mu_2) = W_{2,d_\beta}^2(\mu_1, \mu_2) \leq \beta W_2^2(\mu_1, \mu_2) \\ W_2^2(\mu_1, \mu_2) \leq W_{2,d_\beta}^2(\mu_1, \mu_2) = W_2^2(f_\sharp\mu_1, f_\sharp\mu_2) \end{aligned} \tag{22}$$

and thus also $f_\sharp\mu_n \to f_\sharp\mu, f_\sharp\nu_n \to f_\sharp\nu$ in W_2. Then we can use [6, Proposition 7.1.3] to guarantee the existence of a subsequence of optimal plans $\tilde{\alpha}_{n_k}$ for $W_2(f_\sharp\mu_{n_k}, f_\sharp\nu_{n_k})$ such that $\tilde{\alpha}_{n_k} \to \tilde{\alpha}$ for an optimal plan $\tilde{\alpha}$ for $W_2(f_\sharp\mu, f_\sharp\nu)$. Thus by (21) there exists a subsequence of optimal plans α_{n_k} for $W_{2,d_\beta}(\mu_n, \nu_n)$ such that $\alpha_{n_k} \to \alpha$ for an optimal plan α for $W_{2,d_\beta}(\mu, \nu)$. Note that the W_2 convergence of $\alpha_{n_k} \to \alpha$ follows from a computation similar to (22) for $(f, f)_\sharp$ instead of $f_\sharp$. □

Proof of Proposition 12. Since we are in the compact setting the concept of weak and W_2 convergence coincide which we will use without mentioning in the following [51]. Now by [6, Proposition 7.1.3] and Lemma 17, there exists a subsequence of α_n converging in W_2 to an optimal plan α^{β_k} for $W_{2,d_{\beta_k}}(\mu,\nu)$. Choose n_k monotonely increasing such that $W_2(\alpha^{\beta_k},\alpha_{n_k}) < \frac{1}{k}$. We know by [29, Proposition 3.11] that $\alpha_{\beta_k} \to \alpha$ in W_2 for an optimal plan $\alpha \in \Gamma_Y^4(\mu,\nu)$ for $W_{2,Y}(\mu,\nu)$. Thus, for $\epsilon > 0$, there exists a k such that $\frac{1}{k} + W_2(\alpha^{\beta_k},\alpha) < \epsilon$ and we obtain

$$W_2(\alpha_{n_k},\alpha) \leq W_2(\alpha_{n_k},\alpha^{\beta_k}) + W_2(\alpha^{\beta_k},\alpha) \leq \frac{1}{k} + W_2(\alpha^{\beta_k},\alpha) < \epsilon$$

which proves the claim. □

F Benamou-Brenier like formula for $W_{p,Y}$

Theorem 18. *Let $\mu_1,\mu_2 \in \mathcal{P}_{2,Y}(\mathbb{R}^d \times \mathbb{R}^m)$ and denote by $\mathcal{A}(\mu_1,\mu_2)$ the set of tuples (μ_t, v_t), where*

i) $\mu_t \in \mathcal{P}_{p,Y}$ is a continuous curve,

ii) There exists a family of disintegrations $\mu_t = (\mu_t)_y \otimes P_Y$ such that $(\mu_t)_y$ is continuous P_Y-a.e.,

iii) v_t is a Borel vector field fulfilling $\int_0^1 \|v_t\|_{L^2(\mu_t)}\mathrm{d}t < \infty$ and $(v_t)_j = 0$ for all $j \leq d$ for μ_t a.e. $(y,x) \in \mathbb{R}^{d+m}$,

iv) μ_t fulfills the continuity equation in the sense of distributions for v_t.

Then we have that

$$W_{p,Y}(\mu_1,\mu_2) = \min_{(\mu_t,v_t)\in\mathcal{A}(\mu_1,\mu_2)} \left\{ \int_0^1 \|v_t\|^2_{L^2(\mu_t)}\,\mathrm{d}t \right\}$$

In order to prove Theorem 18, we need the following auxiliary lemma.

Lemma 19. *Let $\mu_t \in P_{p,Y}(\mathbb{R}^d \times \mathbb{R}^m)$ be a solution of the continuity equation with v_t such that $(v_t)_j = 0$ for $j \leq d$ and $\int_0^1 \int_{\mathbb{R}^{d+m}} \|v_t\|\mathrm{d}\mu_t\mathrm{d}t < \infty$. Then the disintegration $(\mu_t)_y$ satisfies the continuity equation for $v_{t,y}$ for P_Y-a.e. $y \in \mathbb{R}^d$.*

Proof. Let $\phi \in C_c^\infty((0,1)\times\mathbb{R}^m)$ be a test function and let $g \in C_c^\infty(\mathbb{R}^d)$. Then for $\psi(t,y,x) = g(y)\phi(t,x)$ we have that $\psi \in C_c^\infty((0,1)\times\mathbb{R}^{d+m})$ is a valid test function which we can insert into the continuity equation, use the theorem of Fubini to change the order of the integrals and obtain

$$\int_{\mathbb{R}^d} g(y) \int_0^1 \int_{\mathbb{R}^m} \partial_t\phi(x,t) + \langle v_{t,y}, \nabla_x\phi\rangle \mathrm{d}(\mu_t)_y(x)\mathrm{d}P_Y(y) = 0$$

Since g was arbitrary we obtain that $\int_0^1 \int_{\mathbb{R}^m} \partial_t \phi(x,t) + \langle v_{t,y}, \nabla_x \phi \rangle \mathrm{d}(\mu_t)_y(x) = 0$ for P_Y a.e. $y \in \mathbb{R}^d$. Since $C_c^\infty((0,1) \times \mathbb{R}^d)$ contains a dense countable subset in the $\|\cdot\|_\infty$ topology we only need to test on countably ϕ we can conclude that $(\mu_t)_y$ is a solution of the continuity equation for $v_{t,y}$ for P_Y a.e. $y \in \mathbb{R}^d$. □

Proof of Theorem 18. First we show " $\leq$ ". By Proposition 19, we have that $(\mu_t)_y$ fulfills the continuity equation for v_{t,x_1} P_Y-a.e. and by assumption $(\mu_t)_y$ is weakly continuous. Thus we have that $W_2((\mu_1)_y, (\mu_2)_y) \leq \int_0^1 \int_{\mathbb{R}^m} \|v_{t,y}\|^2 (\mu_t)_y \mathrm{d}t$. Now we get

$$\begin{aligned} W_{2,Y}(\mu_1, \mu_2) &= \int_{\mathbb{R}^d} W_p((\mu_1)_y, (\mu_2)_y) \, \mathrm{d}P_Y \leq \int_{\mathbb{R}^n} \int_0^1 \int_{\mathbb{R}^m} \|v_{t,y}\|^2 \mathrm{d}(\mu_t)_y \, \mathrm{d}t \mathrm{d}P_Y \\ &= \int_0^1 \int_{\mathbb{R}^d} \int_{\mathbb{R}^m} \|v_{t,y}\|^2 \mathrm{d}(\mu_t)_y \, \mathrm{d}P_Y \mathrm{d}t = \int_0^1 \int_{\mathbb{R}^d} \int_{\mathbb{R}^m} \|v_t\|^2 \mathrm{d}(\mu_t)_y \, \mathrm{d}P_Y \mathrm{d}t \\ &= \int_0^1 \|v_t\|^2_{L^2(\mu_t)} \, \mathrm{d}t, \end{aligned}$$

where we used that $\int \|v_{t,y}\|^2 \mathrm{d}\mu_t = \int \|v_t\|^2 \mathrm{d}\mu_t$, since $(v_t)_j = 0$ for all $j \leq n$ for μ_t a.e. $(y,x) \in \mathbb{R}^{d+m}$.
The direction " $\geq$ " follows from Proposition 6 and Lemma 5ii). □

G Implementation Details

We use a setup similar to [50], using the time dependent U-Net architecture from [42] which are trained using Adam [32]. As in [50] we clip the gradient norm to 1 and use exponential moving averaging with a decay of 0.9999. The differences are we use a constant learning rate of 2e-4, 256 model channels and no dropout. We train using 50k target samples for 300 epochs using a batch size of 500 for the minibatch OT couplings and a batch size of 100 for training the networks. We set the same random seed during training to be able to compare runs for different sources of couplings. The conditional coupling plans are calculated using the Python Optimal Transport package [20]. For inference simulate the corresponding ODEs using the torchiffeq [15] package. To evaluate our results, we use the Fréchet inception distance (FID) [28][1]. We compute the distance on 50k training samples, for which we generate 50k samples given the same labels as the training samples.

Further generated samples for the best performing method i.e $\beta = 100$:

[1]We use the implementation from `https://github.com/mseitzer/pytorch-fid`.

Figure 5: Uncurated samples sorted by class labels of the OT Bayesian Flow matching method with $\beta = 100$.