fall into ML 2025

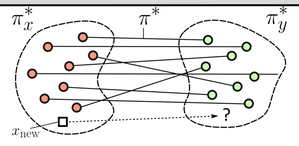
4th conference on machine learning & AI

Inverse Entropic Optimal Transport Solves Semi-supervised Learning via Data Likelihood Maximization

Mikhail Persiianov[1] Arip Asadulaev[2] Nikita Andreev Nikita Starodubcev[3]
Dmitry Baranchuk[3] Anastasis Kratsios[4,5] Evgeny Burnaev[1,6] Alexander Korotin[1,6]

Skoltech [1]
Mohamed bin Zayed University of Artificial Intelligence [2]
Yandex [3]
Vector Institute | Institut Vecteur [4]
McMaster University [5]
AIRI [6]

Domain Translation Setups



Supervised (paired) domain translation setup

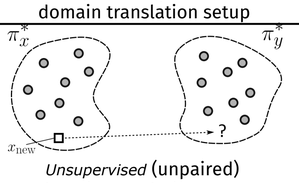
Setup: Requires paired data

$$XY_{\text{paired}} = \{(x_i, y_i)\}_{i=1}^{P} \sim \pi^*,$$

where each input x is matched with its output y.

Goal: Learn $\pi^*(y|x)$ to generate predictions $y|x_{\text{new}}$ for unseen inputs x_{new} that are not present in the training data.

Limitation: Paired datasets are costly and difficult to collect.

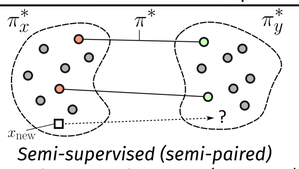


Unsupervised (unpaired) domain translation setup

Setup: No paired examples are available. We observe only unpaired samples from source $X_{\text{unpaired}} = \{x_i\}_{i=1}^{Q} \sim \pi_x^*$ and target $Y_{\text{unpaired}} = \{y_j\}_{j=1}^{R} \sim \pi_y^*$ distributions, respectively.

Challenge: Problem is **ill-posed** → solutions are often ambiguous. Requires additional constraints and regularization.

Relevance: Highly practical, since unpaired data is abundant.



Semi-supervised (semi-paired) domain translation setup (***our focus***)

Setup: Access to both: *paired data* $XY_{\text{paired}} \sim \pi^*$ and additional *unpaired samples*

$$X_{\text{unpaired}} \sim \pi_x^*, \; Y_{\text{unpaired}} \sim \pi_y^*.$$

Idea: Use paired data for precision, exploit unpaired data for abundance and generalization.

Convention: $P \leq Q, R$, with paired samples included among the unpaired ones.

Part I: Data Likelihood Maximization

Goal: Approximate the true distribution π^* by a parametric model π_θ^θ via KL-divergence minimization:

$$\mathrm{KL}\left(\pi^* \| \pi^\theta\right) = \underbrace{\mathrm{KL}\left(\pi_x^* \| \pi_x^\theta\right)}_{\text{Marginal}} + \underbrace{\mathbb{E}_{x\sim\pi_x^*} \mathrm{KL}\left(\pi^*(\cdot|x) \| \pi^\theta(\cdot|x)\right)}_{\text{Conditional}} =$$

$$= \text{const} - \mathbb{E}_{x\sim\pi_x^*} \mathrm{H}\left(\pi^*(\cdot|x)\right) - \mathbb{E}_{x,y\sim\pi^*} \log \pi^\theta(y|x).$$

Resulting minimization objective:

$$\mathcal{L}(\theta) \doteq -\mathbb{E}_{x,y\sim\pi^*} \log \pi^\theta(y|x). \tag{1}$$

Note: Minimizing (1) is equivalent to maximizing the conditional likelihood. **Limitation:** Requires fully paired data; unpaired samples cannot be used. **Goal:** extend to unpaired data.

Part II: Exploiting Unpaired Data

To address the above-mentioned issue and utilize unpaired data, we first use **Gibbs-Boltzmann parametrization**:

$$\pi^\theta(y|x) \doteq \frac{\exp\left(-E^\theta(y|x)\right)}{Z^\theta(x)}, \; Z^\theta(x) \doteq \int_{\mathcal{Y}} \exp\left(-E^\theta(y|x)\right) \mathrm{d}y, \tag{2}$$

where $E^\theta(\cdot|x) : \mathcal{Y} \to \mathbb{R}$ is *the Energy function*, and $Z^\theta(x)$ is the *normalization constant*. Substituting (2) into (1), we obtain:

$$\mathcal{L}(\theta) = \mathbb{E}_{x,y\sim\pi^*} E^\theta(y|x) + \mathbb{E}_{x\sim\pi_x^*} \log Z^\theta(x). \tag{3}$$

Observation: This objective already provides an opportunity to exploit the unpaired samples from the marginal distribution π_x^* to learn the conditional distributions $\pi^\theta(\cdot|x) \approx \pi^*(\cdot|x)$.

Decoupling Cost and Potential

Goal: Incorporate independent samples $y \sim \pi_y^*$ into the objective to enable semi-supervised learning.

We propose, **energy function parameterization:**

$$E^\theta(y|x) \doteq \frac{c^\theta(x,y) - f^\theta(y)}{\varepsilon}, \tag{4}$$

which decouples the cost function $c^\theta(x,y)$ and the potential $f^\theta(y)$.

Note: Changes in $f^\theta(y)$ can be offset by $c^\theta(x,y)$, resulting in the same energy. For example, setting $f^\theta(y) \equiv 0$ and $\varepsilon = 1$ makes the energy fully determined by $c^\theta(x,y)$.

Objective for Paired and Unpaired Data

Substituting (4) into (3), we obtain our **final objective**:

$$\mathcal{L}(\theta) = \underbrace{\varepsilon^{-1}\mathbb{E}_{x,y\sim\pi^*}[c^\theta(x,y)]}_{\text{Joint, requires pairs } (x,y)\sim\pi^*} - \underbrace{\varepsilon^{-1}\mathbb{E}_{y\sim\pi_y^*} f^\theta(y)}_{\text{Marginal, requires } y\sim\pi_y^*} +$$

$$+ \underbrace{\mathbb{E}_{x\sim\pi_x^*} \log Z^\theta(x)}_{\text{Marginal, requires } x\sim\pi_x^*} \to \min_\theta.$$

Dual Potential Parameterization

The most computationally intensive part of optimizing $\mathcal{L}(\theta)$ is computing the integral for the normalization constant $Z^\theta(x)$.

Solution: Introduce a lightweight parameterization that allows close form expressions for each term.

Cost function parameterization:

$$c^\theta(x,y) = -\varepsilon \log \sum_{m=1}^{M} v_m^\theta(x) \exp\left(\frac{\langle a_m^\theta(x), y\rangle}{\varepsilon}\right), \tag{1}$$

where $v_m^\theta(x) : \mathbb{R}^{D_x} \to \mathbb{R}_+$ and $a_m^\theta(x) : \mathbb{R}^{D_x} \to \mathbb{R}^{D_y}$ are parametric functions (e.g., neural networks) with learnable parameters θ_c.

To complement the cost, we propose for the **dual potential** f^θ:

$$f^\theta(y) = \varepsilon \log \sum_{n=1}^{N} w_n^\theta \mathcal{N}(y \mid b_n^\theta, \varepsilon B_n^\theta),$$

where $\theta_f = \{w_n^\theta, b_n^\theta, B_n^\theta\}_{n=1}^N$ are learnable parameters: $w_n^\theta \geq 0$, $b_n^\theta \in \mathbb{R}^{D_y}$, $B_n^\theta \in \mathbb{R}^{D_y \times D_y}$ symmetric positive definite. *Total:* $\theta = \theta_c \cup \theta_f$.

Note: To avoid notation overload, we omit $^\theta$ in the sequel.

Tractable Normalization Constant

Proposition. *Our parameterization of the cost c^θ and potential f^θ*

$$Z^\theta(x) \doteq \sum_{m=1}^{M}\sum_{n=1}^{N} z_{mn}(x), \; \textit{where}$$

$$z_{mn}(x) \doteq w_n v_m(x) \exp\left(\frac{a_m(x)^\top B_n a_m(x) + 2b_n^\top a_m(x)}{2\varepsilon}\right).$$

This expression is essential for *efficiently optimizing* the loss $\mathcal{L}(\theta)$.

Tractable Conditional Distributions

Proposition. *From our parametrization of the cost function and dual potential it follows that the $\pi^\theta(\cdot|x)$ are Gaussian mixtures:*

$$\pi^\theta(y|x) = \frac{1}{Z^\theta(x)} \sum_{m=1}^{M}\sum_{n=1}^{N} z_{mn}(x) \mathcal{N}(y \mid d_{mn}(x), \varepsilon B_n),$$

where $d_{mn}(x) \doteq b_n + B_n a_m(x)$ and $z_{mn}(x)$ is from Proposition above.

Sampling y given x is therefore fast and straightforward.

Universal Approximation of the Method

Question: How expressive is our parametrization of π_θ?

Theorem (Universal conditional distributions). *Under mild assumptions on the joint distribution π^*, for any $\delta > 0$ there exist:*

1. *$N > 0$ and a Gaussian mixture f^θ with N components,*
2. *$M > 0$ and cost c^θ defined via fully-connected neural networks $a_m : \mathbb{R}^{D_x} \to \mathbb{R}^{D_y}$, $v_m : \mathbb{R}^{D_x} \to \mathbb{R}_+$ with ReLU,*

such that the resulting π^θ satisfies

$$\mathrm{KL}\left(\pi^* \| \pi^\theta\right) < \delta.$$

Implication: Our parametrization can approximate *any* conditional distribution under mild regularity and boundedness assumptions.

Neural Parametrization

We used log-sum-exp and Gaussian mixtures for the cost c^θ and dual potential f^θ.

Question: Can we replace them with neural networks? Yes — the loss $\mathcal{L}(\theta)$ still applies.

Problem: For Gaussian mixtures, Z_θ is analytic. For neural networks, Z_θ becomes intractable.

Solution: Even if $\mathcal{L}(\theta)$ is intractable, its gradient $\frac{\partial}{\partial\theta}\mathcal{L}(\theta)$ can still be computed for optimization.

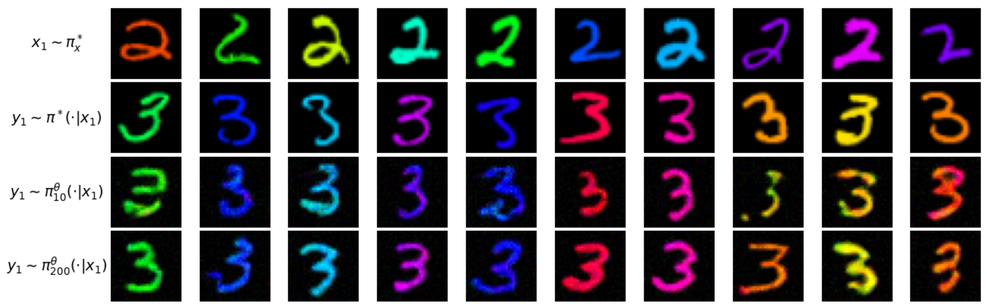
Proposition (Gradient of the loss). *It holds that*

$$\frac{\partial}{\partial\theta}\mathcal{L}(\theta) = \varepsilon^{-1}\left\{\mathbb{E}_{x,y\sim\pi^*}\left[\frac{\partial}{\partial\theta}c^\theta(x,y)\right] - \mathbb{E}_{y\sim\pi_y^*}\left[\frac{\partial}{\partial\theta}f^\theta(y)\right] + \mathbb{E}_{x\sim\pi_x^*}\mathbb{E}_{y\sim\pi^\theta(y|x)}\left[\frac{\partial}{\partial\theta}\left(f^\theta(y) - c^\theta(x,y)\right)\right]\right\}.$$

Colored MNIST (Pixel Space)

Setup. We use the Colored MNIST dataset and modify the task of translating digit 2 to digit 3 using *unpaired* images to demonstrate our method's ability to recover alignment consistent with *pairs*.

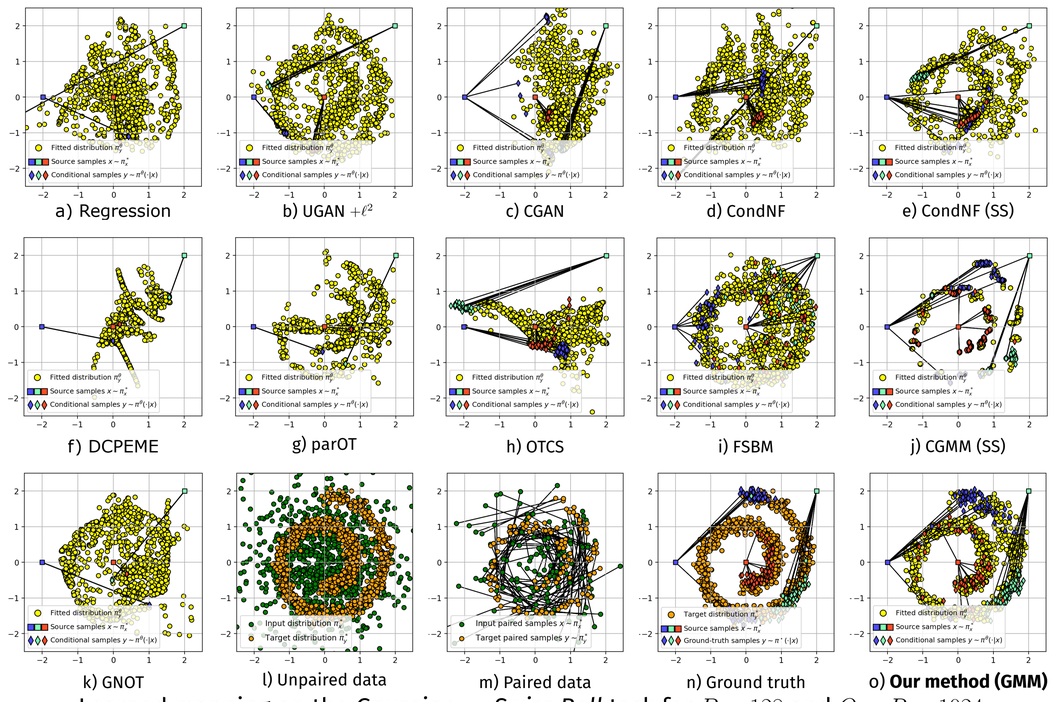
- Synthetic pairs are generated by shifting the hue of source images by $120°$.
- Specifically, for a source hue $h \in [0°, 360°)$, the corresponding target hue is $(h + 120°) \bmod 360°$.



Colored MNIST $2 \to 3$ task with hue shift. *Rows:* source (row 1), ground-truth target (row 2), resulting mapping produced by **our method (NN)** with 10 pairs (row 3), and with 200 pairs (row 4).

Gaussian to Swiss-Roll (2D Space)

Task: transform samples from a Gaussian π_x^* into a Swiss Roll π_y^*. The ground-truth plan π^* is obtained from a mini-batch OT plan with a designed cost to induce bi-modal conditionals $\pi^*(\cdot|x)$.



Learned mapping on the *Gaussian → Swiss Roll* task for $P = 128$ and $Q = R = 1024$.