Entropy-Regularized Optimal Transport

Mikhail Persiianov, Grigoriy Ksenofontov

Moscow, 2026

Overview

Recap: domain translation and OT problem

Entropy-regularized OT

Weak optimal transport

Recap: energy-based models

Energy-guided EOT

Recap: domain translation and OT problem

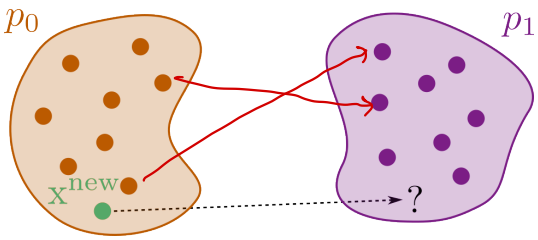
Domain translation: formal problem

Unsupervised setting.

We observe two datasets

$$\{x_i\}_{i=1}^{N} \sim p_0, \qquad \{y_j\}_{j=1}^{M} \sim p_1,$$

with **no paired correspondences**.



Goal: learn a mapping

$$T : \mathcal{X} \to \mathcal{Y}$$

such that

$$T_\sharp p_0 \approx p_1.$$

Key difficulty: the pushforward constraint does not uniquely determine T.

Ambiguity of distribution matching

The constraint

$$T_\sharp p_0 = p_1$$

does **not** uniquely determine the mapping.

Simple example.
If

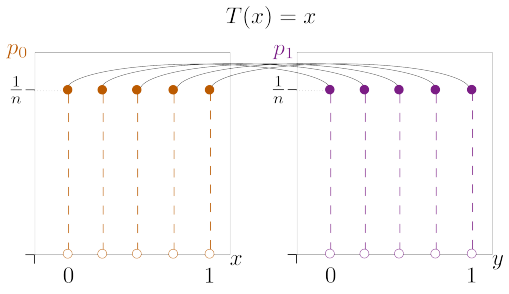
$$p_0 = p_1 = \mathrm{Uniform}[0, 1],$$

then both

$$T(x) = x \qquad \text{and} \qquad T(x) = 1 - x$$

satisfy the pushforward condition.

Distribution matching alone does not enforce semantic alignment.



Why optimal transport?

Since infinitely many mappings satisfy

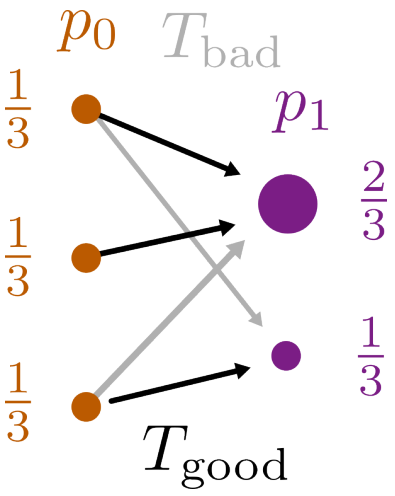
$$T_\sharp p_0 = p_1,$$

we need a principle to select one.

Optimal transport principle:
Among all admissible maps, choose the one minimizing transport cost:

$$\min_{T_\sharp p_0 = p_1} \mathbb{E}_{x \sim p_0}\big[c(x, T(x))\big].$$

OT selects the least distorted mapping.



Earth Mover's Intuition

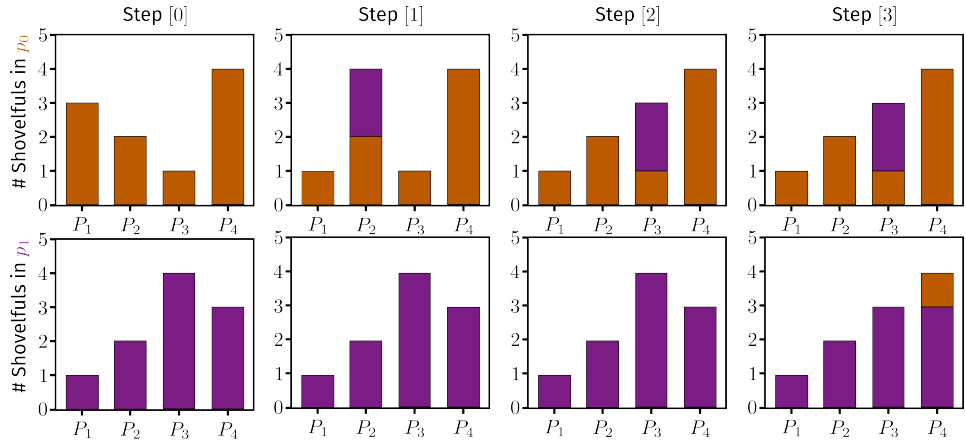


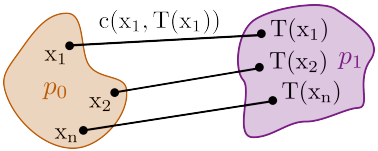
Figure 1: Step-by-step movement of mass from p_0 to p_1.[1]

[1]Lilian Weng (2017). **"From GAN to WGAN".** In: *lilianweng.github.io*. URL: https://lilianweng.github.io/posts/2017-08-20-gan/.

Monge vs. Kantorovich Formulations of Optimal Transport

Monge's formulation

$$\mathrm{OT}_c(p_0, p_1) = \min_{T_\sharp p_0 = p_1} \mathbb{E}_{x\sim p_0}\big[c(x, T(x))\big]$$

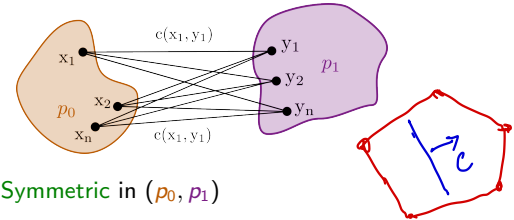


- Asymmetric constraint: $T_\sharp p_0 = p_1$
- No mass splitting $\Rightarrow$ solution may not exist
- Example: $p_0 = \delta_0, \quad p_1 = \frac{1}{2}(\delta_{-1} + \delta_1)$

Kantorovich's formulation

min <C, X>
X: X^T 1 = b X 1 = a

$$\mathrm{OT}_c(p_0, p_1) = \min_{\pi \in \Pi(p_0, p_1)} \mathbb{E}_{(x,x')\sim\pi}\big[c(x, x')\big] \quad (\mathrm{OT})$$



- Symmetric in (p_0, p_1)
- Allows mass splitting
- A minimizer π^* always exists

Kantorovich's formulation is a **relaxation** of Monge's. For many (p_0, p_1), the optimal plan is deterministic: $\pi^*(\cdot|x) = \delta_{T^*(x)}$, and both formulations yield the same $\mathbb{W}_1$.

Practical challenges of unregularized OT

Even though Kantorovich OT is well-defined:

Computational issues:

- Solving OT exactly is expensive
- Discrete problem scales as $\mathcal{O}(n^3)$

Statistical issues:

- Empirical OT can be unstable in high dimension
- Solutions may overfit finite samples

Modeling issues in domain translation:

- Deterministic transport may collapse modes
- No built-in mechanism encouraging diversity

How can we modify OT to achieve this?

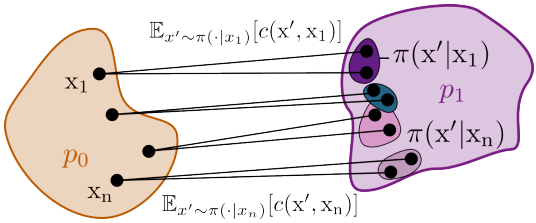
Entropy-regularized OT

Optimal Conditional Plans

Let π^* solve the equation EOT problem. By disintegration:

$$\pi^*(x, x') = \pi^*(x)\,\pi^*(x'|x) = p_0(x)\,\pi^*(x'|x).$$

The family $\{\pi^*(\cdot|x)\}_{x\in\mathcal{X}}$ are called **optimal conditional plans**.



Interpretation:

- For each source point x, we obtain a distribution over targets.
- Transport becomes **stochastic**.
- These conditionals define a generative mechanism: $x \longmapsto y \sim \pi^*(\cdot|x)$.

Entropic Optimal Transport (EOT)

Entropy regularization[2] of equation OT problem guarantees uniqueness and stochastic mappings.

Entropic Optimal Transport (EOT) solves:

$$\mathrm{EOT}_{c,\varepsilon}(p_0, p_1) = \min_{\pi \in \Pi(p_0, p_1)} \mathbb{E}_{x,y\sim\pi}[c(x,y)] - \varepsilon \mathbb{E}_{x\sim p_0} \mathrm{H}(\pi(\cdot|x)), \qquad \text{(EOT)}$$

where $H(\pi)$ is the entropy of π and $\varepsilon > 0$ controls regularization.

$H(\pi) := \int \pi \log \pi \, dx \, dy$

Connection: Equivalent to the *Static Schrödinger Bridge* problem:

$$\pi^* = \arg\min_{\pi \in \Pi(p_0, p_1)} \mathrm{KL}\left(\pi \| \pi^{\mathrm{ref}}\right), \qquad \text{(SB)}$$

where the aim of the problem is to find the transport plan π closest to π^{ref} in terms of the Kullback-Leibler (KL) divergence.

[2] Marco Cuturi (2013). **"Sinkhorn distances: Lightspeed computation of optimal transport".** In: *Advances in neural information processing systems* 26.

Equivalence of EOT Formulations

The proof of this equivalence is straightforward:

$$\min_{\pi \in \Pi(p_0, p_1)} \mathrm{KL}\left(\pi \| \pi^{\mathrm{ref}}\right) = \min_{\pi \in \Pi(p_0, p_1)} \mathbb{E}_{x,y \sim \pi} \log \frac{\pi(x,y)}{\pi^{\mathrm{ref}}(x,y)} = \tag{1}$$

$$\min_{\pi \in \Pi(p_0, p_1)} \left\{ \mathbb{E}_{x,y \sim \pi} \underbrace{\left[-\log \pi^{\mathrm{ref}}(x,y) \right]}_{\stackrel{\mathrm{def}}{=} c(x,y)} - \mathrm{H}(\pi) \right\} = \tag{2}$$

$$\min_{\pi \in \Pi(p_0, p_1)} \left\{ \mathbb{E}_{x,y \sim \pi}[c(x,y)] - \mathrm{H}(\pi) \right\}. \tag{3}$$

Using the equivalence of the entropic formulations we conclude that equation equation SB is equivalent to equation EOT for $\varepsilon = 1$.

Observation: The cost function $c(x, y)$ defines a *reference measure* that determines the mapping we aim to reconstruct in the forward problem equation EOT.

Equivalent Forms of EOT

When p_0, p_1 are absolutely continuous,

$$\mathrm{KL}\left(\pi \| p_0 \otimes p_1\right) = -H(\pi) + H(p_0) + H(p_1).$$

Therefore, the following formulations are equivalent (up to additive constants):

- $\mathbb{E}_\pi[c] + \varepsilon \mathrm{KL}\left(\pi \| p_0 \otimes p_1\right)$
- $\mathbb{E}_\pi[c] - \varepsilon H(\pi)$
- $\mathbb{E}_\pi[c] - \varepsilon \mathbb{E}_{x \sim p_0}[H(\pi(\cdot|x))]$

In the sequel, we mainly use

$$\mathrm{EOT}_{c,\varepsilon}(p_0, p_1) = \min_{\pi \in \Pi(p_0, p_1)} \mathbb{E}_\pi[c] - \varepsilon \mathbb{E}_{x \sim p_0}[H(\pi(\cdot|x))].$$

Equivalent Forms of Entropic OT (I)

Assume p_0, p_1 are absolutely continuous and $\pi \in \Pi(p_0, p_1)$.

Relation between KL and entropy.

$$\mathrm{KL}\left(\pi \| p_0 \otimes p_1\right) = \int \log\left(\frac{d\pi}{d p_0 \, d p_1}\right) d\pi.$$

Expand the logarithm:

$$\log\left(\frac{d\pi}{d p_0 \, d p_1}\right) = \log \frac{d\pi}{dx \, dy} - \log \frac{d p_0}{dx} - \log \frac{d p_1}{dy}.$$

Integrating w.r.t. π gives

$$\mathrm{KL}\left(\pi \| p_0 \otimes p_1\right) = -H(\pi) + H(p_0) + H(p_1),$$

where

$$H(\pi) = -\int \log \frac{d\pi}{dx \, dy} \, d\pi.$$

Since $H(p_0)$ and $H(p_1)$ do *not depend on* π, they are constants in the minimization over π.

Equivalent Forms of Entropic OT (II)

From the previous slide:

$$\mathbb{E}_\pi[c] + \varepsilon \mathrm{KL}\left(\pi \| p_0 \otimes p_1\right) \quad \Longleftrightarrow \quad \mathbb{E}_\pi[c] - \varepsilon H(\pi).$$

Chain rule for entropy.
Because π has marginal p_0,

$$H(\pi) = H(p_0) + \mathbb{E}_{x\sim p_0}\big[H(\pi(\cdot|x))\big].$$

Again, $H(p_0)$ is constant w.r.t. π. Therefore

$$\mathbb{E}_\pi[c] - \varepsilon H(\pi) \quad \Longleftrightarrow \quad \mathbb{E}_\pi[c] - \varepsilon \mathbb{E}_{x\sim p_0}\big[H(\pi(\cdot|x))\big].$$

Conclusion.

All three objectives differ only by additive constants, hence they yield the same minimizer.

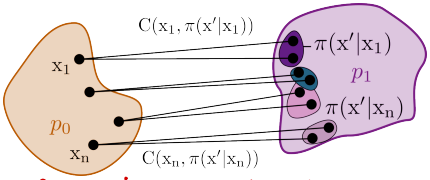
Weak optimal transport

Weak Optimal Transport (WOT)

Weak OT[3] generalizes Kantorovich OT equation OT by penalizing transport to distributions rather than points:

$$\mathrm{WOT}_C(p_0, p_1) = \min_{\pi \in \Pi(p_0, p_1)} \mathbb{E}_{x \sim p_0} C(x, \pi(\cdot|x)), \quad \text{(p-WOT)}$$

where $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ is a *weak cost* evaluating the displacement of x into a distribution $\pi(\cdot|x)$.



- Mass can be spread arbitrarily across $\pi(\cdot|x)$
- **Kantorovich OT as WOT:** equation OT is a special case of equation p-WOT with

$$C(x, \pi(\cdot|x)) = \mathbb{E}_{y \sim \pi}[c(x, y)]$$

- **Entropic OT as WOT:** equation EOT is a special case of equation p-WOT with

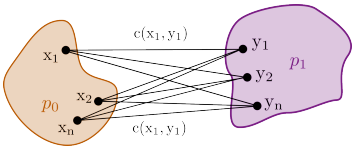
$$C_{\mathrm{EOT}}(x, \pi(\cdot|x)) = \mathbb{E}_{y \sim \pi}[c(x, y)] - \varepsilon H(\pi(\cdot|x)). \tag{4}$$

[4] Nathael Gozlan et al. (2017). **"Kantorovich duality for general transport costs and applications".** In: *Journal of Functional Analysis* 273.11, pp. 3327–3405

Kantorovich OT vs. Weak OT

Kantorovich's formulation (Strong OT)

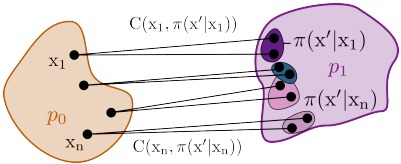
$$\min_{\pi \in \Pi(p_0, p_1)} \mathbb{E}_{(x,x') \sim \pi} \left[c(x, x') \right]$$



- Cost depends on pairs (x, x')
- $\pi(x, x')$ is a transport plan
- Allows mass splitting

Weak OT formulation

$$\min_{\pi \in \Pi(p_0, p_1)} \mathbb{E}_{x \sim p_0} \left[C\left(x, \pi(\cdot|x)\right) \right]$$



- Cost depends on x and conditional $\pi(\cdot|x)$
- Penalizes the *distribution* of transported mass
- Explicitly stochastic transport

Strong OT measures the cost of individual moves.
Weak OT measures the cost of distributions of moves.

Weak OT Duality

Under mild assumptions, equation p-WOT admits a *weak semi-dual* representation[4]:

$$\mathrm{WOT}_C(p_0, p_1) = \max_{f \in \mathcal{C}(\mathcal{Y})} \Big\{ \mathbb{E}_{x \sim p_0} f^C(x) + \mathbb{E}_{y \sim p_1} f(y) \Big\}, \tag{sd-WOT}$$

where the *weak C-transform* defined as

$$f^C(x) = \min_{\mu \in \mathcal{P}(\mathcal{Y})} \Big\{ C(x, \mu) - \mathbb{E}_{y \sim \mu} f(y) \Big\}. \tag{5}$$

Relation to classical OT. If $C(x, \mu) = \mathbb{E}_{y \sim \mu}[c(x, y)]$, then

$$f^C(x) = \min_{y \in \mathcal{Y}} \{ c(x, y) - f(y) \},$$

and we recover the Kantorovich dual formulation[5]:

$$\min_{\pi \in \Pi(p_0, p_1)} \mathbb{E}_{(x,x') \sim \pi}[c(x, x')] = \max_{f} \left\{ \mathbb{E}_{x \sim p_0}[f^c(x)] + \mathbb{E}_{x' \sim p_1}[f(x')] \right\}.$$

[4] Julio Backhoff-Veraguas, Mathias Beiglböck, and Gudmun Pammer (2019). **"Existence, duality, and cyclical monotonicity for weak transport costs".** In: *Calculus of Variations and Partial Differential Equations* 58.6, p. 203.

[5] Cédric Villani et al. (2008). **Optimal transport: old and new.** Vol. 338. Springer.

Why Not a Min–Max Reformulation?

Corollary (Maximin reformulation[6])

$$W_{\varepsilon}OT_{c,\varepsilon}(p_0, p_1) = \max_{f} \min_{T} \mathcal{L}(f, T),$$

$$\mathcal{L}(f, T) = \int_{\mathcal{Y}} f(y) dp_1(y) + \int_{\mathcal{X}} \Big(C\big(x, T(x, \cdot)_{\#} p_0\big) - \int_{\mathcal{Z}} f(T(x, z)) dp_0(z) \Big) dp_0(x),$$

where $T(x, z)$ is a stochastic transport map and $z \sim p_0$ is auxiliary noise (e.g. $\mathcal{N}(0, 1)$).

Why this formulation is problematic:

- Introduces a **min–max saddle-point problem** (unstable training).
- Requires learning an additional generator T.
- The extracted T^* is **not guaranteed** to be an optimal stochastic OT map.
- For strong costs, this may lead to **conditional collapse**.

[6] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev (2023). **"Neural Optimal Transport".** In: *The Eleventh International Conference on Learning Representations*.

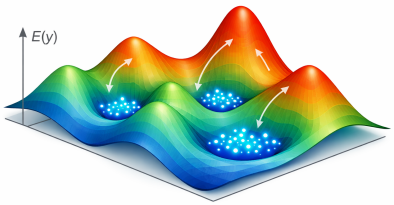
Recap: energy-based models

Energy-Based Models

Energy-Based Models (EBMs) are a classical class of generative models[7].

They parameterize a distribution $p_{\text{data}} \in \mathcal{P}(\mathcal{Y})$ via the Gibbs density:

$$p_{\text{data}}(y) = \frac{1}{Z_{\text{data}}} \exp\big(-E_{\text{data}}(y)\big).$$



- $E_{\text{data}} : \mathcal{Y} \to \mathbb{R}$ — energy function
- $Z_{\text{data}} = \int_{\mathcal{Y}} \exp(-E_{\text{data}}(y))\,\mathrm{d}y$ — partition function

Low energy $\Rightarrow$ *high probability.*

[7] Yann LeCun et al. (2006). **"A tutorial on energy-based learning".** In: *Predicting structured data* 1.0.

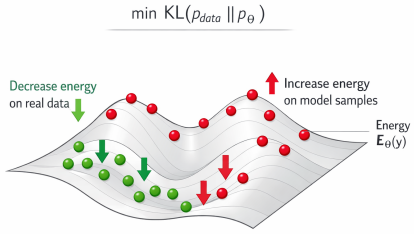
Learning in EBMs

Let:

- p_{data} — true data distribution,
- $p_{\text{data}} \approx p_\theta$ — parametric EBM with energy E_θ.

We fit the model by minimizing

$$\min_\theta \mathrm{KL}\left(p_{\text{data}} \| p_\theta\right).$$



Gradient of the KL divergence[8]:

$$\frac{\partial}{\partial\theta}\mathrm{KL}\left(p_{\text{data}} \| p_\theta\right) = \underbrace{\mathbb{E}_{y\sim p_{\text{data}}}\left[\partial_\theta E_\theta(y)\right]}_{\text{Can be estimated by Monte Carlo}} - \underbrace{\mathbb{E}_{y\sim p_\theta}\left[\partial_\theta E_\theta(y)\right]}_{\text{Requires samples from the model}}.$$

- Decrease energy on real data,
- Increase energy on model samples.

[9] Jianwen Xie et al. (2016). **"A theory of generative convnet".** In: *International Conference on Machine Learning*. PMLR, pp. 2635–2644

Sampling from an EBM

To estimate the model expectation $\mathbb{E}_{y \sim p_\theta}[\cdot]$, we must generate samples from p_θ.

Standard approach: **Unadjusted Langevin Algorithm (ULA)**[9]

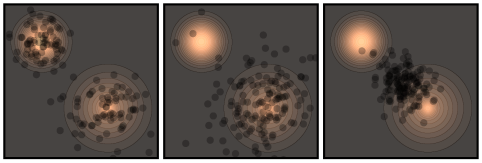
Discretized Langevin dynamics:

$$Y_{l+1} = Y_l - \frac{\eta}{2}\nabla_y E_\theta(Y_l) + \sqrt{\eta}\,\xi_l, \qquad \xi_l \sim \mathcal{N}(0, I).$$

Algorithm parameters:

- Step size $\eta > 0$
- Initialization $Y_0 \sim p_0$
- Number of steps L

Deterministic drift toward low-energy regions + stochastic noise for exploration.



Left: target distribution
Middle: step size too large → biased distribution
Right: step size too small → very slow mixing

[10] Gareth O Roberts and Richard L Tweedie (1996). **"Exponential convergence of Langevin distributions and their discrete approximations".** In: *Bernoulli*, pp. 341–363

Practical Aspects & Limitations of EBM

The main limitation of EBM is the necessity to sample from unnormalized densities both during training and inference. Several technical tricks are used to address this:

1. **Replay buffer**[10]: Maintain a collection of previously sampled points from the model and use them as an initialization of ULA in subsequent training stages.
2. **Decreasing learning rate**[11]: Gradually reduce η from η_{max} to η_{min} (sometimes to $\eta_{\text{min}} = 0$) to improve convergence to better samples in the local modes of p_θ.
3. **Noising training data**[9]: Perturb samples with $x_n \leftarrow x_n + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ for small $\sigma \approx 0$, to prevent overfitting to train data.

[10] Yilun Du and Igor Mordatch (2019). **"Implicit generation and modeling with energy based models".** In: *Advances in Neural Information Processing Systems* 32.

[11] Erik Nijkamp et al. (2020). **"On the anatomy of mcmc-based maximum likelihood learning of energy-based models".** In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04, pp. 5272–5280.

Energy-guided EOT

Optimizer of the weak C_{EOT}-transform[12]

Weak C_{EOT}-transform. For the entropic weak cost C_{EOT}, the transform is

$$f^{C_{\text{EOT}}}(x) = \min_{\mu \in \mathcal{P}(\mathcal{Y})} \Big\{ \mathbb{E}_{y\sim\mu}[c(x,y)] - \varepsilon \mathrm{H}(\mu) - \mathbb{E}_{y\sim\mu}[f(y)] \Big\}.$$

Question: What is the optimal distribution μ for fixed (x, f)?

Theorem (Closed-form optimizer)

For $f \in \mathcal{C}(\mathcal{Y})$ and $x \in \mathcal{X}$, the inner problem admits a **unique** *minimizer*

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

where $Z(f,x) = \int_{\mathcal{Y}} \exp\left(\frac{f(y)-c(x,y)}{\varepsilon}\right) dy$.

The optimizer is a Gibbs distribution.

[12] Petr Mokrov et al. (2024). **"Energy-guided Entropic Neural Optimal Transport".** In: *The Twelfth International Conference on Learning Representations*.

Weak semi-dual EOT in closed form

Closed-form weak C_{EOT}-transform. Substituting the Gibbs optimizer gives

$$f^{C_{\mathrm{EOT}}}(x) = -\varepsilon \log \left(\int_{\mathcal{Y}} \exp\left(\frac{f(y) - c(x,y)}{\varepsilon} \right) dy \right).$$

Plug into the weak semi-dual equation sd-WOT:

$$\mathrm{EOT}_{c,\varepsilon}(p_0, p_1) = \sup_f \left\{ -\varepsilon \int_{\mathcal{X}} \log Z(f,x)\, dp_0(x) + \int_{\mathcal{Y}} f(y)\, dp_1(y) \right\}. \qquad \text{(sd-WEOT)}$$

Interpretation.

- For each x, we obtain a Gibbs distribution $\mu_x^f(y) \propto \exp\left(\frac{f(y)-c(x,y)}{\varepsilon}\right)$.
- $-\varepsilon \log Z(f,x)$ is a log-partition function.
- The dual becomes an **energy learning problem**.

Entropy-regularized OT can be viewed as learning an energy function f.

Learning the Weak Semi-Dual EOT Potential

Parameterize the dual potential $f \in \mathcal{C}(\mathcal{Y}) \to f_\theta$, $\theta \in \Theta$ using a neural network.

From the reformulated weak semi-dual EOT objective equation sd-WEOT, we obtain the loss:

$$\max_{f_\theta} L(\theta) = \max_{f_\theta} -\varepsilon \mathbb{E}_{x \sim p_0} \log Z(f_\theta, x) + \mathbb{E}_{y \sim p_1} f_\theta(y).$$

Theorem (Gradient of the weak semi-dual loss $L(\theta)$[13])

It holds true that

$$\partial_\theta L(\theta) = -\mathbb{E}_{x \sim p_0} \mathbb{E}_{y \sim \mu_x^{f_\theta}} \left[\partial_\theta f_\theta(y)\right] + \mathbb{E}_{y \sim p_1} \left[\partial_\theta f_\theta(y)\right].$$

- Increase f_θ on target samples.
- Decrease f_θ on samples from $\mu_x^{f_\theta}$.

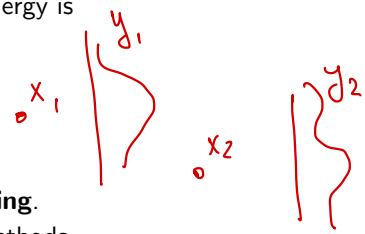
Same positive/negative structure as EBMs — but conditionally on x.

[13] Petr Mokrov et al. (2024). **"Energy-guided Entropic Neural Optimal Transport".** In: *The Twelfth International Conference on Learning Representations*.

Weak Dual EOT as Conditional Energy Learning

Although f_θ itself is *not* an energy function, the pair (c, f_θ) induces conditional Gibbs distributions: $\mu_x^{f_\theta}(y) \propto \exp\left(\frac{f_\theta(y) - c(x,y)}{\varepsilon}\right)$. Their corresponding energy is

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



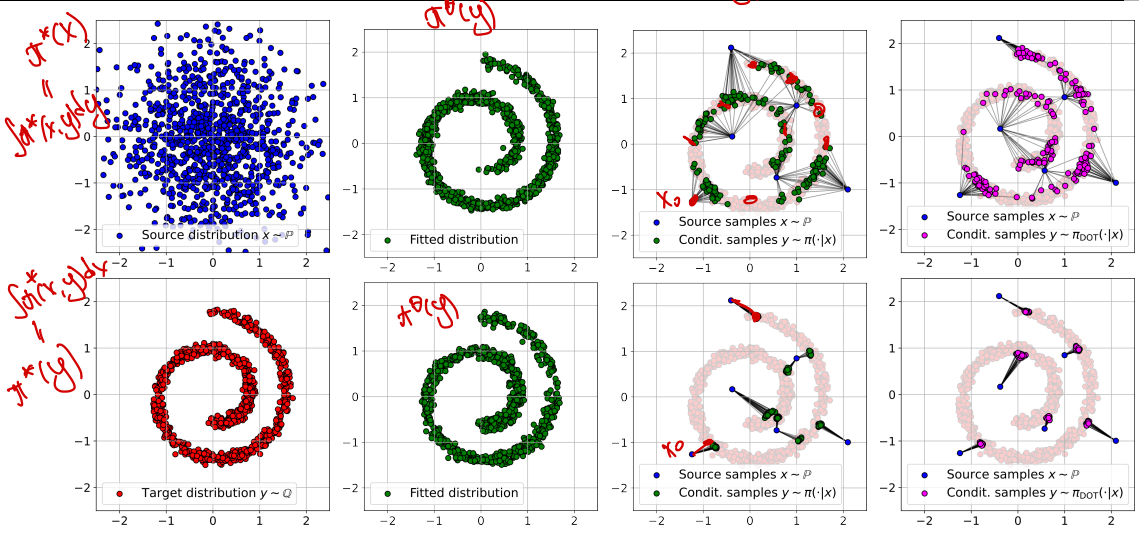
- Each x defines a **conditional EBM** over y.
- Weak dual optimization becomes **conditional energy learning**.
- Sampling from $\mu_x^{f_\theta}$ can be done via ULA or other MCMC methods.

Practical gradient estimation. We need samples from $\pi^{f_\theta}(x,y) = \mu_x^{f_\theta}(y)\, p_0(x)$.
Two-stage sampling:

1. Sample $x_1, \ldots, x_N \sim p_0$
2. For each x_i, sample $y_i \sim \mu_{x_i}^{f_\theta}$ (e.g., Langevin dynamics)

Entropy-regularized OT reduces to source-conditioned EBM training.

Gaussian → Swissroll (2D)



(a) Source and target distributions p_0 and p_1

(b) Fitted distributions p_θ; (up) $\varepsilon = 10^{-1}$, $\varepsilon = 10^{-3}$

(c) Fitted conditional plans $\pi_\theta(\cdot|x)$; $\varepsilon = 10^{-1}$, $\varepsilon = 10^{-3}$.

(d) Discrete conditional plans; (up) $\varepsilon = 10^{-1}$, $\varepsilon = 10^{-3}$.

Unpaired AFHQ Cat → Dog Translation

Figure 3: AFHQ 512 × 512 *Cat→Dog* unpaired translation by Energy-guided EOT solver[14] applied in the latent space of StyleGAN2-ADA. *Left:* source samples; *right:* translated samples.

[14] Petr Mokrov et al. (2024). **"Energy-guided Entropic Neural Optimal Transport".** In: *The Twelfth International Conference on Learning Representations*.

Conclusion & Takeaways

Starting point: Unsupervised domain translation is ill-posed — distribution matching alone is not enough.

Optimal Transport: Provides a principled way to select mappings via a cost.

Entropy-Regularized OT:

- Ensures uniqueness and stochastic transport
- Equivalent to a KL projection (Schrödinger Bridge)
- Admits a weak semi-dual formulation

Main insight: Weak semi-dual EOT $\Longleftrightarrow$ Conditional Energy-Based Learning.

- Dual potential f_θ induces conditional Gibbs plans
- Training reduces to source-conditioned EBM optimization
- Sampling via Langevin dynamics

Entropy-regularized Weak OT can be solved by learning an energy function.