Statistical Divergences for Learning and Inference: A Non-Asymptotic Viewpoint

Lang Liu

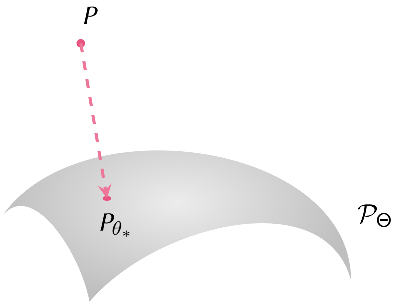
University of Washington

September 22, 2022

Committee: Zaid Harchaoui (Chair), Soumik Pal (Co-Chair)
Thomas Richardson, Kevin Jamieson, Hanna Hajishirzi (GSR)

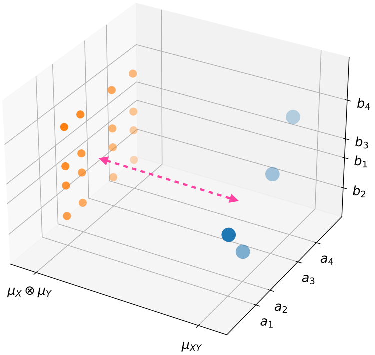
Motivating Examples: Statistical Estimation

- **Data** $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P$.
- **Parametric family** $\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$, where Θ is convex and compact.
- **Goal:** identify $\theta_\star$ so that $P_{\theta_\star}$ is "closest" to P.



Motivating Examples: Independence Testing

- **Data** $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} \mu_{XY}$ with marginals μ_X and μ_Y.
- **Goal:** determine whether X is independent of Y.
- **Strategy:** measure the "distance" between μ_{XY} and $\mu_X \otimes \mu_Y$.



Motivating Examples: Generative Model Comparison†

Model 1 generation

Model 2 generation

Real images

†Liu et al. In *NeurIPS*, 2021.

Statistical Estimation with the KL Divergence

- Kullback-Leibler (KL) divergence

$$\mathrm{KL}(P\|Q) := \int \log\left(\mathrm{d}P/\mathrm{d}Q\right)\mathrm{d}P.$$

- Minimum KL estimation

$$\theta_\star := \arg\min_{\theta\in\Theta} \mathrm{KL}(P\|P_\theta) = \arg\min_{\theta\in\Theta} \left\{\mathbb{E}[\log P(Z)] - \mathbb{E}[\log P_\theta(Z)]\right\}.$$

Statistical Estimation with the KL Divergence

- Kullback-Leibler (KL) divergence

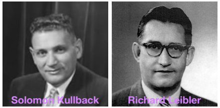
$$\mathrm{KL}(P\|Q) := \int \log\left(\mathrm{d}P/\mathrm{d}Q\right)\mathrm{d}P.$$

- Minimum KL estimation (maximum likelihood estimation)

$$\theta_\star := \arg\min_{\theta\in\Theta} \mathrm{KL}(P\|P_\theta) = \arg\min_{\theta\in\Theta} \left\{ \mathbb{E}[-\log P_\theta(Z)] =: \underbrace{L(\theta)}_{\text{Risk}} \right\}.$$

Statistical Estimation with the KL Divergence

- Kullback-Leibler (KL) divergence



$$\mathrm{KL}(P \| Q) := \int \log\,(\mathrm{d}P/\mathrm{d}Q)\mathrm{d}P.$$

- Minimum KL estimation (maximum likelihood estimation)

$$\theta_\star := \operatorname*{arg\,min}_{\theta\in\Theta} \mathrm{KL}(P \| P_\theta) = \operatorname*{arg\,min}_{\theta\in\Theta} \Big\{ \mathbb{E}[-\log P_\theta(Z)] =: \underbrace{L(\theta)}_{\text{Risk}} \Big\}.$$

- Maximum likelihood estimator (MLE)

$$\theta_n := \operatorname*{arg\,min}_{\theta\in\Theta} \Big\{ -\frac{1}{n}\sum_{i=1}^{n} \log P_\theta(Z_i) =: \underbrace{L_n(\theta)}_{\text{Empirical risk}} \Big\}.$$

Statistical Estimation with the KL Divergence

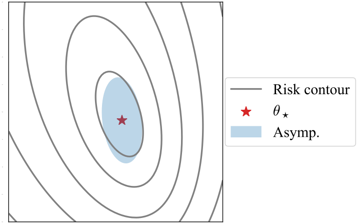
Asymptotic theory

- $\sqrt{n}(\theta_n - \theta_\star) \to_d \mathcal{N}(0, \Sigma)$.

Statistical Estimation with the KL Divergence

Asymptotic theory

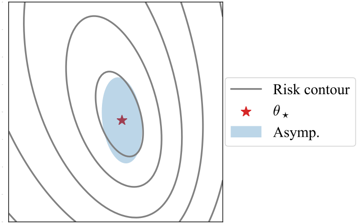
- $n\|\Sigma_n^{-1/2}(\theta_n - \theta_\star)\|_2^2 \to_d \chi_d^2$.
- Slutsky's Lemma.
- Asymptotically tight.
- Valid for $n \to \infty$ and fixed d.



Statistical Estimation with the KL Divergence

Asymptotic theory

- $n \| \Sigma_n^{-1/2} (\theta_n - \theta_\star) \|_2^2 \to_d \chi_d^2$.
- Slutsky's Lemma.
- Asymptotically tight.
- Valid for $n \to \infty$ and fixed d.



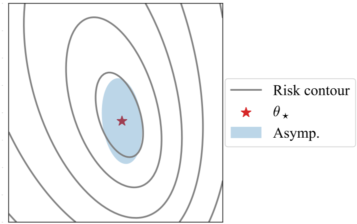
Non-asymptotic theory

- $L(\theta_n) - L(\theta_\star) \leq O\big(n^{-1}\big)$.

Statistical Estimation with the KL Divergence

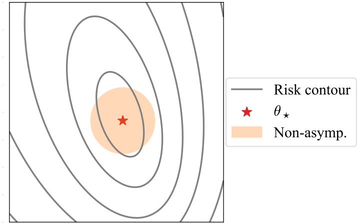
Asymptotic theory

- $n\|\Sigma_n^{-1/2}(\theta_n - \theta_\star)\|_2^2 \to_d \chi_d^2$.
- Slutsky's Lemma.
- Asymptotically tight.
- Valid for $n \to \infty$ and fixed d.



Non-asymptotic theory

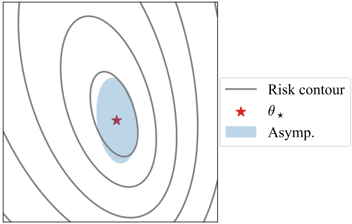
- $\|\theta_n - \theta_\star\|_2^2 \leq O(n^{-1})$.
- Strong convexity.
- Conservative.
- Valid for all n and d.



Statistical Estimation with the KL Divergence

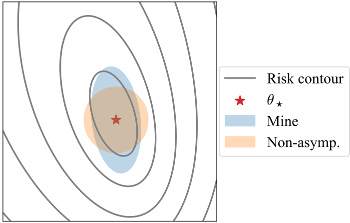
Asymptotic theory

- $n\|\Sigma_n^{-1/2}(\theta_n - \theta_\star)\|_2^2 \to_d \chi_d^2$.
- Slutsky's Lemma.
- Asymptotically tight.
- Valid for $n \to \infty$ and fixed d.



My contribution

- $\|\Sigma_n^{-1/2}(\theta_n - \theta_\star)\|_2^2 \le O(n^{-1})$.
- Pseudo self-concordance.
- Conservative.
- Valid for $n > O(d)$.

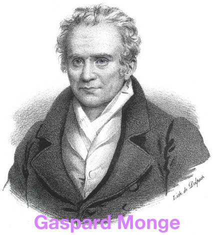


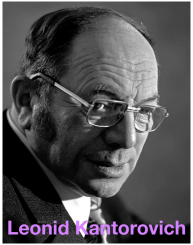
Independence Testing with Entropy Regularized Optimal Transport

Monge-Kantorovich optimal transport

$$S(P, Q) := \min_{\gamma \in \mathrm{CP}(P,Q)} \int c \, d\gamma.$$

- $c \geq 0$ cost function.
- $\mathrm{CP}(P, Q)$ set of couplings.





Independence Testing with Entropy Regularized Optimal Transport

Entropy regularized optimal transport (EOT)

$$S_\varepsilon(P, Q) := \min_{\gamma \in \mathrm{CP}(P,Q)} \left[\int c \mathrm{d}\gamma + \varepsilon \mathrm{KL}(\gamma \| P \otimes Q) \right].$$

Plug-in estimator $S_\varepsilon(P_n, Q_n)$.

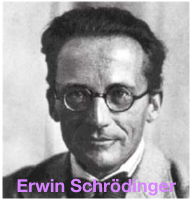
- **Faster rate of convergence**: $O(n^{-1/2})$ rather than $O(n^{-2/d})$.
- **Faster algorithm**: $O(n^2)$ time rather than $O(n^3)$ time.

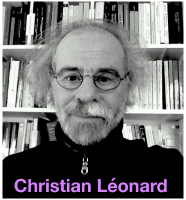
Independence Testing with Entropy Regularized Optimal Transport

Entropy regularized optimal transport (EOT)

$$\operatorname*{arg\,min}_{\gamma \in \mathrm{CP}(P,Q)} \left[\int c \, \mathrm{d}\gamma + \varepsilon \mathrm{KL}(\gamma \| P \otimes Q) \right] = \operatorname*{arg\,min}_{\gamma \in \mathrm{CP}(P,Q)} \mathrm{KL}(\gamma \| R_\varepsilon).$$

- $R_\varepsilon(z, z') \propto \exp(-c(z, z')/\varepsilon)$.
- Schrödinger bridge problem.
- Information projection.





Independence Testing with Entropy Regularized Optimal Transport‡

Two-sample testing

- Sinkhorn algorithm: $O(n^2)$ time.
- Finite-sample bounds.
- Empirical process theory

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - \mathbb{E}[f(Z)] \right|,$$

$\{Z_i\}_{i=1}^n$ i.i.d. copies of Z.

‡Sinkhorn '67, Cuturi '13, van de Vaart and Wellner '96.

Independence Testing with Entropy Regularized Optimal Transport‡

Two-sample testing

- Sinkhorn algorithm: $O(n^2)$ time.
- Finite-sample bounds.
- Empirical process theory

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(Z_i) - \mathbb{E}[f(Z)] \right|,$$

$\{Z_i\}_{i=1}^n$ i.i.d. copies of Z.

Independence testing

- Sinkhorn algorithm: $O(n^4)$ time.
- No theoretical guarantee.

‡Sinkhorn ’67, Cuturi ’13, van de Vaart and Wellner ’96.

Independence Testing with Entropy Regularized Optimal Transport§

Two-sample testing

- Sinkhorn algorithm: $O(n^2)$ time.
- Finite-sample bounds.
- Empirical process theory

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}[f(Z)] \right|,$$

$\{Z_i\}_{i=1}^n$ i.i.d. copies of Z.

My contribution

- Efficient algorithm: $O(n^2)$ time.
- Finite-sample bounds.
- U-process theory

$$\sup_{g\in\mathcal{G}} \left| \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} g(X_i, Y_j) - \mathbb{E}[g(X, Y')] \right|,$$

$\{(X_i, Y_i)\}_{i=1}^n$ i.i.d. copies of (X, Y).

§Sinkhorn '67, Cuturi '13, van de Vaart and Wellner '96, de la Peña and Giné '99.

Outline

Part I. Non-asymptotics of the minimum Kullback-Leibler divergence estimation.

- A non-asymptotic viewpoint of classical asymptotic theory.
- A finite-sample confidence set adapted to the risk landscape.
- Extension to semi-parametric estimation.

Outline

Part I. Non-asymptotics of the minimum Kullback-Leibler divergence estimation.

- A non-asymptotic viewpoint of classical asymptotic theory.
- A finite-sample confidence set adapted to the risk landscape.
- Extension to semi-parametric estimation.

Part II. Independence testing with the entropy regularized optimal transport.

- A new independence criterion and the associated test.
- Non-asymptotic bounds for the empirical estimator.
- Efficient algorithm for the test statistic.

Outline

Part I. Non-asymptotics of the minimum Kullback-Leibler divergence estimation.

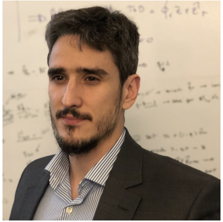
- A non-asymptotic viewpoint of classical asymptotic theory.
- A finite-sample confidence set adapted to the risk landscape.
- Extension to semi-parametric estimation.

Part II. Independence testing with the entropy regularized optimal transport.

- A new independence criterion and the associated test.
- Non-asymptotic bounds for the empirical estimator.
- Efficient algorithm for the test statistic.

Part III. Future directions.

Part I. Non-Asymptotics of the Minimum Kullback-Leibler Divergence Estimation



Carlos Cinelli

Zaid Harchaoui

To be submitted @ AISTATS 2023

@ COLT 2022

Minimum Kullback-Leibler Divergence Estimation

- **Data** $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P$.
- **Parametric family** $\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$.
- **Target parameter**

$$\theta_\star := \arg\min_{\theta\in\Theta} \mathrm{KL}(P \| P_\theta) = \arg\min_{\theta\in\Theta} \Big\{ \mathbb{E}[-\log P_\theta(Z)] := \mathbb{E}[\underbrace{\ell(\theta; Z)}_{\text{Loss function}}] := \underbrace{L(\theta)}_{\text{Risk}} \Big\}.$$

- **Maximum likelihood estimator** (MLE)

$$\theta_n := \arg\min_{\theta\in\Theta} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; Z_i) := \underbrace{L_n(\theta)}_{\text{Empirical risk}} \right\}.$$

Related Work: Asymptotic Theory[¶]

Well-specified model: $P \in \mathcal{P}_\Theta$

$$\sqrt{n}(\theta_n - \theta_\star) \to_d \mathcal{N}(0, H_\star^{-1}),$$

where $H_\star := H(\theta_\star) := \nabla^2 L(\theta_\star)$.

[¶] Cramér '46, Huber '74, Ibragimov and Has'minskii '81, van der Vaart '00.

Related Work: Asymptotic Theory[¶]

Well-specified model: $P \in \mathcal{P}_\Theta$

$$\sqrt{n}(\theta_n - \theta_\star) \to_d \mathcal{N}(0, H_\star^{-1}),$$

where $H_\star := H(\theta_\star) := \nabla^2 L(\theta_\star)$.

Mis-specified model: $P \notin \mathcal{P}_\Theta$

$$\sqrt{n}(\theta_n - \theta_\star) \to_d \mathcal{N}(0, H_\star^{-1} G_\star H_\star^{-1}),$$

where $G_\star := G(\theta_\star) := \mathbb{E}[\nabla \ell(\theta_\star; Z) \nabla \ell(\theta_\star; Z)^\top]$.

[¶] Cramér '46, Huber '74, Ibragimov and Has'minskii '81, van der Vaart '00.

Related Work: Non-Asymptotic Theory

Specific models

- Gaussian regression (Baraud '04).
- Ridge regression (Hsu et al '14).
- Logistic regression (Bach '10).

Related Work: Non-Asymptotic Theory

Specific models

- Gaussian regression (Baraud ’04).
- Ridge regression (Hsu et al ’14).
- Logistic regression (Bach ’10).

General approaches

- Empirical process (Spokoiny ’12).
- Convex optimization (Ostrovskii and Bach ’21).

Non-Asymptotic Theory with Strong Convexity

Non-asymptotic theory: with high probability,

$$\underbrace{L(\theta_n) - L(\theta_\star)}_{\text{Excess risk}} \leq O(n^{-1}).$$

Non-Asymptotic Theory with Strong Convexity

Non-asymptotic theory: with high probability,

$$\underbrace{\nabla L(\theta_\star)(\theta_n - \theta_\star)}_{0} + \frac{1}{2}(\theta_n - \theta_\star)^\top H(\bar{\theta})(\theta_n - \theta_\star) = \underbrace{L(\theta_n) - L(\theta_\star)}_{\text{Excess risk}} \leq O(n^{-1}).$$

Non-Asymptotic Theory with Strong Convexity

Non-asymptotic theory: with high probability,

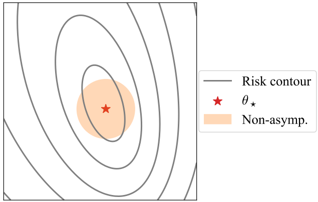
$$\underbrace{\nabla L(\theta_\star)(\theta_n - \theta_\star)}_{0} + \frac{1}{2}(\theta_n - \theta_\star)^\top H(\bar{\theta})(\theta_n - \theta_\star) = \underbrace{L(\theta_n) - L(\theta_\star)}_{\text{Excess risk}} \le O(n^{-1}).$$

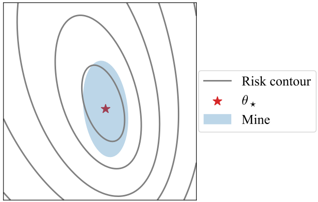
Strong convexity $H(\theta) \succeq \lambda I$

$$\lambda \|\theta_n - \theta_\star\|_2^2 \le O(n^{-1}).$$

Self-Concordance $H(\bar{\theta}) \approx H_n(\theta_n)$

$$\|H_n(\theta_n)^{1/2}(\theta_n - \theta_\star)\|_2^2 \le O(n^{-1}).$$

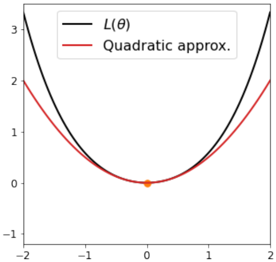




Strong Convexity versus Self-Concordance

Strong convexity

- Globally lower bounded Hessian.
- No control on how Hessian varies.



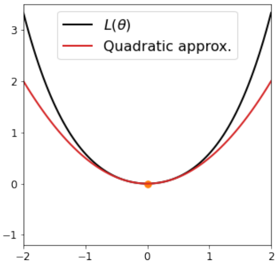
Strong Convexity versus Self-Concordance

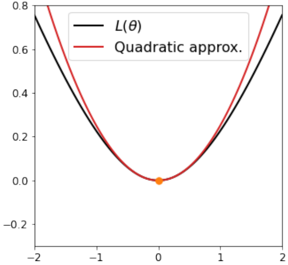
Strong convexity

- Globally lower bounded Hessian.
- No control on how Hessian varies.

Self-concordance

- No global lower bound.
- Slowly varying Hessian.





Self-Concordance

Define $\mathrm{D}f(x)[u] := \frac{\mathrm{d}}{\mathrm{d}t} f(x + tu)|_{t=0}$ and $\mathrm{D}^2 f(x)[u, u] := \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(x + tu)|_{t=0}$.

Definition 1 (Nesterov and Nemirovskii '94)

Let f be closed and convex. We say f is *self-concordant* with parameter $R > 0$ if

$$\left|\mathrm{D}^3 f(x)[u, u, u]\right| \leq R \left|\mathrm{D}^2 f(x)[u, u]\right|^{3/2}.$$

Self-Concordance

Define $\mathrm{D}f(x)[u] := \frac{\mathrm{d}}{\mathrm{d}t} f(x+tu)|_{t=0}$ and $\mathrm{D}^2 f(x)[u,u] := \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(x+tu)|_{t=0}$.

Definition 1 (Nesterov and Nemirovskii '94)

Let f be closed and convex. We say f is *self-concordant* with parameter $R > 0$ if

$$\left|\mathrm{D}^3 f(x)[u,u,u]\right| \leq R \left|\mathrm{D}^2 f(x)[u,u]\right|^{3/2}.$$

- Newton's method.
- Interior point methods.
- Most non-quadratic loss functions are not self-concordant.

Pseudo Self-Concordance

Definition 2 (Bach '10)

Let f be closed and convex. We say f is *pseudo self-concordant* with parameter $R > 0$ if

$$\left|\mathrm{D}^3 f(x)[u, u, u]\right| \leq R\|u\|_2 \mathrm{D}^2 f(x)[u, u].$$

Pseudo Self-Concordance

Definition 2 (Bach '10)

Let f be closed and convex. We say f is *pseudo self-concordant* with parameter $R > 0$ if

$$\left|\mathrm{D}^3 f(x)[u, u, u]\right| \leq R\|u\|_2 \mathrm{D}^2 f(x)[u, u].$$

- **Hessian approximation**:

$$e^{-R\|y-x\|_2} \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq e^{R\|y-x\|_2} \nabla^2 f(x).$$

- **Localization**: $x_\star := \arg\min_x f(x)$ satisfies

$$\|x_\star - x\|_{\nabla^2 f(x)} \lesssim \|\nabla f(x)\|_{\nabla^2 f(x)^{-1}},$$

where $\|u\|_A := \sqrt{u^\top A u}$.

Effective Dimension

Effective dimension $d_\star := \mathbf{Tr}(H_\star^{-1/2} G_\star H_\star^{-1/2})$

- **Well-specified model**: $d_\star = d$.
- **Mis-specified model**:
  - Problem-specific characterization of the complexity of Θ.
  - The sandwich covariance is the limiting covariance of $\sqrt{n} H_\star^{1/2}(\theta_n - \theta_\star)$.

Effective Dimension

Effective dimension $d_\star := \mathbf{Tr}(H_\star^{-1/2} G_\star H_\star^{-1/2})$

- **Well-specified model**: $d_\star = d$.
- **Mis-specified model**:
  - Problem-specific characterization of the complexity of Θ.
  - The sandwich covariance is the limiting covariance of $\sqrt{n} H_\star^{1/2}(\theta_n - \theta_\star)$.

		Poly-Poly	**Poly-Exp**	**Exp-Poly**	**Exp-Exp**
Eigendecay	$G_\star$	$i^{-\alpha}$	$i^{-\alpha}$	$e^{-\mu i}$	$e^{-\mu i}$
	$H_\star$	$i^{-\beta}$	$e^{-\nu i}$	$i^{-\beta}$	$e^{-\nu i}$
Ratio	$d_\star / d$	$d^{(\beta-\alpha)\vee(-1)}$	$d^{-\alpha} e^{\nu d}$	d^{-1}	1 if $\mu = \nu$ d^{-1} if $\mu > \nu$ $d^{-1} e^{(\nu-\mu)d}$ if $\mu < \nu$

Main Results

Theorem 3 (Informal)

Under the ***pseudo self-concordance*** *assumption and other assumptions, whenever*

$$n \gtrsim O(d + d_\star),$$

with probability at least $1 - \delta$, *the MLE* θ_n *uniquely exists and satisfies*

$$n \|\theta_n - \theta_\star\|_{H_\star}^2 \lesssim \log(1/\delta) d_\star .$$

Main Results

Theorem 3 (Informal)

*Under the **pseudo self-concordance** assumption and other assumptions, whenever*

$$n \gtrsim O(d + d_\star),$$

with probability at least $1 - \delta$*, the MLE* θ_n *uniquely exists and satisfies*

$$n \|\theta_n - \theta_\star\|_{H_\star}^2 \lesssim \log(1/\delta) d_\star.$$

- Recall $\sqrt{n} H_\star^{1/2}(\theta_n - \theta_\star) \to_d \mathcal{N}(0, H_\star^{-1/2} G_\star H_\star^{-1/2}) \Rightarrow n \|\theta_n - \theta_\star\|_{H_\star}^2 \approx d_\star$.
- Characterize the critical sample size.
- **Localization**: $\|\theta_n - \theta_\star\|_{H_n(\theta_\star)}^2 \lesssim \|\nabla L_n(\theta_\star)\|_{H_n(\theta_\star)^{-1}}^2$.

Main Results

Confidence bound

- Approximate $H_\star$ by $H_n(\theta_n)$ (**Hessian approximation**).
- Approximate $d_\star$ by $d_n := \mathbf{Tr}(H_n(\theta_n)^{-1/2} G_n(\theta_n) H_n(\theta_n)^{-1/2})$.

Theorem 4 (Informal)

Under the ***pseudo self-concordance*** *assumption and other assumptions, whenever*

$$n \gtrsim O(d \log n + d_\star),$$

with probability at least $1-\delta$*, the MLE* θ_n *uniquely exists and satisfies*

$$n \left\|\theta_n - \theta_\star\right\|^2_{H_n(\theta_n)} \lesssim \log(1/\delta) d_n.$$

Semi-Parametric Estimation

- **Nuisance parameter** $g_0 \in (\mathcal{G}, \|\cdot\|_{\mathcal{G}})$.
- **Population risk** $L(\theta, g) := \mathbb{E}[\ell(\theta, g; Z)]$.
- **Two-step learning procedure based on sample-splitting**[‖]
  - Obtain a nonparametric estimator $\hat{g}$ on one sub-sample.
  - Estimate $\theta_\star$ via empirical risk minimization on another sub-sample:

$$\theta_n = \arg\min_{\theta \in \Theta} L_n(\theta, \hat{g}).$$

Example 5 (Robinson '88)

Let Y outcome, D treatment, and X control. Consider

$$Y = D\theta_\star + g_0(X) + U.$$

[‖]Chernozhukov et al '18, Foster and Syrgkanis '20, Chaudhuri et al '07.

Semi-Parametric Estimation

Theorem 6 (Informal)

*Under the **pseudo self-concordance** and other assumptions, with probability at least* $1 - \delta$*,*

$$\|\theta_n - \theta_\star\|_{H_\star}^2 \lesssim \frac{d_\star}{n} \log(1/\delta) + \|\hat{g} - g_0\|_{\mathcal{G}}^2 .$$

- If g_0 is p-smooth, it can be estimated at rate $O(n^{-p/(2p+d)})$.
- The term $\|\hat{g} - g_0\|_{\mathcal{G}}^2$ **cannot** achieve the $O(n^{-1})$ rate.

Semi-Parametric Estimation

Neyman orthogonality (Neyman '79)

$$\mathrm{D}_g \nabla_\theta L(\theta_\star, g_0)[g - g_0] = 0.$$

Theorem 7 (Informal)

Under the ***pseudo self-concordance****, Neyman orthogonality, and other assumptions, with probability at least* $1 - \delta$*,*

$$\|\theta_n - \theta_\star\|_{H_\star}^2 \lesssim \frac{d_\star}{n} \log(1/\delta) + \|\hat{g} - g_0\|_{\mathcal{G}}^4 .$$

- If g_0 is p-smooth, it can be estimated at rate $O(n^{-p/(2p+d)})$.
- The term $\|\hat{g} - g_0\|_{\mathcal{G}}^4$ **can** achieve the $O(n^{-1})$ rate as long as $p \geq d/2$.

Part II. Independence Testing with Entropy Regularized Optimal Transport

Soumik Pal

Zaid Harchaoui

@ AISTATS 2022 (Oral)

Independence Testing

Problem:

- Let $(X, Y) \sim \mu_{XY}$ on $\mathcal{X} \times \mathcal{Y}$ with marginals μ_X and μ_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 $T_n(X, Y) > t_n(\alpha)$.

Related Work

Independence criteria:

- Classical independence criterion (Hoeffding '48, Kruskal '58, Lehmann '66)
  - Pearson's correlation coefficient.
  - Spearman's ρ.
  - Kendall's τ.

Related Work

Independence criteria:

- Classical independence criterion (Hoeffding ’48, Kruskal ’58, Lehmann ’66)
  - Pearson’s correlation coefficient.
  - Spearman’s ρ.
  - Kendall’s τ.
- Distance-based independence criterion.
  - Distance covariance (dCov) (Székely et al. ’07).
  - Hilbert-Schmidt independence criterion (HSIC) (Gretton et al. ’05).

Related Work

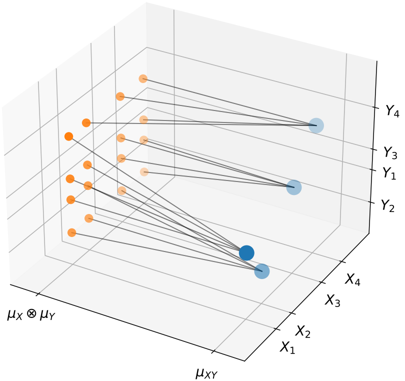
Independence criteria:

- Classical independence criterion (Hoeffding '48, Kruskal '58, Lehmann '66)
  - Pearson's correlation coefficient.
  - Spearman's ρ.
  - Kendall's τ.
- Distance-based independence criterion.
  - Distance covariance (dCov) (Székely et al. '07).
  - Hilbert-Schmidt independence criterion (HSIC) (Gretton et al. '05).
- Optimal transport based independence criterion.
  - Wasserstein correlation coefficient (Wiesel '21, Mordant and Segers '21, Nies et al. '21).
  - Rank-based independence criterion (Shi et al. '20, Deb & Sen '21).

Entropy Regularized Optimal Transport Independence Criterion

ETIC—define $T(X, Y)$ by

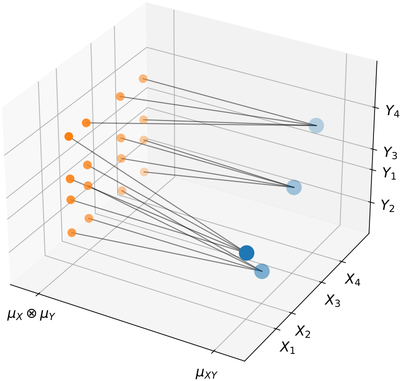
$$\bar{S}_\varepsilon(\mu_{XY}, \mu_X \otimes \mu_Y) := S_\varepsilon(\mu_{XY}, \mu_X \otimes \mu_Y) - S_\varepsilon(\mu_{XY}, \mu_{XY})/2 - S_\varepsilon(\mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y)/2.$$



Entropy Regularized Optimal Transport Independence Criterion

ETIC—define $T(X, Y)$ by

$$\bar{S}_\varepsilon(\mu_{XY}, \mu_X \otimes \mu_Y) := S_\varepsilon(\mu_{XY}, \mu_X \otimes \mu_Y) - S_\varepsilon(\mu_{XY}, \mu_{XY})/2 - S_\varepsilon(\mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y)/2.$$



Statistical Properties of ETIC

- **Test statistic** $T_n(X, Y) := \bar{S}_\varepsilon(\hat{\mu}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y)$.
- **Absolute error** $|T_n(X, Y) - T(X, Y)|$.
- **Upper bound via duality**

$$\underbrace{\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i, Y_i) - \mathbb{E}[f(X, Y)] \right|}_{\text{Empirical process theory}} + \underbrace{\sup_{f \in \mathcal{F}} \left| \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f(X_i, Y_j) - \mathbb{E}[f(X, Y')] \right|}_{\text{U-process theory}},$$

where $\mathcal{F}$ is some smooth function class.

Statistical Properties of ETIC

Theorem 8

Assume that μ_X and μ_Y are supported on a bounded domain with radius R. Then we have, with probability at least $1-\delta$,

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

Statistical Properties of ETIC

Theorem 8

Assume that μ_X and μ_Y are supported on a bounded domain with radius R. Then we have, with probability at least $1-\delta$,

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

Remark 1

- *Rate of convergence* $O(n^{-1/2})$.
- *The choice of* $\varepsilon = R^2$ *gives* $C_d \sqrt{\log(6/\delta)} R^2 / \sqrt{n}$.

Statistical Properties of ETIC

Theorem 8

Assume that μ_X and μ_Y are supported on a bounded domain with radius R. Then we have, with probability at least $1-\delta$,

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

Remark 2

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 $n \to \infty$.*

Computational Aspects of ETIC

The information projection formulation

$$\min_{\gamma \in \mathrm{CP}(\hat{\mu}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y)} \mathrm{KL}(\gamma \| R_\varepsilon).$$

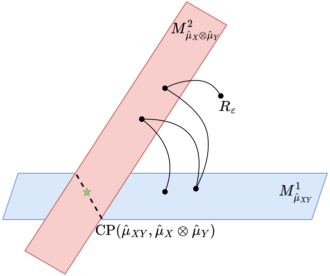
- $\mathrm{CP}(\hat{\mu}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y) = M^1_{\hat{\mu}_{XY}} \cap M^2_{\hat{\mu}_X \otimes \hat{\mu}_Y}$.
- $M^1_{\hat{\mu}_{XY}} := \{\gamma : \text{the first marginal is } \hat{\mu}_{XY}\}$.
- $M^2_{\hat{\mu}_X \otimes \hat{\mu}_Y} := \{\gamma : \text{the second marginal is } \hat{\mu}_X \otimes \hat{\mu}_Y\}$.

Computational Aspects of ETIC

The information projection formulation

$$\min_{\gamma \in \mathrm{CP}(\hat\mu_{XY}, \hat\mu_X \otimes \hat\mu_Y)} \mathrm{KL}(\gamma \| R_\varepsilon).$$

- $\mathrm{CP}(\hat\mu_{XY}, \hat\mu_X \otimes \hat\mu_Y) = M^1_{\hat\mu_{XY}} \cap M^2_{\hat\mu_X \otimes \hat\mu_Y}$.
- Deming and Stephan '40.
- Sinkhorn '64.

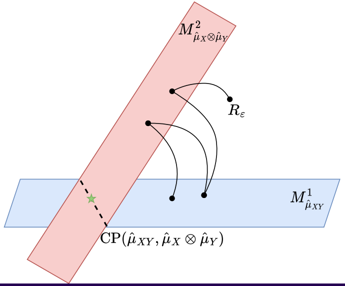


Computational Aspects of ETIC

The information projection formulation

$$\min_{\gamma \in \mathrm{CP}(\hat{\mu}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y)} \mathrm{KL}(\gamma \| R_\varepsilon).$$

- $\mathrm{CP}(\hat{\mu}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y) = M^1_{\hat{\mu}_{XY}} \cap M^2_{\hat{\mu}_X \otimes \hat{\mu}_Y}$.
- **Sinkhorn algorithm**: $O(n^4)$ time and $O(n^4)$ space.
- **Our algorithm**: $O(n^2)$ time and $O(n)$ space.



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

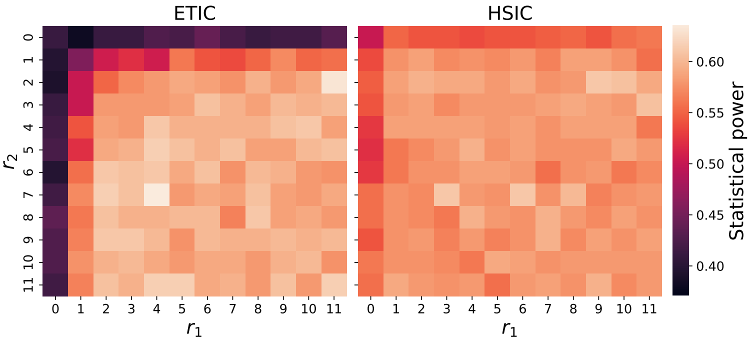
- Parallel European Parliament corpus (Koehn '05).
- Randomly select $n = 64$ English documents and a paragraph in each document.
- (English paragraph, random paragraph in the same document in French).
- Feature embeddings of dimension 768 with LaBSE (Feng et al. '20).

Independence tests

- HSIC with Gaussian kernels.
- ETIC with the weighted quadratic cost and same parameters.
- Hyper-parameters: $r_1, r_2 \in [0.25, 4]$.

Independence Testing on Bilingual Text

ETIC outperforms HSIC for many values of r_1 and r_2.



Part III. Future Directions

Higher Order Orthogonality in Semi-Parametric Estimation

- Partially linear model (PLM) with non-Gaussian residual.
  - The two-stage estimator has large bias.
  - Need more robustness!

Higher Order Orthogonality in Semi-Parametric Estimation

- Partially linear model (PLM) with non-Gaussian residual.
  - The two-stage estimator has large bias.
  - Need more robustness!
- *k*-**orthogonality** (Mackey et al '18)

$$\mathrm{D}_g^t \nabla_\theta L(\theta_\star, g_0)[\underbrace{g - g_0, \ldots, g - g_0}_{t}] = 0, \quad \forall t \leq k.$$

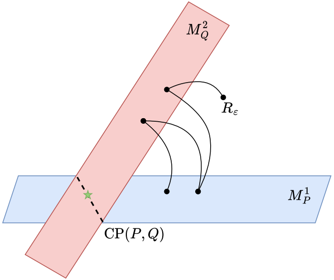
- **Robustness**: g_0 only needs to be estimated at rate $O(n^{-1/(2k+2)})$.
- **Feasibility**: we can construct a 2-orthogonal risk for the PLM with non-Gaussian residual.

Alternating Procedures in Statistics

Entropy regularized optimal transport

$$\operatorname*{arg\,min}_{\gamma \in \mathrm{CP}(P,Q)} \mathrm{KL}(\gamma \| R_\varepsilon),$$

where $\mathrm{CP}(P, Q) = M_P^1 \cap M_Q^2$.

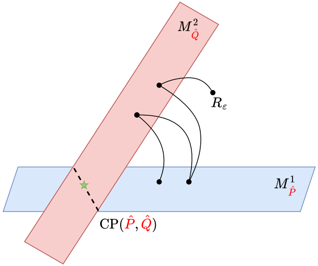


Alternating Procedures in Statistics

Entropy regularized optimal transport

$$\arg\min_{\gamma \in \mathrm{CP}(\hat{P},\hat{Q})} \mathrm{KL}(\gamma \| R_\varepsilon),$$

where $\hat{P}$ and $\hat{Q}$ are estimated from data.

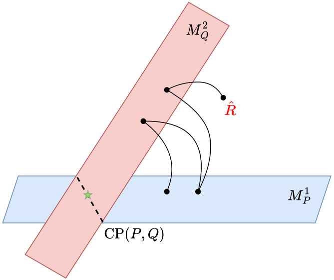


Alternating Procedures in Statistics

Iterative proportional fitting (raking)

$$\underset{\gamma \in \mathrm{CP}(P,Q)}{\arg\min}\ \mathrm{KL}(\gamma \| \hat{R}),$$

where P and Q are known, and $\hat{R}$ is estimated from data.



Alternating Procedures in Statistics

Alternating conditional expectations

$$f(X,Y) - \underset{h(X,Y)\in(H_1+H_2)^{\perp}}{\arg\min}\ \mathbb{E}[(f(X,Y)-h(X,Y))^2],$$

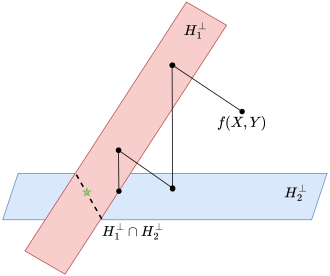
where $H_1 := \{h_1(X) \in \mathbf{L}^2\}$ and $H_2 := \{h_2(Y) \in \mathbf{L}^2\}$.

Alternating Procedures in Statistics

Alternating conditional expectations

$$f(X,Y) - \underset{h(X,Y)\in H_1^{\perp}\cap H_2^{\perp}}{\arg\min} \mathbb{E}[(f(X,Y) - h(X,Y))^2],$$

where $H_1 := \{h_1(X) \in \mathbf{L}^2\}$ and $H_2 := \{h_2(Y) \in \mathbf{L}^2\}$.

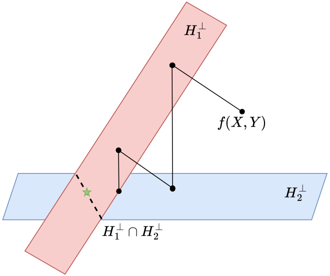


Alternating Procedures in Statistics

Alternating conditional expectations**

$$f(X_{1:n}, Y_{1:n}) - \operatorname*{arg\,min}_{h(X_{1:n}, Y_{1:n}) \in H_1^\perp \cap H_2^\perp} \mathbb{E}[(f(X_{1:n}, Y_{1:n}) - h(X_{1:n}, Y_{1:n}))^2],$$

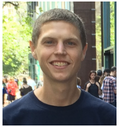
where $H_1 := \{\sum_{i=1}^n h_1(X_i) \in \mathbf{L}^2\}$ and $H_2 := \{\sum_{i=1}^n h_2(Y_i) \in \mathbf{L}^2\}$.

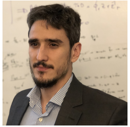
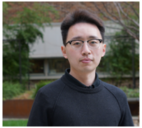


**Harchaoui, Liu, and Pal. *Under review*, 2022.

Thank You




W NSF IFDS Institute for Foundations of Data Science

Schrödinger's Lazy Gas Experiment

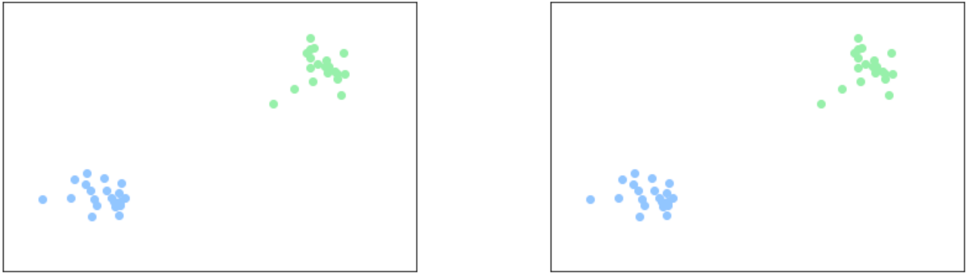


Figure: **Left**: high temperature; **Right**: low temperature.

The Schrödinger Bridge

The Schrödinger bridge (Föllmer ’88, Léonard ’12)

- A particle L making jumps according to

$$f_\varepsilon(y \mid x) \propto \exp\left(-\frac{1}{\varepsilon}\|x-y\|^2\right).$$

- Observe initial and terminal configurations $L_0 \sim P$ and $L_1 \sim Q$.
- What is the most likely joint distribution (or coupling) between L_0 and L_1?

$$P \xrightarrow{f_\varepsilon(y \mid x)} Q$$

The Schrödinger Bridge Problem and Entropy-Regularized OT

The Schrödinger bridge (Föllmer '88, Léonard '12)

- Consider a Markov chain with initial distribution P and transition probability f_ε.
- The joint distribution is

$$R_\varepsilon(x, y) := P(x) f_\varepsilon(y \mid x).$$

- Conditioned on the initial and terminal configurations being P and Q,

$$\mu_{\mathrm{SB}} := \operatorname*{arg\,min}_{\gamma \in \mathrm{CP}(P,Q)} \mathrm{KL}(\gamma \| R_\varepsilon). \tag{1}$$

Partially Linear Model

Let Y outcome, D treatment, and X control. Consider

$$Y = D\theta_0 + \alpha_0(X) + U$$
$$D = \beta_0(X) + V.$$

- Partialling out the effect of X

$$Y = (D - \beta_0(X))\theta_0 + \gamma_0(X) + U$$

- Reparameterization $g_0 = (\beta_0, \gamma_0)$.
- Neyman orthogonal risk

$$L(\theta, g) := \mathbb{E}\left[(Y - \gamma(X) - (D - \beta(X))\theta)^2\right].$$

Proof Sketch for the OSL Estimation Bound

By Taylor's theorem,

$$
\begin{aligned}
0 &\geq L_n(\theta_n, \hat{g}) - L_n(\theta_\star, \hat{g}) \\
&= \nabla_\theta L_n(\theta_\star, \hat{g})^\top (\theta_n - \theta_\star) + \|\theta_n - \theta_\star\|^2_{H_n(\bar{\theta}, \hat{g})}/2 \\
&= [\nabla_\theta L_n(\theta_\star, \hat{g}) - \nabla_\theta L(\theta_\star, \hat{g})]^\top (\theta_n - \theta_\star) + \nabla_\theta L(\theta_\star, \hat{g})^\top (\theta_n - \theta_\star) + \|\theta_n - \theta_\star\|^2_{H_n(\bar{\theta}, \hat{g})}/2 \\
&\geq \|\nabla_\theta L_n(\theta_\star, \hat{g}) - \nabla_\theta L(\theta_\star, \hat{g})\|_{H_\star^{-1}} \|\theta_n - \theta_\star\|_{H_\star} + \nabla_\theta L(\theta_\star, \hat{g})^\top (\theta_n - \theta_\star) + \|\theta_n - \theta_\star\|^2_{H_n(\bar{\theta}, \hat{g})}/2 \\
&\gtrsim - \left[\sqrt{d_\star / n} + \|\hat{g} - g_0\|^2_{\mathcal{G}} \right] \|\theta_n - \theta_\star\|_{H_\star} + \|\theta_n - \theta_\star\|^2_{H_\star}.
\end{aligned}
$$

Properties of ETIC

Proposition 1 (Informal)

*Let $\mathcal{X}$ and $\mathcal{Y}$ be compact equipped with Lipschitz costs c_1 and c_2. Assume that $k_i := \exp(-c_i/\varepsilon)$ are universal for $i = 1, 2$. Then ETIC with $c := c_1 \oplus c_2$ is a **valid independence criterion**.*

Proposition 2 (Informal)

*Let $p = \Omega(\log n/\tau^2)$ be the number of **random features**. Then the random feature approximation of ETIC is of accurate with error at most τ.*

Properties of ETIC

Hilbert-Schmidt independence criterion (HSIC)

- Two kernels k and l.

$$\mathrm{HSIC}(X, Y) = \mathbb{E}[k(X_1, X_2)l(Y_1, Y_2)] + \mathbb{E}[k(X_1, X_2)l(Y_3, Y_4)] - \frac{1}{2}\mathbb{E}[k(X_1, X_2)l(Y_1, Y_3)]$$

Proposition 3 (Informal)

Under appropriate assumptions, we have

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

and

$$T_\varepsilon(X, Y) \to OT(\mu_{XY}, \mu_X \otimes \mu_Y), \quad \text{as } \varepsilon \to 0.$$

Computational Aspects of ETIC

Sinkhorn

- **Inputs**: $a, b \in \mathbb{R}^{n^2}$ and $K \in \mathbb{R}^{n^2 \times n^2}$.
- **Initialization**: $u, v \in \mathbb{R}^{n^2}$.
- **Update**:

$$u \leftarrow a \oslash Kv$$
$$v \leftarrow b \oslash K^\top u.$$

- Time $O(n^4)$ and space $O(n^4)$.

Tensor Sinkhorn

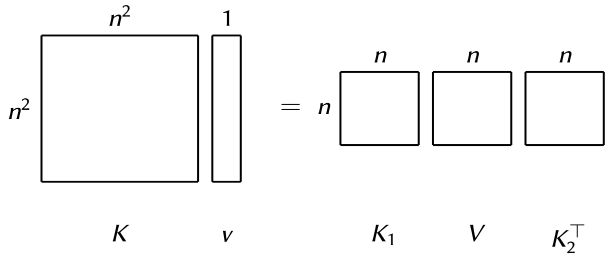
- **Inputs**: $A, B, K_1, K_2 \in \mathbb{R}^{n \times n}$.
- **Initialization**: $U, V \in \mathbb{R}^{n \times n}$.
- **Update**:

$$U \leftarrow A \oslash (K_1 V K_2^\top)$$
$$V \leftarrow B \oslash (K_1^\top U K_2).$$

- Time $O(n^3)$ and space $O(n^2)$.

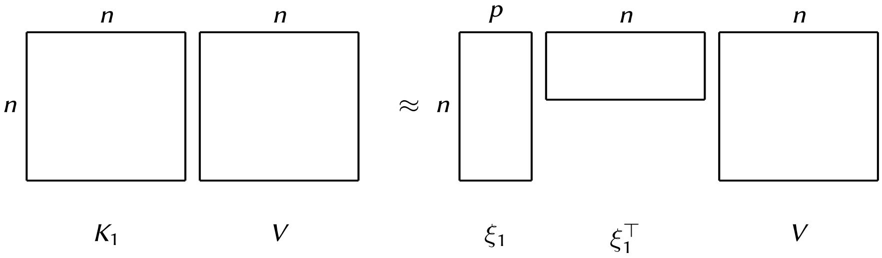
Computational Aspects of ETIC

Algorithm	Strategy	Basic operation	Time	Space
Sinkhorn	Alternative projection	Kv	$O(n^4)$	$O(n^4)$
Tensor Sinkhorn (TS)	$K = K_1 \otimes K_2$	$K_1 V K_2^\top$	$O(n^3)$	$O(n^2)$



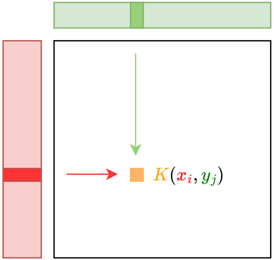
Computational Aspects of ETIC

Algorithm	Strategy	Basic operation	Time	Space
Sinkhorn	Alternative projection	Kv	$O(n^4)$	$O(n^4)$
Tensor Sinkhorn (TS)	$K = K_1 \otimes K_2$	$K_1 V K_2^\top$	$O(n^3)$	$O(n^2)$
TS + Random Features (TS-RF)	$K_i \approx \xi_i \xi_i^\top$	$\xi_1 \xi_1^\top V \xi_2 \xi_2^\top$	$O(pn^2)$	$O(n^2)$



Computational Aspects of ETIC

Algorithm	Strategy	Basic operation	Time	Space
Sinkhorn	Alternative projection	Kv	$O(n^4)$	$O(n^4)$
Tensor Sinkhorn (TS)	$K = K_1 \otimes K_2$	$K_1 V K_2^\top$	$O(n^3)$	$O(n^2)$
TS + Random Features (TS-RF)	$K_i \approx \xi_i \xi_i^\top$	$\xi_1 \xi_1^\top V \xi_2 \xi_2^\top$	$O(pn^2)$	$O(n^2)$
Large scale TS-RF (LS-RF)	Symbolic matrices	$\xi_1 \xi_1^\top V \xi_2 \xi_2^\top$	$O(pn^2)$	$O(pn)$



Computational Aspects of ETIC

The large-scale implementation is efficient in both time and memory.