Energy-guided Entropic Neural Optimal Transport
(Second project status report at Selected Topics of DS course)

Petr Mokrov
Skolkovo Institute of Science and Technology
Moscow, Russia
petr.mokrov@skoltech.ru

Abstract

Energy-Based Models (EBMs) are known in the Machine Learning community for the decades. Since the seminal works devoted to EBMs dating back to the noughties there have been appearing a lot of efficient methods which solve the generative modelling problem by means of energy potentials (unnormalized likelihood functions). In contrast, the realm of Optimal Transport (OT) and, in particular, neural OT solvers is much less explored and limited by few recent works (excluding WGAN based approaches which utilize OT as a loss function and do not model OT maps themselves). In our work, we bridge the gap between EBMs and Entropy-regularized OT. We present the novel methodology which allows utilizing the recent developments and technical improvements of the former in order to enrich the latter. We validate the applicability of our method on toy 2D scenarios as well as standard unpaired image-to-image translation problems. For the sake of simplicity, we choose simple short- and long- run EBMs as a backbone of our Energy-guided Entropic OT method, leaving the application of more sophisticated EBMs for future research.

1 Introduction

The computational Optimal Transport (OT) field is an emergent and fruitful area in the Machine Learning community which finds its applications in generative modelling [3, 25, 12], domain adaptation [44, 60, 66], unpaired image-to-image translation [71, 31], datasets manipulation [1], population dynamics [40, 67], gradient flows modelling [2, 43], barycenter estimation [33, 16]. The majority of the applications listed above utilize OT as a loss function, e.g., have WGAN-like objectives which compare the generated (fake) and true data distributions. However, for some practical use cases, e.g., unpaired image to image translation [35], it is worth to model the OT maps or plans by themselves. These plans minimize the average movement of probability mass from the source distribution to the target distribution with respect to the so-called transport *cost* function. From the perspectives of unpaired data translation, such optimal plans could be considered as a desirable solution (source to target transform) given a properly specified *cost* function which penalizes data content modification.

The existing approaches which recover OT plans are based on various theoretically-advised techniques. Some of them [41, 32] utilize the specific form of cost function, e.g., squared Euclidean distance. The others [71, 39] modify GAN objectives with additional OT regularizer. However, there are some indications (see Theorem 1 from [20]), that such methods are *biased*. Of the special interest are [15, 35, 53] which take the advantage of dual OT problem formulation and treat the adversarial "critic" appearing in the proposed objective as a Lagrangian multiplier which ensures the generative distribution coincides with the target one. Moreover, it seems that the method from these works is the only approach leveraging *unbiased* large-scale continuous OT with general cost functions. At the same time, [34] notes, that the method from [15, 35, 53] may yield *fake* solutions due to convexity-related issues of the optimization problem they consider. The authors of [34] propose to use strictly

convex regularizers which guarantee the uniqueness of the recovered OT plans. And one popular choice which has been extensively studied both in discrete [9, 22] and continuous [22, 8, 44] settings is the **Entropy**. The well-studied methodological choices for modelling Entropy-regularized Optimal Transport (EOT) include (a) stochastic dual maximization approach which prescribes alternating optimization of dual potentials [59, 11] and (b) dynamic setup having connection to Schrödinger bridge problem [6, 26, 7]. In contrast to the methods presented in the literature, we come up with the approach for solving EOT built upon EBMs.

Contributions. We propose a novel energy-based view on the EOT problem. From the theoretical point of view, we characterize the EOT as a constrained optimization problem, the approach inspired by Energy-Based GANs [10], and connect the dual Lagrangian multiplier which originates due to this characterization with standard dual OT Kantorovich potentials (§3) From the technical point of view, we propose theoretically-grounded yet easy-to-implement modifications to the standard EBMs training objective and procedure which makes them capable to recover the EOT plans (§3.1). From the practical point of view, we validate our approach on toy 2D settings as well as Colored MNIST data transformation task (§5).

Notations. Throughout the paper, $\mathcal{X}$ and $\mathcal{Y}$ are compact subsets of Euclidean space, i.e., $\mathcal{X} \subset \mathbb{R}^{D_x}$ and $\mathcal{Y} \subset \mathbb{R}^{D_y}$. Practically, we will assume that $D_x = D_y$ yet our analysis holds for $D_x \neq D_y$. Taking into account real world applications, the compactness of $\mathcal{X}$ and $\mathcal{Y}$ is not a restrictive assumption, since the typical real data on hand, e.g., images, has bounded support. From the theoretical point of view, this property much simplifies the technical proofs, which contributes to the clearness and rigorousness of the paper. Nevertheless, we look forward to lifting of this assumption, which is a possible avenue for future work.

The continuous functions on $\mathcal{X}$ and $\mathcal{Y}$ are denoted as $\mathcal{C}(\mathcal{X})$ and $\mathcal{C}(\mathcal{Y})$, respectively. In turn, $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ are the sets of Borel probability distributions on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Given distributions $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ and $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$, $\Pi(\mathbb{P}, \mathbb{Q})$ designates the set of *couplings* between the distributions $\mathbb{P}$ and $\mathbb{Q}$, i.e., probability distributions on the product space $\mathcal{X} \times \mathcal{Y}$ with the first and second marginals given by $\mathbb{P}$ and $\mathbb{Q}$, respectively. We use $\Pi(\mathbb{P})$ to denote the set of probability distributions on the product space $\mathcal{X} \times \mathcal{Y}$ with the first marginal given by $\mathbb{P}$. The absolutely continuous probability distributions on $\mathcal{X}$ and $\mathcal{Y}$ are $\mathcal{P}_{\text{ac}}(\mathcal{X})$ and $\mathcal{P}_{\text{ac}}(\mathcal{Y})$, respectively. For $\mathbb{P} \in \mathcal{P}_{\text{ac}}(\mathcal{X})$ and $\mathbb{Q} \in \mathcal{P}_{\text{ac}}(\mathcal{Y})$ we use $\frac{\mathrm{d}\mathbb{P}(x)}{\mathrm{d}x}$ and $\frac{\mathrm{d}\mathbb{Q}(y)}{\mathrm{d}y}$ to denote the corresponding probability density functions, i.e., Radon-Nikodym derivatives with respect to the standard Lebesgue measures on $\mathcal{X}$ and $\mathcal{Y}$. Given distributions μ and ρ defined on a set $\mathcal{Z}$, $\mu \ll \rho$ means that μ is absolutely continuous with respect to ρ. We recall the definitions of the Kullback-Leibler divergence (KL) and the Entropy (H) below:

$$\text{KL}\left(\mu \| \rho\right) = \begin{cases} \int_{\mathcal{Z}} \log \frac{\mathrm{d}\mu(z)}{\mathrm{d}\rho(z)} \mathrm{d}\mu(z) & \text{, if } \mu \ll \rho \\ +\infty & \text{, else} \end{cases}$$

$$H(\mu) = \begin{cases} -\int_{\mathcal{Z}} \log \frac{\mathrm{d}\mu(z)}{\mathrm{d}z} \mathrm{d}\mu(z) & \text{, if } \mu \in \mathcal{P}_{\text{ac}}(\mathcal{Z}) \\ -\infty & \text{, else} \end{cases}$$

Both KL and $-H$ are known to be lower-semicontinuous (w.r.t. the weak topology) and *strictly* convex [57, 48, 47].

2 Background

2.1 Optimal Transport

The introduction of OT and EOT in this subsection generally follows [26]. For the specific details regarding EOT, we refer the readers to [22]. The theory behind OT could be found in [65, 57].

Let $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ and $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$. The primal OT problem due to Kantorovich [65] is as follows:

$$\text{OT}_c(\mathbb{P}, \mathbb{Q}) \stackrel{\text{def}}{=} \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int\limits_{\mathcal{X} \times \mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y). \tag{1}$$

In the equation above $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a *cost* function which reflects a practitioner's knowledge on how data from the source and target distribution should be aligned. Typically, the cost function

is chosen to be Euclidean norm $c(x, y) = \|x - y\|_2$ yielding the 1-Wasserstain distance ($\mathbb{W}_1$) or halved squared Euclidean norm $c(x, y) = \frac{1}{2}\|x - y\|_2^2$ yielding the square of 2-Wasserstein distance ($\mathbb{W}_2^2$). When delivering the theory, we assume the cost function to be continuous, and in the practical sections of our work, we set the cost to be the half squared Euclidean norm.

The distributions $\pi^* \in \Pi(\mathbb{P}, \mathbb{Q})$ which minimize objective (1) are called the *Optimal Transport plans*. Problem (1) may have several OT plans [50, Remark 2.3] and in order to impose the uniqueness and obtain more tractable optimization problem, a common trick is to regularize (1) with strictly convex (w.r.t. distribution π) functionals $\mathcal{R} : \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ [4, §2].

Entropy-regularized Optimal Transport. In our work, we utilize the popular Entropic regularization [9] which has found its applications in various existing works [62, 58, 54]. This is mainly because of amenable sample complexity [21, §3] and tractable dual representation of the Entropy-regularized OT problem which can be leveraged by solid computational procedures, e.g., Sinkhorn's algorithm [9, 64]. Besides, the EOT objective is known to be strictly convex [22] which greatly contributes to the theoretical properties like convergence of the practical procedures and uniqueness of the OT plan.

Let $\varepsilon > 0$. The EOT problem can be formulated in the following ways:

$$\text{EOT}_{c,\varepsilon}^{(1)}(\mathbb{P}, \mathbb{Q}) \stackrel{\text{def}}{=} \min_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) + \varepsilon \text{KL}\left(\pi \| \mathbb{P} \times \mathbb{Q}\right), \tag{2}$$

$$\text{EOT}_{c,\varepsilon}^{(2)}(\mathbb{P}, \mathbb{Q}) \stackrel{\text{def}}{=} \min_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) - \varepsilon H(\pi), \tag{3}$$

$$\text{EOT}_{c,\varepsilon}(\mathbb{P}, \mathbb{Q}) \stackrel{\text{def}}{=} \min_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) - \varepsilon \int_{\mathcal{X}} H(\pi(y|x)) \underbrace{\mathrm{d}\mathbb{P}(x)}_{=\mathrm{d}\pi(x)}, \tag{4}$$

These formulations are equivalent when $\mathbb{P}$ and $\mathbb{Q}$ are absolutely continuous w.r.t. the corresponding standard Lebesgue measures. The equivalence holds true thanks to $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$. It could be seen from the following equation:

$$\text{KL}\left(\pi \| \mathbb{P} \times \mathbb{Q}\right) = \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{\mathrm{d}\pi(x, y) / \mathrm{d}\left[x, y\right]}{\mathrm{d}\mathbb{P}(x)/\mathrm{d}x \cdot \mathrm{d}\mathbb{Q}(y)/\mathrm{d}y} \right) \mathrm{d}\pi(x, y) =$$

$$\underbrace{\int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{\overbrace{\mathrm{d}\pi(x, y) / \mathrm{d}\left[x, y\right]}^{= \mathrm{d}\pi(y|x)/\mathrm{d}y}}{\mathrm{d}\mathbb{P}(x)/\mathrm{d}x} \right) \overbrace{\mathrm{d}\pi(x, y)}^{= \mathrm{d}\pi(y|x)\mathrm{d}\mathbb{P}(x)}}_{= \int_{\mathcal{X}} H(\pi(y|x))\mathrm{d}\mathbb{P}(x)} - \underbrace{\int_{\mathcal{X} \times \mathcal{Y}} \log \frac{\mathrm{d}\mathbb{Q}(y)}{\mathrm{d}y} \mathrm{d}\pi(x, y)}_{= -H(\mathbb{Q})} =$$

$$\underbrace{\int_{\mathcal{X} \times \mathcal{Y}} \log \frac{\mathrm{d}\pi(x, y)}{\mathrm{d}\left[x, y\right]} \mathrm{d}\pi(x, y)}_{= H(\pi)} - \underbrace{\int_{\mathcal{X} \times \mathcal{Y}} \log \frac{\mathrm{d}\mathbb{P}(x)}{\mathrm{d}x} \mathrm{d}\pi(x, y)}_{= -H(\mathbb{P})} - \underbrace{\int_{\mathcal{X} \times \mathcal{Y}} \log \frac{\mathrm{d}\mathbb{Q}(y)}{\mathrm{d}y} \mathrm{d}\pi(x, y)}_{= -H(\mathbb{Q})}.$$

In other words, $\text{KL}\left(\pi \| \mathbb{P} \times \mathbb{Q}\right) = \int_{\mathcal{X}} H(\pi(y|x))\mathrm{d}\mathbb{P}(x) + H(\mathbb{Q}) = H(\pi) + H(\mathbb{Q}) + H(\mathbb{P})$ and the equations (2), (3) and (4) are the same up to additive constants. In the remaining paper, we will primarily work with the EOT formulation (4), and, henceforth, we will additionally assume $\mathbb{P} \in \mathcal{P}_{\text{ac}}(\mathcal{X})$, $\mathbb{Q} \in \mathcal{P}_{\text{ac}}(\mathcal{Y})$ when necessary.

Let $\pi^* \in \Pi(\mathbb{P}, \mathbb{Q})$ be the solution of EOT problem. The measure disintegration theorem yields:

$$\mathrm{d}\pi^*(x, y) = \mathrm{d}\pi^*(y|x)\mathrm{d}\pi^*(x) = \mathrm{d}\pi^*(y|x)\mathrm{d}\mathbb{P}(x).$$

The distributions $\pi^*(\cdot|x)$ will play an important role in our further analysis. Actually, they constitute the only ingredient needed to (stochastically) transform a source point $x \in \mathcal{X}$ to the target samples $y_1, y_2, \cdots \in \mathcal{Y}$ w.r.t. EOT plan. We say that the distributions $\{\pi^*(\cdot|x)\}_{x \in \mathcal{X}}$ are the *optimal conditional plans*.

Dual formulation of the EOT problem. The primal EOT problem (2) permits the dual reformulation [22]. Let $u \in \mathcal{C}(\mathcal{X})$ and $v \in \mathcal{C}(\mathcal{Y})$. Define the dual functional by

$$F_\varepsilon(u, v) \stackrel{\text{def}}{=} \int_{\mathcal{X}} u(x)\mathrm{d}\mathbb{P}(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\mathbb{Q}(y) - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right) \mathrm{d}\left[\mathbb{P} \times \mathbb{Q}\right](x, y), \tag{5}$$

then the strong duality holds [22, Proposition 2.1]:

$$\mathrm{EOT}^{(1)}_{c,\varepsilon}(\mathbb{P}, \mathbb{Q}) = \max_{u\in\mathcal{C}(\mathcal{X}), v\in\mathcal{C}(\mathcal{Y})} F_\varepsilon(u, v). \tag{6}$$ [1]

The potentials u^*, v^* which constitute a solution of the (6) are called the (Entropic) Kantorovich potentials. The optimal transport plan π^* which solves the primal problem (2) could be recovered from a pair of Kantorovich potentials (u^*, v^*) as

$$\mathrm{d}\pi^*(x, y) = \exp\left(\frac{u^*(x) + v^*(y) - c(x, y)}{\varepsilon}\right) \mathrm{d}\mathbb{P}(x)\mathrm{d}\mathbb{Q}(y).$$

From the practical viewpoint, the dual objective (6) is an unconstrained maximization problem which can be solved by conventional optimization procedures. The existing methods based on (6) as well as their limitations and drawbacks will be discussed in related works section, §4.2.

Semi-dual formulation of the EOT problem. The objective (6) is a convex optimization problem. By fixing a function $v \in \mathcal{C}(\mathcal{Y})$ and applying the first-order optimality conditions for the marginal optimization problem $\max_u F_\varepsilon(u, v)$, one can recover the solution in the closed-form:

$$v^{c,\varepsilon}(x) \stackrel{\text{def}}{=} \arg\max_{u\in\mathcal{C}(\mathcal{X})} F_\varepsilon(u, v) = -\varepsilon \log\left(\int_{\mathcal{Y}} \exp\left(\frac{v(y) - c(x, y)}{\varepsilon}\right) \mathrm{d}\mathbb{Q}(y)\right). \tag{7}$$

By substituting the argument $v \in \mathcal{C}(\mathcal{X})$ of the $\max$ with $v^{c,\varepsilon}$ in equation (6) and performing several simplifications, one can recover the objective

$$\mathrm{EOT}^{(1)}_{c,\varepsilon}(\mathbb{P}, \mathbb{Q}) = \max_{v\in\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} v^{c,\varepsilon}(x)\mathrm{d}\mathbb{P}(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\mathbb{Q}(y), \tag{8}$$

called *semi-dual* formulation of EOT problem [21, §4.3]. It is not so popular as the classical dual problem (6) since the estimation of $v^{c,\varepsilon}$ is non-trivial, yet we recall it as formulation (8) will play a certain role in our further analysis.

EOT problem as a weak OT problem. The reformulated primal EOT problem (4) can be understood as the so-called *weak* OT problem [24, 5]. Given a *weak* transport cost $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ which penalizes the displacement of a point $x \in \mathcal{X}$ into a distribution $\pi(\cdot|x) \in \mathcal{P}(\mathcal{Y})$, the weak OT problem is defined as follows:

$$\mathrm{WOT}_C(\mathbb{P}, \mathbb{Q}) \stackrel{\text{def}}{=} \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \int_{\mathcal{X}} C(x, \pi(\cdot|x)) \underbrace{\mathrm{d}\pi(x)}_{=\mathrm{d}\mathbb{P}(x)}. \tag{9}$$

By considering

$$C_{\mathrm{EOT}}(x, \pi(\cdot|x)) = \int_{\mathcal{Y}} c(x, y)\mathrm{d}\pi(y|x) - \varepsilon H(\pi(y|x)), \tag{10}$$

the EOT formulation (4) can be seen as a particular case of weak OT problem (9). Note, that if the weak cost C is *strictly* convex and lower semicontinuous, as in the case with C_{EOT}, the solution for (9) exists and unique [5].

[1] Strictly speaking, there should be $\sup$ here, not $\max$. The maximum could be attained when considering the potentials u, v in natural spaces $L^{\exp}_\varepsilon$ [42, Def. 2.1]. However, slightly abusing the notations, following [22, 21], and noting, that continuous functions are dense in $L^{\exp}_\varepsilon$ w.r.t. proper topology, here, and later we write $\max$.

Weak dual formulation of the EOT problem. Similar to the case of classical Kantorovich OT (1), the weak OT problem permits the dual representation. Let $f \in \mathcal{C}(\mathcal{Y})$. Following [5, Eq. (1.3)] one introduces *weak* C-transform $f^C : \mathcal{X} \to \mathbb{R}$ by

$$f^C(x) \stackrel{\text{def}}{=} \inf_{\mu \in \mathcal{P}(\mathcal{Y})} \left\{ C(x, \mu) - \int_{\mathcal{Y}} f(y) \mathrm{d}\mu(y) \right\}. \tag{11}$$

For our particular case of EOT-advised weak OT cost (10), equation (11) reads as

$$f^{C_{\text{EOT}}}(x) = \min_{\mu \in \mathcal{P}(\mathcal{Y})} \left\{ \int_{\mathcal{Y}} c(x, y) \mathrm{d}\mu(y) - \varepsilon H(\mu) - \int_{\mathcal{Y}} f(y) \mathrm{d}\mu(y) \right\} \stackrel{\text{def}}{=} \min_{\mu \in \mathcal{P}(\mathcal{Y})} \mathcal{G}_{x,f}(\mu). \tag{12}$$

Note, that the existence and uniqueness of the minimizer for (12) follows from Weierstrass theorem[2] [57, Box 1.1.] and strict convexity of $\mathcal{G}_{x,f}$ in μ. The dual weak functional $F^w_C : \mathcal{C}(\mathcal{Y}) \to \mathbb{R}$ for primal WOT problem (9) is

$$F^w_C(f) \stackrel{\text{def}}{=} \int_{\mathcal{X}} f^C(x) \mathrm{d}\mathbb{P}(x) + \int_{\mathcal{Y}} f(y) \mathrm{d}\mathbb{Q}(y).$$

Thanks to the compactness of the basic spaces $\mathcal{X}$ and $\mathcal{Y}$, for the cost of interest (10) there is the strong duality [24, Thm. 9.5]:

$$\text{EOT}_{c,\varepsilon}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{C}(\mathcal{Y})} \left\{ \int_{\mathcal{X}} \min_{\mu_x \in \mathcal{P}(\mathcal{Y})} \mathcal{G}_{x,f}(\mu_x) \mathrm{d}\mathbb{P}(x) + \int_{\mathcal{Y}} f(y) \mathrm{d}\mathbb{Q}(y) \right\} = \sup_{f \in \mathcal{C}(\mathcal{Y})} F^w_{C_{\text{EOT}}}(f). \tag{13}$$

We say that (13) is *weak dual objective*. Two questions arise with formulation (13).

- Does it exist an optimal potential f^*, on which the supremum is attained?
- Given an optimal f^*, can we recover the optimal plan which solves the primal EOT problem (4)?

The answers to both questions are positive. The existence of an optimal potential f^* will follow directly from our further analysis[3], yet this fact can be found in previous works, see [49, 4] for reference. Concerning the second question, in order to recover the optimal transport plan for an optimal potential $f^* \in \arg\sup_f F^w_{C_{\text{EOT}}}(f)$, consider distributions $\mu^*_x \in \mathcal{P}(\mathcal{Y})$ solving the internal minimization problem for f^*:

$$\mu^*_x \stackrel{\text{def}}{=} \arg\min_{\mu_x \in \mathcal{P}(\mathcal{Y})} \mathcal{G}_{x,f^*}(\mu_x).$$

Then, distribution $\pi^*(x, y)$, such that

$$\mathrm{d}\pi^*(x, y) = \mathrm{d}\mu^*_x(y) \mathrm{d}\mathbb{P}(x),$$

is the optimal plan for (4). In other words, each *"saddle"* point $(f^*, x \mapsto \mu^*_x)$ of objective (13) yields the optimal plan for (4). It is a natural consequence of the *strict* convexity the entropy regularizer possess. Similar results are already known in the literature, see [34, Thm. 2]. For the completeness of exposition, we provide the proof in Appendix A.

Finishing the EOT-related survey, we emphasize, that although weak dual objective (13) and semi-dual one (8) resembles each other, they are **not** the same. This is because the corresponding primal problems $\text{EOT}^{(1)}_{c,\varepsilon}(\mathbb{P}, \mathbb{Q})$ and $\text{EOT}_{c,\varepsilon}(\mathbb{P}, \mathbb{Q})$ are equivalent but **not** the identical. We will see similarities and differences between these problems, as well as practical consequences of the differences, in subsequent sections.

[2]The Weierstrass theorem could be applied, since $\mathcal{P}(\mathcal{Y})$ is compact and the equation under $\min$ in (12) is lower semicontinuous w.r.t. input measures μ.

[3]As in the case of standard dual EOT, the optimal pontential f^* can be **not** continuous. Yet, we leave the abusing notations for the reasons, explained in the footnote 1

2.2 Energy-based models

The EBMs are a fundamental class of deep Generative Modelling techniques [38, 55] which parameterize distributions of interest $\mu \in \mathcal{P}(\mathcal{Y})$ by means of the Gibbs-Boltzmann distribution density:

$$\frac{\mathrm{d}\mu(y)}{\mathrm{d}y} = \frac{1}{Z}\exp\left(-E(y)\right). \tag{14}$$

In the equation above $E : \mathcal{Y} \to \mathbb{R}$ is the *Energy function* (unnormalized log-likelihood), and $Z = \int_{\mathcal{Y}} \exp(-E(y))\mathrm{d}y$ is the normalization constant, known as the partition function.

Let $\mu \in \mathcal{P}(\mathcal{Y})$ be a true data distribution which is accessible by samples and $\mu_\theta(y), \theta \in \Theta$ be a parametric family of distributions approximated using, e.g., a deep Neural Network E_θ which imitates the Energy function in (14). The problem of generative modelling is to pick up a parameter $\theta \in \Theta$ bringing the parametric distribution $\mu_\theta(y)$ as close as possible to the reference distribution μ, which is typically done in virtue of a pre-specified divergence between the distributions. One popular choice for EBMs is the Kullback-Leibler divergence. The minimization of $\mathrm{KL}\left(\mu\|\mu_\theta\right)$ is done via gradient descent, thanks to the fact that the gradient of the loss function is well-known [38, 70]:

$$\frac{\partial}{\partial\theta}\mathrm{KL}\left(\mu\|\mu_\theta\right) = \int\limits_{\mathcal{Y}} \frac{\partial}{\partial\theta}E_\theta(y)\mathrm{d}\mu(y) - \int\limits_{\mathcal{Y}}\left[\frac{\partial}{\partial\theta}E_\theta(y)\right]\mathrm{d}\mu_\theta(y). \tag{15}$$

The expectations which constitute the right-hand side of the (15) are estimated by Monte-Carlo which requires samples from the corresponding distributions μ an μ_θ. While the samples from μ are given, the sampling from μ_θ is carried out via Unadjusted Langevin Algorithm (ULA) [52], which iterates the discretized Langevin dynamics

$$Y_{l+1} = Y_l - \frac{\eta}{2}\frac{\partial}{\partial y}E_\theta(Y_l) + \sqrt{\eta}\xi_l\,, \quad \xi_l \sim \mathcal{N}(0,1) \tag{16}$$

starting from a simple prior distribution $Y_0 \sim \mu_0$, for L steps, with a small discretization step $\eta > 0$. In practice, there have been developed a lot of methods, which improve or circumvent the aforementioned procedure by informative initialization [28, 14], more sophisticated MCMC approaches [37, 51, 45], regularizations [13, 36], explicit auxiliary generators [69, 72, 27, 18]. The application of these improved EBM methods for the EOT problem we consider is a fruitful avenue for future work.

3 Method

We start our analysis by taking a close look at objective (13). The following proposition characterizes the inner $\min_{\mu_x}$ optimization problem arising in (13).

Proposition 1 (Optimizer of weak C_{EOT}-transform)**.** *Let $f \in \mathcal{C}(\mathcal{Y})$ and $x \in \mathcal{X}$. Then inner weak dual objective $\min_{\mu\in\mathcal{P}(\mathcal{Y})}\mathcal{G}_{x,f}(\mu)$ permits unique minimizer μ_x^f given by*

$$\frac{\mathrm{d}\mu_x^f(y)}{\mathrm{d}y} \stackrel{def}{=} \frac{1}{Z(f,x)}\exp\left(\frac{f(y)-c(x,y)}{\varepsilon}\right),$$

where $Z(f,x) \stackrel{def}{=} \int_{\mathcal{Y}}\exp\left(\frac{f(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}y$ is the conditional partition function.

Proof. In what follows, we will analyse the optimizers of the objective $\min_{\mu\in\mathcal{P}(\mathcal{Y})}\mathcal{G}_{x,f}(\mu)$.

$$\begin{aligned}
\mathcal{G}_{x,f}(\mu) &= \int_{\mathcal{Y}} c(x,y)\mathrm{d}\mu(y) + \varepsilon\int_{\mathcal{Y}}\log\frac{\mathrm{d}\mu(y)}{\mathrm{d}y}\mathrm{d}\mu(y) - \int_{\mathcal{Y}} f(y)\mathrm{d}\mu(y)\\
&= \varepsilon\int_{\mathcal{Y}}\left(\frac{c(x,y)-f(y)}{\varepsilon} + \log\frac{\mathrm{d}\mu(y)}{\mathrm{d}y}\right)\mathrm{d}\mu(y)\\
&= \varepsilon\int_{\mathcal{Y}}\Bigg(\underbrace{-\log Z(f,x)}_{\text{doesn't depend on } y} + \underbrace{\log Z(f,x) - \frac{f(y)-c(x,y)}{\varepsilon}}_{=-\log\frac{\mathrm{d}\mu_x^f(y)}{\mathrm{d}y}} + \log\frac{\mathrm{d}\mu(y)}{\mathrm{d}y}\Bigg)\mathrm{d}\mu(y)\\
&= -\varepsilon\log Z(f,x) + \varepsilon\int_{\mathcal{Y}}\left(-\log\frac{\mathrm{d}\mu_x^f(y)}{\mathrm{d}y} + \log\frac{\mathrm{d}\mu(y)}{\mathrm{d}y}\right)\mathrm{d}\mu(y)\\
&= -\varepsilon\log Z(f,x) + \varepsilon\mathrm{KL}\left(\mu\|\mu_x^f\right). \qquad (17)
\end{aligned}$$

The last equality holds true thanks to the fact, that $\mu_x^f \in \mathcal{P}_{\text{ac}}(\mathcal{Y})$ and $\forall y\in\mathcal{Y}: \frac{\mathrm{d}\mu_x^f(y)}{\mathrm{d}y} > 0$. This leads to the conclusion, that the absolute continuity of μ ($\mu \ll \lambda$, where λ is the Lebesgue measure on $\mathcal{Y}$) is equivalent to the absolute continuity of μ w.r.t. μ_x^f ($\mu \ll \mu_x^f$). In particular, if $\mu \notin \mathcal{P}_{\text{ac}}(\mathcal{Y})$, then the last equality in the derivations above reads as $+\infty = +\infty$. From (17) we conclude, that $\mu_x^f = \arg\min_{\mu\in\mathcal{P}(\mathcal{Y})}\mathcal{G}_{x,f}(\mu)$. □

As we can see from equation (17), functional $\mathcal{G}_{x,f}$ at optimal argument distribution μ_x^f takes the form

$$\mathcal{G}_{x,f}(\mu_x^f) = -\varepsilon\log Z(f,x) = -\varepsilon\log\left(\int_{\mathcal{Y}}\exp\left(\frac{f(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}y\right), \qquad (18)$$

resembling c,ε-transform (7). The difference is the integration measure, which is $\mathbb{Q}$ in case of (7) and standard Lebesgue one in our case (18).

From the *theoretical* point of view, such dissimilarity is not significant, that is why semi-dual (8) and weak dual (13) optimization problems are expected to share such properties as convergence, existence of optimizers and so on. In particular, there is the direct correspondence between Kantorovich potentials u^*, v^* solving dual EOT problem (6) and optimizers of weak dual objective $f^* \in \arg\sup_f F^w_{C_{\text{EOT}}}(f)$. Indeed, the optimal conditional plan $\pi^*(y|x)$ could be expressed through (u^*, v^*) by:

$$\mathrm{d}\pi^*(y|x) = \exp\left(\frac{u^*(x)+v^*(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\mathbb{Q}(y).$$

Let $\mathbb{Q}\in\mathcal{P}_{\text{ac}}(\mathcal{Y})$, and $E_{\mathbb{Q}}:\mathcal{Y}\to\mathbb{R}$ be the energy function of $\mathbb{Q}$, i.e. $\frac{\mathrm{d}\mathbb{Q}(y)}{\mathrm{d}y}\propto\exp\left(-E_{\mathbb{Q}}(y)\right)$. Then,

$$\frac{\mathrm{d}\pi^*(y|x)}{\mathrm{d}y} \propto \exp\left(\frac{v^*(y)-c(x,y)}{\varepsilon} - E_{\mathbb{Q}}(y)\right),$$

where $u^*(x)$ is omitted since it is hidden in the partition function. Therefore,

$$f^*_{v^*} = v^* - \varepsilon E_{\mathbb{Q}}$$

with the corresponding point-to-distribution map $x\mapsto\mu_x^{f^*_{v^*}}$ recovers the optimal plan, i.e $\mathrm{d}\pi^*(x,y) = \mathrm{d}\mu_x^{f^*_{v^*}}(y)\mathrm{d}\mathbb{P}(x)$ solves primal problem (4). Intuitively, this leads to the conclusion, that $f^*_{v^*}$ solves weak dual problem (13). This is indeed true thanks to direct accordance between potentials f and retrieved conditional distributions μ_x^f, see Proposition 1. Different f result in different distributions

$$\mathrm{d}\pi^f(x,y) \stackrel{\text{def}}{=} \mathrm{d}\mu_x^f(y)\mathrm{d}\mathbb{P}(x),$$

and only $f^*_{v^*}$ results in the optimal plan.

From the *practical* point of view, the difference between (18) and (7) is much more influential. Actually, it is the particular form of internal weak dual problem solution (18) that will allow us to utilize EBMs.

Thanks to (18), weak dual objective (13) permits the following reformulation:

$$\mathrm{EOT}_{c,\varepsilon}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{C}(\mathcal{Y})}\left\{-\varepsilon\int\limits_{\mathcal{X}}\log Z(f,x)\mathrm{d}\mathbb{P}(x) + \int\limits_{\mathcal{Y}} f(y)\mathrm{d}\mathbb{Q}(y)\right\}. \tag{19}$$

3.1 Optimization procedure

Following the standard machine learning practices, we parameterize functions $f \in \mathcal{C}(\mathcal{Y})$ as neural networks f_θ with parameters $\theta \in \Theta$ and derive the loss function corresponding to (19) by:

$$L(\theta) \stackrel{\text{def}}{=} -\varepsilon\int\limits_{\mathcal{X}}\log Z(f_\theta,x)\mathrm{d}\mathbb{P}(x) + \int\limits_{\mathcal{Y}} f_\theta(y)\mathrm{d}\mathbb{Q}(y). \tag{20}$$

The conventional way to optimize loss functions such as (20) is the stochastic gradient ascent. In the following proposition, we derive the gradient of $L(\theta)$ w.r.t. θ.

Proposition 2 (Gradient of the weak dual loss $L(\theta)$)**.** *It holds true that:*

$$\frac{\partial}{\partial\theta}L(\theta) = -\int\limits_{\mathcal{X}}\int\limits_{\mathcal{Y}}\left[\frac{\partial}{\partial\theta}f_\theta(y)\right]\mathrm{d}\mu_x^{f_\theta}(y)\mathrm{d}\mathbb{P}(x) + \int\limits_{\mathcal{Y}}\frac{\partial}{\partial\theta}f_\theta(y)\mathrm{d}\mathbb{Q}(y). \tag{21}$$

Proof. The direct derivations read as follows:

$$\begin{aligned}
\frac{\partial}{\partial\theta}L(\theta) = -\varepsilon\int\limits_{\mathcal{X}}\frac{\partial}{\partial\theta}\log Z(f_\theta,x)\mathrm{d}\mathbb{P}(x) + \int\limits_{\mathcal{Y}}\frac{\partial}{\partial\theta}f_\theta(y)\mathrm{d}\mathbb{Q}(y) \quad &= \\
-\varepsilon\int\limits_{\mathcal{X}}\frac{1}{Z(f_\theta,x)}\left\{\frac{\partial}{\partial\theta}\int\limits_{\mathcal{Y}}\exp\left(\frac{f_\theta(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}y\right\}\mathrm{d}\mathbb{P}(x) + \int\limits_{\mathcal{Y}}\frac{\partial}{\partial\theta}f_\theta(y)\mathrm{d}\mathbb{Q}(y) \quad &= \\
-\varepsilon\int\limits_{\mathcal{X}}\left\{\frac{1}{Z(f_\theta,x)}\int\limits_{\mathcal{Y}}\left[\frac{\frac{\partial}{\partial\theta}f_\theta(y)}{\varepsilon}\right]\exp\left(\frac{f_\theta(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}y\right\}\mathrm{d}\mathbb{P}(x) + \int\limits_{\mathcal{Y}}\frac{\partial}{\partial\theta}f_\theta(y)\mathrm{d}\mathbb{Q}(y) \quad &= \\
-\int\limits_{\mathcal{X}}\left\{\int\limits_{\mathcal{Y}}\left[\frac{\partial}{\partial\theta}f_\theta(y)\right]\underbrace{\frac{1}{Z(f_\theta,x)}\exp\left(\frac{f_\theta(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}y}_{=\mathrm{d}\mu_x^{f_\theta}(y)}\right\}\mathrm{d}\mathbb{P}(x) + \int\limits_{\mathcal{Y}}\frac{\partial}{\partial\theta}f_\theta(y)\mathrm{d}\mathbb{Q}(y) \quad &= \\
-\int\limits_{\mathcal{X}}\int\limits_{\mathcal{Y}}\left[\frac{\partial}{\partial\theta}f_\theta(y)\right]\mathrm{d}\mu_x^{f_\theta}(y)\mathrm{d}\mathbb{P}(x) + \int\limits_{\mathcal{Y}}\frac{\partial}{\partial\theta}f_\theta(y)\mathrm{d}\mathbb{Q}(y), &
\end{aligned}$$

which finishes the proof. □

Formula (21) resembles the gradient of Energy-based loss, formula (15). This allows us to look at the weak dual EOT problem (4) from the perspectives of EBMs. In order to emphasize the *novelty* of our approach, and, at the same time, establish the deep connection between the optimization of weak dual objective in the form (19) and EBMs, below we characterize the similarities and differences between standard EBMs optimization procedure and our proposed EOT-encouraged gradient ascent following $\partial L(\theta)/\partial\theta$.

Differences. In contrast to the case of EBMs, potential f_θ, optimized by means of loss function L, does not represent an energy function by itself. However, the tandem of cost function c and f_θ helps to recover the Energy functions of *conditional* distributions $\mu_x^{f_\theta}$:

$$E_{\mu_x^{f_\theta}}(y) = \frac{c(x,y)-f_\theta(y)}{\varepsilon}.$$

Therefore, one can sample from distributions $\mu_x^{f_\theta}$ following ULA (16) or using more advanced MCMC approaches [23, 61, 30, 56]. In practice, when estimating (21), we need samples $(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)$ from the distribution:

$$\mathrm{d}\pi^{f_\theta}(x, y) = \mathrm{d}\mu_x^{f_\theta}(y)\mathrm{d}\mathbb{P}(x).$$

They could be derived through the simple two-stage procedure:

1. Sample $x_1, \dots x_N \sim \mathbb{P}$,i.e., derive random batch from the source dataset.
2. Sample $y_1|x_1 \sim \mu_{x_1}^{f_\theta}, \dots, y_N|x_N \sim \mu_{x_N}^{f_\theta}$, e.g., performing Langevin steps (16).

Similarities. Besides a slightly more complicated two-stage procedure for sampling from generative distribution π^{f_θ}, the gradient ascent optimization with (21) is similar to the gradient descent with (15). This allows a practitioner to adopt the existing practically-efficient and fine-tuned architectures of EBMs, e.g., [14, 13, 19, 74], in order to solve EOT.

We summarize our findings and detail our optimization procedure in the Algorithm 1. This procedure is *basic*, i.e. for the sake of simplicity, we specially remove all technical subtleties and tricks which are typically used when optimizing EBMs. In particular, in our experiments, we additionally enrich the Algorithm 1 with persistent replay buffer [63, 46] keeping and occasionally updating samples (x, y) from π^{f_θ}. Particular implementation details are given in the Experiments section, §5.

In the conclusion of the section, we want to underline that our theoretical and practical setup allows performing theoretically-grounded **truly conditional** data generation by means of EBMs, which unlocks the data-to-data translation applications for EBM community. Note, that the existing approaches leveraging such applications with Energy-inspired methodology lack theoretical interpretability, see discussions in related works §4.1.

Algorithm 1: Entropic Optimal Transport via Energy-Based Modelling

Input : Source and target distributions $\mathbb{P}$ and $\mathbb{Q}$, accessible by samples;
Entropy regularization coefficient $\varepsilon > 0$, cost function $c(x, y) : \mathbb{R}^{D_x} \times \mathbb{R}^{D_y} \to \mathbb{R}$;
number of Langevin steps $K > 0$, Langevin discretization step size $\eta > 0$;
basic noise std $\sigma_0 > 0$; potential network $f_\theta : \mathbb{R}^{D_y} \to \mathbb{R}$, batch size $N > 0$.

Output : trained potential network f_{θ^*} recovering optimal conditional EOT plans

for $i = 1, 2, \dots$ **do**

- Derive batches $\{x_n\}_{n=1}^N = X \sim \mathbb{P}, \{y_n\}_{n=1}^N = Y \sim \mathbb{Q}$ of sizes N;
- Sample basic noise $Y^{(0)} \sim \mathcal{N}(0, \sigma_0)$ of size N;
- **for** $k = 1, 2, \dots, K$ **do**
  - Sample $Z^{(k)} = \{z_n^{(k)}\}_{n=1}^N$, where $z_n^{(k)} \sim \mathcal{N}(0, 1)$;
  - Obtain $Y^{(k)} = \{y_n^{(k)}\}_{n=1}^N$ with Langevin step:
  - $y_n^{(k)} \leftarrow y_n^{(k-1)} + \frac{\eta}{2\varepsilon} \cdot \text{stop\_grad}\Big(\frac{\partial}{\partial y} \left[f_\theta(y) - c(x_n, y) \right] \big|_{y = y_n^{(k-1)}} \Big) + \sqrt{\eta} z_n^{(k)}$
- $\widehat{L} \leftarrow -\frac{1}{N} \Bigg[\sum\limits_{y_n^{(K)} \in Y^{(K)}} f_\theta \left(y_n^{(K)} \right) \Bigg] + \frac{1}{N} \Bigg[\sum\limits_{y_n \in Y} f_\theta \left(y_n \right) \Bigg]$;
- Perform a gradient step over θ by using $\frac{\partial \widehat{L}}{\partial \theta}$;

4 Related works

In this section, we look over existing works which are relevant to our proposed method. We divide our survey into two main parts. At first, we discuss the Energy-Based approaches which introduces the ideas bearing resemblance to ours or tackles the similar practical problem setups. Secondly, we perform an overview of Optimal Transport solvers related to Entropy-regularized formulation of OT.

4.1 Energy-Based Models

Unpaired data-to-data translation. The world of EBMs is rich and full of various curious and interesting ideas. In particular, the problem in which one need to transform a point $x \in \mathcal{X}$ from a

source domain to corresponding points $y_1^x, y_2^x, \dots \subset \mathcal{Y}$ from a target domain, i.e., unpaired data-to-data translation [75], is also in the EBMs practitioner's area of interests. Since the major application of our proposed Energy-guided methodology is exactly unpaired data-to-data translation (treated as EOT problem), it is important to reveal the key ideas and principles underlining EBMs-advised studies dealing with the same problem. We found, that although the existing works [73, 74] demonstrate plausibly-looking practical results, they **lack theoretical verification**. Let the source and target domains of images are given by $\mathcal{X}$ and $\mathcal{Y}$ with the corresponding data distributions $\mathbb{P}$ and $\mathbb{Q}$. The authors of [73, 74] propose to learn the energy function of $\mathbb{Q}$ following the standard EBMs practices with the MCMC procedure (16), initialized by source samples $x \sim \mathbb{P}$. The theoretical properties of this approach remain unclear. Furthermore, being passed through MCMC, the obtained samples may lose the conditioning on the source samples. This may lead to degenerate image-to-image maps, where the information about the source data is fully neglected. In contrast, our proposed approach is free from the aforementioned problems. Besides, the entropy regularization coefficient ε, which is the hyperparameter of our method, can be tuned to reach the desired tradeoff between the conditioning power ($\varepsilon \to 0$ means almost deterministic $x \mapsto y$ mapping) and data variability ($\varepsilon \to +\infty$ makes the generated y to be independent of source datum x).

Energy function as the Lagrangian multiplier. In this paragraph, we speculate on the role potential f plays in objective (13). By rearranging $\min$ and integral sign $\int_{\mathcal{X}}$ in the right-hand side of (13), the trick introduced in [35], one can derive the $\max\min$ formulation of the dual weak problem by:

$$\max_{f}\min_{\pi\in\Pi(\mathbb{P})} \underbrace{\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(y,x) - \varepsilon\int_{\mathcal{X}} H(\pi(\cdot|x))\mathrm{d}\mathbb{P}(x)}_{\text{part I}} \underbrace{-\int_{\mathcal{X}}\int_{\mathcal{Y}} f(y)\mathrm{d}\pi(y|x)\mathrm{d}\mathbb{P}(x)}_{\text{part II}} \underbrace{+\int_{\mathcal{Y}} f(y)\mathrm{d}\mathbb{Q}(y)}_{\text{part III}}. \tag{22}$$

In the equation above, for each point $x \in \mathcal{X}$, the argument distribution μ_x arising in the objective (13) is treated as $\pi(\cdot|x)$. The functional under $\max\min$ operator in eqution (22) cosists of three parts. The first part is exactly the primal EOT functional $\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi - \varepsilon\int_{\mathcal{X}} H(\pi(\cdot|x))\mathrm{d}\mathbb{P}$, whose minimization w.r.t. $\pi \in \Pi(\mathbb{P},\mathbb{Q})$ unfolds primal EOT (4). The last two terms constitute the linear functional of f:

$$f \mapsto \int_{\mathcal{Y}} f(y)\mathrm{d}\big(\pi_y - \mathbb{Q}\big)(y),$$

and characterize the discrepancy (corresponding to f) between the target distribution $\mathbb{Q}$ and the projection of $\pi \in \Pi(\mathbb{P}) \subset \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ on $\mathcal{Y}$, which we denote as π_y. In other words, f resembles standard Lagrangian multiplier from constrained optimization theory. Actually, it lifts the target distribution constraint in (4).

It turns out, that this idea of constrained optimization and corresponding dual Lagrangian formulation was already caught by EBMs researchers [10] and finds it application in the Inverse Reinforcement Learning [29, 68]. In order to force the discriminators of GANs storing meaningful information about generative distribution, the authors of [10] propose to *calibrate* the GAN objective with a convex functional $K : \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$. They come up with constrained optimization problem [10, Eq. 2]:

$$\min_{\mu_\theta\in\mathcal{P}(\mathcal{Y})} K(\mu_\theta), \text{ such that } \mu_\theta = \mathbb{Q}$$

with the corresponding Lagrangian $\mathcal{L}(\mu_\theta, f) \stackrel{\text{def}}{=} K(\mu_\theta) - \int_{\mathcal{Y}} f(y)\mathrm{d}\mu_\theta(y) + \int_{\mathcal{Y}} f(y)\mathrm{d}\mathbb{Q}(y)$, maximized by f and minimized by parameters of generator μ_θ:

$$\max_{f}\min_{\mu_\theta} \mathcal{L}(\mu_\theta, f). \tag{23}$$

When considering K to be the Entropy $K(\mu) = H(\mu)$, the authors of [10] show, that potential f^* from the optimal saddle point (f^*, μ_θ^*) of the objective above is exactly the Energy function of optimal generative distribution μ_θ^* ($= \mathbb{Q}$). For this Entropy case it is interesting to note, that (23) can be recovered from our dual weak reformulation (22) when $c(x,y) = 0$ (or, equivalently, $\varepsilon \to +\infty$), i.e., when conditional distributions $\pi(y|x)$ become independent of source points $x \in \mathcal{X}$. Therefore, for $K = H$ our objective (13) is the generalization of (23), since it includes additional controllable OT-guided conditioning.

Concerning the Reinforcement Learning applications [29, 68], the authors of these papers work with Entropy case and treat Lagrangian multiplier f in (23) as the parametric reward function which should be learned to be in concordance with real agent's scenarios. This setup is quite different from what we consider, and we do not get into details of these researches.

4.2 Optimal Transport

Discrete OT. The Discrete OT is the specific domain in OT research area, which deals with distributions supported on finite discrete sets. In this case, the source and target distributions could be represented as normalized vectors with non-negative elements and a recovered (Entropic) OT plan is nothing more than a matrix. There have been developed various methods for solving Discrete OT problems [50], the most efficient of which deals with Discrete EOT [9]. In spite of good theoretically-grounded convergence guaranties, it is hard to adopt the DOT solvers for out-of-distribution sampling and mapping, which limits their applicability in real world scenarios.

Continuous OT. In the continuous setting, the discrete support assumption is eliminated and the source and target distributions become accessible only by samples from (limited) budgets. In this case, OT plans are typically parameterized with neural networks. They are typically optimized with the help of SGD-like methods by deriving random batches from the datasets. The approaches dealing with such practical setup are called *continuous OT solvers*.

There exists a lot of continuous OT solvers [41, 32, 15, 71, 53, 20] which differentiate by underlining theoretical principles (e.g., consider primal or dual OT formulation, utilize properties of specific cost functions), technical implementations, ultimate problem formulations, and range of OT-related tasks to be solved by corresponding solvers. In particular, the majority of the methods deals with Monge's formulation of OT problem rather then Kantorovich's one (1) This means that they search for OT plan given by deterministic source to target map $T : \mathcal{X} \to \mathcal{Y}$:

$$\mathrm{d}\pi(x, y) = \mathrm{d}\mathbb{P}(x)\delta[y = T(x)].$$

Only a limited number of approaches are capable of leveraging OT problems which require stochastic mapping, and, therefore, potentially applicable for our EOT case. Note, that in this paragraph we exclude methods designed *specifically* for EOT and cover them in further narrative. The recent line of works [35, 34, 4] considers dual formulation of weak OT problem [24] and comes up with $\max\min$ objective similar to (22), but for various weak [35, 34] and even general [4] cost functions. However, their proposed methodology requires estimation of weak cost by samples, which complicates its application for EOT case. An alternative concept [71] is designed for classical Kantorovich OT, but can be easily adopted for weak OT. The authors of [71] work with primal OT formulation (1) and lift boundary source and target distribution constraints by WGAN losses. Apart from the fact, that their method may result in biased OT plans, it also utilizes sample estimation of corresponding functionals. Therefore, this approach also can not be directly adapted for our EOT setup.

Entropy-regularized OT. To begin with, we underline that, to the best of our knowledge, all continuous EOT solvers are based either on KL-guided formulation (2) or unconditional entropic formulation (3) with its connection to Schrödinger bridge problem. Our proposed approach seems to be the first which take the advantage of conditional entropic formulation (4). Historically, one of the first large-scale technique for EOT is presented in [22] and developed in [59]. It is based on dual form of (2). The authors of [59] propose to treat the optimization problem (6) by parameterizing the dual potentials $u = u_\theta$ and $v = v_\phi$ as deep neural networks and alternating stochastic gradient ascent steps with respect to the parameters θ and ϕ by using random batches from $\mathbb{P}$ and $\mathbb{Q}$. However, there is an important concern with this approach. Even if one managed to learn an optimal pair of Kantorovich potentials (u^*, v^*), there is no direct way to model, e.g., sample from, optimal conditional plans $\pi^*(y|x)$:

$$\mathrm{d}\pi^*(y|x) = \exp\left(\frac{u^*(x) + v^*(y) - c(x, y)}{\varepsilon}\right)\mathrm{d}\mathbb{Q}(y) \propto \exp\left(\frac{v^*(y) - c(x, y)}{\varepsilon}\right)\mathrm{d}\mathbb{Q}(y),$$

which is typically needed in data-to-data translation tasks. The problem is that target distribution $\mathbb{Q}$ is unknown and given only by dataset samples. In order to leverage this issue, [11] proposes to employ separate score-based model approximating $\mathbb{Q}$. Indeed, having neural network approximations (u^*_θ, v^*_ϕ) of dual Kantorovich potentials (u^*, v^*) and score function approximation $s_\mathbb{Q}(y) \approx \boldsymbol{\nabla} \log \frac{\mathrm{d}\mathbb{Q}(y)}{\mathrm{d}y}$, the sampling from $\pi^*(y|x)$ can be done by, e.g., performing Langevin steps (16). The gradient of Energy

function in this case is the combination of $s_{\mathbb{Q}}$, v_ϕ^*, and cost function c. The utilization of MCMC at inference makes the work [11] to be the closest to ours. The main practical differences between our method and [11] and disadvantages of the latter which we try to overcome are stated below:

1. As well as [59], the authors of [11] optimize dual potentials (u, v) following (6). This procedure is unstable for small ε, since it requires the exponentiation of large numbers, which are of order ε^{-1}. At the same time, small ε regime is practically important for some downstream applications when one need a close-to-deterministic plan between $\mathcal{X}$ and $\mathcal{Y}$ domains. Unlike alternating optimization of u and v (6), our Energy-based approach does not require exponent computation and can be adapted for a small ε by proper adjusting of ULA (16) parameters (step size, number of steps, temperature).
2. In order to sample from conditional plans $\pi(y|x)$, [11] needs *three* models, including a third-party score-based model. In contrast, our algorithm results in *single* potential f_θ which captures all the necessary information about point-to-distribution map $x \mapsto \pi(\cdot|x)$.

The alternative line of EOT solvers [6, 26, 7] is based on the connection between primal EOT (2) and the *Schrödinger bridge* problem. The idea is to model the EOT plan as a time-dependent stochastic process with learnable drift and diffusion terms, starting from $\mathbb{P}$ an initial time and approaching $\mathbb{Q}$ at a final time. This process is optimized to be close to standard Brownian motion, see [7, Problem 4.1]. Although the developed practical procedures show good results on large-scale data domains, see [6, Fig. 7], [26, Fig. 4], [7, Fig. 4], they typically require simulating the learned stochastic process via, e.g., Euler-Maruyama algorithm, and *store* the whole simulation history. This is because each intermediate sample contributes to final loss function. In other words, the optimization is rather **costly** in terms of computational resources. Moreover, the methods [6, 26, 7] work primarily with the cost function given by squared l_2 norm and hardly could be accommodated for the more general case. Anyway, our approach much differs from aforementioned methods from both practical and theoretical viewpoints. And we emphasize that in our work we pave a principled connection between EBMs and EOT problem. We believe, that further studies will manage to empower EOT research domain with recent Energy models, capable to efficiently sort out truly large-scale setups [13, 19, 74].

5 Experiments

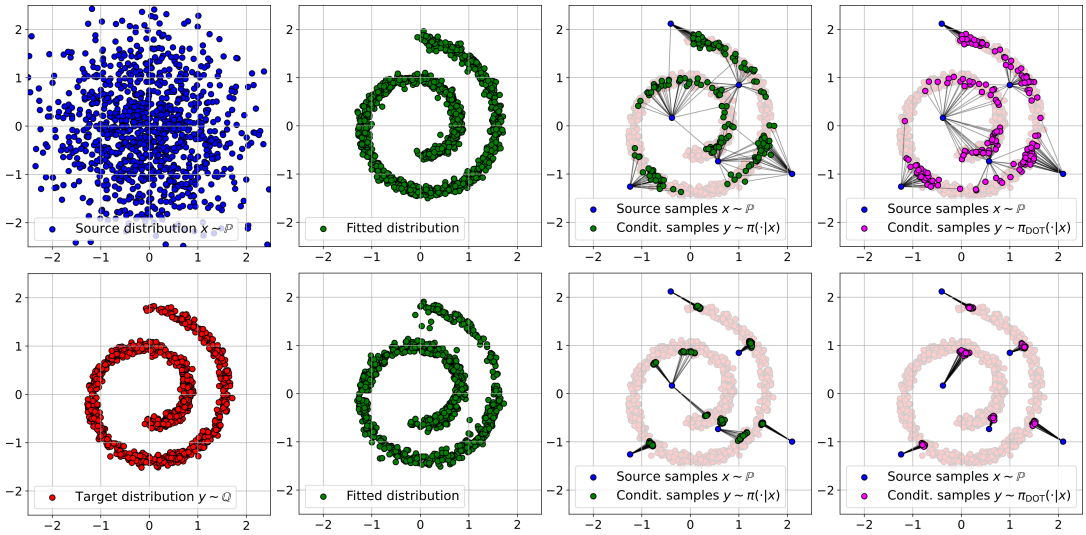
In what follows, we demonstrate the performance of our method on toy 2D scenario and Colored MNIST images transformation benchmark [26]. Note, that our aim here is to show that the proposed Energy-guided EOT methodology actually works. We leave technically-saturated adjustment of our approach for practically-important large-scale applications for future work. Both in 2D and image cases the cost function is chosen to be squared halved l_2 norm: $c(x, y) = \frac{1}{2}\|x - y\|_2^2$.

Our code is written in PyTorch. The experiments are conducted on a single GTX 1080Ti. The actual neural network architectures as well as practical training setups are disclosed in the corresponding subsections. In all our experiments, we take the advantage of replay buffer $\mathcal{B}$ with the parameters similar to [14]. When training, the ULA algorithm is initialized by samples from $\mathcal{B}$ with probability $p = 0.95$ and from uniform noise with probability $1 - p = 0.05$.

5.1 Toy 2D experiment

We apply our method for 2D *Gaussian→Swissroll* distributions modification task and demonstrate the qualitative results on Figure 1 for Entropy regularization coefficients $\varepsilon = 0.1, 0.001$. For this experiment, we parameterize the potential f_θ as MLP with two hidden layers and *LeakyReLU*(`negative_slope`$= 0.2$) as the activation function. Each hidden layer has 256 neurons. The metaparameters of the Algorithm 1 as follows: $K = 100, \sigma_0 = 1, N = 1024$, see the meaning of each particular variable in the algorithm listing. The Langevin discretization steps are $\eta = 0.05$ for $\varepsilon = 0.1$ and $\eta = 0.005$ for $\varepsilon = 0.001$. The reported numbers are chosen for reasons of training stability.

Figure 1b shows that our method succeeds in transforming source distribution $\mathbb{P}$ to target distribution $\mathbb{Q}$ for both Entropy regularization coefficients. In order to ensure that our approach learns optimal conditional plans $\pi^*(y|x)$ well, and correctly solves EOT problem, we provide Figures 1c and 1d. On these images, we pick six points $x \in \mathcal{X}$ and demonstrate samples from the conditional plans



(a) Input and target distributions $\mathbb{P}$ and $\mathbb{Q}$.

(b) **Our** fitted distributions; $\varepsilon = 10^{-1}$ (above), $\varepsilon = 10^{-3}$ (below).

(c) **Our** fitted conditional plans $\pi(\cdot|x)$; $\varepsilon = 10^{-1}$ (above), $\varepsilon = 10^{-3}$ (below)

(d) Discrete conditional plans $\pi_{\text{DOT}}(\cdot|x)$; $\varepsilon = 10^{-1}$ (above), $\varepsilon = 10^{-3}$ (below)

Figure 1: Performance of Energy-guided EOT on *Gaussian→Swissroll* 2D setup.

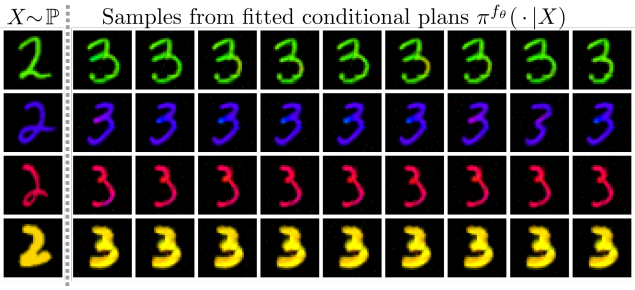


Figure 2: Performance of Energy-guided EOT on Colored Mnist "2"→"3" adaptation

$\pi(\cdot|x)$, obtained either by our method ($\pi(\cdot|x) = \mu_x^{f_\theta}$) or by discrete EOT solver [17]. Note that, in contrast to our approach, the samples generated by the discrete EOT solver come solely from the training dataset.

5.2 Colored MNIST

In this subsection, we consider Colored MNIST, proposed in [26, Sec. 5.3]. Following [26], we set source and target distributions $\mathbb{P}$ and $\mathbb{Q}$ to be colored handwritten images of digits “2” and “3” accordingly. For solving the corresponding EOT problem, we adopt the technical choices from [14]. In particular, learned potential f_θ is build upon ResNet architecture. Actually, our code is the minimal adaptation of `https://github.com/rosinality/igebm-pytorch`. We do nothing but embed the cost function gradient computation when performing Langeving steps, leaving all the training hyperparameters unchanged.

In this experiment, we choose $\varepsilon = 2.25 \cdot 10^{-5}$. For now we observe training instabilities when experimenting with Entropy regularization coefficients $\tilde{\varepsilon} \gg \varepsilon$, which are probably due to improper hyperparameters adjustment. In the future versions of the manuscript, we look forward to unraveling all these technical tuning businesses.

The qualitative results are presented in the Figure 2. As we can see, our learned EOT successfully preserves color and geometry of the transformed images.

References

[1] David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020.

[2] David Alvarez-Melis, Yair Schiff, and Youssef Mroueh. Optimizing functionals on the space of probabilities with input convex neural networks. *Transactions on Machine Learning Research*, 2022.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[4] Arip Asadulaev, Alexander Korotin, Vage Egiazarian, and Evgeny Burnaev. Neural optimal transport with general cost functionals. *arXiv preprint arXiv:2205.15403*, 2022.

[5] Julio Backhoff-Veraguas, Mathias Beiglböck, and Gudmun Pammer. Existence, duality, and cyclical monotonicity for weak transport costs. *Calculus of Variations and Partial Differential Equations*, 58(6):203, 2019.

[6] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[7] Tianrong Chen, Guan-Horng Liu, and Evangelos Theodorou. Likelihood training of schrödinger bridge using forward-backward SDEs theory. In *International Conference on Learning Representations*, 2022.

[8] Christian Clason, Dirk A Lorenz, Hinrich Mahler, and Benedikt Wirth. Entropic regularization of continuous optimal transport problems. *Journal of Mathematical Analysis and Applications*, 494(1):124432, 2021.

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

[10] Zihang Dai, Amjad Almahairi, Philip Bachman, Eduard Hovy, and Aaron Courville. Calibrating energy-based generative adversarial networks. In *International Conference on Learning Representations*, 2017.

[11] Max Daniels, Tyler Maunu, and Paul Hand. Score-based generative neural networks for large-scale optimal transport. *Advances in neural information processing systems*, 34:12955–12965, 2021.

[12] Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3483–3491, 2018.

[13] Yilun Du, Shuang Li, B. Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy based models. In *Proceedings of the 38th International Conference on Machine Learning (ICML-21)*, 2021.

[14] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

[15] Jiaojiao Fan, Shu Liu, Shaojun Ma, Yongxin Chen, and Hao-Min Zhou. Scalable computation of monge maps with general costs. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

[16] Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen Chen. Scalable computations of wasserstein barycenter via input convex neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1571–1581. PMLR, 18–24 Jul 2021.

[17] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[18] Ruiqi Gao, Erik Nijkamp, Diederik P Kingma, Zhen Xu, Andrew M Dai, and Ying Nian Wu. Flow contrastive estimation of energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7518–7528, 2020.

[19] Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P Kingma. Learning energy-based models by diffusion recovery likelihood. In *International Conference on Learning Representations*, 2021.

[20] Milena Gazdieva, Litu Rout, Alexander Korotin, Andrey Kravchenko, Alexander Filippov, and Evgeny Burnaev. An optimal transport perspective on unpaired image super-resolution. *arXiv preprint arXiv:2202.01116*, 2022.

[21] Aude Genevay. *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres (ComUE), 2019.

[22] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

[23] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[24] Nathael Gozlan, Cyril Roberto, Paul-Marie Samson, and Prasad Tetali. Kantorovich duality for general transport costs and applications. *Journal of Functional Analysis*, 273(11):3327–3405, 2017.

[25] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[26] Nikita Gushchin, Alexander Kolesov, Alexander Korotin, Dmitry Vetrov, and Evgeny Burnaev. Entropic neural optimal transport via diffusion processes. *arXiv preprint arXiv:2211.01156*, 2022.

[27] Tian Han, Erik Nijkamp, Xiaolin Fang, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Divergence triangle for joint training of generator model, energy-based model, and inferential model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8670–8679, 2019.

[28] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[29] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

[30] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[31] Weining Hu, Meng Li, and Xiaomeng Ju. Improved cyclegan for image-to-image translation.

[32] Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. In *International Conference on Learning Representations*, 2021.

[33] Alexander Korotin, Vage Egiazarian, Lingxiao Li, and Evgeny Burnaev. Wasserstein iterative networks for barycenter estimation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[34] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Kernel neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023.

[35] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023.

[36] Rithesh Kumar, Sherjil Ozair, Anirudh Goyal, Aaron Courville, and Yoshua Bengio. Maximum entropy generators for energy-based models. *arXiv preprint arXiv:1901.08508*, 2019.

[37] John Lawson, George Tucker, Bo Dai, and Rajesh Ranganath. Energy-inspired models: Learning with sampler-induced distributions. *Advances in Neural Information Processing Systems*, 32, 2019.

[38] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[39] Guansong Lu, Zhiming Zhou, Jian Shen, Cheng Chen, Weinan Zhang, and Yong Yu. Large-scale optimal transport via adversarial training with cycle-consistency. *arXiv preprint arXiv:2003.06635*, 2020.

[40] Shaojun Ma, Shu Liu, Hongyuan Zha, and Haomin Zhou. Learning stochastic behaviour from aggregate data. In *International Conference on Machine Learning*, pages 7258–7267. PMLR, 2021.

[41] Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

[42] Simone Di Marino and Augusto Gerolin. An optimal transport approach for the schrödinger bridge problem and convergence of sinkhorn algorithm. *Journal of Scientific Computing*, 85(2):27, 2020.

[43] Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 34:15243–15256, 2021.

[44] Tuan Nguyen, Trung Le, He Zhao, Quan Hung Tran, Truyen Nguyen, and Dinh Phung. Most: Multi-source domain adaptation via optimal transport for student-teacher learning. In *Uncertainty in Artificial Intelligence*, pages 225–235. PMLR, 2021.

[45] Erik Nijkamp, Ruiqi Gao, Pavel Sountsov, Srinivas Vasudevan, Bo Pang, Song-Chun Zhu, and Ying Nian Wu. MCMC should mix: Learning energy-based model with neural transport latent space MCMC. In *International Conference on Learning Representations*, 2022.

[46] Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5272–5280, 2020.

[47] Tomohiro Nishiyama. Convex optimization on functionals of probability densities. *arXiv preprint arXiv:2002.06488*, 2020.

[48] Marcel Nutz. Introduction to entropic optimal transport. *Lecture notes, Columbia University*, 2021.

[49] François-Pierre Paty and Marco Cuturi. Regularized optimal transport is ground cost adversarial. In *International Conference on Machine Learning*, pages 7532–7542. PMLR, 2020.

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

[51] Yixuan Qiu, Lingsong Zhang, and Xiao Wang. Unbiased contrastive divergence algorithm for training energy-based latent variable models. In *International Conference on Learning Representations*, 2020.

[52] Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

[53] Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. In *International Conference on Learning Representations*, 2022.

[54] Giorgi Rukhaia. *A FreeForm Optics Application of Entropic Optimal Transport*. PhD thesis, Université Paris sciences et lettres, 2021.

[55] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798, 2007.

[56] Sergey Samsonov, Evgeny Lagutin, Marylou Gabrié, Alain Durmus, Alexey Naumov, and Eric Moulines. Local-global MCMC kernels: the best of both worlds. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

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

[58] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[59] Vivien Seguy, Bharath Bhushan Damodaran, Remi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large scale optimal transport and mapping estimation. In *International Conference on Learning Representations*, 2018.

[60] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[61] Øivind Skare, Erik Bølviken, and Lars Holden. Improved sampling-importance resampling and reduced bias importance sampling. *Scandinavian Journal of Statistics*, 30(4):719–737, 2003.

[62] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

[63] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071, 2008.

[64] Francisco Vargas, Pierre Thodoroff, Austen Lamacraft, and Neil Lawrence. Solving schrödinger bridges via maximum likelihood. *Entropy*, 23(9):1134, 2021.

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

[66] Rui Wang, Ruiyi Zhang, and Ricardo Henao. Wasserstein uncertainty estimation for adversarial domain matching. *Frontiers in big Data*, 5, 2022.

[67] Yisen Wang, Bo Dai, Lingkai Kong, Sarah Monazam Erfani, James Bailey, and Hongyuan Zha. Learning deep hidden nonlinear dynamics from aggregate data. *arXiv preprint arXiv:1807.08237*, 2018.

[68] Huang Xiao, Michael Herman, Joerg Wagner, Sebastian Ziesche, Jalal Etesami, and Thai Hong Linh. Wasserstein adversarial imitation learning. *arXiv preprint arXiv:1906.08113*, 2019.

[69] Jianwen Xie, Yang Lu, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. Cooperative training of descriptor and generator networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):27–45, 2018.

[70] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *International Conference on Machine Learning*, pages 2635–2644. PMLR, 2016.

[71] Yujia Xie, Minshuo Chen, Haoming Jiang, Tuo Zhao, and Hongyuan Zha. On scalable and efficient computation of large scale optimal transport. In *International Conference on Machine Learning*, pages 6882–6892. PMLR, 2019.

[72] Xuwang Yin, Shiying Li, and Gustavo K Rohde. Learning energy-based models with adversarial training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 209–226. Springer, 2022.

[73] Yang Zhao and Changyou Chen. Unpaired image-to-image translation via latent energy transport. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16418–16427, 2021.

[74] Yang Zhao, Jianwen Xie, and Ping Li. Learning energy-based generative models via coarse-to-fine expanding and sampling. In *International Conference on Learning Representations*, 2021.

[75] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

A Restoration of the optimal EOT plan through the weak dual objective

In what follows, we establish the possibility to recover the optimal transport plan π^* which solves the primal EOT problem (4) using the optimizers of weak dual objective (13). We note, that our analysis can be seen as the combination of [35, Lemma 4] and [34, Thm. 2], yet we don't parameterize the conditional distributions μ_x arising in objective (13) as stochastic pushforward maps $T(x,\cdot)_\sharp\mathbb{S}$, as it was done in [35, 34]. We start by the following

Lemma 1 (The optimal EOT plan is the solution of the weak dual problem)**.** *Let* $\pi^* \in \Pi(\mathbb{P},\mathbb{Q})$ *solves objective* (4) *and* $f^* \in \mathcal{C}(\mathcal{Y})$ *be a maximizer of* (13)*. Then* $\pi^*(\cdot|x) = \arg\min_{\mu_x} \mathcal{G}_{x,f^*}(\mu_x)$ $\mathbb{P}$*-almost surely.*

Proof. Since f^* is the optimal potential, (13) reads as:

$$\text{EOT}_{c,\varepsilon}(\mathbb{P},\mathbb{Q}) = \int\limits_{\mathcal{X}} \min_{\mu_x\in\mathcal{P}(\mathcal{Y})} \mathcal{G}_{x,f^*}(\mu_x)\mathrm{d}\mathbb{P}(x) + \int\limits_{\mathcal{Y}} f^*(y)\mathrm{d}\mathbb{Q}(y). \tag{24}$$

On the other hand, since π^* is optimal, then:

$$\text{EOT}_{c,\varepsilon}(\mathbb{P},\mathbb{Q}) = \int\limits_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi^*(x,y) - \varepsilon\int\limits_{\mathcal{X}} H(\pi^*(y|x))\mathrm{d}\mathbb{P}(x) =$$

$$\int\limits_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi^*(x,y) - \varepsilon\int\limits_{\mathcal{X}} H(\pi^*(y|x))\mathrm{d}\mathbb{P}(x) - \underbrace{\int\limits_{\mathcal{X}\times\mathcal{Y}} f^*(y)\mathrm{d}\pi^*(x,y)}_{=\int_{\mathcal{Y}} f^*(y)\mathrm{d}\mathbb{Q}(y)} + \int\limits_{\mathcal{Y}} f^*(y)\mathrm{d}\mathbb{Q}(y) =$$

$$\int\limits_{\mathcal{X}}\left\{\int\limits_{\mathcal{Y}} c(x,y)\mathrm{d}\pi^*(y|x) - \varepsilon H(\pi^*(y|x)) - \int\limits_{\mathcal{Y}} f^*(y)\mathrm{d}\pi^*(y|x)\right\}\mathrm{d}\mathbb{P}(x) + \int\limits_{\mathcal{Y}} f^*(y)\mathrm{d}\mathbb{Q}(y) =$$

$$\int\limits_{\mathcal{X}} \mathcal{G}_{x,f^*}(\pi^*(\cdot|x))\mathrm{d}\mathbb{P}(x) + \int\limits_{\mathcal{Y}} f^*(y)\mathrm{d}\mathbb{Q}(y). \tag{25}$$

The expressions (24) and (25) are equal to each other, which results in the following:

$$\int\limits_{\mathcal{X}} \min_{\mu_x\in\mathcal{P}(\mathcal{Y})} \mathcal{G}_{x,f^*}(\mu_x)\mathrm{d}\mathbb{P}(x) = \int\limits_{\mathcal{X}} \mathcal{G}_{x,f^*}(\pi^*(\cdot|x))\mathrm{d}\mathbb{P}(x). \tag{26}$$

For each point $x\in\mathcal{X}$ the functional $\mu \mapsto \mathcal{G}_{x,f^*}(\mu) = \int_{\mathcal{Y}} c(x,y)\mathrm{d}\mu(y) - \varepsilon H(\mu) - \int_{\mathcal{Y}} f^*(y)\mathrm{d}\mu(y)$ is strictly convex as the combination of linear terms (w.r.t. μ) and strictly convex negative entropy term. Therefore, the minimizer of $\mathcal{G}_{x,f^*}$ is unique, $\mu_x^* \stackrel{\text{def}}{=} \arg\min_{\mu_x} \mathcal{G}_{x,f^*}(\mu_x)$, and

$$\forall \mu_x \in \mathcal{P}(\mathcal{Y}),\ \mu_x \neq \mu_x^* :\quad \mathcal{G}_{x,f^*}(\mu_x) > \mathcal{G}_{x,f^*}(\mu_x^*) = \min_{\mu_x\in\mathcal{P}(\mathcal{Y})} \mathcal{G}_{x,f^*}(\mu_x). \tag{27}$$

The equality (26) combined with (27) leads to the conclusion, that $\mu_x^* = \pi^*(\cdot|x)$ holds $\mathbb{P}$-a.s., which finishes the proof. □

Lemma 1 results in the following:

Corollary 1 (The optimizers of weak dual objective recover the optimal EOT plan)**.** *Let $f^* \in \mathcal{C}(\mathcal{Y})$ is a maximizer of* (13)*. For each point $x \in \mathcal{X}$ consider the distribution $\mu_x^* = \arg\min_{\mu_x} \mathcal{G}_{x,f^*}(\mu_x)$. Then the distribution $\pi^* \in \Pi(\mathbb{P})$ defined by*

$$\mathrm{d}\pi^*(x, y) = \mathrm{d}\mu_x^*(y)\mathrm{d}\mathbb{P}(x)$$

is the optimal transport plan.