Entropy Regularized Optimal Transport Independence Criterion

Lang Liu, Soumik Pal, Zaid Harchaoui

University of Washington

March 30, 2022

Institute for Foundations of Data Science

Team

Lang Liu

Soumik Pal

Zaid Harchaoui

Statistical Test of Independence

Problem:

- Let $(X, Y) \sim P_{XY}$ on $\mathcal{X} \times \mathcal{Y}$ with marginals P_X and P_Y.
- Let $\{(X_i, Y_i)\}_{i=1}^n$ be i.i.d. copies of (X, Y).

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

Statistical Test of Independence

Problem:

- Let $(X, Y) \sim P_{XY}$ on $\mathcal{X} \times \mathcal{Y}$ with marginals P_X and P_Y.
- Let $\{(X_i, Y_i)\}_{i=1}^n$ be i.i.d. copies of (X, Y).

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

Strategy:

- Define an independence criterion $T(X, Y)$ such that
  - $T(X, Y) \geq 0$,
  - $T(X, Y) = 0$ iff X and Y are independent.
- Estimate the criterion from data $T_n(X, Y)$.
- Choose a critical value $t_n(\alpha)$ and reject $\mathbf{H}_0$ if $T_n(X, Y) > t_n(\alpha)$.

Entropy Regularized Optimal Transport Independence Criterion

ETIC—define $T(X, Y)$ by

$$\bar{S}_\lambda(P_{XY}, P_X \otimes P_Y) := S_\lambda(P_{XY}, P_X \otimes P_Y) - \frac{1}{2} S_\lambda(P_{XY}, P_{XY}) - \frac{1}{2} S_\lambda(P_X \otimes P_Y, P_X \otimes P_Y),$$

where $S_\lambda(P, Q)$ the cost of entropy regularized optimal transport (EOT)

$$\min_{\gamma \in \mathrm{CP}(P,Q)} \left[\int c(z, z') \mathrm{d}\gamma(z, z') + \lambda \mathrm{KL}(\gamma \| P \otimes Q) \right].$$

Entropy Regularized Optimal Transport Independence Criterion

ETIC—define $T(X, Y)$ by

$$\bar{S}_\lambda(P_{XY}, P_X \otimes P_Y) := S_\lambda(P_{XY}, P_X \otimes P_Y) - \frac{1}{2} S_\lambda(P_{XY}, P_{XY}) - \frac{1}{2} S_\lambda(P_X \otimes P_Y, P_X \otimes P_Y),$$

where $S_\lambda(P, Q)$ the cost of entropy regularized optimal transport (EOT)

$$\min_{\gamma \in \mathrm{CP}(P,Q)} \left[\int c(z, z') \mathrm{d}\gamma(z, z') + \lambda \mathrm{KL}(\gamma \| P \otimes Q) \right].$$

- Consider an *additive cost*:

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

- ETIC is a **valid independence criterion** with appropriate c_1 and c_2.
- Assumptions on c_i via kernels $k_i := \exp\{-c_i/\lambda\}$ for $i \in \{1, 2\}$.

Computational Aspects of ETIC

ETIC test statistic:

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

where $\hat{P}_{XY} = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_i)}$, $\hat{P}_X = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, and $\hat{P}_Y = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$.

Computational Aspects of ETIC

ETIC test statistic:

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

where $\hat{P}_{XY} = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_i)}$, $\hat{P}_X = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, and $\hat{P}_Y = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$.

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

	Cost matrix	**Computation per iteration**
Sinkhorn	$n^2 \times n^2$	$n^2 \times n^2$ and $n^2 \times 1$

Computational Aspects of ETIC

ETIC test statistic:

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

where $\hat{P}_{XY} = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Y_i)}$, $\hat{P}_X = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, and $\hat{P}_Y = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$.

- Naïve Sinkhorn algorithm: $\tilde{O}(n^4)$ time and $O(n^4)$ space.
- *Tensor Sinkhorn algorithm*: $\tilde{O}(n^3)$ time and $O(n^2)$ space.

	Cost 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 $n \times n$

Computational Aspects of ETIC

ETIC test statistic:

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

where $\hat{P}_{XY} = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_i)}$, $\hat{P}_X = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, and $\hat{P}_Y = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$.

- Naïve Sinkhorn algorithm: $\tilde{O}(n^4)$ time and $O(n^4)$ space.
- *Tensor Sinkhorn algorithm*: $\tilde{O}(n^3)$ time and $O(n^2)$ space.
- Tensor Sinkhorn with *random feature approximation*: $\tilde{O}(pn^2)$ time and $O(n^2)$ space.

	Cost 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 $n \times p$

Statistical Properties of ETIC

Theorem (Liu et al. '22)

*Assume that c is the **weighted quadratic cost** and P_X and P_Y are supported on a **bounded domain with radius** R. Then we have, with probability at least $1 - \delta$,*

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

Statistical Properties of ETIC

Theorem (Liu et al. '22)

Assume that c is the ***weighted quadratic cost*** *and P_X and P_Y are supported on a* ***bounded domain with radius*** *R. Then we have, with probability at least $1-\delta$,*

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

Remark

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

Statistical Properties of ETIC

Theorem (Liu et al. '22)

Assume that c is the ***weighted quadratic cost*** *and P_X and P_Y are supported on a* ***bounded domain with radius*** *R. Then we have, with probability at least $1-\delta$,*

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

Remark

Theorem 2 implies that the power of the ETIC test is asymptotically one.

- *Under $\mathbf{H}_0$, $T(X,Y) = 0$ and thus the critical value $t_n(\alpha)$ should be of order $O(n^{-1/2})$.*
- *Under $\mathbf{H}_1$, $T(X,Y) > 0$ and thus $T_n(X,Y)$ will alway exceed $t_n(\alpha)$ as $n \to \infty$.*

Independence Testing on Bilingual Text

Bilingual text

- Parallel European Parliament corpus (Koehn '05).
- Randomly select $n = 64$ English documents and a paragraph in each document.
- (English paragraph, random paragraph in the same document in French).
- Feature embeddings of dimension 768 with LaBSE (Feng et al. '20).

Independence Testing on Bilingual Text

Bilingual text

- Parallel European Parliament corpus (Koehn '05).
- Randomly select $n = 64$ English documents and a paragraph in each document.
- (English paragraph, random paragraph in the same document in French).
- Feature embeddings of dimension 768 with LaBSE (Feng et al. '20).

Independence tests

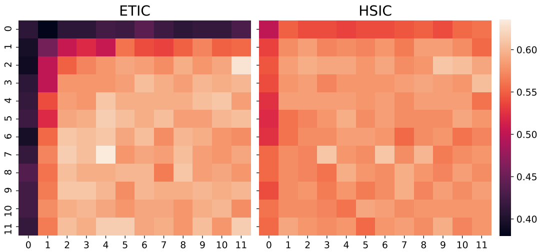
- ETIC with a weighted quadratic cost inducing Gaussian kernels

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

- Hilbert-Schmidt independence criterion (HSIC) with the same kernels.
- Median heuristic: $\sigma_1 = r_1 M_x$ and $\sigma_2 = r_2 M_y$ with $r_1, r_2 \in [0.25, 4]$.

Independence Testing on Bilingual Text

ETIC outperforms HSIC for many values of r_1 and r_2.



Conclusion

- A new independence criterion ETIC and the associated test.
- An efficient algorithm to compute empirical ETIC.
- Amenable to gradient backpropagation.
- Finite-sample guarantees for its statistical properties.
- Higher power with a large range of hyperparameters.

Code

SCAN ME

Appendix

The Schrödinger Bridge Problem and Entropy Regularized OT

The Schrödinger bridge (SCB) problem

- Schrödinger's lazy gas experiment (Schrödinger '32).
- The SCB problem in continuum (Föllmer '88, Léonard '12).
- Survey on SCB (Léonard '14, Chen et al. '21).

Discrete entropy regularized optimal transport (EOT)

- Discrete EOT (Cuturi '13, Ferradans et al. '14).
- Limit laws (Bigot et al. '19, Klatt et al. '20).
- Finite-sample bounds (Genevay et al. '19, Mena and Weed '19).

Discrete Schrödinger bridge

- Discrete SCB for a particular cost (Pal and Wong '20).
- Discrete SCB for general costs (HLP '20).

Properties of ETIC

Proposition (Liu et al. '22)

Let c be a continuous cost function. If either c is bounded or P_{XY} and $P_X \otimes P_Y$ have compact support, it holds that

$$T_\lambda(X, Y) \to \begin{cases} 0 & \text{if } c = c_1 \oplus c_2 \\ -\frac{1}{2} HSIC_{c_1,c_2}(X, Y) & \text{if } c = c_1 \otimes c_2, \end{cases} \quad \text{as } \lambda \to \infty.$$

Moreover, if both P_{XY} and $P_X \otimes P_Y$ are densities (or discrete measures), then

$$T_\lambda(X, Y) \to C_{OT}(P_{XY}, P_X \otimes P_Y), \quad \text{as } \lambda \to 0.$$

Statistical Properties of ETIC

Empirical Sinkhorn divergence

$$C_{\mathrm{SD}}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{U_i}, \frac{1}{n}\sum_{i=1}^{n}\delta_{V_i}\right).$$

Empirical ETIC

$$C_{\mathrm{SD}}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{(X_i,Y_i)}, \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{(X_i,Y_j)}\right).$$

	First marginnal	Second marginal	Independent marginals?
SD	Sum of i.i.d. point masses	Sum of i.i.d. point masses	Yes
ETIC	Sum of i.i.d. point masses	**Sum of dependent point masses**	**No**

Statistical Properties of ETIC

Theorem (Liu et al. '22)

Assume that c_1 and c_2 are quadratic costs and P_X and P_Y are sub-Gaussian with parameter σ^2. Then we have

$$\mathbb{E}\,|T_n(X,Y) - T(X,Y)| \leq C_d \left(1 + \frac{\sigma^{\lceil 5d/2 \rceil + 6}}{\lambda^{\lceil 5d/4 \rceil + 3}}\right) \frac{\lambda}{\sqrt{n}}.$$

Remark

When $\lambda := \lambda_n = o(1)$ is chosen such that $\lambda_n = \omega(n^{-1/(\lceil 5d/2 \rceil + 4)})$, we have

$$T_n(X,Y) \to_{\mathbf{L}^1} C_{OT}(P_{XY}, P_X \otimes P_Y) = \mathrm{W}_2^2(P_{XY}, P_X \otimes P_Y).$$

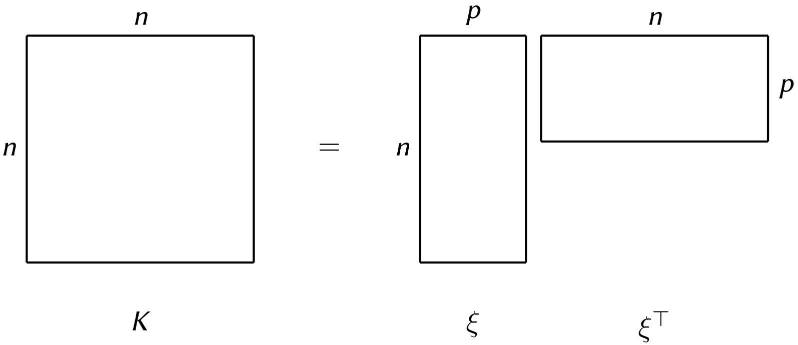
The Tensor Sinkhorn Algorithm

- A and B distributions on $\{x_i\}_{i=1}^n \times \{y_i\}_{i=1}^n$.
- $K_1 = \exp(-C_1/\lambda)$ and $K_2 = \exp(-C_2/\lambda)$.
- Compute $C_{\text{EOT}}(A, B)$.

Algorithm 1 Tensor Sinkhorn Algorithm

1: **Input:** A, B, K_1, and K_2.
2: Initialize $U \leftarrow \mathbf{1_{n\times n}}$ and $V \leftarrow \mathbf{1_{n\times n}}$.
3: **while** not converge **do**
4: $\quad U \leftarrow A \oslash (K_1 V K_2^\top)$ and $V = B \oslash (K_1^\top U K_2)$.
5: **end while**
6: **Output:** $\langle \varepsilon \log U, A \rangle_{\mathbf{F}} + \langle \varepsilon \log V, B \rangle_{\mathbf{F}}$.

Random Feature Approximation



ETIC with Random Features

- Consider Gibbs kernels of the form

$$k_1(x, x') = \int \varphi(x, u)^\top \varphi(x', u) d\rho_1(u) \quad \text{and} \quad k_2(y, y') = \int \psi(y, v)^\top \psi(y', v) d\rho_2(v).$$

- Obtain an i.i.d. sample $\boldsymbol{u} := \{u_k\}_{k=1}^p$ and approximate $k_1(x, x')$ by

$$k_{1,\boldsymbol{u}}(x, x') := \frac{1}{p} \sum_{k=1}^{p} \varphi(x, u_k)^\top \varphi(x', u_k).$$

- Obtain an i.i.d. sample $\boldsymbol{v} := \{v_k\}_{k=1}^p$ and approximate $k_2(y, y')$ by

$$k_{2,\boldsymbol{v}}(y, y') := \frac{1}{p} \sum_{k=1}^{p} \psi(y, v_k)^\top \psi(y', v_k).$$

ETIC with Random Features

Approximate $c((x, y), (x', y'))$ by

$$c_{\boldsymbol{u},\boldsymbol{v}}((x, y), (x', y')) := -\lambda \log k_{1,\boldsymbol{u}}(x, x') - \lambda \log k_{2,\boldsymbol{v}}(y, y').$$

Proposition (Liu et al. '22)

Let $p = \Omega(\tau^{-2} \log(n/\delta))$. Under appropriate assumptions, it holds that, with probability at least $1 - \delta$,

$$\left| C_{EOT, c_{\boldsymbol{u},\boldsymbol{v}}}(A, B) - C_{EOT,c}(A, B) \right| \leq \tau.$$

ETIC-Based Tests

The ETIC test with regularization parameter λ:

$$\psi(\alpha) := \mathbb{1}\{T_{n,\lambda}(X, Y) > t_{n,\lambda}(\alpha)\},$$

where α is the significance level and $H_{n,\lambda}(\alpha)$ is the critical value.

The adaptive ETIC test:

$$\psi_a(\alpha) := \mathbb{1}\left\{\max_{\lambda\in\Lambda} \bar{T}_{n,\lambda}(X, Y) > t_{n,\Lambda}(\alpha)\right\}$$

- Λ a finite set of regularization parameters.
- $\bar{T}_{n,\lambda}(X, Y) = [T_{n,\lambda}(X, Y) - \mathbb{E}[T_{n,\lambda}(X, Y)]]/\mathrm{Sd}(T_{n,\lambda}(X, Y))$ the studentized version.