UNIVERSITY of WASHINGTON

Entropy Regularized Optimal Transport Independence Criterion

NSF

Lang Liu[1], Soumik Pal[2], Zaid Harchaoui[1]

[1] Department of Statistics, University of Washington

[2] Department of Mathematics, University of Washington

Overview

- Propose a new **entropy regularized optimal transport independence criterion (ETIC)** and associated independence test.
- Develop an **efficient algorithm** to compute the test statistic which **scales quadratically** in both time and space.
- The ETIC can fit into a **differentiable programming** framework and is amenable to **reverse mode automatic differentiation**.
- We establish **non-asymptotic bounds** for the test statistic, characterizing the type I error and power of ETIC.

Statistical Test of Independence

Problem. Let $(X, Y) \sim P_{XY}$ on $\mathcal{X} \times \mathcal{Y}$ with marginals P_X and P_Y. Given an i.i.d. sample $\{(X_i, Y_i)\}_{i=1}^n$, the goal is to test for

$$\mathbf{H}_0 : X \text{ and } Y \text{ are independent} \leftrightarrow \mathbf{H}_1 : X \text{ and } Y \text{ are dependent.}$$

Steps to design an independence test.

1. Define a valid **independence criterion** $T(X, Y)$ such that $T(X, Y) \geq 0$ and $T(X, Y) = 0$ iff $X \perp\!\!\!\perp Y$.
2. Estimate the criterion from data $T_n(X, Y)$ (**test statistic**).
3. Choose a **critical value** $t_n(\alpha)$ and reject $\mathbf{H}_0$ if $T_n(X, Y) > t_n(\alpha)$.

ETIC

ETIC. Define $T(X, Y) := \bar{S}_\lambda(P_{XY}, P_X \otimes P_Y)$ as

$$S_\lambda(P_{XY}, P_X \otimes P_Y) - \frac{1}{2} S_\lambda(P_{XY}, P_{XY}) - \frac{1}{2} S_\lambda(P_X \otimes P_Y, P_X \otimes P_Y).$$

- **Sinkhorn distance** (Cuturi '13, Ferradans et al. '14)

$$S_\lambda(\varphi, \psi) := \min_{\gamma \in \mathsf{CP}(\varphi,\psi)} \left[\int c \, \mathrm{d}\gamma + \lambda \mathsf{KL}(\gamma \| \varphi \otimes \psi) \right].$$

- **Sinkhorn divergence** (Ramdas et al. '17, Genevay et al. '18, Feydy et al. '19) $\bar{S}_\lambda(\varphi, \psi)$.

Additive cost. Consider

$$c((x, y), (x', y')) := c_1(x, x') + c_2(y, y').$$

- ETIC is a **valid independence criterion** with proper c_1 and c_2.
- Assumptions on c_i via kernels $k_i := \exp\{-c_i/\lambda\}$.

ETIC with Weighted Quadratic Cost

Weighted quadratic cost. Consider

$$c((x, y), (x', y')) := w_1 \|x - x'\|^2 + w_2 \|y - y'\|^2.$$

- The objective in $S_\lambda(P_{XY}, P_X \otimes P_Y)$ becomes

$$w_1 \int \|x - x'\|^2 \mathrm{d}\gamma_1 + w_2 \int \|y - y'\|^2 \mathrm{d}\gamma_2 + \lambda \mathsf{KL}(\gamma \| P_{XY} \otimes P_X \otimes P_Y).$$

- Induces a valid independence criterion.

Computational Aspects of ETIC

ETIC test statistic. Let $\hat{P}_{XY}, \hat{P}_X, \hat{P}_Y$ be empirical distributions.

$$T_n(X, Y) := \bar{S}_\lambda(\hat{P}_{XY}, \hat{P}_X \otimes \hat{P}_Y).$$

Sinkhorn algorithm. $\tilde{O}(n^4)$ time and $O(n^4)$ space.

- $\hat{P}_X \otimes \hat{P}_Y$ is supported on n^2 atoms.
- **Gibbs matrix** $K = \exp(-C/\lambda) \in \mathbb{R}^{n^2 \times n^2}$.

Tensor Sinkhorn algorithm. $\tilde{O}(n^3)$ time and $O(n^2)$ space.

- A and B random matrices on $\{x_i\}_{i=1}^n \times \{y_i\}_{i=1}^n$.
- $K \to K_1$ and K_2 where $K_i = \exp(-C_i/\lambda) \in \mathbb{R}^{n \times n}$ for $i \in \{1, 2\}$.
- Amenable to **gradient backpropagation**.

Algorithm 1 Tensor Sinkhorn Algorithm

```
1: Input: A, B, K_1, and K_2.
2: Initialize U ← 1_{n×n} and V ← 1_{n×n}.
3: while not converge do
4:     U ← A ⊘ (K_1 V K_2^T) and V = B ⊘ (K_1^T U K_2).
5: end while
6: Output: ⟨λ log U, A⟩_F + ⟨λ log V, B⟩_F.
```

Random feature approximation. $\tilde{O}(pn^2)$ time and $O(n^2)$ space.

- Approximate $K_i \approx \xi_i \xi_i^\top$ for $i \in \{1, 2\}$.
- $\xi_i \in \mathbb{R}^{n \times p}$ **random feature matrix**.

	Gibbs matrix	Computation per iteration
Sinkhorn	$n^2 \times n^2$	$n^2 \times n^2$ and $n^2 \times 1$
Tensor Sinkhorn	Two $n \times n$	$n \times n$ and $n \times n$
Tensor Sinkhorn with RF	Two $n \times p$	$n \times n$ and $n \times p$

Main Results

High probability bound. Let c be the **weighted quadratic cost**. Assume P_X and P_Y have **bounded supports** of radius R. With probability at least $1 - \delta$,

$$|T_n(X, Y) - T(X, Y)| \leq C_d \left(\lambda + \frac{R^{5d+16}}{\lambda^{5d/2+7}} \sqrt{\log \frac{6}{\delta}} \right) \frac{1}{\sqrt{n}}.$$

- The **rate of convergence** is $O(n^{-1/2})$.
- The choice of $\lambda = R^2$ gives $C_d \sqrt{\log(6/\delta)} R^2 n^{-1/2}$.
- $T(X, Y) = T_\lambda(X, Y) \to 0$ as $\lambda \to \infty$.

Power analysis. The power of ETIC is **asymptotically one**.

- Under $\mathbf{H}_0$, $T(X, Y) = 0$ and thus $t_n(\alpha) = O(n^{-1/2})$.
- Under $\mathbf{H}_1$, $T(X, Y) > 0$ and thus $T_n(X, Y) > t_n(\alpha)$ as $n \to \infty$.

Real Data Experiment

Bilingual text. Parallel European Parliament corpus (Koehn '05).

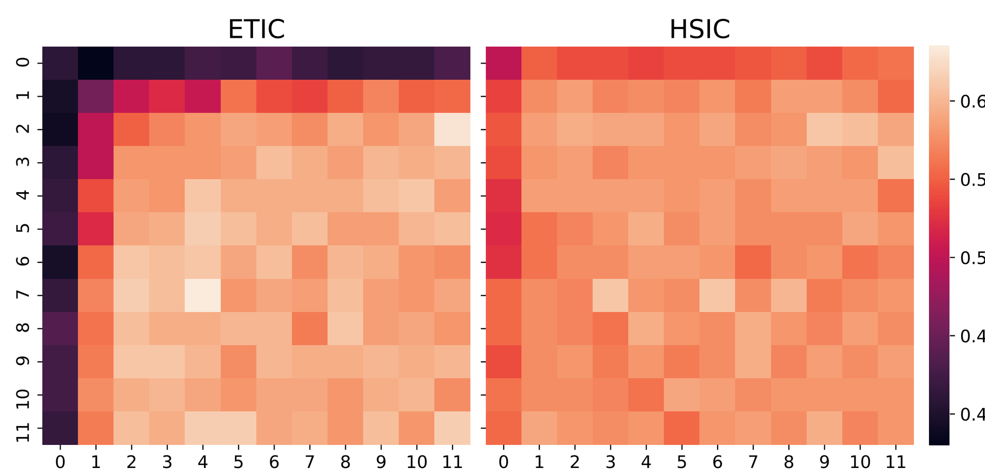
- **Random paragraph** in each of 64 random documents in English.
- (English paragraph, **random paragraph** in the **same document** in French).
- Feature embedding of dimension 768 with LaBSE (Feng et al. '20).

Independence tests.

- **ETIC** with weighted quadratic cost inducing Gaussian kernels

$$k_1(x, x') = e^{-\|x - x'\|^2/\sigma_1} \quad \text{and} \quad k_2(y, y') = e^{-\|y - y'\|/\sigma_2}.$$

- **Hilbert-Schmidt independence criterion (HSIC)** with the same kernels.
- **Median heuristic:** $\sigma_i = r_i M_i$ with $r_i \in [0.25, 4]$ for $i \in \{1, 2\}$.



Code available at *https://github.com/langliu95/etic-experiments*.