Estimating Barycenters of Distributions with Neural Optimal Transport

Alexander Kolesov [*1] **Petr Mokrov** [*1] **Igor Udovichenko** [1] **Milena Gazdieva** [1] **Gudmund Pammer** [2]
Evgeny Burnaev [1 3] **Alexander Korotin** [1 3]

Abstract

Given a collection of probability measures, a practitioner sometimes needs to find an "average" distribution which adequately aggregates reference distributions. A theoretically appealing notion of such an average is the Wasserstein barycenter, which is the primal focus of our work. By building upon the dual formulation of Optimal Transport (OT), we propose a new scalable approach for solving the Wasserstein barycenter problem. Our methodology is based on the recent Neural OT solver: it has bi-level adversarial learning objective and works for general cost functions. These are key advantages of our method since the typical adversarial algorithms leveraging barycenter tasks utilize tri-level optimization and focus mostly on quadratic cost. We also establish theoretical error bounds for our proposed approach and showcase its applicability and effectiveness in illustrative scenarios and image data setups. Our source code is available at `https://github.com/justkolesov/NOTBarycenters`.

1. Introduction

Generative Modeling encompasses a set of tools designed for manipulating probability distributions. Among them, a special place is occupied by the Wasserstein barycenter problem. This problem consists in finding a distribution which minimizes the average of specific divergences to some pre-defined distributions. In the context of Wasserstein barycenters, the divergences are chosen to be Optimal Transport (OT) costs due to attractive geometrical properties and clear intuition behind them. Starting with the seminal work (Agueh & Carlier, 2011), the Wasserstein barycenter problem has been consistently gaining significant attention in the research community. This is partially due to the compelling mathematical theory and an extensive list of applications: Bayesian inference (Srivastava et al., 2018), Geometric modelling (Solomon et al., 2015), Style Transfer (Mroueh, 2020), Texture mixing (Lacombe et al., 2023), Reinforcement Learning (Likmeta et al., 2023), Federated Learning (Singh & Jaggi, 2020), etc. Reflecting the practitioners' interest, the computational OT literature is also inspired by the Wasserstein barycenter problem and proposes several methods for solving this task. Early works, e.g., (Cuturi & Doucet, 2014; Anderes et al., 2016), deal with a discrete learning setup, i.e., they assume that the distributions of interest are discrete. Unfortunately, the resulting solvers mostly lack certain beneficial properties a practitioner may require, see §2.3. Our work is in line with the alternative continuous learning setup. To tackle this challenge, existing barycenter solvers for the continuous learning setup typically

a) resort to complex algorithmic procedures, e.g.: tri-level adversarial learning objectives (Fan et al., 2021; Korotin et al., 2022), Langevin sampling (Kolesov et al., 2023).

b) consider only specific formulations of the Wasserstein barycenter problem to make it more tractable, e.g.: deal with quadratic cost (Noble et al., 2023), utilize entropic/quadratic regularization (Li et al., 2020).

Contribution. We take a step forward and propose a pioneering approach for solving the Wasserstein barycenter problem which **a)** permits conventional bi-level (max-min) adversarial objective and **b)** can be adapted to *various* formulations of the barycenter problem, e.g., with general cost functions, w/wo regularizations. In particular:

1. We combine recent Neural OT method (Korotin et al., 2023b) with the congruence condition (§4.1) and derive a novel practical barycenter algorithm (§4.2, §4.3).
2. We consider several formulations of the barycenter problem, and for each particular formulation, we obtain quality bounds for the recovered solutions (§4.1).
3. We showcase the performance of our method on moderate- and high dimensional data (§5), e.g., in image space and the latent space of a pre-trained StyleGAN.

[*]Equal contribution [1]Skolkovo Institute of Science and Technology, Moscow, Russia [2]Department of Mathematics, ETH Zürich, Zürich, Switzerland [3]Artificial Intelligence Research Institute, Moscow, Russia. Correspondence to: Alexander Kolesov <a.kolesov@skoltech.ru>, Petr Mokrov <petr.mokrov@skoltech.ru>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

We acknowledge that our proposed approach borrows a significant number of theoretical and technical ideas from (Kolesov et al., 2023). Compared to the majority of other works, in practice, we do not limit ourselves to the quadratic cost functions and demonstrate an intriguing *shape-color* setup (§5.2). To the best of our knowledge, this setup paves a new principled way for creating generative distributions which aggregate some practically desired features gathered from multi-source data. We believe that the development of this idea will be crucial for solving real-world tasks (Kolesov et al., 2023, §B.2).

Notations. We write $\overline{K} = \{1, 2, \dots, K\}$ for $K \in \mathbb{N}$. For objects $o_1, o_2, \dots$ indexed by natural numbers, we use $o_{1:K}$ to denote the tuplet $(o_1, o_2, \dots, o_K)$. Throughout the paper, $\mathcal{X} \subset \mathbb{R}^{D'}, \mathcal{Y} \subset \mathbb{R}^D$ and $\mathcal{X}_k \subset \mathbb{R}^{D_k}$ are compact subsets of Euclidean spaces. The set of real-valued continuous functions on $\mathcal{X}$ is denoted by $\mathcal{C}(\mathcal{X})$. We use $\mathcal{P}(\mathcal{X})$ for the set of probability distributions on $\mathcal{X}$ and $\Pi(\mathbb{P}) \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ for the set of probability distributions on $\mathcal{X} \times \mathcal{Y}$ with fixed first marginal $\mathbb{P} \in \mathcal{P}(\mathcal{X})$. All probability distributions on $\mathcal{X} \times \mathcal{Y}$ with the marginals $\mathbb{P}$ and $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ are denoted by $\Pi(\mathbb{P}, \mathbb{Q}) \subset \Pi(\mathbb{P})$ (a.k.a. transport plans). For $\pi \in \mathbb{P}(\mathcal{X} \times \mathcal{Y})$, $\pi(\cdot|x)$ denotes the conditional distribution given the $\mathcal{X}$-coordinate, and $\pi^{\mathcal{Y}} \in \mathcal{P}(\mathcal{Y})$ denotes the second marginal distribution of π. For a measurable map T, we write $T_{\#}$ for the corresponding push-forward operator.

2. Background

To begin with, we recall the OT problem and its semi-dual formulation (§2.1). Then we proceed to formulate the corresponding OT barycenter problem (§2.2). Finally, we specify our learning setup (§2.3). Throughout the paper, when we refer to the OT problem we mean both the *classical* (strong) formulation, and its *weak* generalization.

Entropies and energy distances. We use H and KL to denote the differential entropy and the Kullback-Leibler divergence, respectively, see (Nutz, 2021, Def. 1.1). Let $\ell \in \mathcal{C}(\mathcal{Y} \times \mathcal{Y})$ be a *semimetric of negative type* (Sejdinovic et al., 2013, §2.1), e.g., $\ell(y, y') = \|y - y'\|_2$. For $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{Y})$, the square of the energy distance w.r.t. ℓ (Sejdinovic et al., 2013, Eq.2.5) is given by:

$$\mathcal{E}^2_\ell(\mu_1, \mu_2) \stackrel{\text{def}}{=} 2 \mathop{\mathbb{E}}_{\substack{y_1 \sim \mu_1 \\ y_2 \sim \mu_2}} \ell(y_1, y_2) - \mathop{\mathbb{E}}_{\substack{y_1 \sim \mu_1 \\ y_1' \sim \mu_1}} \ell(y_1, y_1') - \mathop{\mathbb{E}}_{\substack{y_2 \sim \mu_2 \\ y_2' \sim \mu_2}} \ell(y_2, y_2'),$$

where y_1, y_1', y_2, y_2' are pairwise independent. The energy distance $\mathcal{E}_\ell$ is a metric on $\mathcal{P}(\mathcal{Y})$ (Klebanov et al., 2005). It is deeply connected to MMDs (Sejdinovic et al., 2013) and induces a distance on $\Pi(\mathbb{P})$ via

$$\rho_\ell(\pi_1, \pi_2) \stackrel{\text{def}}{=} \sqrt{\mathop{\mathbb{E}}_{x \sim \mathbb{P}} \mathcal{E}^2_\ell(\pi_1(\cdot|x), \pi_2(\cdot|x))}, \tag{1}$$

where $\pi_1, \pi_2 \in \Pi(\mathbb{P})$. Indeed, ρ_ℓ defines a metric and was introduced in (Asadulaev et al., 2024, Appendix D).

2.1. Classical and Weak Optimal Transport

Let $\mathbb{P} \in \mathcal{P}(\mathcal{X}), \mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ and consider $c \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$, called the *ground* cost function. The classical OT problem (Kantorovitch, 1958) between $\mathbb{P}$ and $\mathbb{Q}$ consists in:

$$\text{OT}_c(\mathbb{P}, \mathbb{Q}) \stackrel{\text{def}}{=} \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \Big\{ \mathop{\mathbb{E}}_{(x,y) \sim \pi} c(x, y) \Big\}. \tag{2}$$

The specific choice $c(x, y) = \frac{1}{2}\|x - y\|_2^2$ in (2) yields $\mathbb{W}_2^2(\mathbb{P}, \mathbb{Q}) \stackrel{\text{def}}{=} \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{(x,y) \sim \pi} \frac{1}{2}\|x - y\|_2^2$, commonly known as the (squared) Wasserstein-2 distance.

Weak OT is a generalization of classical OT (2), where *weak* cost functions $C \colon \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ are employed. The weak OT problem (Gozlan et al., 2017) then is defined as follows:

$$\text{OT}_C(\mathbb{P}, \mathbb{Q}) \stackrel{\text{def}}{=} \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \mathop{\mathbb{E}}_{x \sim \mathbb{P}} C\big(x, \pi(\cdot|x)\big). \tag{3}$$

A typical example of a weak cost function C is:

$$C(x, \mu) = \mathop{\mathbb{E}}_{y \sim \mu} c(x, y) + \gamma \mathcal{R}(\mu),$$

where $c(x, y)$ is a ground cost function, $\gamma > 0$ is the regularization parameter, and $\mathcal{R} : \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ is a *regularizer*. Introducing regularization into the classical OT formulation (2) has proven to be beneficial, leading to the development of efficient optimization algorithms (Cuturi, 2013; Blondel et al., 2018) or the improvement of theoretical properties (Korotin et al., 2023a; Asadulaev et al., 2024). Table 1 presents some popular examples explored in recent OT literature. If a weak cost function C is *appropriate*[1], then weak OT (3) admits a minimizer π^* called the *OT plan* (Backhoff-Veraguas et al., 2019, Th. 1.2). Given that $\mathcal{X}, \mathcal{Y}$ are compact and c is continuous, the weak cost functions from Table 1 are *appropriate*, see (Backhoff-Veraguas et al., 2019; Korotin et al., 2023a).

Weak OT duality. For an appropriate weak cost function C, the weak OT problem (3) satisfies the duality

$$\text{OT}_C(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{C}(\mathcal{Y})} \Big\{ \mathop{\mathbb{E}}_{x \sim \mathbb{P}} f^C(x) + \mathop{\mathbb{E}}_{y \sim \mathbb{Q}} f(y) \Big\}, \tag{4}$$

where $f^C(x) \stackrel{\text{def}}{=} \inf_{\mu \in \mathcal{P}(\mathcal{Y})} \big\{ C(x, \mu) - \mathbb{E}_{y \sim \mu} f(y) \big\}$ is the *weak C-transform*, see (Gozlan et al., 2017), (Backhoff-Veraguas et al., 2019, Th. 1.3) for further details.

2.2. Classical and Weak OT Barycenter

Let $\mathbb{P}_k \in \mathcal{P}(\mathcal{X}_k)$ be given distributions and $C_k : \mathcal{X}_k \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ be appropriate weak cost functions, $k \in \overline{K}$. For

[1] Lower-boundedness, convexity in the second argument and joint lower semicontinuity on $\mathcal{X} \times \mathcal{P}(\mathcal{Y})$.

Cost function	$C(x, \mu) =$
classical (Fan et al., 2023; Rout et al., 2022)	$\mathbb{E}_{y\sim\mu}c(x, y)$
ϵ-entropic (Mokrov et al., 2024)	$\mathbb{E}_{y\sim\mu}c(x, y) - \epsilon H(\mu)$
γ-weak (kernel) quadratic (Korotin et al., 2023b;a)	$\mathbb{E}_{y\sim\mu}\|x - y\|^\alpha - \frac{\gamma}{2}\mathbb{E}_{y,y'\sim\mu}\|y - y'\|^\alpha, \alpha \in [1, 2]$

Table 1: Popular instances of weak cost functions.

positive weights $\lambda_k, \sum_{k=1}^K \lambda_k = 1$, the *weak OT barycenter* problem consists in finding a distribution that minimizes the (weighted) sum of OT problems with fixed first marginals in $\mathbb{P}_{1:K}$:

$$\mathcal{L}^* \stackrel{\text{def}}{=} \inf_{\mathbb{Q}\in\mathcal{P}(\mathcal{Y})} \sum_{k=1}^K \lambda_k \text{OT}_{C_k}(\mathbb{P}_k, \mathbb{Q}). \tag{5}$$

This equation encompasses various OT barycenter formulations explored in computational OT literature (Fan et al., 2021; Li et al., 2020; Cazelles et al., 2021). The utilization of certain weak OT problems in (5) instead of classical OT (Agueh & Carlier, 2011) permits us to derive theoretical guarantees for the recovered barycenter solutions, see Theorem 4.2 as well as (Li et al., 2020; Kolesov et al., 2023). Problem (5) admits a minimizer $\mathbb{Q}^* \in \mathcal{P}(\mathcal{Y})$, which is unique provided that the weak cost functions C_k are strictly convex w.r.t. the second argument. These assertions follow from standard measure-theoretic arguments, cf. (Backhoff-Veraguas et al., 2019), and have been observed in previous works, e.g., (Kolesov et al., 2023, §2.2).

Considered weak cost functions. The weak OT barycenter task (5) is overly general and thus has to be instantiated. To demonstrate the versatility of our approach, we stick in our experiments (§5) to the following families of cost functions, adapted from Table 1:

- Classical: $C(x, \mu) = \mathbb{E}_{y\sim\mu}c(x, y)$. (6)
- ϵ-KL: $C(x, \mu) = \mathbb{E}_{y\sim\mu}c(x, y) + \epsilon\text{KL}(\mu\|\mu_0)$. (7)
  Distribution μ_0 is a given prior, e.g., Normal; $\epsilon > 0$.
- γ-Energy: $C(x, \mu) = \mathbb{E}_{y\sim\mu}c(x, y) + \gamma\mathcal{E}_\ell^2(\mu, \mu_0)$. (8)
  Distribution μ_0 is a given prior; $\gamma > 0$.

Here, $\mu_0 \in \mathcal{P}(\mathcal{Y})$ can be seen as a prior distribution, reflecting the prior knowledge about the barycenter distribution $\mathbb{Q}^*$, which is precisely the way we use it in our experiments in the latent space.

The weak cost functions introduced above are *appropriate*. Indeed, they are lower-bounded (due to the compactness of $\mathcal{X}, \mathcal{Y}$). Convexity and lower semicontinuity of KL are well-known (Nutz, 2021), and the same properties also hold true for $\mathcal{E}_\ell^2$, see, e.g., (Asadulaev et al., 2024, Appendix A.2). For the classical cost function, these properties follow directly from its linearity (integrated against measures) and continuity of ground cost function c on the compact $\mathcal{X} \times \mathcal{Y}$. Meanwhile, the cost functions (6), (7), (8) have different theoretical properties which affect the analysis, see, e.g., Theorem 4.2.

2.3. Computational setup

In practice, the distributions $\mathbb{P}_k$ are not explicitly available, which causes ambiguity on how to adopt (5) for real-world problems. To avoid confusion, we describe below in detail our **learning setup**. Let π_k^*, $k \in \overline{K}$, be the (unknown) weak OT plans between the distributions $\mathbb{P}_k$ and the weak OT barycenter $\mathbb{Q}^*$ for given (known) cost functions C_k. We assume empirical samples (datasets) $X_k = \{x_k^i\}_{i=1}^{N_k} \sim \mathbb{P}_k$ are given. Our goal is to find approximations $\widehat{\pi}_k \in \Pi(\mathbb{P}_k)$ of the true OT plans π_k^*. These approximations are assumed to realize the *conditional* sampling procedure, i.e., taking a sample $x_k \sim \mathbb{P}_k$ as input and producing samples from $\widehat{\pi}_k(\cdot|x_k)$ as output. We stress that the input samples x_k are not necessarily from the training datasets X_k. The setup described above is called **continuous** (Li et al., 2020; Kolesov et al., 2023). Alternatively, the discrete setup (Peyré et al., 2019) aims at solving the OT barycenter problem between empirical measures $\widehat{\mathbb{P}}_k = {}^1/_{N_k}\sum_{n=1}^{N_k}\delta_{x_k^n}$. The resulting OT plan approximations operate exclusively with samples presented in the datasets X_k, which then requires extra effort to adapt for new samples (De Lara et al., 2021).

3. Related works

In this section, we give an overview of related works. We start with adversarial Neural OT methods, which form the basis of our approach. Then we discuss competitive OT barycenter methods that follow the continuous setup (§2.3).

Maximin Neural OT solvers leverage the dual problem (4) through optimization of adversarial-like max-min objectives. Among them, (Henry-Labordere, 2019; Rout et al., 2022; Gazdieva et al., 2022; Fan et al., 2023) focus on classical OT (2). On the other hand, (Korotin et al., 2023a;b) address specifically regularized cost functions, see Table 1. Recent works consider more exotic OT formulations, e.g., general OT cost *functionals* (Asadulaev et al., 2024), diffusion-based Entropic OT (Gushchin et al., 2023), etc.

Continuous OT barycenter methods differentiate by the two key characteristics: which problem they solve (general/particular form of the barycenter problem) and which computational algorithm they use.

In terms of the first characteristic, the majority of methods (Korotin et al., 2021c; Fan et al., 2021; Korotin et al., 2022; Noble et al., 2023) deal with quadratic (Euclidean) ground cost functions, i.e., $c_k(x_k, y) = {}^1/_2\|x_k - y\|_2^2$. A limited number of works consider general c_k (Li et al., 2020; Chi et al., 2023; Kolesov et al., 2023), but require additional entropic (quadratic) regularization. In contrast, our proposed approach provides greater flexibility, as it permits a wide range of admissible weak OT cost functions, including those

introduced in §2.2, but not limited to them.

Now we discuss existing computational procedures for solving the OT barycenter task. A branch of works (Fan et al., 2021; Korotin et al., 2022) employs tri-level (min-max-min) adversarial objectives, which may be hard to optimize. Other works manage to develop non-adversarial minimization objectives, but with limitations: (Li et al., 2020) *requires* a fixed prior distribution, (Noble et al., 2023; Kolesov et al., 2023) utilize time-consuming sampling procedures, (Korotin et al., 2021c) relies on Input Convex NNs (Amos et al., 2017) which have scalability issues (Korotin et al., 2021b). In contrast, our proposed approach uses a bi-level (max-min) computational algorithm (§4.3) widely embraced in generative modeling, specifically, in the field of Neural OT. Noteworthy, (Chi et al., 2023) also proposes a bi-level objective but utilizes the *reinforce* procedure. The latter is tricky, e.g., it requires variance reduction techniques. A comprehensive comparison of our proposed method and (Chi et al., 2023) can be found in Appendix E.

4. Proposed Method

In §4.1, we derive our novel max-min optimization objective for learning weak OT barycenters and establish error bounds for approximate solutions. After that, in §4.2, we describe the parametrization of plans $\Pi(\mathbb{P})$ as (stochastic) maps. The described techniques are crucial when adapting our derived objective to practice. In §4.3, we present the resulting computational algorithm. All the proofs can be found in Appendix A. An additional conceptual derivation of our main result (Theorem 4.1) is in Appendix A.3.

4.1. Maximin Dual Weak OT Barycenter formulation

Let $f_k \in \mathcal{C}(\mathcal{Y})$, $\pi_k \in \Pi(\mathbb{P}_k)$. We introduce the functionals:

$$\mathcal{V}(f_{1:K}, \pi_{1:K}) \stackrel{\text{def}}{=} \sum_{k=1}^{K} \lambda_k \Big\{ \mathop{\mathbb{E}}_{x_k \sim \mathbb{P}_k} C_k\big(x_k, \pi(\cdot|x_k)\big) - \mathop{\mathbb{E}}_{x_k \sim \mathbb{P}_k} \Big(\mathop{\mathbb{E}}_{y \sim \pi_k(\cdot|x_k)} f_k(y) \Big) \Big\}; \quad (9)$$

$$\mathcal{L}(f_{1:K}) \stackrel{\text{def}}{=} \inf_{\substack{\pi_k \in \Pi(\mathbb{P}_k), \\ (k \in \overline{K})}} \mathcal{V}(f_{1:K}, \pi_{1:K}). \quad (10)$$

Theorem 4.1 presents our main theoretical result. It states that if the potentials $f_{1:K}$ satisfy the **congruence** condition $\sum_{k=1}^{K} \lambda_k f_k \equiv 0$, then the optimization of (10) yields the optimal value to the OT barycenter objective (5).

Theorem 4.1 (max-min formulation for OT barycenter)**.** *The optimal value $\mathcal{L}^*$ of the OT barycenter problem (5) is given by the following* max-min *objective:*

$$\mathcal{L}^* = \sup_{\sum \lambda_k f_k = 0} \mathcal{L}(f_{1:K}) = \sup_{\sum \lambda_k f_k = 0} \inf_{\substack{\pi_k \in \Pi(\mathbb{P}_k), \\ (k \in \overline{K})}} \mathcal{V}(f_{1:K}, \pi_{1:K}). \quad (11)$$

Comparing with the literature, we establish in Theorem 4.1 duality for the barycenter problems with **general weak** cost functions. Notably, existing alternatives in the literature primarily address specific cases, such as squared Euclidean ground cost (Agueh & Carlier, 2011), entropic and quadratic regularization (Li et al., 2020; Kolesov et al., 2023). At the same time, the proof of our Theorem 4.1 is an easy generalization of (Kolesov et al., 2023, Thm. 4.1).

Quality bounds. Recall that our ultimate goal is to approximate true OT plans π_k^* between reference distributions $\mathbb{P}_k$ and the barycenter $\mathbb{Q}^*$, see §2.3. Let $(\widehat{f}_{1:K}, \widehat{\pi}_{1:K})$ be a tuple which approximately solves (11). Theorem 4.2 below characterizes proximity of the recovered plans $\widehat{\pi}_{1:K}$ to the true plans $\pi^*_{1:K}$, covering the cases of classical (6), ϵ − KL (7) and γ-energy (8). The reported quality bounds are based on duality gaps, i.e., errors for solving outer (sup) and internal (inf) optimization problems w.r.t. the functional $\mathcal{V}$.

Theorem 4.2 (Quality bounds for recovered plans)**.** *Let $\widehat{f}_{1:K}$ be congruent potentials and $\widehat{\pi}_k \in \Pi(\mathbb{P}_k)$, $k \in \overline{K}$ be OT plan approximations. Consider the duality gaps:*

$$\delta_1(\widehat{f}_{1:K}, \widehat{\pi}_{1:K}) \stackrel{\text{def}}{=} \mathcal{V}(\widehat{f}_{1:K}, \widehat{\pi}_{1:K}) - \mathcal{L}(\widehat{f}_{1:K}); \quad (12)$$

$$\delta_2(\widehat{f}_{1:K}) \stackrel{\text{def}}{=} \mathcal{L}^* - \mathcal{L}(\widehat{f}_{1:K}). \quad (13)$$

The following statements hold true:

1. *Let C_k be classical (6) cost functions, and $\mathcal{Y}$ be a convex set. Assume that the maps $y \mapsto c_k(x_k, y) - \widehat{f}_k(y)$ are β-strongly convex for each $k \in \overline{K}$, $x_k \in \mathcal{X}_k$. Then,*
$$\sum_{k=1}^{K} \lambda_k \mathop{\mathbb{E}}_{x \sim \mathbb{P}_k} \mathbb{W}_2^2\big(\widehat{\pi}_k(\cdot|x), \pi_k^*(\cdot|x)\big) \leq \frac{2}{\beta}\Big(\delta_1 + \delta_2\Big),$$

2. *Let C_k be ϵ-KL (7) cost functions, $\epsilon > 0$, and the prior μ_0 has positive continuous Lebesgue density on $\mathcal{Y}$. Then,*
$$\sum_{k=1}^{K} \lambda_k \rho_{TV}(\widehat{\pi}_k, \pi_k^*)^2 \leq \frac{1}{\epsilon}\Big(\delta_1 + \delta_2\Big),$$
where ρ_{TV} is the total variation distance.

3. *Let C_k be γ-Energy cost functions, $\gamma > 0$. Then,*
$$\sum_{k=1}^{K} \lambda_k \rho_\ell(\widehat{\pi}_k, \pi_k^*)^2 \leq \frac{2}{\gamma}\Big(\delta_1 + \delta_2\Big).$$

From the quality bounds, we deduce that when the pair $(\widehat{f}_{1:K}, \widehat{\pi}_{1:K})$ solves the inner and outer optimization in (11) relatively well, e.g., our Algorithm 1 below which optimizes (11) converged nearly to an optimum, then the recovered plans $\widehat{\pi}_{1:K}$ are **close** to the true OT plans. One problem may occur with the strong convexity assumptions for the maps $y \mapsto c_k(x_k, y) - \widehat{f}_k(y)$ in the case of classical cost functions. Due to the difficulties of imposing (strong) convexity constraints (Korotin et al., 2021b), we can not guarantee them

in practice. Note that previous works (Fan et al., 2023; Rout et al., 2022) also did not care much about them. Nevertheless, the experiments (§5) demonstrate a good performance of our OT barycenter solver with classical cost functions in various setups. At the same time, we stress that the usage of regularized cost functions (7), (8) completely eliminates the need for imposing strong convexity constraints in Theorem 4.2. This is an important motivation for considering weak OT (3) in addition to the classical one (2) (Korotin et al., 2023a) when solving OT barycenter problems.

Relation to prior works. Theorem 4.2 encompasses and generalizes existing error analysis results for recovered OT plans, and makes them applicable for the OT barycenter problem. Prior works which establish quality bounds when solving OT-related problems from dual perspectives are: (Makkuva et al., 2020, Th. 3.6), (Fan et al., 2023, Th. 4), (Rout et al., 2022, Th. 4.3) (classical cost); (Gushchin et al., 2023, Th. 4.3), (Mokrov et al., 2024, Th. 2), (Kolesov et al., 2023, Th. 2) (entropic); (Asadulaev et al., 2024, Th. 3) (general strongly convex cost).

4.2. Parameterizing transport plans via stochastic maps

Objective (11) prescribes optimization over plans $\pi_k \in \Pi(\mathbb{P}_k)$. Direct optimization over probability distributions is non-trivial, and we adopt the parameterization of plans with functions. We pursue the following approaches.

Stochastic maps. The approach is similar to (Korotin et al., 2023b, §4.1). We introduce an auxiliary space $\mathcal{S} \subset \mathbb{R}^{D_s}$, an atomless distribution $\mathbb{S} \in \mathcal{P}(\mathcal{S})$, and consider measurable maps $T : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$. Every plan $\pi \in \Pi(\mathbb{P})$ can be implicitly represented by T_π s.t. $\pi(\cdot|x) = T_\pi(x, \cdot)_\# \mathbb{S}$. Particularly, given $x \sim \mathbb{P}$ and $s \sim \mathbb{S}$, the pair $(x, T_\pi(x, s))$ is distributed as π. In turn, every measurable $T : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$ implicitly specifies a plan $\pi_T \in \Pi(\mathbb{P})$. Taking the advantage of stochastic maps parameterization, (11) allows for an alternative objective:

$$\mathcal{L}^* = \sup_{\sum \lambda_k f_k = 0} \inf_{T_{1:K}} \widetilde{\mathcal{V}}(f_{1:K}, T_{1:K}); \tag{14}$$

$$\widetilde{\mathcal{V}}(f_{1:K}, T_{1:K}) \stackrel{\text{def}}{=} \sum_{k=1}^{K} \lambda_k \Big\{ \mathop{\mathbb{E}}_{x_k \sim \mathbb{P}_k} C_k\big(x_k, T_k(x_k, \cdot)_\# \mathbb{S}\big) - \mathop{\mathbb{E}}_{x_k \sim \mathbb{P}_k} \mathop{\mathbb{E}}_{s \sim \mathbb{S}} f_k\big(T_k(x_k, s)\big) \Big\}. \tag{15}$$

Optimization (14) is generic and can be implemented in practice as far as it is possible to estimate the weak cost function by samples. This is the case for classical (6) and γ-energy (8) cost functions.

Gaussian model. We parameterize conditional plans $\pi(\cdot|x)$ as Gaussian distributions $\mathcal{N}(\cdot|\mu(x), \text{diag}(\sigma^2(x)))$. The method is widely used, e.g., in the celebrated VAEs (Kingma & Welling, 2014). Technically, the optimization w.r.t. transport plans boils down to optimization of functions $\mu(x), \sigma(x)$ given by, e.g., encoding NNs. The proposed Gaussian model is the particular instance of stochastic maps model with $\mathbb{S} = \mathcal{N}(0, I_{D_s})$, $T(x, s) = \mu(x) + \sigma(x) \cdot s$. Note that the Gaussian parameterization allows explicitly computing the KL regularization term in (7). Theoretically, the substitution of $\pi(\cdot|x) \in \mathcal{P}(\mathcal{Y})$ with Gaussians is not very accurate and should be seen as a technical trick. To alleviate the imprecision of this approximation, we utilize Gaussian model exclusively in the latent space, see §5.

Deterministic maps. We consider measurable maps $T : \mathcal{X} \to \mathcal{Y}$ and model conditional plans as deterministic distributions with delta functions $\pi(\cdot|x) = \delta_{T(x)}(\cdot)$. The method is a particular case of stochastic maps parameterization with removed stochasticity. This can be naturally applied to classical OT cost. In this case, modelling with deterministic maps has a direct relation to the seminal formulation of OT problem due to Monge (Monge, 1781) and has found its application in a number of works, e.g., (Makkuva et al., 2020; Korotin et al., 2021a; Rout et al., 2022; Fan et al., 2023). Such an extensive practical utilization combined with our encouraging numerical validation (§5) makes the deterministic maps model meaningful for solving the OT barycenter problem.

4.3. Computational barycenter algorithm

In this subsection, we develop a practical optimization procedure for solving (14). We parameterize (stochastic) maps $T_{1:K}$ and potentials $f_{1:K}$ as neural networks. The parameter space for the maps is $\Phi = \Phi_1 \times \Phi_2 \cdots \times \Phi_K$; the NN maps are $T_{k,\phi} : \mathbb{R}^{D_k} \times \mathbb{R}^{D_s} \to \mathbb{R}^D$, $\phi = (\phi_1, \dots \phi_K) \in \Phi$. When modeling deterministic maps, we omit stochastic dimensions $\mathbb{R}^{D_s}$. The parameter space for the potentials is $\Theta = \Theta_1 \times \Theta_2 \dots \times \Theta_K$. To ensure the congruence condition, we parameterize the potentials $f_{k,\theta}$, $\theta = (\theta_1, \dots, \theta_K) \in \Theta$ with help of auxiliary NNs $g_{\theta_k} : \mathbb{R}^D \to \mathbb{R}$ as follows: $f_{k,\theta} \stackrel{\text{def}}{=} g_{\theta_k} - \sum_{k'=1}^{K} \lambda_{k'} g_{\theta_{k'}}$. This trick is used in (Li et al., 2020; Kolesov et al., 2023). Below, $T_{1:K,\phi}$ and $f_{1:K,\theta}$ denote the tuplets of the NN maps and potentials.

Training. To optimize max-min objective (14) with NNs we utilize stochastic gradient ascent-descent algorithm by performing several minimization steps w.r.t. $T_{1:K,\phi}$ per each maximization step w.r.t. congruent potentials $f_{1:K,\theta}$. The functional $\widetilde{\mathcal{V}}$ in (14) is estimated from samples via Monte-Carlo. We derive random batches from reference distributions $\mathbb{P}_k$ and (optionally) random batches from auxiliary $\mathbb{S}$. This allows us to straightforwardly estimate the second term $\mathbb{E}_{x \sim \mathbb{P}_k} \mathbb{E}_{s \sim \mathbb{S}} f_{k,\theta}(T_{k,\phi}(x_k, s))$ in (15). The details on approximating weak cost functions $\widehat{C}$ are below. These approximations are averaged over batches derived from $\mathbb{P}_k$ when computing the estimator for the 1st term in (15).

- **Classical**: Sample auxiliary batch $S \sim \mathbb{S}$, use mean:

$$\widehat{C}(x, T(x,S)) = \sum_{s\in S} \frac{c(x,T(x,s))}{|S|}.$$

For deterministic maps model: $\widehat{C}(x,T(x)) = c(x,T(x))$.

- **ϵ-KL**: We utilize *Gaussian model* for stochastic maps, i.e, T is defined by means and deviations μ, σ, see §4.2. Distributions $\mathbb{S}$, μ_0 are assumed to be Gaussian. Then,

$$\widehat{C}(x, T(x,S)) = \sum_{s\in S} \frac{c(x,T(x,s))}{|S|},\ S\sim\mathbb{S} \quad \text{\# sample mean}$$
$$+\epsilon \mathrm{KL}(\mathcal{N}(\mu(x), \sigma(x)) \| \mu_0). \quad \text{\# computed analytically}$$

- **γ-Energy**: Sample auxiliary ($S\sim\mathbb{S}$) and prior ($Y_0\sim\mu_0$) batches. Then compute:

$$\widehat{C}(x, T(x,S), Y_0) = \sum_{s\in S} \frac{c(x,T(x,s))}{|S|} +$$
$$\gamma \sum_{s\in S} \left(2 \sum_{y\in Y_0} \frac{\ell(T(x,s),y)}{|S||Y_0|} - \sum_{s'\in S\setminus\{s\}} \frac{\ell(T(x,s),T(x,s'))}{|S|(|S|-1)} \right),$$

which is the estimator of (8) up to T-independent constant (Gretton et al., 2012, Lemma 6).

All ingredients for computing the sample estimator of $\widetilde{\mathcal{V}}$ are ready, and we proceed to our main barycenter Algorithm 1.

Algorithm 1 OT Barycenter via Neural Optimal Transport

Input: Distributions $\mathbb{P}_{1:K}, \mathbb{S}$ accessible by samples; NN maps $T_{k,\phi} : \mathbb{R}^{D_k}\times\mathbb{R}^{D_s}\to\mathbb{R}^D$ and NN congruent potentials $f_{k,\theta}:\mathbb{R}^D\to\mathbb{R}$, $k\in\overline{K}$; number of inner iterations M_T; weak cost function estimator $\widehat{C}$; batch sizes; prior distribution μ_0.

Output: Learned (stochastic) maps $T_{1:K,\phi^*}$ representing OT plans between $\mathbb{P}_k$ and barycenter $\mathbb{Q}^*$.

repeat

  Sample batches $X_k\sim\mathbb{P}_k$, $k\in\overline{K}$;

  For each $x_k\in X_k$ sample auxiliary batch $S[x_k]\sim\mathbb{S}$;

  $\mathcal{V}_f \leftarrow \sum_{k\in\overline{K}} \lambda_k \Big\{ \sum_{x_k\in X_k} \sum_{s_k\in S[x_k]} \frac{f_{k,\theta}(T_{k,\phi}(x_k,s_k))}{|X_k||S[x_k]|} \Big\}$;

  Update θ by using $\frac{\partial\mathcal{V}_f}{\partial\theta}$;

  for $m_T = 1, 2, \dots, M_T$ **do**

    Sample batches $X_k\sim\mathbb{P}_k$; For each $x_k\in X_k$**:**

      Sample auxiliary batch $S[x_k]\sim\mathbb{S}$;

      (Optionally) sample prior batch $Y_0[x_k]\sim\mu_0$;

    $\mathcal{V}_{T_k} \leftarrow \sum_{x\in X_k} \Big\{ \frac{\widehat{C}(x, T_{k,\phi}(x,S[x]), Y_0[x])}{|X_k|}$

    $- \sum_{s\in S[x]} \frac{f_{k,\theta}(T_{k,\phi}(x,s))}{|X_k||S[x]|} \Big\}$; $\quad \mathcal{V}_T \leftarrow \sum_{k\in\overline{K}} \lambda_k \mathcal{V}_{T_k}$;

    Update ϕ by using $\frac{\partial\mathcal{V}_T}{\partial\phi}$;

until *not converged*;

5. Experimental Illustrations

In this section, we showcase the performance of our proposed method. We have an illustrative experiment in 2D (§5.1), a benchmark experiment with CelebA images (§5.3) and an impressive experiment with shape- and color-preserving cost functions (§5.2). Additional practical test cases with multidimensional Gaussians and MNIST digits are in Appendix B; the experimental details are in Appendix C. Our source code is available at `https://github.com/justkolesov/NOTBarycenters`.

OT barycenters constrained to image manifolds. Following (Kolesov et al., 2023), in a number of our experiments, we restrict the support of the desired barycenter distribution to some image manifold $\mathcal{M}$. In our experiments, these manifolds are given by pre-trained StyleGAN generators G, i.e., $\mathcal{M} = G(\mathcal{Z})$, where $\mathcal{Z} = \mathbb{R}^{D_z}$ is the latent space of StyleGAN. In order to make our Algorithm 1 learn the barycenter on $\mathcal{M}$ we use a specific parameterization of NN maps $T_{k,\phi}$. At first, they push input points to latent space and then recover target points with G. The details of the specific StyleGAN models in use can be found in the respective subsections. Note that the manifold-constrained barycenter task is equivalent to *unconstrained* OT barycenter problem in the latent space with cost functions $c_{k,G}(x_k, z) \stackrel{\text{def}}{=} c_k(x_k, G(z))$, which are *principially* non-quadratic. The latter makes the manifold case impractical for the majority of existing barycenter solvers.

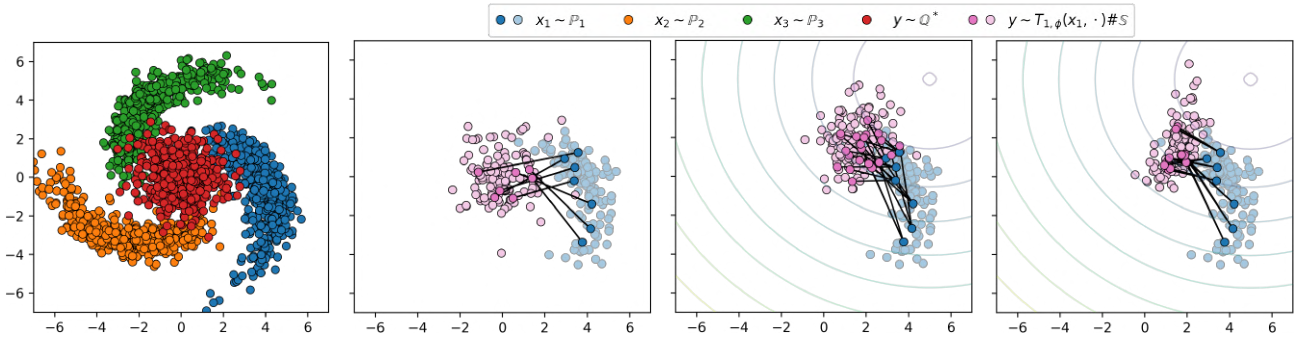
5.1. Illustrative 2D barycenters (Twister)

The experiment follows **2D Twister** setup from (Kolesov et al., 2023, §5.1). In particular, we introduce a map $u:\mathbb{R}^2\to\mathbb{R}^2$ which rotates input points x by angles proportional to $\|x\|$. Then for distributions $\mathbb{P}_{1:3}$ depicted in Fig. 1a, classical cost functions $c_k(x_k, y) = 1/2\|u(x_k) - u(y)\|_2^2$ and weights $\lambda_k = 1/3$ the (unregularized) ground truth barycenter $\mathbb{Q}^*$ is known, see (Kolesov et al., 2023, Appendix C.1). It is the zero-centered Gaussian, see Fig. 1a. Our approach without regularizations, i.e., with classical costs (6), successfully recovers $\mathbb{Q}^*$, see Fig. 1b. Since our setup is symmetric, we demonstrate stochastic OT mapping only from $\mathbb{P}_1$. In the other runs (Figs. 1c, 1d) we demonstrate how the obtained barycenters are influenced by regularizations. Specifically, we use ϵ-KL (7) and γ-Energy (8) costs with Gaussian prior $\mu_0 = \mathcal{N}\big((5,5), I_2\big)$. As expected, the recovered barycenters do not coincide with $\mathbb{Q}^*$ and 'tend' to μ_0.

5.2. Shape-Color Experiment

The majority of continuous barycenter solvers (Korotin et al., 2022; Fan et al., 2021; Noble et al., 2023) work only with quadratic cost functions in data space, i.e., fit barycenters that are direct pixel-wise averages of samples from inputs $\mathbb{P}_k$, e.g., see Fig. 5 in Appendix B. However, such barycenters are poorly useful in practice. Our proposed approach works for general cost functions, and we conduct an experiment that demonstrates learning *more reasonable* barycenters. Below we introduce the components of our setup.

We consider ϵ-KL OT barycenter problem (7) with weights



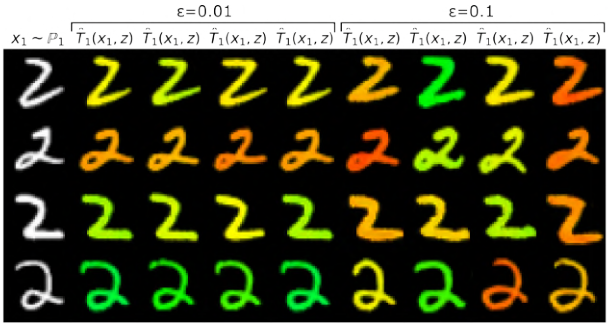
(a) Input distributions $\mathbb{P}_{1:3}$ and GT barycenter $\mathbb{Q}^*$ for the twisted cost.

(b) **Our** unregularized (classical cost functions) barycenter, map from $\mathbb{P}_1$.

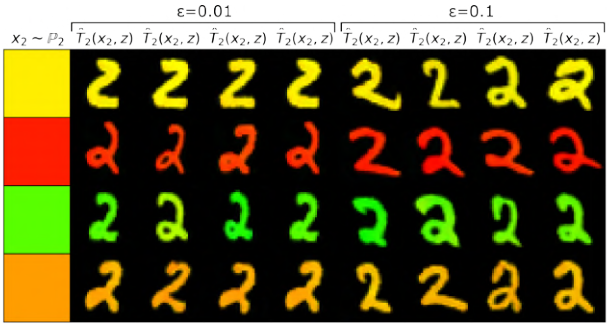
(c) **Our** regularized barycenter, $\epsilon = 1$-KL cost functions, map from $\mathbb{P}_1$. The learned barycenter tends to prior μ_0.

(d) **Our** regularized barycenter, $\gamma = 1$-Energy costs, map from $\mathbb{P}_1$. The learned barycenter tends to prior μ_0.

Figure 1: 2D Twister experiment (§5.1). Contours represent the prior distribution μ_0.



(a) Source ($x_1 \sim \mathbb{P}_1$) and transformed ($y = T_1(x_1, z)$) samples for input $\mathbb{P}_1$ (first reference distribution).



(b) Source ($x_2 \sim \mathbb{P}_2$) and transformed ($y = T_2(x_2, z)$) samples for input $\mathbb{P}_2$ (second reference distribution).

Figure 2: **Our** learned stochastic maps to the OT barycenter in the Shape-Color experiment (§5.2).

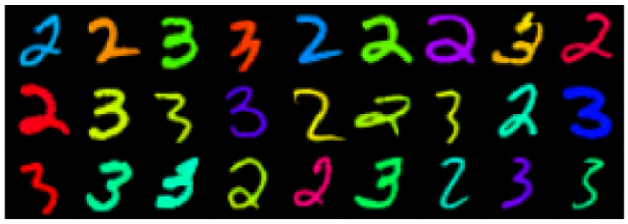


Figure 3: Samples from the StyleGAN G (which represents manifold $\mathcal{M}$) trained on colored MNIST digits "2" &"3". $(\frac{1}{2}, \frac{1}{2})$ for distributions $\mathbb{P}_1$ and $\mathbb{P}_2$. We test $\epsilon = 0.01, 0.1$.

Shape distribution ($\mathbb{P}_1$). This distribution is composed of grayscale (non-colored) images of MNIST digits '2'. This is a distribution on space $\mathcal{X}_1 \stackrel{\text{def}}{=} [0,1]^{32\times32}$ of images.

Color distribution ($\mathbb{P}_2$). This is a distribution of HSV (hue, saturation, value) vectors (colors) on $\mathcal{X}_2 = [0,1]^3$. Saturation and value are set to 1 while hue $\sim U[0, 0.5]$. Thus, $\mathbb{P}_2$ is composed of red, yellow and green colors.

Manifold $\mathcal{M}$. We search for the barycenter on the manifold $\mathcal{M}$ of digits "2" and "3" of **all** colors (red, orange, yellow, green, cyan, blue, purple). It is represented by the StyleGAN G trained on colored images '2' and '3' (Figure 3).

Transport costs. The transport cost for samples from distribution $\mathbb{P}_1$ is $c_1(x_1, z) \stackrel{\text{def}}{=} \frac{1}{2}||x_1 - H_g(G(z))||^2$, where $H_g : \mathbb{R}^{3\times32\times32} \to \mathbb{R}^{32\times32}$ is decolorization operator. This cost compares shapes and does not take the color into account. The other transport cost is $c_2(x_2, z) \stackrel{\text{def}}{=} \frac{1}{2}||x_2 - H_c(G(z))||^2$, where $H_c : \mathbb{R}^{3\times32\times32} \to \mathbb{R}^3$ transforms generated digit to three-dimensional HSV vector of its color, see Appendix C.

Setup summary. Manifold $\mathcal{M}$ contains both digits "2" and "3" of all colors. Meanwhile, the barycenter of $\mathbb{P}_1$ and $\mathbb{P}_2$ is not expected to be supported by the *whole* $\mathcal{M}$. Indeed, it should contain only shapes as in $\mathbb{P}_1$ and only colors as in $\mathbb{P}_2$, i.e., *only* red, yellow and green colors.

Results. *Our method indeed recovers the expected barycenter.* The learned map from $\mathbb{P}_1$ to the barycenter preserves the shape of input digit "2" but paints it with colors from $\mathbb{P}_2$, see Fig. 2. At the same time, the map from $\mathbb{P}_2$ generates images of digit "2" with a given color. Furthermore, the diversity of generated images can be controlled by the parameter ϵ.

5.3. Ave, celeba! Barycenter Benchmark Dataset

To quantitatively evaluate our proposed method, we utilize Ave, celeba! barycenter benchmark as proposed in (Korotin et al., 2022). It contains 3 distributions $(\mathbb{P}_0, \mathbb{P}_1, \mathbb{P}_2)$ of transformed CelebA faces. These transformations are such that the barycenter $\mathbb{Q}^*$ of the reference distributions with

(a) input, transformed samples ($\mathbb{P}_0$). **TRUE** shows GT mapping $\mathbb{P}_0 \to \mathbb{Q}^*$.

(b) input, transformed samples ($\mathbb{P}_1$). **TRUE** shows GT mapping $\mathbb{P}_1 \to \mathbb{Q}^*$.

(c) input, transformed samples ($\mathbb{P}_2$). **TRUE** shows GT mapping $\mathbb{P}_2 \to \mathbb{Q}^*$.

Figure 4: Learned (stochastic) maps to the OT barycenter by different solvers; Ave, Celeba! experiment (§5.3).

Space	Solver	FID↓		
		$k=1$	$k=2$	$k=3$
Data space	SCWB	56.7	53.2	58.8
	WIN	49.3	46.9	61.5
	OURS	**39.0**	**38.6**	**39.8**
Mani-fold	EgBary	**8.4**	**8.7**	**10.2**
	OURS	30.7	31.0	31.7
	OURS (ϵ)	34.5	34.9	35.7
	OURS (γ)	38.3	37.8	37.6

Table 2: Quantitative comparison of barycenter solvers on the Ave, celeba! benchmark dataset.

weights $(1/4, 1/2, 1/4)$ and classical quadratic cost functions coincides with CelebA distribution. Moreover, the ground truth quadratic OT maps (T_0^*, T_1^*, T_2^*) from the reference distributions to $\mathbb{Q}^*$ are also known by construction. All that remains is to compare the recovered barycenter and mappings with the true ones.

In our experiments, we consider both conventional data-space setup and *manifold-constrained* setup. In the latter case, similar to (Kolesov et al., 2023), we utilize StyleGAN generator trained on CelebA dataset. The extended information on the training/performance of our method and baselines on Ave, celeba! benchmark can be found in Appendix D.

Data-space. Our competitors are: WIN (Korotin et al., 2022), SCWB (Fan et al., 2021). These methods are good baselines for continuous OT barycenter setup. Similar to the competitors, we use unregularized cost functions. From Table 2 we see that our method achieves **the best** FID.

Manifold-constrained. We run our method with both classic (6) and regularized (7), (8) cost functions, see the samples from recovered barycenters in Fig. 4. Additional samples for different regularization strengths γ, ϵ are in the Appendix, Figs. 8, 9. The competitive method is EgBary (Kolesov et al., 2023), it demonstrates better results than ours, see Table 2. We leave the detailed analysis of this fact to future research, but note that EgBary utilizes MCMC (namely, Langevin sampling) at training and inference. Probably, this procedure better suits latent-space setups, but is rather time-consuming. For completeness, we place a detailed experimental comparison of our method and EgBary in Appendix D.1.

6. Discussion

Potential impact. The main feature of our new OT barycenter solver is flexibility. Specifically, it is adjustable to dif-

ferent variants of the barycenter problem and can recover both deterministic and stochastic transport plans between data distributions. At the same time, we do not use exotic computational procedures: the core of our algorithm is a conventional bi-level adversarial game. Due to these properties, our method seems to be a good candidate to become a *standard* tool for solving the OT barycenter problem.

Another impactful byproduct of our paper is a new promising practical setup (§5.2). It provides the ability to create barycenter distributions with desired characteristics by proper selection of *non-Euclidean* cost functions. We believe that this finding will be useful in industrial tasks.

Limitations. (a) When dealing with classical cost functions, our approach is not guaranteed to recover OT plans π_k^*. In particular, the quality bounds for the recovered solutions in this case were established only under specific assumptions, see our Theorem 4.2, Statement 1. **(b)** The utilization of Gaussian models (§4.2) for transport plan parameterization is a simplification and may result in biased solutions. **(c)** Adversarial optimization procedures like those used in our paper may be prone to instabilities and necessitate hyperparameter tuning. Nevertheless, our method seems to work well for different costs and parameterizations (§5). We conjecture that the concerns from above are avoidable or not so important in practice.

Acknowledgements

Skoltech was supported by the Analytical center under the RF Government (subsidy agreement 000000D730321P5Q0002, Grant No. 70-2021-00145 02.11.2021).

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning, namely, generative modeling and computational optimal transport. Potential broader impact of our research is the same as that of most of the other generative modeling researches. In fact, there are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Agueh, M. and Carlier, G. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2): 904–924, 2011.

Álvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J., and Matrán, C. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

Amos, B., Xu, L., and Kolter, J. Z. Input convex neural networks. In *International Conference on Machine Learning*, pp. 146–155. PMLR, 2017.

Anderes, E., Borgwardt, S., and Miller, J. Discrete wasserstein barycenters: Optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84:389–409, 2016.

Asadulaev, A., Korotin, A., Egiazarian, V., Mokrov, P., and Burnaev, E. Neural optimal transport with general cost functionals. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gIiz7tBtYZ.

Backhoff-Veraguas, J., Beiglböck, M., and Pammer, G. Existence, duality, and cyclical monotonicity for weak transport costs. *Calculus of Variations and Partial Differential Equations*, 58(6):203, 2019.

Bertsekas, D. and Shreve, S. E. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.

Blondel, M., Seguy, V., and Rolet, A. Smooth and sparse optimal transport. In *International conference on artificial intelligence and statistics*, pp. 880–889. PMLR, 2018.

Cazelles, E., Tobar, F., and Fontbona, J. A novel notion of barycenter for probability distributions based on optimal weak mass transport. *Advances in Neural Information Processing Systems*, 34:13575–13586, 2021.

Chi, J., Yang, Z., Li, X., Ouyang, J., and Guan, R. Variational wasserstein barycenters with c-cyclical monotonicity regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7157–7165, 2023.

Choi, J., Choi, J., and Kang, M. Generative modeling through the semi-dual formulation of unbalanced optimal transport. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=7WQt1J13ex.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Cuturi, M. and Doucet, A. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pp. 685–693. PMLR, 2014.

Daniels, M., Maunu, T., and Hand, P. Score-based generative neural networks for large-scale optimal transport. *Advances in neural information processing systems*, 34: 12955–12965, 2021.

De Lara, L., González-Sanz, A., and Loubes, J.-M. A consistent extension of discrete optimal transport maps for machine learning applications. *arXiv preprint arXiv:2102.08644*, 2021.

Fan, J., Taghvaei, A., and Chen, Y. Scalable computations of wasserstein barycenter via input convex neural networks. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1571–1581. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/fan21d.html.

Fan, J., Liu, S., Ma, S., Zhou, H.-M., and Chen, Y. Neural monge map estimation and its applications. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=2mZSlQscj3. Featured Certification.

Gazdieva, M., Rout, L., Korotin, A., Kravchenko, A., Filippov, A., and Burnaev, E. An optimal transport perspective on unpaired image super-resolution. *arXiv preprint arXiv:2202.01116*, 2022.

Gozlan, N., Roberto, C., Samson, P.-M., and Tetali, P. Kantorovich duality for general transport costs and applications. *Journal of Functional Analysis*, 273(11):3327–3405, 2017.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Gushchin, N., Kolesov, A., Korotin, A., Vetrov, D., and Burnaev, E. Entropic neural optimal transport via diffusion processes. In *Advances in Neural Information Processing Systems*, 2023.

Henry-Labordere, P. (martingale) optimal transport and anomaly detection with neural networks: A primal-dual algorithm. *arXiv preprint arXiv:1904.04546*, 2019.

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

Kechris, A. *Classical descriptive set theory*, volume 156. Springer Science & Business Media, 2012.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Klebanov, L. B., Beneš, V., and Saxl, I. *N-distances and their applications*. Charles University in Prague, the Karolinum Press Prague, Czech Republic, 2005.

Kolesov, A., Mokrov, P., Udovichenko, I., Gazdieva, M., Pammer, G., Burnaev, E., and Korotin, A. Energy-guided continuous entropic barycenter estimation for general costs. *ArXiv*, abs/2310.01105, 2023. URL https://api.semanticscholar.org/CorpusID:263605771.

Korotin, A., Egiazarian, V., Asadulaev, A., Safin, A., and Burnaev, E. Wasserstein-2 generative networks. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=bEoxzW_EXsa.

Korotin, A., Li, L., Genevay, A., Solomon, J. M., Filippov, A., and Burnaev, E. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *Advances in Neural Information Processing Systems*, 34:14593–14605, 2021b.

Korotin, A., Li, L., Solomon, J., and Burnaev, E. Continuous wasserstein-2 barycenter estimation without minimax optimization. In *International Conference on Learning Representations*, 2021c. URL https://openreview.net/forum?id=3tFAs5E-Pe.

Korotin, A., Egiazarian, V., Li, L., and Burnaev, E. Wasserstein iterative networks for barycenter estimation. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=GiEnzxTnaMN.

Korotin, A., Selikhanovych, D., and Burnaev, E. Kernel neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=Zuc_MHtUma4.

Korotin, A., Selikhanovych, D., and Burnaev, E. Neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=d8CBRlWNkqH.

Lacombe, J., Digne, J., Courty, N., and Bonneel, N. Learning to generate wasserstein barycenters. *Journal of Mathematical Imaging and Vision*, 65(2):354–370, 2023.

Li, L., Genevay, A., Yurochkin, M., and Solomon, J. M. Continuous regularized wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33:17755–17765, 2020.

Likmeta, A., Sacco, M., Metelli, A. M., and Restelli, M. Wasserstein actor-critic: Directed exploration via optimism for continuous-actions control. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37 (7):8782–8790, Jun. 2023. doi: 10.1609/aaai.v37i7.

26056. URL https://ojs.aaai.org/index.php/AAAI/article/view/26056.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 6672–6681. PMLR, 2020.

Mokrov, P., Korotin, A., Kolesov, A., Gushchin, N., and Burnaev, E. Energy-guided entropic neural optimal transport. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=d6tUsZeVs7.

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

Mroueh, Y. Wasserstein style transfer. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 842–852. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/mroueh20a.html.

Noble, M., Bortoli, V. D., Doucet, A., and Durmus, A. Tree-based diffusion schrödinger bridge with applications to wasserstein barycenters. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=H2SuXHbFIn.

Nutz, M. Introduction to entropic optimal transport. *Lecture notes, Columbia University*, 2021.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Rout, L., Korotin, A., and Burnaev, E. Generative modeling with optimal transport maps. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=5JdLZg346Lw.

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

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. Large scale optimal transport and mapping estimation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1zlp1bRW.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pp. 2263–2291, 2013.

Singh, S. P. and Jaggi, M. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 33:22045–22055, 2020.

Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

Srivastava, S., Li, C., and Dunson, D. B. Scalable bayes via barycenter in wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer New York, NY, 2009. doi: https://doi.org/10.1007/b13794.

A. Proofs

In this section, we give the proofs for our Theorems in §4. In our calculations and formulas below, we sometimes substitute the expectations written as $\mathbb{E}_{\cdot\sim\mu}(*)$ with their integral representations $\int(*)d\mu(\cdot)$. This is done for ease of comprehension. In the last subsection (§A.3), we give a high-level conceptual derivation of our bi-level objective (11).

A.1. Proof of Theorem 4.1

Proof. Before jumping into the proof, observe that any coupling in $\pi \in \Pi(\mathbb{P}_k)$ can be written, by the disintegration theorem, as $\pi(dx, dy) = \mathbb{P}_k(dx)\, K(x; dy)$, where $K : \mathcal{X}_k \to \mathcal{P}(\mathcal{Y})$ is a measurable function. Therefore, we have

$$\begin{aligned}\inf_{\pi_k\in\Pi(\mathbb{P}_k)} \int_{\mathcal{X}_k\times\mathcal{Y}} \left\{C_k(x_k, \pi_k(\cdot|x_k)) - f_k(y)\right\} d\pi_k(x_k, y) &= \inf_{\substack{K:\mathcal{X}_k\to\mathcal{P}(\mathcal{Y})\\ \text{measurable}}} \int_{\mathcal{X}_k} \left\{C_k(x_k, K(x_k)) - \int_{\mathcal{Y}} f_k(y)\, K(x_k; dy)\right\} d\mathbb{P}_k(x_k)\\ &= \int_{\mathcal{X}_k} \inf_{\mu\in\mathcal{P}(\mathcal{Y})} \left\{C_k(x_k, \mu) - \int f_k(y)\, d\mu(y)\right\} d\mathbb{P}_k(x_k) \qquad (16)\\ &= \mathbb{E}_{x_k\sim\mathbb{P}_k} f_k^{C_k}(x_k),\end{aligned}$$

where the second equality is justified by the following reasoning: the map $(x_k, \mu) \mapsto C(x_k, \mu) - \int f_k(y)\, d\mu(y)$ is measurable and bounded from below, thus, we can invoke (Bertsekas & Shreve, 1996, Proposition 7.27 and Proposition 7.50) to find, for every $\epsilon > 0$, a measurable function $K^\epsilon : \mathcal{X}_k \to \mathcal{P}(\mathcal{Y})$ such that

$$C_k(x_k, K^\epsilon(x_k)) - \int_{\mathcal{Y}} f_k(y)\, dK^\epsilon(x_k; dy) \le \inf_{\mu\in\mathcal{P}(\mathcal{Y})} C_k(x_k, \mu) - \int_{\mathcal{Y}} f_k(y)\, d\mu(y) + \epsilon = f_k^{C_k}(x_k) + \epsilon \quad \mathbb{P}_k\text{-a.e. } x_k.$$

From this inequality follows readily the equality in (16), as ϵ was arbitrary.

Next, we turn our attention towards the barycenter problem. Using the weak optimal transport duality (4), we have for every $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ that

$$\sum_{k=1}^K \lambda_k \mathrm{OT}_{C_k}(\mathbb{P}_k, \mathbb{Q}) = \sup_{f_{1:K}\in C(\mathcal{Y})^K} \sum_{k=1}^K \lambda_k \left(\mathbb{E}_{x_k\sim\mathbb{P}_k} f_k^{C_k}(x_k) + \mathbb{E}_{y\sim\mathbb{Q}} f_k(y)\right).$$

Notice that the map

$$\mathcal{P}(\mathcal{Y}) \times \mathcal{C}(\mathcal{Y})^K \ni (\mathbb{Q}, f_{1:K}) \mapsto \sum_{k=1}^K \lambda_k \left\{\mathbb{E}_{x_k\sim\mathbb{P}_k} f_k^{C_k}(x_k) + \mathbb{E}_{y\sim\mathbb{Q}} f_k(y)\right\}$$

is continuous and convex in $\mathbb{Q}$ and, thanks to (Kolesov et al., 2023, Proposition 1), concave in $f_{1:K}$. As $\mathcal{Y}$ is compact, the same holds true for $\mathcal{P}(\mathcal{Y})$, hence, we can apply Sion's minimax theorem (Sion, 1958, Theorem 3.4), which permits us to swap $\inf$ with $\sup$ and get

$$\begin{aligned}\mathcal{L}^* &= \inf_{\mathbb{Q}\in\mathcal{P}(\mathcal{Y})} \sup_{f_{1:K}\in\mathcal{C}(\mathcal{Y})^K} \sum_{k=1}^K \lambda_k \left(\mathbb{E}_{x_k\sim\mathbb{P}_k} f_k^{C_k}(x_k) + \mathbb{E}_{y\sim\mathbb{Q}} f_k(y)\right)\\ &= \sup_{f_{1:K}\in\mathcal{C}(\mathcal{Y})^K} \inf_{\mathbb{Q}\in\mathcal{P}(\mathcal{Y})} \sum_{k=1}^K \lambda_k \left(\mathbb{E}_{x_k\sim\mathbb{P}_k} f_k^{C_k}(x_k) + \mathbb{E}_{y\sim\mathbb{Q}} f_k(y)\right)\\ &= \sup_{f_{1:K}\in\mathcal{C}(\mathcal{Y})^K} \left\{m(f_{1:K}) + \sum_{k=1}^K \lambda_k \mathbb{E}_{x_k\sim\mathbb{P}_k} f_k^{C_k}(x_k)\right\},\end{aligned}$$

where $m(f_{1:K}) \stackrel{\text{def}}{=} \inf_{y\in\mathcal{Y}} \left\{\sum_{k=1}^K \lambda_k f_k(y)\right\}$. Since $(f_1 - a)^{C_1} = f_1^{C_1} + a$ for all $a \in \mathbb{R}$, cf. (Kolesov et al., 2023, Proposition 1 (ii)), we can translate f_1 by $-\frac{m(f_{1:K})}{\lambda_1}$ without changing the value. Consequently,

$$\mathcal{L}^* = \sup_{\substack{f_{1:K}\in\mathcal{C}(\mathcal{Y})^K\\ m(f_{1:K})=0}} \sum_{k=1}^K \lambda_k \mathbb{E}_{x_k\sim\mathbb{P}_k} f_k^{C_k}(x_k).$$

Pick a vector $f_{1:K} \in \mathcal{C}(\mathcal{Y})^K$ with $m(f_{1:K}) = 0$. Writing $\tilde{f}_1 := -\sum_{k=2}^K \frac{\lambda_k}{\lambda_1} f_k$, we have

$$0 = m(f_{1:K}) \leq \sum_{k=1}^K \lambda_k f_k = f_1 - \tilde{f}_1.$$

Due to monotonicity of the C-transform (cf. (Kolesov et al., 2023, Proposition 1 (i))), we find that $f_1^{C_1} \leq \tilde{f}_1^{C_1}$. Thus, by replacing $(f_1, \dots, f_K)$ with $(\tilde{f}_1, f_2, \dots, f_K) \in \mathcal{C}(\mathcal{Y})^K$, we improve the value. We obtain

$$\begin{aligned}
\mathcal{L}^* &= \sup_{\substack{f_{1:K} \in \mathcal{C}(\mathcal{Y})^K, \\ \sum_{k=1}^K \lambda_k f_k = 0}} \sum_{k=1}^K \lambda_k \mathbb{E}_{x_k \sim \mathbb{P}_k} f_k^{C_k}(x_k) \\
&= \sup_{\substack{f_{1:K} \in \mathcal{C}(\mathcal{Y})^K, \\ \sum_{k=1}^K \lambda_k f_k = 0}} \sum_{k=1}^K \lambda_k \left\{ \inf_{\pi_k \in \Pi(\mathbb{P}_k)} \int_{\mathcal{Y}} \left[C_k(x_k, \pi_k(\cdot|x_k)) - f_k(y) \right] d\pi_k(x_k, y) \right\} \\
&= \sup_{\substack{f_{1:K} \in \mathcal{C}(\mathcal{Y})^K, \\ \sum_{k=1}^K \lambda_k f_k = 0}} \inf_{\substack{\pi_k \in \Pi(\mathbb{P}_k), \\ (k \in \overline{K})}} \mathcal{V}(f_{1:K}, \pi_{1:K}),
\end{aligned}$$

where the second equality follows from (16), and conclude the proof. □

A.2. Proof of Theorem 4.2

Proof. We subsequently prove our statements for classical (6), ϵ-KL (7) and γ-Energy cost functions. Our analysis of specific cost cases will follow a generic template explained below. For a potential $f_k \in \mathcal{C}(\mathcal{Y})$ and a plan $\pi_k \in \Pi(\mathbb{P}_k)$ we define:

$$\mathcal{V}_k(f_k, \pi_k) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{x_k \sim \mathbb{P}_k} C_k\big(x_k, \pi_k(\cdot|x_k)\big) - \mathop{\mathbb{E}}_{x_k \sim \mathbb{P}_k} \Big(\mathop{\mathbb{E}}_{y \sim \pi_k(\cdot|x_k)} f_k(y) \Big). \tag{17}$$

In what follows, for given congruent potentials $\widehat{f}_{1:K}$ and plans $\widehat{\pi}_{1:K}$, we express (9) and (10) with help of (17). The functional $\mathcal{V}$, cf. (9), can be expressed by the following weighted sum:

$$\mathcal{V}(\widehat{f}_{1:K}, \widehat{\pi}_{1:K}) = \sum_{k=1}^K \lambda_k \mathcal{V}_k(\widehat{f}_k, \widehat{\pi}_k). \tag{18}$$

In turn, the inner minimization problem (10) can be split into separate optimization problems over plans $\pi_k \in \Pi(\mathbb{P}_k)$:

$$\mathcal{L}_k(\widehat{f}_k) \stackrel{\text{def}}{=} \inf_{\pi_k \in \Pi(\mathbb{P}_k)} \mathcal{V}_k(\widehat{f}_k, \pi_k). \tag{19}$$

When considering the specific cases for different cost functions, we will manage to explicitly or implicitly recover the minimizers of functionals $\mathcal{V}_k$:

$$\pi_k^f \in \mathop{\arg\inf}_{\pi_k \in \Pi(\mathbb{P}_k)} \mathcal{V}_k(\widehat{f}_k, \pi_k).$$

The arguments regarding the existence (and possible uniqueness) of plans π_k^f will be given correspondingly. Note that it is better to write $\pi_k^{\widehat{f}_k}$ instead of π_k^f. Nevertheless, abusing the notations, we will use exactly π_k^f in the text.

Given $\pi_k^f \in \Pi(\mathbb{P}_k)$, equation for functional $\mathcal{L}$ (10) can be written as follows:

$$\mathcal{L}(\widehat{f}_{1:K}) = \sum_{k=1}^K \lambda_k \mathcal{L}_k(\widehat{f}_k) = \sum_{k=1}^K \lambda_k \mathcal{V}_k(\widehat{f}_k, \pi_k^f). \tag{20}$$

The difference between (18) and (20) is exactly the first gap δ_1 (12):

$$\delta_1 = \mathcal{V}(\widehat{f}_{1:K}, \widehat{\pi}_{1:K}) - \mathcal{L}(\widehat{f}_{1:K}) = \sum_{k=1}^{K} \lambda_k \Big\{ \mathcal{V}_k(\widehat{f}_k, \widehat{\pi}_k) - \mathcal{V}_k(\widehat{f}_k, \pi_k^f) \Big\} = \sum_{k=1}^{K} \lambda_k \delta_{1,k}(\widehat{f}_k, \widehat{\pi}_k)\,;$$
$$\delta_{1,k}(\widehat{f}_k, \widehat{\pi}_k) \stackrel{\text{def}}{=} \mathcal{V}_k(\widehat{f}_k, \widehat{\pi}_k) - \mathcal{V}_k(\widehat{f}_k, \pi_k^f). \tag{21}$$

The detailed analysis of quantities $\delta_{1,k}(\widehat{f}_k, \widehat{\pi}_k)$ will be provided for each particular cost function case.

Now we move on the the analysis of the second gap δ_2 (13) and derive the similar factorization for this quantity. Recall that the optimal value $\mathcal{L}^*$ of OT barycenter problem (5) is:

$$\mathcal{L}^* = \sum_{k=1}^{K} \lambda_k \text{OT}_{C_k}(\mathbb{P}_k, \mathbb{Q}^*) = \sum_{k=1}^{K} \lambda_k \int\limits_{\mathcal{X}_k} C_k(x_k, \pi_k^*(\cdot|x_k)) d\mathbb{P}_k(x_k), \tag{22}$$

where $\mathbb{Q}^*$ is the barycenter distribution and $\pi_k^* \in \Pi(\mathbb{P}_k, \mathbb{Q}^*)$, $k \in \overline{K}$, are (weak) OT plans between $\mathbb{P}_k$ and $\mathbb{Q}^*$. Note that the second marginal distribution of π_k^* for each k is $\mathbb{Q}^*$. Thanks to the congruence condition for potentials $\widehat{f}_{1:K}$, we have:

$$\sum_{k=1}^{K} \lambda_k \int\limits_{\mathcal{X}_k} \int\limits_{\mathcal{Y}} \widehat{f}_k(y) d\pi_k^*(y|x_k) d\mathbb{P}_k(x_k) = \sum_{k=1}^{K} \lambda_k \int\limits_{\mathcal{Y}} \widehat{f}_k(y) \underbrace{d(\pi_k^*)^{\mathcal{Y}}(y)}_{=d\mathbb{Q}^*(y)} = \int\limits_{\mathcal{Y}} \underbrace{\Big\{ \sum_{k=1}^{K} \lambda_k \widehat{f}_k(y) \Big\}}_{=0} d\mathbb{Q}^*(y) = 0. \tag{23}$$

The combination of (22) and (23) yields the following:

$$\mathcal{L}^* = \sum_{k=1}^{K} \lambda_k \int\limits_{\mathcal{X}_k} C_k(x_k, \pi_k^*(\cdot|x_k)) d\mathbb{P}_k(x_k) - \underbrace{\sum_{k=1}^{K} \lambda_k \int\limits_{\mathcal{X}_k} \int\limits_{\mathcal{Y}} \widehat{f}_k(y) d\pi_k^*(y|x_k) d\mathbb{P}_k(x_k)}_{=0 \text{ from (23)}}$$
$$= \sum_{k=1}^{K} \lambda_k \Big\{ \int\limits_{\mathcal{X}_k} C_k\big(x_k, \pi_k^*(\cdot|x_k)\big) - \int\limits_{\mathcal{X}_k} \int\limits_{\mathcal{Y}} \widehat{f}_k(y) d\pi_k^*(y|x_k) d\mathbb{P}_k(x_k) \Big\} = \sum_{k=1}^{K} \lambda_k \mathcal{V}_k(\widehat{f}_k, \pi_k^*).$$

By summarizing the expression for $\mathcal{L}^*$ and (20) we derive the factorization for δ_2:

$$\delta_2 = \mathcal{L}^* - \mathcal{L}(\widehat{f}_{1:K}) = \sum_{k=1}^{K} \lambda_k \Big\{ \mathcal{V}_k(\widehat{f}_k, \pi_k^*) - \mathcal{V}_k(\widehat{f}_k, \pi_k^f) \Big\} = \sum_{k=1}^{K} \lambda_k \delta_{2,k}(\widehat{f}_k);$$
$$\delta_{2,k}(\widehat{f}_k) = \mathcal{V}_k(\widehat{f}_k, \pi_k^*) - \mathcal{V}_k(\widehat{f}_k, \pi_k^f). \tag{24}$$

Now we are ready to consider the specific cost functions. Following the outline above, for each case we will:

- Establish the existence of plans π_k^f which optimize functionals $\mathcal{L}_k$. Derive some properties of these plans needed for further analysis.
- Analyze the gaps $\delta_{1,k} = \mathcal{V}_k(\widehat{f}_k, \widehat{\pi}_k) - \mathcal{V}_k(\widehat{f}_k, \pi_k^f)$ and $\delta_{2,k} = \mathcal{V}_k(\widehat{f}_k, \pi_k^*) - \mathcal{V}_k(\widehat{f}_k, \pi_k^f)$. Using $\delta_{1,k}$ and $\delta_{2,k}$, upper bound some discrepancies between $\widehat{\pi}_k, \pi_k^f$ and π_k^*, π_k^f.
- Aggregate the obtained upper bounds for each $k \in \overline{K}$. Derive the ultimate bound utilizing the identity:

$$\delta_1 + \delta_2 = \sum_{k=1}^{K} \lambda_k (\delta_{1,k} + \delta_{2,k}).$$

Proof of statement 1 (classical cost function). We consider cost functions C_k, $k \in \overline{K}$ given by

$$C_k(x_k, \mu) = \int_{\mathcal{Y}} c_k(x_k, y) d\mu(y), \tag{25}$$

where $x_k \in \mathcal{X}_k$, $\mu \in \mathcal{P}(\mathcal{Y})$. Here, the functional $\mathcal{V}_k$, cf. (17), takes the following form

$$\mathcal{V}_k(\widehat{f}_k, \pi_k) = \int_{\mathcal{X}_k} \int_{\mathcal{Y}} \{c_k(x_k, y) - \widehat{f}_k(y)\} d\pi_k(y|x_k) d\mathbb{P}_k(x_k). \tag{26}$$

For convenience, we introduce the function $g_k(x_k, y) \stackrel{\text{def}}{=} c_k(x_k, y) - \widehat{f}_k(y)$, $x_k \in \mathcal{X}_k$, $y \in \mathcal{Y}$. Note that g_k is β-strongly convex in the second argument by the Theorem assumption. Define the maps $T_k^f : \mathcal{X}_k \to \mathcal{Y}$ as follows:

$$T_k^f(x_k) \stackrel{\text{def}}{=} \arg\inf_{y \in \mathcal{Y}} g_k(x_k, y).$$

Note that the $\arg\inf$ above exists and is unique since $\mathcal{Y}$ is compact and g_k is strongly convex in y. Furthermore, since $x \mapsto \inf_{y \in \mathcal{Y}} g_k(x, y)$ and g_k are both continuous, we have that the graph

$$\text{graph}(T_k^f) = \left\{ (x_k, y) \in \mathcal{X}_k \times \mathcal{Y} : g_k(x_k, y) = \inf_{\tilde{y} \in \mathcal{Y}} g_k(x_k, \tilde{y}) \right\}$$

is a closed subset of $\mathcal{X}_k \times \mathcal{Y}$ and therefore Borel. The latter is by (Kechris, 2012, Theorem 4.12) equivalent to T_k^f being Borel. Consequently, it induces a transport plan π_k^f via

$$(\text{Id}, T_k^f)_{\#} \mathbb{P}_k \stackrel{\text{def}}{=} \pi_k^f \in \Pi(\mathbb{P}_k).$$

We continue to study the functional $\mathcal{V}_k$ as defined in (26). For any $\pi_k \in \Pi(\mathbb{P}_k)$ we have

$$\begin{aligned}
\mathcal{V}_k(\widehat{f}_k, \pi_k) &= \int_{\mathcal{X}_k} \int_{\mathcal{Y}} g_k(x_k, y) d\pi_k(y|x_k) d\mathbb{P}_k(x_k) \geq \int_{\mathcal{X}_k} g_k(x_k, T_k^f(x_k)) d\mathbb{P}_k(x_k) \\
&= \int_{\mathcal{X}_k} \int_{\mathcal{Y}} g_k(x_k, y) d\pi_k^f(y|x_k) d\mathbb{P}_k(x_k) = \mathcal{V}_k(\widehat{f}_k, \pi_k^f),
\end{aligned}$$

i.e., π_k^f indeed minimizes (26) over $\Pi(\mathbb{P}_k)$ and solves (19) for classical cost functions of the form (25).

Now we take a closer look at the gaps $\delta_{1,k}$ defined in (21). For $k \in \overline{K}$ we compute

$$\begin{aligned}
\delta_{1,k} &= \mathcal{V}_k(\widehat{f}_k, \widehat{\pi}_k) - \mathcal{V}_k(\widehat{f}_k, \pi_k^f) = \int_{\mathcal{X}_k} \left\{ \int_{\mathcal{Y}} g_k(x_k, y) d\widehat{\pi}_k(y|x_k) - g_k\big(x_k, T_k^f(x_k)\big) \right\} d\mathbb{P}_k(x_k) \\
&= \int_{\mathcal{X}_k} \int_{\mathcal{Y}} \left\{ g_k(x_k, y) - g_k\big(x_k, T_k^f(x_k)\big) \right\} d\widehat{\pi}_k(y|x_k) d\mathbb{P}_k(x_k) \qquad (27) \\
&\geq \frac{\beta}{2} \int_{\mathcal{X}_k} \int_{\mathcal{Y}} \|y - T_k^f(x_k)\|^2 \underbrace{d\widehat{\pi}_k(y|x_k) d\mathbb{P}_k(x_k)}_{= d\widehat{\pi}_k(x_k, y)} = \frac{\beta}{2} \int_{\mathcal{X}_k \times \mathcal{Y}} \|y - T_k^f(x_k)\|^2 d\widehat{\pi}_k(x_k, y). \qquad (28)
\end{aligned}$$

To transition from (27) to (28) we utilize the inequality $\frac{\beta}{2}\|T_k^f(x_k) - y\|^2 \leq g_k(x_k, y) - g_k\big(x_k, T_k^f(x_k)\big)$, which directly follows from β-strong convexity (recall that $T_k^f(x_k)$ is the minimizer of $y \mapsto g_k(x_k, y)$).

Concerning the gaps $\delta_{2,k}$, cf. (24), we can conduct the same analysis as for $\delta_{1,k}$ (eqs. (27), (28)) and derive:

$$\delta_{2,k} \geq \frac{\beta}{2} \int_{\mathcal{X}_k \times \mathcal{Y}} \|y - T_k^f(x_k)\|^2 d\pi_k^*(x_k, y). \tag{29}$$

We are left to summarize inequalities for $\delta_{1,k}$ (28) and $\delta_{2,k}$ (29):

$$\delta_{1,k} + \delta_{2,k} \geq \frac{\beta}{2} \left\{ \int_{\mathcal{X}_k \times \mathcal{Y}} \|y - T_k^f(x_k)\|^2 \underbrace{d\widehat{\pi}_k(x_k, y)}_{= d\widehat{\pi}_k(y|x_k) d\mathbb{P}_k(x_k)} + \int_{\mathcal{X}_k \times \mathcal{Y}} \|y - T_k^f(x_k)\|^2 \underbrace{d\pi_k^*(x_k, y)}_{d\pi_k^*(y|x_k) d\mathbb{P}_k(x_k)} \right\} =$$

$$\frac{\beta}{2}\int_{\mathcal{X}_k}\Big\{\int_{\mathcal{Y}}\|y-T_k^f(x_k)\|^2 d\widehat{\pi}_k(y|x_k)+\int_{\mathcal{Y}}\|y'-T_k^f(x_k)\|^2 d\pi_k^*(y'|x_k)\Big\}d\mathbb{P}_k(x_k)=$$

$$\frac{\beta}{2}\int_{\mathcal{X}_k}\Big\{\int_{\mathcal{Y}\times\mathcal{Y}}\|y-T_k^f(x_k)\|^2 d\widehat{\pi}_k(y|x_k)\underbrace{d\pi_k^*(y'|x_k)}_{\text{integrates to 1}}+\int_{\mathcal{Y}\times\mathcal{Y}}\|y'-T_k^f(x_k)\|^2\underbrace{d\widehat{\pi}_k(y|x_k)}_{\text{integrates to 1}}d\pi_k^*(y'|x_k)\Big\}d\mathbb{P}_k(x_k)=$$

$$\frac{\beta}{2}\int_{\mathcal{X}_k}\Big\{\int_{\mathcal{Y}\times\mathcal{Y}}\big(\|y-T_k^f(x_k)\|^2+\|y'-T_k^f(x_k)\|^2\big)d\widehat{\pi}_k(y|x_k)d\pi_k^*(y'|x_k)\Big\}d\mathbb{P}_k(x_k)\geq \tag{30}$$

$$\frac{\beta}{2}\int_{\mathcal{X}_k}\Big\{\int_{\mathcal{Y}\times\mathcal{Y}}\frac{\|y-y'\|^2}{2}d\widehat{\pi}_k(y|x_k)d\pi_k^*(y'|x_k)\Big\}d\mathbb{P}_k(x_k)\geq \tag{31}$$

$$\frac{\beta}{2}\int_{\mathcal{X}_k}\Big\{\inf_{\gamma_{x_k}\in\Pi\left(\widehat{\pi}_k(\cdot|x_k),\pi_k^*(\cdot|x_k)\right)}\int_{\mathcal{Y}\times\mathcal{Y}}\frac{\|y-y'\|^2}{2}d\gamma_{x_k}(y,y')\Big\}d\mathbb{P}_k(x_k)=\frac{\beta}{2}\int_{\mathcal{X}_k}\mathbb{W}_2^2\big(\widehat{\pi}_k(\cdot|x_k),\pi_k^*(\cdot|x_k)\big)d\mathbb{P}_k(x_k).$$

The transition from (30) to (31) follows from simple consideration:

$$\|y-T_k^f(x_k)\|^2+\|y'-T_k^f(x_k)\|^2\geq 1/2\big(\|y-T_k^f(x_k)\|+\|T_k^f(x_k)-y'\|\big)^2\overset{\text{Triang. ineq.}}{\geq} 1/2\|y-y'\|^2.$$

Aggregating the inequalities for $\delta_{1_k}+\delta_{2,k}$ with weights λ_k, $k\in\overline{K}$ finishes the proof of statement 1:

$$\delta_1+\delta_2=\sum_{k=1}^K\lambda_k(\delta_{1,k}+\delta_{2,k})\geq\frac{\beta}{2}\sum_{k=1}^K\lambda_k\int_{\mathcal{X}_k}\mathbb{W}_2^2\big(\widehat{\pi}_k(\cdot|x),\pi_k^*(\cdot|x)\big)d\mathbb{P}_k(x).$$

Proof of statement 2 (ϵ-KL weak cost function). Recall that the ϵ-KL cost functions are given by

$$C_k(x_k,\mu)=\int_{\mathcal{Y}}c_k(x_k,y)d\mu(y)+\epsilon\text{KL}(\mu\|\mu_0)=\int_{\mathcal{Y}}c_k(x_k,y)d\mu(y)-\epsilon H(\mu)-\epsilon\int_{\mathcal{Y}}\log\frac{\mathrm{d}\mu_0(y)}{\mathrm{d}y}d\mu(y).$$

The term $\frac{\mathrm{d}\mu_0(y)}{\mathrm{d}y}$ denotes the density function of the distribution μ_0 w.r.t. the Lebesgue measure at point y. The functional $\mathcal{V}_k$ (17) takes for a given potential $\widehat{f}_k$ and an arbitrary $\pi_k\in\Pi(\mathbb{P}_k)$ the form:

$$\mathcal{V}_k(\widehat{f}_k,\pi_k)=$$

$$\int_{\mathcal{X}_k}\Big\{\int_{\mathcal{Y}}c_k(x_k,y)d\pi_k(y|x_k)-\epsilon H\big(\pi_k(\cdot|x_k)\big)-\epsilon\int_{\mathcal{Y}}\log\frac{\mathrm{d}\mu_0(y)}{\mathrm{d}y}d\pi_k(y|x_k)\Big\}d\mathbb{P}_k(x_k)-\int_{\mathcal{X}_k}\int_{\mathcal{Y}}\widehat{f}_k(y)d\pi_k(y|x_k)d\mathbb{P}_k(x_k)=$$

$$\int_{\mathcal{X}_k}\Big\{\int_{\mathcal{Y}}c_k(x_k,y)d\pi_k(y|x_k)-\epsilon H\big(\pi_k(\cdot|x_k)\big)-\int_{\mathcal{Y}}\underbrace{\big[\epsilon\log\frac{\mathrm{d}\mu_0(y)}{\mathrm{d}y}+\widehat{f}_k(y)\big]}_{\overset{\text{def}}{=}\widetilde{f}_k(y)}d\pi_k(y|x_k)\Big\}d\mathbb{P}_k(x_k)=$$

$$\int_{\mathcal{X}_k}\Big\{\int_{\mathcal{Y}}c_k(x_k,y)d\pi_k(y|x_k)-\epsilon H\big(\pi_k(\cdot|x_k)\big)-\int_{\mathcal{Y}}\widetilde{f}_k(y)d\pi_k(y|x_k)\Big\}d\mathbb{P}_k(x_k). \tag{32}$$

The expression under curly brackets in (32) appeared in previous works, see (Mokrov et al., 2024, Equation 8). Following (Mokrov et al., 2024), we introduce $\mathcal{G}_{x_k,\widetilde{f}_k}:\mathcal{P}(\mathcal{Y})\to\mathbb{R}$ as follows:

$$\mathcal{G}_{x_k,\widetilde{f}_k}(\mu)\overset{\text{def}}{=}\int_{\mathcal{Y}}c_k(x_k,y)d\mu(y)-\epsilon H(\mu)-\int_{\mathcal{Y}}\widetilde{f}_k(y)d\mu(y). \tag{33}$$

The functional (33) is minimized by distribution $\mu_{x_k}^{\widetilde{f}_k}\in\mathcal{P}(\mathcal{Y})$ with density (Mokrov et al., 2024, Theorem 1):

$$\frac{\mathrm{d}\mu_{x_k}^{\widetilde{f}_k}(y)}{\mathrm{d}y}=\frac{1}{Z(\widetilde{f}_k,x_k)}\exp\left(\frac{\widetilde{f}_k(y)-c_k(x_k,y)}{\epsilon}\right)=\frac{1}{Z(\widetilde{f}_k,x_k)}\frac{\mathrm{d}\mu_0(y)}{\mathrm{d}y}\exp\left(\frac{\widehat{f}_k(y)-c_k(x_k,y)}{\epsilon}\right),$$

where $Z(\widetilde{f}_k, x_k) = \int_{\mathcal{Y}} \exp\big(\frac{\widetilde{f}_k(y) - c_k(x_k, y)}{\epsilon}\big) dy = \int_{\mathcal{Y}} \exp\big(\frac{\widehat{f}_k(y) - c_k(x_k, y)}{\epsilon}\big) d\mu_0(y)$ is the normalizing constant (a.k.a. partition function). Define the plan:

$$d\pi_k^f(x_k, y) \stackrel{\text{def}}{=} d\mathbb{P}_k(x_k) d\mu_{x_k}^{\widetilde{f}_k}(y).$$

Note that $\pi_k^f(\cdot|x_k) = \mu_{x_k}^{\widetilde{f}_k}$. For arbitrary $\pi_k \in \Pi(\mathbb{P}_k)$ we have:

$$\mathcal{V}_k(\widehat{f}_k, \pi_k) = \int_{\mathcal{X}_k} \underbrace{\mathcal{G}_{x_k, \widetilde{f}_k}(\pi_k(\cdot|x_k))}_{\geq \mathcal{G}_{x_k, \widetilde{f}_k}(\mu_{x_k}^{\widetilde{f}_k})} d\mathbb{P}_k(x_k) \geq \int_{\mathcal{X}_k} \mathcal{G}_{x_k, \widetilde{f}_k}\big(\mu_{x_k}^{\widetilde{f}_k}\big) d\mathbb{P}_k(x_k) = \mathcal{V}_k(\widehat{f}_k, \pi_k^f),$$

i.e., distribution π_k^f indeed minimizes $\mathcal{V}_k$ for ϵ-KL weak cost. Similar to (Mokrov et al., 2024, Equation 14) we derive

$$\mathcal{V}_k(\widehat{f}_k, \pi_k^f) = \int_{\mathcal{X}_k} \mathcal{G}_{x_k, \widetilde{f}_k}\big(\pi_k^f(\cdot|x_k)\big) d\mathbb{P}_k(x_k) = -\epsilon \int_{\mathcal{X}_k} \log Z(\widetilde{f}_k, x_k) \mathbb{P}_k(x_k).$$

Now we are ready to analyze the gaps $\delta_{1,k}$ (21) and $\delta_{2,k}$ (24). Note that our further derivations are similar to the proof of (Mokrov et al., 2024, Theorem 2).

$$\delta_{1,k} = \mathcal{V}_k(\widehat{f}_k, \widehat{\pi}_k) - \mathcal{V}_k(\widehat{f}_k, \pi_k^f) =$$

$$\int_{\mathcal{X}_k} \Big\{ \int_{\mathcal{Y}} c_k(x_k, y) d\widehat{\pi}_k(y|x_k) - \epsilon H\big(\widehat{\pi}_k(\cdot|x_k)\big) - \int_{\mathcal{Y}} \widetilde{f}_k(y) d\widehat{\pi}_k(y|x_k) \Big\} d\mathbb{P}_k(x_k) + \epsilon \int_{\mathcal{X}_k} \log Z(\widetilde{f}_k, x_k) d\mathbb{P}_k(x_k) =$$

$$-\epsilon \int_{\mathcal{X}_k} \int_{\mathcal{Y}} \underbrace{\frac{\widetilde{f}_k(y) - c_k(x_k, y)}{\epsilon}}_{= \log \frac{\mathrm{d}\pi_k^f(y|x_k)}{\mathrm{d}y} + \log Z(\widetilde{f}_k, x_k)} d\widehat{\pi}_k(y|x_k) d\mathbb{P}_k(x_k) + \epsilon \int_{\mathcal{X}_k} \log Z(\widetilde{f}_k, x_k) d\mathbb{P}_k(x_k) - \epsilon \int_{\mathcal{X}_k} H(\widehat{\pi}_k(\cdot|x_k)) d\mathbb{P}_k(x_k) = \tag{34}$$

$$-\epsilon \int_{\mathcal{X}_k} \int_{\mathcal{Y}} \log \frac{\mathrm{d}\pi_k^f(y|x_k)}{\mathrm{d}y} d\widehat{\pi}_k(y|x_k) d\mathbb{P}_k(x_k) + \epsilon \int_{\mathcal{X}_k} \int_{\mathcal{Y}} \log \frac{\mathrm{d}\widehat{\pi}_k(y|x_k)}{\mathrm{d}y} d\widehat{\pi}_k(y|x_k) d\mathbb{P}_k(x_k) = \tag{35}$$

$$\epsilon \int_{\mathcal{X}_k} \int_{\mathcal{Y}} \Big[\log \frac{\mathrm{d}\widehat{\pi}_k(y|x_k)}{\mathrm{d}y} - \log \frac{\mathrm{d}\pi_k^f(y|x_k)}{\mathrm{d}y} \Big] d\widehat{\pi}_k(y|x_k) d\mathbb{P}_k(x_k) = \epsilon \int_{\mathcal{X}_k} \mathrm{KL}\big(\widehat{\pi}_k(\cdot|x_k) \| \pi_k^f(\cdot|x_k)\big) d\mathbb{P}_k(x_k). \tag{36}$$

In (35) and (36) we implicitly assume that $\widehat{\pi}_k(\cdot|x_k)$ is absolutely continuous. We remark that if this is not the case, then the derivations above still holds true, since in this case both negative entropy in (34) and KL divergence in (36) are equal to $+\infty$. Thanks to the fact that $\widehat{\pi}_k, \pi_k^f \in \Pi(\mathbb{P}_k)$, i.e., the first marginal of these distributions equals to $\mathbb{P}_k$, (36) can be further rewritten, see (Mokrov et al., 2024, Appendix B.2):

$$\delta_{1,k} = \epsilon \int_{\mathcal{X}_k} \mathrm{KL}\big(\widehat{\pi}_k(\cdot|x_k) \| \pi_k^f(\cdot|x_k)\big) d\mathbb{P}_k(x_k) = \epsilon \mathrm{KL}(\widehat{\pi}_k \| \pi_k^f). \tag{37}$$

Regarding the gaps $\delta_{2,k}$ (24), a similar analysis as for $\delta_{1,k}$ yields:

$$\delta_{2,k} = \mathcal{V}_k(\widehat{f}_k, \pi_k^*) - \mathcal{V}_k(\widehat{f}_k, \pi_k^f) = \epsilon \mathrm{KL}(\pi_k^* \| \pi_k^f). \tag{38}$$

To derive the final bounds, we recall *Pinsker's* inequality (Tsybakov, 2009, Lemma 2.5). Given distributions π^a and π^b, we have:

$$2\rho_{\mathrm{TV}}(\pi^a, \pi^b)^2 \leq \mathrm{KL}(\pi^a \| \pi^b),$$

where ρ_{TV} is the total variation distance, see (Tsybakov, 2009, Definition 2.4). Using Pinsker's inequality, we find:

$$\delta_1 + \delta_2 = \sum_{k=1}^K \lambda_k (\delta_{1,k} + \delta_{2,k}) \stackrel{\text{Eqs. (37),(38)}}{=} \epsilon \sum_{k=1}^K \lambda_k \big(\mathrm{KL}(\widehat{\pi}_k \| \pi_k^f) + \mathrm{KL}(\pi_k^* \| \pi_k^f) \big)$$

$$\overset{\text{Pinsker's ineq.}}{\geq} \epsilon \sum_{k=1}^{K} \lambda_k \big(2\rho_{\text{TV}}(\widehat{\pi}_k, \pi_k^f)^2 + 2\rho_{\text{TV}}(\pi_k^*, \pi_k^f)^2\big) \geq \epsilon \sum_{k=1}^{K} \lambda_k \big(\rho_{\text{TV}}(\widehat{\pi}_k, \pi_k^f) + \rho_{\text{TV}}(\pi_k^*, \pi_k^f)\big)^2$$

$$\overset{\text{Triangle ineq.}}{\geq} \epsilon \sum_{k=1}^{K} \lambda_k \rho_{\text{TV}}(\widehat{\pi}_k, \pi_k^*)^2,$$

which finishes the proof of statement 2.

Proof of statement 3 (γ-Energy weak cost function case). Let $\pi_k \in \Pi(\mathbb{P}_k)$ be arbitrary and $\widehat{f}_k$ be a given potential. Below we take a closer look at functional $\mathcal{V}_k$ (17) for γ-Energy cost function:

$$\begin{aligned}\mathcal{V}_k(\widehat{f}_k, \pi_k) &= \int_{\mathcal{X}_k} \Big\{ \int_{\mathcal{Y}} c_k(x_k, y) d\pi_k(y|x_k) + \gamma \mathcal{E}_\ell^2\big(\pi_k(\cdot|x_k), \mu_0\big) \Big\} d\mathbb{P}_k(x_k) - \int_{\mathcal{X}_k} \int_{\mathcal{Y}} \widehat{f}_k(y) d\pi_k(y|x_k) d\mathbb{P}_k(x_k) \\ &= \int_{\mathcal{X}_k} \hat{C}_k(x_k, \pi_k(y|x_k)) d\mathbb{P}_k(x_k),\end{aligned}$$

where

$$\hat{C}_k(x_k, \mu) \overset{\text{def}}{=} \int_{\mathcal{Y}} \big(c_k(x_k, y) - \hat{f}_k(y)\big)\, d\mu(y) + \gamma \mathcal{E}_\ell^2\big(\mu, \mu_0\big).$$

To see the existence of a minimizer for $\mathcal{V}_k$, observe that $\Pi(\mathbb{P}_k)$ is compact w.r.t. weak convergence of measures. This follows easily from $\Pi(\mathbb{P}_k)$ being a closed subset of the compact $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ (where the latter is due to $\mathcal{X} \times \mathcal{Y}$ being compact). The function $\hat{C}_k$ is bounded from below (since $c_k, \hat{f}_k$ as well as $\mathcal{E}_\ell^2$ are bounded from below), continuous w.r.t. weak convergence of measures (since $c_k, \hat{f}_k$ and ℓ are continuous), and convex in the second argument (as the integral is linear in μ and $\mathcal{E}_\ell^2(\cdot, \mu_0)$ is convex). Hence, there exists by (Backhoff-Veraguas et al., 2019, Theorem 2.9) a minimizer $\pi_k^f \in \Pi(\mathbb{P}_k)$ for $\mathcal{V}_k(\widehat{f}_k, \cdot)$.

Similarly, we have

$$\mathcal{V}_k(\widehat{f}_k, \pi_k) = \int_{\mathcal{X}_k \times \mathcal{Y}} \big(c_k(x_k, y) - \hat{f}_k(y)\big)\, d\pi_k(x_k, y) + \gamma \rho_\ell^2(\pi_k, \pi_0),$$

where $\pi_0 \overset{\text{def}}{=} \mathbb{P}_k \otimes \mu_0$, from where it is evident that $\mathcal{V}_k$ is 2γ-strongly convex on $\Pi(\mathbb{P}_k)$ w.r.t. the metric ρ_ℓ (1). For the definition of strong convexity, see (Asadulaev et al., 2024, Appendix A, Definition 2).

Then the inequalities for the gaps $\delta_{1,k}$ (21) and $\delta_{2,k}$ (24) directly follow from strong convexity of $\mathcal{V}_k$, see (Asadulaev et al., 2024, Appendix A, Lemma 1):

$$\begin{aligned}\delta_{1,k} &= \mathcal{V}_k(\widehat{f}_k, \widehat{\pi}_k) - \mathcal{V}_k(\widehat{f}_k, \pi_k^f) \overset{\substack{\text{Lemma 1,} \\ \text{Asadulaev et. al.}}}{\geq} \gamma \rho_\ell(\pi_k^f, \widehat{\pi}_k)^2, \\ \delta_{2,k} &= \mathcal{V}_k(\widehat{f}_k, \pi_k^*) - \mathcal{V}_k(\widehat{f}_k, \pi_k^f) \geq \gamma \rho_\ell(\pi_k^f, \pi_k^*)^2.\end{aligned}$$

To finish the proof, we summarize the inequalities above over $k \in \overline{K}$ with weights λ_k:

$$\begin{aligned}\delta_1 + \delta_2 &= \sum_{k=1}^{K} \lambda_k (\delta_{1,k} + \delta_{2,k}) \geq \sum_{k=1}^{K} \lambda_k \big(\gamma \rho_\ell(\pi_k^f, \widehat{\pi}_k)^2 + \gamma \rho_\ell(\pi_k^f, \pi_k^*)^2\big) \\ &\geq \gamma \sum_{k=1}^{K} \lambda_k \frac{\big(\rho_\ell(\pi_k^f, \widehat{\pi}_k) + \rho_\ell(\pi_k^f, \pi_k^*)\big)^2}{2} \overset{\text{Triang. ineq.}}{\geq} \frac{\gamma}{2} \sum_{k=1}^{K} \lambda_k \rho_\ell(\widehat{\pi}_k, \pi_k^*)^2. \qquad \square\end{aligned}$$

A.3. Intuitive derivation of max-min OT barycenter objective (11)

In this subsection, we give an intuition behind our proposed bi-level objective (11). In particular, we explain how we managed to avoid min-max-min optimization and whence the congruence condition arises.

To derive our proposed formulation (11) we borrow some ideas from the existing literature:

1. The first idea stems from the NOT paper (Korotin et al., 2023b). The authors take advantage of the dual formulation of Weak OT problem (4) and come up with the following $\max\text{-}\min$ objective:

$$\text{OT}_C(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{C}(\mathcal{Y})} \inf_{\pi \in \Pi(\mathbb{P})} \left\{ \mathbb{E}_{x \sim \mathbb{P}} C(x, \pi(\cdot|x)) - \mathbb{E}_{y \sim \pi^{\mathcal{Y}}} f(y) + \mathbb{E}_{y \sim \mathbb{Q}} f(y) \right\}. \tag{NOT}$$

Recall that in the formula above $\pi^{\mathcal{Y}}$ is the second marginal of the plan π (projection to $\mathcal{Y}$). Note that the **direct extension** of the (NOT) objective to the barycenter problem **will result in** a $\min\text{-}\max\text{-}\min$ **problem**, because we will have to optimize not only w.r.t. to f_k and π_k, but also w.r.t. the barycenter distribution $\mathbb{Q}$. This is approximately how the method from WIN paper (Korotin et al., 2022) is constructed (it is one of the baselines we consider in our paper). In contrast, we proceed in a smarter way and combine the (NOT) objective with another brilliant idea which ultimately allows us to avoid $\min\text{-}\max\text{-}\min$:

2. **Congruent** potentials. Let $\mathbb{P}_{1:K}$ be the reference distributions for which we seek the barycenter. Let $\mathcal{H}_k(f, \pi, \mathbb{Q})$ be the functional under the $\sup_f \inf_\pi$ in (NOT), i.e.:

$$\text{OT}_{C_k}(\mathbb{P}_k, \mathbb{Q}) = \sup_{f \in \mathcal{C}(\mathcal{Y})} \inf_{\pi \in \Pi(\mathbb{P}_k)} \mathcal{H}_k(f, \pi, \mathbb{Q}).$$

The barycenter problem (5) with weights $\lambda_{1:K}$ then can be formulated as follows:

$$\mathcal{L}^* = \inf_{\mathbb{Q} \in \mathcal{P}(\mathcal{Y})} \sum_{k=1}^{K} \sup_{f_k \in \mathcal{C}(\mathcal{Y})} \inf_{\pi_k \in \Pi(\mathbb{P})} \lambda_k \mathcal{H}_k(f_k, \pi_k, \mathbb{Q}). \tag{NOTB NAIVE}$$

This is exactly what we called the *direct* $\min\text{-}\max\text{-}\min$ *extension of the* (NOT) *objective to the barycenter problem* in the paragraph above. So, the question is, how to avoid tri-level adversariality? We propose to exploit the **optimality condition** of the (NOTB NAIVE) objective w.r.t. the learned barycenter distribution $\mathbb{Q}$. In simpler words, we equate **the (Frechet) derivative** of the (NOTB NAIVE) objective w.r.t. $\mathbb{Q}$ to zero at the optimum, i.e., where $\mathbb{Q} = \mathbb{Q}^*$. In the mathematical literature such "derivative" is called first variation, see (Santambrogio, 2015, §7.2). Omitting some details, throwing away $\sup_f, \inf_\pi$ and being mathematically non-rigorous, the optimality condition reads as follows:

$$\nabla_{\mathbb{Q}} \left\{ \sum_{k=1}^{K} \lambda_k \mathbb{E}_{y \sim \mathbb{Q}} f_k^*(y) \right\} \Big|_{\mathbb{Q}=\mathbb{Q}^*} = 0, \tag{OPTIMALITY}$$

where f_k^* are the optimal potentials of the (NOT) objectives between $\mathbb{P}_k$ and $\mathbb{Q}^*$. Indeed, each expression of the (NOT) objective consists of three terms, and only the last term depends on $\mathbb{Q}$. When we take the derivative $\nabla_{\mathbb{Q}}$, the first two terms vanish. The analysis of the (OPTIMALITY) problem is already an easy task, since it is known that $\nabla_{\mathbb{Q}} \big[\mathbb{E}_{y \sim \mathbb{Q}} f(y)\big] = f$. Thanks to this observation, the (OPTIMALITY) reads as $\sum_{k=1}^{K} \lambda_k f_k = 0$, which is exactly the **congruence** condition that we utilize in our paper.

In fact, eliminating the outer $\inf_{\mathbb{Q}}$ optimization and incorporating the **congruence** condition into the (NOTB NAIVE) objective is exactly the method we propose in our paper. Of course, our explanation above is far from being mathematically rigorous. But we hope that our ideological derivation in this subsection sheds some light on the magic behind our approach and makes it more intuitive.

B. Extended experiments

B.1. Barycenters for Gaussians

In this experiment, we consider the OT barycenter problem with *Gaussian* input distributions $\mathbb{P}_k$. In the Gaussian case, it is known that the ground-truth OT barycenter for the classical OT cost functions $c_k(x_k, y) = 1/2\|x_k - y\|^2$ is also Gaussian and can be computed using the fixed point iteration procedure (Álvarez-Esteban et al., 2016). As the baselines, we take two recent solvers: EgBary (Kolesov et al., 2023) and WIN (Korotin et al., 2022). The authors of the first paper solve the Entropic OT (EOT) barycenter problem, see Table 1, and we consider their approach for small $\epsilon = 0.01$ to reduce the bias. The second paper learns unregularized OT barycenters with quadratic cost functions through an iterative procedure.

For assessment of our solver and WIN, we use the unexplained variance percentage metrics defined by $\mathcal{L}_2\text{-UVP}(\hat{T}) = 100 \cdot [\|\hat{T} - T^*\|_{\mathbb{P}}^2 / \text{var}(\mathbb{Q}^*)]\%$, where $\mathbb{Q}^*$ is the ground truth OT barycenter, see (Korotin et al., 2021a, §5.1). For evaluation

of EgBary which learns the EOT plans, we consider their barycentric projections, see (Kolesov et al., 2023, Appendix C4). Our results are presented in Table 3, the reported values of $\mathcal{L}_2$-UVP are averaged over input distributions $\mathbb{P}_k$ (w.r.t. barycenter weights λ_k).

Method/Dim	2	4	8	16	64
Ours	**0.01**	**0.02**	**0.04**	**0.04**	**0.08**
EgBary	0.02	0.05	0.06	0.09	0.84
WIN	0.03	0.08	0.13	0.25	0.75

Table 3: $\mathcal{L}_2$-UVP for our method, EgBary ($\epsilon = 0.01$) and WIN (Korotin et al., 2022), $D = 2, 4, 8, 16, 64$.

We see that our approach gives better results than its competitors for all the dimensions. The difference is especially visible for the biggest dimension $D = 64$. These results are expected, since EgBary approach is designed to learn the regularized barycenter, while WIN algorithm utilizes the iterative procedure leading to accumulation of the error with the increase of dimension.

B.2. Barycenters for MNIST digits ("0" and "1")

In this experiment, we seek for the barycenter of two image distributions: grayscaled handwritten digits "0" ($\mathbb{P}_1$) and grayscaled handwritten digits "1" ($\mathbb{P}_2$). We use images from the MNIST dataset. The weights for the barycenter problem are $\lambda_k = 1/2$, the cost functions are classical quadratic Euclidean. The described task is popular in previous continuous OT barycenter research (Fan et al., 2021; Korotin et al., 2022; Noble et al., 2023; Kolesov et al., 2023).

Considered data setups. It is known that the support of the recovered barycenter is contained in the manifold $\mathcal{M}$ of weighted pixel-wise blends $1/2 \cdot x_1 + 1/2 \cdot x_2$ of input images $x_1 \sim \mathbb{P}_1$, $x_2 \sim \mathbb{P}_2$. This allows us to use our idea with an auxiliary StyleGAN model trained on the pixel-wise combinations of digits. In this case, we learn OT barycenters in the latent space of this StyleGAN with *non-quadratic* cost functions. At the same time, we do not limit ourselves to the latent space and also consider the conventional (ambient) data-space setup without auxiliary generative models.

Evaluation. Our results are presented in Fig. 5. In the data-space setup, we launch our Algorithm 1 with classical quadratic cost functions and utilize deterministic maps parameterization, see §4.2, i.e., we recover the *unregularized* OT barycenter. As the baselines for visual comparison, we take WIN (Korotin et al., 2022) and SCWB (Fan et al., 2021) which are specifically designed for solving the unregularized barycenter problem with quadratic cost functions. Qualitatively, our obtained samples from the barycenter are in concordance with these competitors. We also include EgBary (Kolesov et al., 2023) method as the baseline in the data-space setup. Its samples are noised due to Entropic regularization. This shows that the usage of regularized cost functions, e.g., (7) and (8), is more reasonable in the manifold-constrained setup. The corresponding samples can be found in Fig. (5). For ϵ-KL and γ-Energy cost function cases, we use $\epsilon = 0.01$ and $\gamma = 100$. More samples from our recovered barycenters with different regularizations can be found in Fig. 6 and Fig. 7.

C. Experimental details

We aggregate the hyper-parameters of our Algorithm 1 for different experiments in Table 4. For our experiments with manifold-constrained barycenter learning (Shape-Color §5.2; Ave, celeba! §5.3; MNIST 0/1 §B.2) we use StyleGAN2-ada model from the official repository:

`https://github.com/NVlabs/stylegan2-ada-pytorch`.

For the details of the baseline solvers, see (Kolesov et al., 2023, Appendix C). We utilize exactly the same hyper-parameters. To implement EgBary solver (Kolesov et al., 2023), we use their publicly available source code.

Color preserving cost. The operation H_c which extracts the color from the image is designed as follows. We transform all the pixels in the given image to HSV scale. Then we pick only those whose value (V) is greater than 0.8. We take the mean values of these pixels and obtain a vector in $[0, 1]^3$.

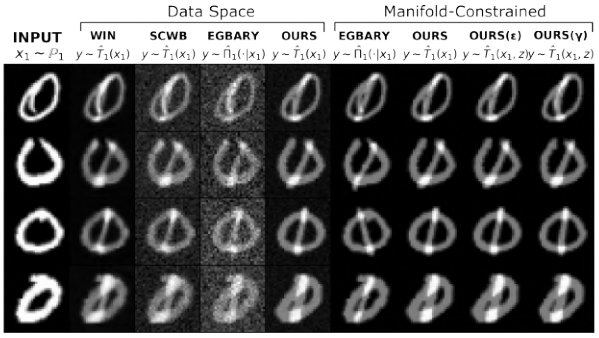
D. Ave Celeba! experimental insights

In this section, we provide further details on Ave Celeba! experiment (§5.3). In particular, we demonstrate training curves for our method, reveal time consumptions and carry out extended comparisons with the competing approaches.

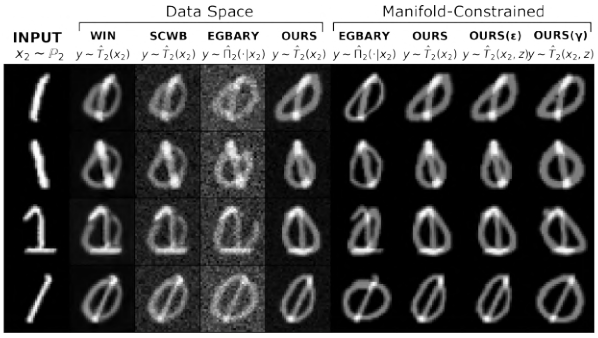
Comparison metrics details. When comparing the methods, we utilize the following metrics: FID, $\mathcal{L}_2$-UVP and transport

Experiment	D	K	ϵ	γ	*batch size*	M_T	λ_1	λ_2	λ_3	$f_{k,\theta}$	$T_{k,\phi}$	$lr_{f_{k,\theta}}$	$lr_{T_{k,\phi}}$	*# of epochs*
Toy 2D	2	3	1	-	2^{10}	3	1/3	1/3	1/3	MLP	MLP	1e-3	1e-3	1200
Toy 2D	2	3	-	1	2^{10}	3	1/3	1/3	1/3	MLP	MLP	1e-3	1e-3	1200
Gaussians	2−64	2	-	-	2^{10}	3	0.25	0.25	0.5	MLP	MLP	1e-3	1e-3	1200
MNIST 0/1	1x32x32	2	-	-	64	10	0.5	0.5	-	ResNet	UNET	2e-4	2e-4	20K
MNIST 0/1	512	2	-	-	64	10	0.5	0.5	-	ResNet	ResNet	2e-4	2e-4	20K
MNIST 0/1	512	2	0.1	-	64	10	0.5	0.5	-	ResNet	ResNet	2e-4	2e-4	20K
MNIST 0/1	512	2	-	10K	64	10	0.5	0.5	-	ResNet	ResNet	2e-4	2e-4	20K
Ave Celeba	3x64x64	3	-	-	64	10	0.25	0.5	0.25	ResNet	UNET	2e-4	2e-4	40K
Ave Celeba	512	3	-	-	64	10	0.25	0.5	0.25	ResNet	ResNet	2e-4	2e-4	40K
Ave Celeba	512	3	0.1	-	64	10	0.25	0.5	0.25	ResNet	ResNet	2e-4	2e-4	40K
Ave Celeba	512	3	-	10K	64	10	0.25	0.5	0.25	ResNet	ResNet	2e-4	2e-4	40K
Shape Color	512	2	0.1	-	64	10	0.5	0.5	-	ResNet	ResNet	2e-4	2e-4	20K

Table 4: Hyper-parameters of Algorithm 1 for different experiments.

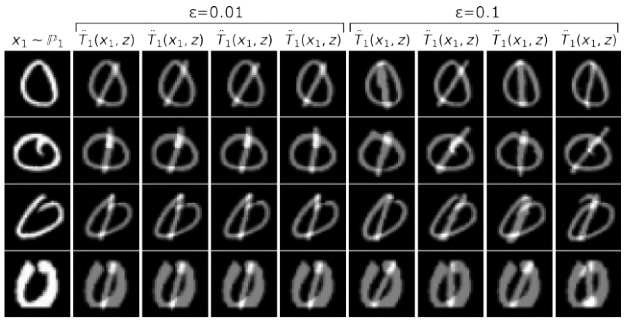


(a) Source and transformed samples for input $\mathbb{P}_1$.

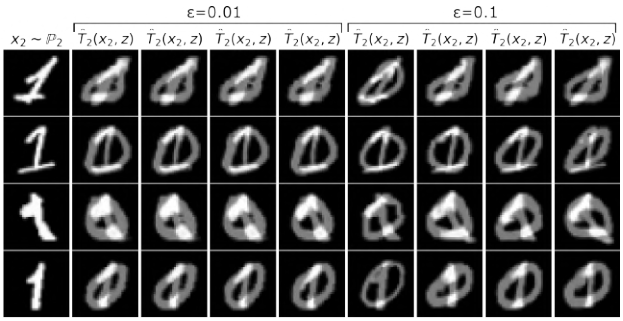


(b) Source and transformed samples for input $\mathbb{P}_2$.

Figure 5: Learned (stochastic) maps to the OT barycenter by different solvers; MNIST 0/1 experiment (§B.2).



(a) Source and transformed samples for input $\mathbb{P}_1$.



(b) Source and transformed samples for input $\mathbb{P}_2$.

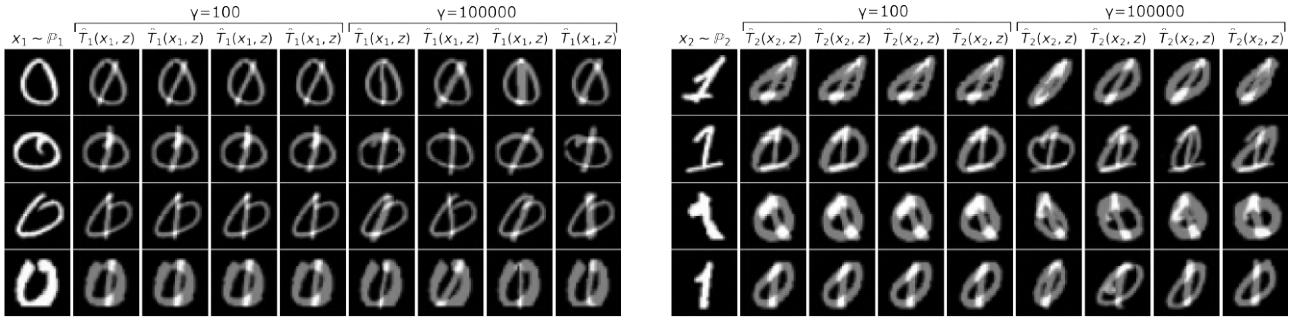
Figure 6: Additional samples from **our** learned stochastic OT barycenter maps. Manifold-constrained data setup; ϵ-KL weak cost functions with different regularization strengths ϵ; MNIST 0/1 experiment (§B.2).

cost (W_2^2). The definitions of these metrics are as follows:

$$\mathcal{L}_2\text{-UVP}(\widehat{T}) = \frac{100\%}{\text{Var}(\mathbb{Q}^*)} \cdot \mathbb{E}_{x\sim\mathbb{P}}\left[\|\widehat{T}(x) - T^*(x)\|^2\right]; \qquad W_2^2(\widehat{T}) = \mathbb{E}_{x\sim\mathbb{P}}\left[\|\widehat{T}(x) - x\|^2\right].$$

In these formulas, $\mathbb{P}$ is a reference distribution ($\mathbb{P} \in \mathbb{P}_{1:3}$); T^* is the ground truth squared Euclidean OT mapping between $\mathbb{P}$ and the GT barycenter $\mathbb{Q}^*$ (which are known by construction); $\widehat{T}$ is a learned mapping. If the learned mapping $\widehat{T}$ is stochastic, i.e., it permits additional noise or it is represented by energy potential, $\mathcal{L}_2$-UVP (W_2^2) are additionally averaged over this stochasticity. If $\widehat{T}$ maps to latent space, its output is additionally fed to StyleGAN encoder before computing the metrics. Note that $\mathcal{L}_2$-UVP directly compares the learned transport map with the true OT map to the barycenter.

Convergence curves for our method. We provide the detailed learning curves of our proposed method (in data/latent space,



(a) Source and transformed samples for input $\mathbb{P}_1$.

(b) Source and transformed samples for input $\mathbb{P}_2$.

Figure 7: Additional samples from **our** learned stochastic OT barycenter maps. Manifold-constrained data setup; γ-Energy weak cost functions with different regularization strengths γ; MNIST 0/1 experiment (§B.2).

(a) input, transformed samples ($\mathbb{P}_0$).

(b) input, transformed samples ($\mathbb{P}_1$).

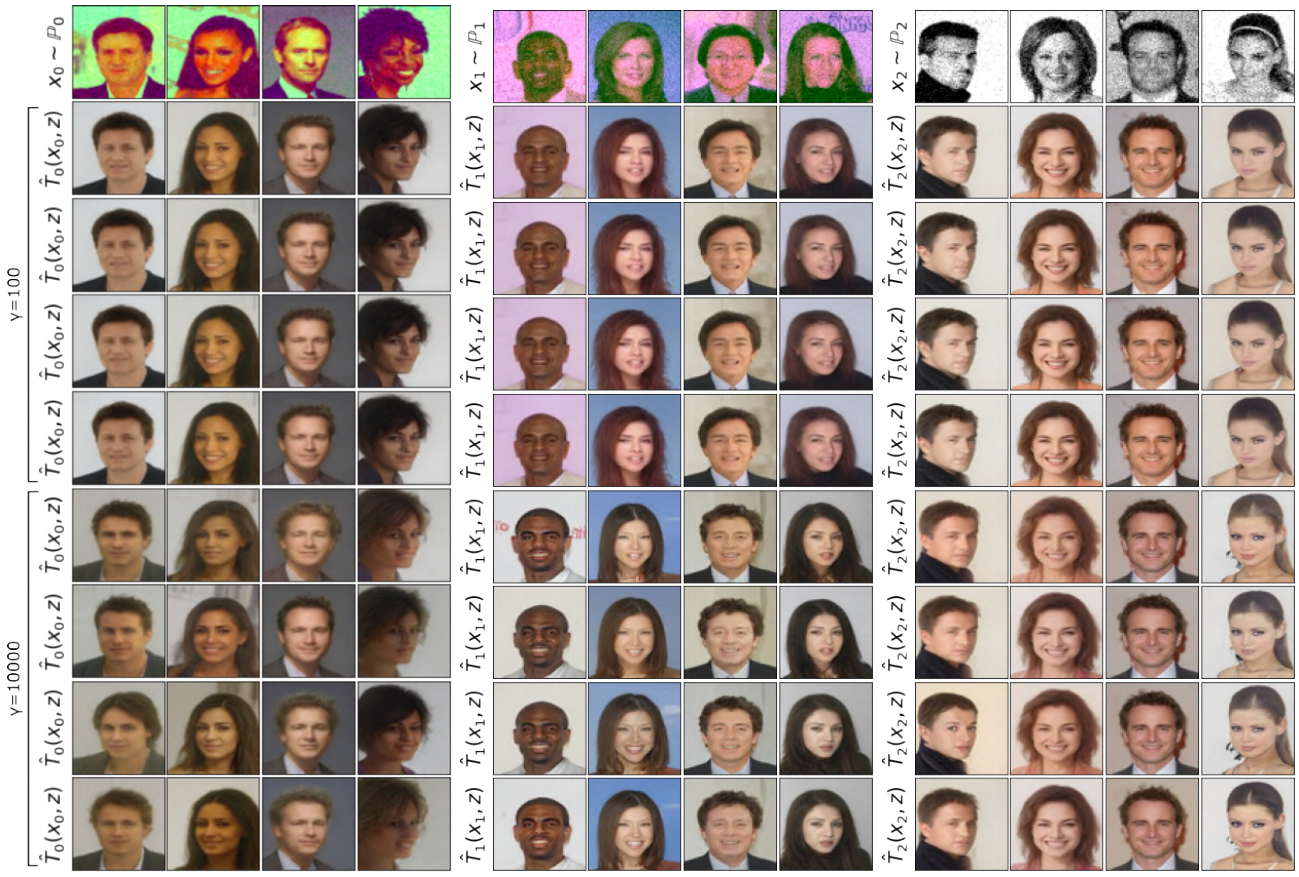
(c) input, transformed samples ($\mathbb{P}_2$).

Figure 8: Additional samples from **our** learned stochastic OT barycenter maps. Manifold-constrained data setup; ϵ-KL weak cost functions with different regularization strengths ϵ; Ave, Celeba! experiment (§5.3).

with different costs), to demonstrate its convergence capabilities, see Figure 10. For the reasonable metrics which we track during the training, we choose: (a) FID 10a; (b) $\mathcal{L}_2$-UVP 10b; (c) Transport cost (W_2^2) 10c.

Interestingly, our approach experiences some local instabilities when learning the third barycenter mapping $\mathbb{P}_3 \to \mathbb{Q}^*$. This is because the third distribution represents *grayscaled* images (see, e.g., Figure 4(c) from our paper), which makes the corresponding mapping much more difficult compared to the others.

Training/inference times. In Table 5, we provide the **training** and **inference** times on Ave Celeba! setup ($k = 3$) for the



(a) input, transformed samples ($\mathbb{P}_0$). (b) input, transformed samples ($\mathbb{P}_1$). (c) input, transformed samples ($\mathbb{P}_2$).

Figure 9: Additional samples from **our** learned stochastic OT barycenter maps. Manifold-constrained data setup; γ-Energy weak cost functions with different regularization strengths γ; Ave, Celeba! experiment (§5.3).

Method	OURS(Data)	OURS(latent)	OURS(ϵ)	OURS(γ)	WIN	EgBary(Data)	EgBary(latent)
Training	30 h	40 h	40 h	40 h	$\sim$ 100 h	100 h	30 h
Inference	0.01 s	0.05 s	0.05 s	0.08 s	0.02 s	175 s	150 s

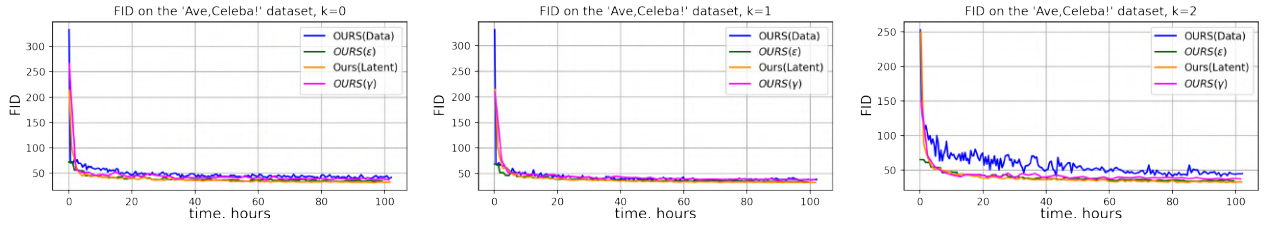
Table 5: Training/inference times of different barycenter solvers; Ave Celeba! experiment.

considered competing solvers. When training the competing solvers, we choose their recommended numbers of training iterations. The training times correspond to the checkpoints for which we provide FID metrics in Table 2. As we can see, while the training times are not very different, at the inference our approach is more competitive.

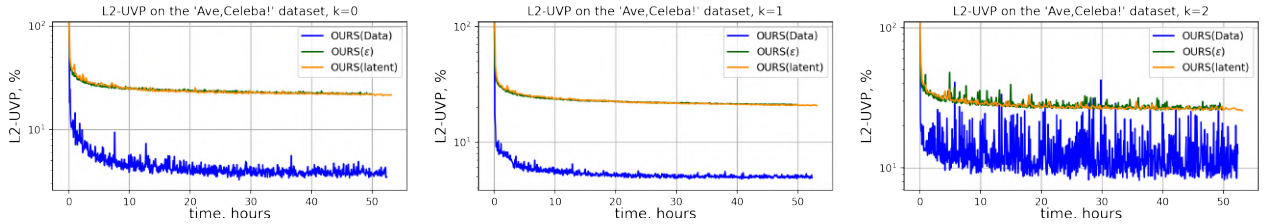
Comparison of our method and WIN (Korotin et al., 2022). We provide the detailed comparison of the training curves of our (bi-level) approach vs. WIN baseline in the data space setup, see Figure 11. Note that WIN represents the most promising tri-level OT barycenter solver. In the chart we demonstrate the evolution of the transport cost during the training. We emphasize that the **curve for WIN is much less stable** (we even trim the plots from the above) which is natural due to the reliance of WIN on tri-level adversariality.

D.1. Detailed comparison of Our method and EgBary (Kolesov et al., 2023)

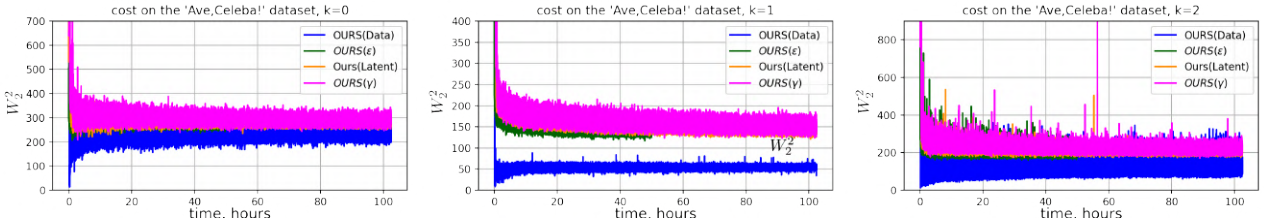
In our experiments with Ave, celeba! Benchmark dataset (§5.3) the competitive method, EgBary (Kolesov et al., 2023), demonstrates better FID scores compared to ours. In what follows, we provide a detailed comparison of our approach with this baseline. As for the setup, we choose Ave Celeba! experiment (§5.3). For the purpose of completeness and clearness of our comparison, we additionally train EgBary in the data space. For EgBary, we use the same Neural Network architectures as we do for our approach (discriminators $f_{k,\theta}$), see our Table 4. The hyperparameters for their method are chosen for



(a) FID vs. time.



(b) $\mathcal{L}_2$-UVP vs. time.



(c) Transport cost vs. time.

Figure 10: Training curves for OUR proposed method (different costs, data/latent setups); Ave Celeba!

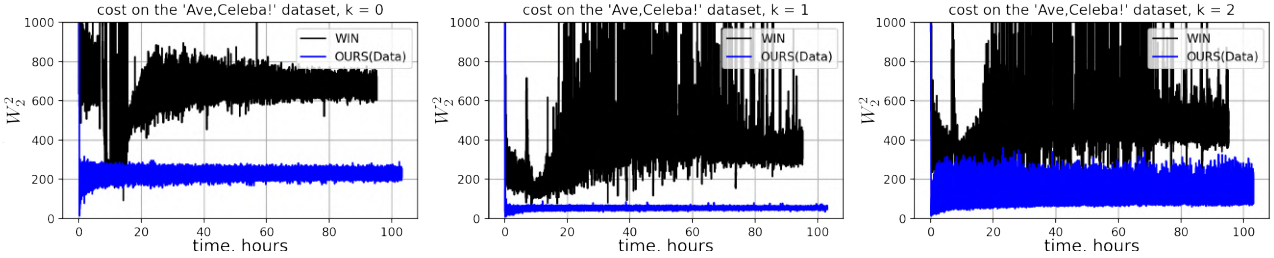


Figure 11: Transport cost w.r.t training time; our method (classical cost) vs. WIN; Data space.

quality reasons and following the guidelines from their paper: $\epsilon = 10^{-2}$, $lr = 10^{-4}$, $iter = 5000$, $\sqrt{\eta} = 0.1$, $L = 500$, $S = 64$.

Comparison at inference. In the provided tables below we demonstrate the inference times, achieved FID and $\mathcal{L}_2$-UVP for

- Our approach (classical cost) and EgBary in data space (Table 7);
- Our approach with different costs (classical, ϵ-KL and γ-Energy) and EgBary in latent StyleGAN space (Table 6).

Some comments:

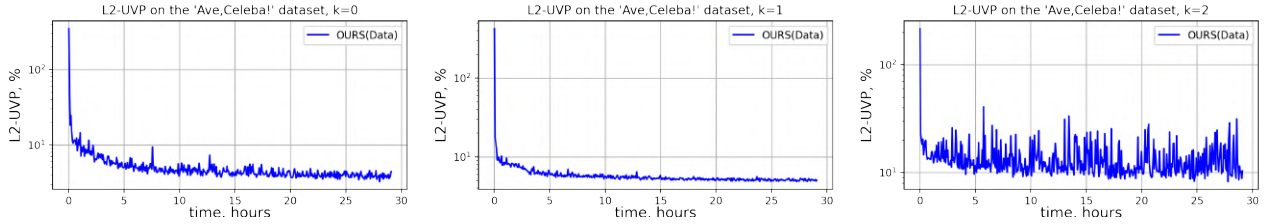
1. EgBary relies on the Langevin sampling. In order to provide further details on EgBary performance, in the tables we report the metrics for different numbers of utilized Langevin steps. Fewer steps - the inference is faster, while the quality metrics are not that good and vice versa.
2. The distribution represented by StyleGAN (learned on CelebA) and the original CelebA distribution are similar but slightly different, which introduces additional biases for the computed metrics. That is why *latent* methods and *data* methods should be compared independently.

Method	FID			L2-UVP, %			Langevin steps	t, sec
	k=0	**k=1**	**k=2**	**k=0**	**k=1**	**k=2**		
EgBary	15.8	15.3	18.3	46	45	48	50	15
	11.3	11.2	14.3	37	36	40	150	45
	8.4	8.7	10.2	35	33	37	250	75
	8.3	8.2	9.9	32	32	34	500	150
	8.2	8.2	9.8	32	31	33	1000	300
OURS	30.7	31.0	31.7	21	21	26	Not Applicable	0.05
OURS(ϵ)	34.5	34.9	35.7	22	21	21		0.05
OURS(γ)	38.3	37.8	37.6	22	22	23		0.08

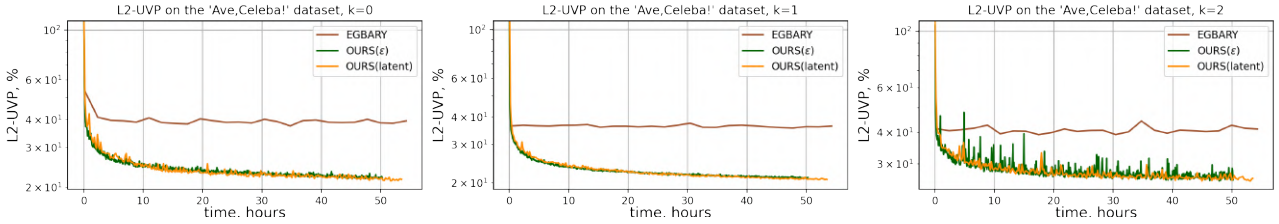
Table 6: OUR method vs. EgBary; Ave Celeba!; Latent space.

Method	FID			L2-UVP,%			Langevin steps	t, sec
	k=0	**k=1**	**k=2**	**k=0**	**k=1**	**k=2**		
EgBary	125.3	122.3	187.6	53	39	46	10	7
	119.4	120.0	169.7	28	27	31	150	105
	118.5	120.4	168.8	26	26	31	250	175
	118.3	120.1	168.9	25	24	30	500	350
	118.8	121.2	170.2	25	24	30	1000	700
OURS	39.0	38.6	39.8	3	4	9	N/A	0.01

Table 7: OUR method vs. EgBary; Ave Celeba!; Data space.



(a) Data space; OURS (classical cost).



(b) Latent space; OURS (classical, ϵ-KL costs), EgBary.

Figure 12: Behaviour of $\mathcal{L}_2$-UVP (w.r.t. to elapsed time) for our approach and EgBary in different setups; Ave Celeba!

Conclusions. At first, we can see that EgBary method is much more time-consuming at the inference stage. This is expected since it uses MCMC sampling. Secondly, in the data space their results (FID) are not that good. In particular, in the data space, our approach demonstrates the best FID. Thirdly, even in the latent space our method demonstrates better results according to the $\mathcal{L}_2$-UVP. At the same time, we acknowledge that in the latent space, their FID score is better than ours.

Comparison in training. We provide the behaviour of $\mathcal{L}_2$-UVP (w.r.t. to elapsed time) when training our approach and EgBary in (i) StyleGAN latent space, see Figure 12b; (ii) data space, see Figure 12a. For the latter, we provide training statistics only for our method.

Conclusions (latent space). EgBary's latent learning curve achieves its optimum (gets stuck) rather quickly and stops improving. Notably, overall **our $\mathcal{L}^2$-UVP score is always much better than their score**. This is probably due to the fact

(a) input, transformed samples ($\mathbb{P}_0$). (b) input, transformed samples ($\mathbb{P}_1$). (c) input, transformed samples ($\mathbb{P}_2$).

Figure 13: Samples from (EgBary) learned stochastic OT barycenter plans; Data space; Ave, Celeba! experiment (§5.3).

that Entropy regularization "blows up" the barycenter distribution and introduces extra bias. Besides, MCMC sampling adds inconvenient noise to the learning process. In particular, in data space, this is clearly seen from the additional noise on the resulting transported samples, see Figure 13.

Conclusions (data space). We provide qualitative examples of EgBary's data space performance in Figure 13. This EgBary's behaviour is analogous to the one that we have already shown in MNIST 0/1 experiment in the data space (Figure 5 in the Appendix). In contrast, our method in the data space works incomparably better both in terms of FID (~ 30 vs. ≥ 100) and $\mathcal{L}^2$-UVP ($\sim 5\%$ vs. $\sim 25\%$).

E. Extended related works. Detailed comparison of our method with (Chi et al., 2023)

Similar to ours, (Chi et al., 2023) utilizes a bi-level optimization objective. To contextualize the novelty of our proposed method, in this section we provide a detailed discussion of this research, compare their approach with ours, and distinguish our contributions compared to their work.

1. Origin of bi-level objective. While both, our objective and the objective by (Chi et al., 2023), are bi-level, the origins of these objectives are rather different. Our considered optimization problem (11) stems from the semi-dual formulation of the weak OT problem (4). This formulation can be used with a diverse set of admissible weak cost functions, in particular, with *unregularized* classical cost function, as we demonstrate in our paper. In turn, the work (Chi et al., 2023) is based on the conventional dual OT problem introduced in (Cuturi, 2013). One limitation of this formulation is that it deals exclusively with regularized OT problems, i.e., seeks for barycenter w.r.t. regularized OT cost functions. Moreover, only two reasonable choices of regularization is known for this formulation: entropic and quadratic. For some attempts to consider more general regularizations coupled with conventional dual OT, see (Blondel et al., 2018).

2. What is optimized in the bi-level objective? The entities that are obtained in the output of our method and the method by (Chi et al., 2023) are different. In our algorithm, we optimize w.r.t. to plans π_k, parameterized as stochastic or deterministic mappings from the reference distributions $\mathbb{P}_k$ to the (implicitly) learned barycenter distribution $\mathbb{Q}$. In contrast, the approach from (Chi et al., 2023) explicitly learns the generative distribution for $\mathbb{Q}$ but **does not** recover the mappings from reference distributions to the recovered barycenter. At the same time, the knowledge of these mappings is important in some applications of barycenter problem, see, e.g., our Shape-Color experiment, §5.2.

3. Is it true that the objective by (Chi et al., 2023) is bi-level? One of the "levels" in the bi-level objective from (Chi et al., 2023) is double $\sup_\phi \sup_\psi$ optimization w.r.t. pair of dual OT potentials. While we acknowledge that no adversariality appears here, we note that typically such sup-sup problems are optimized via an *alternating* procedure, see (Seguy et al., 2018; Daniels et al., 2021).

4. Barycenter distribution parameterization in (Chi et al., 2023). When (explicitly) parameterizing the barycenter distribution, the authors of (Chi et al., 2023) propose the use of Gaussians or Mixtures of Gaussians (MoG) with learnable parameters, see their "Introducing a Variational Distribution" section. While MoG is known to satisfy some universal approximation properties, the performance of such models in high dimensions is questionable.

5. Practical considerations. The authors of (Chi et al., 2023) **only** demonstrate the applicability of their approach in low dimensions (< 9) and in moderate dimensions ($D = 128, 256$) for **Gaussian case**. We are not sure if their approach scales to high dimensions, e.g., to the image domain. In contrast, for our approach, we demonstrate its scalability to the image domain. Furthermore, our proposed procedure is based on well-established architectural, technical and numerical practices from existing research, e.g., (Korotin et al., 2023b; Choi et al., 2023), proven to scale well with dimensions.