Entropy Regularized Optimal Transport Independence Criterion

Lang Liu
University of Washington

Soumik Pal
University of Washington

Zaid Harchaoui
University of Washington

Abstract

We introduce an independence criterion based on entropy regularized optimal transport. Our criterion can be used to test for independence between two samples. We establish non-asymptotic bounds for our test statistic and study its statistical behavior under both the null hypothesis and the alternative hypothesis. The theoretical results involve tools from U-process theory and optimal transport theory. We also offer a random feature type approximation for large-scale problems, as well as a differentiable program implementation for deep learning applications. We present experimental results on existing benchmarks for independence testing, illustrating the interest of the proposed criterion to capture both linear and nonlinear dependencies in synthetic data and real data.

1 INTRODUCTION

Statistical independence measures have been widely used in machine learning and statistics, ranging from independent component analysis (Bach and Jordan, 2002; Gretton et al., 2005) to causal inference (Pfister et al., 2018; Chakraborty and Zhang, 2019), and recently in self-supervised learning (Li et al., 2021) and representation learning (Ozair et al., 2019). Classical dependence measures such as Pearson's correlation coefficient, Spearman's ρ, and Kendall's τ (Hoeffding, 1948; Kruskal, 1958; Lehmann, 1966) focus on real-valued one dimensional random variables and thus are not suitable for high dimensional data; see also (Schweizer and Wolff, 1981; Nikitin, 1995).

One popular choice of independence measures in high dimension is the Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005). This criterion was used to test independence by Gretton et al. (2007b). Several extensions of HSIC are available, such as a relative dependency measure (Bounliphone et al., 2015) and a joint independence measure among multiple random elements (Pfister et al., 2018). Another choice is the distance covariance (dCov) of Székely et al. (2007). dCov was originally developed in Euclidean spaces using characteristic functions and later generalized to metric spaces (Lyons, 2013). In fact, in their most general form, HSIC and dCov are equivalent as shown by Sejdinovic et al. (2013).

A different line of research explored optimal transport to measure dependence. The Wasserstein distance naturally defines a dependence measure when it is used to quantify the dissimilarity between the joint distribution and the product of marginals; see, e.g., (Cifarelli and Regazzini, 2017). The normalized version—the so-called Wasserstein correlation coefficient—has recently gained attention in (Wiesel, 2021; Mordant and Segers, 2021; Nies et al., 2021). Following the classical rank-based tests such as Pearson's ρ, optimal transport is also used to define multivariate ranks and the subsequent independence tests (Shi et al., 2020; Deb and Sen, 2021). However, these tests can suffer from the curse of dimensionality or high computational complexity, limiting their practical usefulness; see (Peyré and Cuturi, 2019) for a discussion.

A remedy to this challenge is to use the entropy regularized formulation of optimal transport. This is particularly attractive from both a computational viewpoint (Cuturi, 2013) and a statistical viewpoint (Rigollet and Weed, 2018). Moreover the empirical counterpart of entropy regularized optimal transport enjoys as an estimator a parametric rate of convergence and thus appears to overcome the curse of dimensionality (Genevay et al., 2019; Mena and Weed, 2019). The Sinkhorn divergence (Feydy et al., 2019), its centered version, defines a semi-metric on probability measures which metrizes weak convergence. Ramdas et al. (2017) used it for two-sample testing and Genevay et al. (2018) for generative modeling; see also (Salimans et al., 2018; Sanjabi et al., 2018). Our ap-

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

proach to independence testing shall build upon this recent progress on entropic regularization.

Outline. In Section 2, we introduce the entropy regularized optimal transport independence criterion (ETIC) and discuss its key properties. We propose the Tensor Sinkhorn algorithm with a random feature approximation to compute ETIC, which admits a quadratic scaling in time and space. We also show how to approximate ETIC using random features, and how to differentiate through ETIC in a framework of differentiable programming. In Section 3, we give our main theoretical results, i.e., non-asymptotic bounds, characterizing the statistical behavior of the empirical estimator of ETIC under both the null and alternative hypotheses. In Section 4, we compare the empirical behavior of ETIC with HSIC on both synthetic and real data. The Appendix contains detailed proofs as well as additional experiments.

Related Work. Statistical metrics on the space of probability measures form the backbone of many dependence measures. On the machine learning side, distributions are compared by embedding them into reproducing kernel Hilbert spaces (Gretton et al., 2007a, 2012). The Hilbert-Schmidt independence criterion (HSIC) uses Hilbertian embeddings of probability distributions to compare the joint distribution and the product of marginals (Gretton et al., 2005, 2007b). On the statistics side, distributions defined on Euclidean spaces are compared via their characteristic functions, leading to the so-called energy distance (Székely and Rizzo, 2004).

A closely related dependence measure is the distance covariance (Székely et al., 2007). These distances were later generalized to general metric spaces of negative type by Lyons (2013), unifying the two notions via the Barycenter map—a quantity similar to the feature map in kernel methods. In fact, the kernel-based and distance-based approaches are equivalent (Sejdinovic et al., 2013). Their corresponding empirical estimators all admit a U-statistics expression, and enjoy a convergence rate that is independent of the dimension. These results can be established using tools from U-statistics theory; see, e.g., (Serfling, 1980).

On the other hand, Wasserstein distances provide a class of metrics on the space of probability measures with nice geometric properties (Ambrosio et al., 2005). However, it is known that its empirical estimate suffers from the curse of dimensionality (Dudley, 1969; Fournier and Guillin, 2015; Weed and Bach, 2019; Lei, 2020), limiting their usage in high-dimension problems. A remedy to this issue is to introduce the entropic regularization. Genevay et al. (2019) showed that the plug-in estimator of the entropy regularized optimal transport cost possesses a parametric rate of convergence when the measures are compactly supported. Their results can be extended to sub-Gaussian distributions (Mena and Weed, 2019) with tools from empirical process theory.

The independence criterion we propose uses entropy regularized optimal transport to compare the joint distribution and the product of marginals. The empirical counterpart involves a product of two empirical measures, leading to a two-sample U-process on paired samples. The resulting U-process requires a sophisticated analysis of its statistical behavior; common tools from empirical processes are ineffective here. Using the decoupling technique from Peña and Giné (1999) and duality theory (Peyré and Cuturi, 2019), we prove a rate of convergence roughly $O(\sigma^{3d} n^{-1/2})$, where n is the sample size, d is the ambient dimension, and σ is the sub-Gaussian parameter, recovering previous results for two sample statistics.

2 ENTROPY REGULARIZED OPTIMAL TRANSPORT INDEPENDENCE CRITERION

We introduce in this section the entropy regularized optimal transport independence criterion (ETIC) and discuss its key properties. We design an independence test based on ETIC and develop an efficient algorithm—the Tensor Sinkhorn algorithm—to compute its test statistic. We also provide an approximation using random features for large-scale problems, as well as a differentiable program implementation for deep learning applications.

Notation. For a Euclidean space $\mathcal{Z}$ equipped with the Borel σ-algebra, let $\mathcal{M}_1(\mathcal{Z})$ be the set of probability measures defined on $\mathcal{Z}$. Let (X, Y) be a pair of random vectors with respective dimension d_1 and d_2 following some joint distribution $P_{XY} \in \mathcal{M}_1(\mathbb{R}^d)$ where $d := d_1 + d_2$. Denote $P_X \in \mathcal{M}_1(\mathbb{R}^{d_1})$ and $P_Y \in \mathcal{M}_1(\mathbb{R}^{d_2})$ the marginal distributions of X and Y, respectively. Given $Q \in \mathcal{M}_1(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2})$ and a real-valued function f on the same domain, we denote $Q[f]$ the expectation $\mathbb{E}_{(X,Y)\sim Q}[f(X,Y)]$. We adopt the notation from the empirical process theory and write $\|Q\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |Q[f]|$ for a real-valued function class $\mathcal{F}$. We say Q is sub-Gaussian with parameter σ^2, denoted as subG(σ^2), if $\mathbb{E}_Q\, e^{\|(X,Y)\|^2/(2d\sigma^2)} \le 2$; see, e.g., Vershynin (2018). We write $\mathbb{E} := \mathbb{E}_{P_{XY}}$ for short.

ETIC. Let $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ be a continuous cost function satisfying $c((x, y), (x', y')) = 0$ iff $(x, y) = (x', y')$. We introduce the entropy regularized optimal transport independence criterion (ETIC):

$$T(X,Y) := T_\varepsilon(X,Y) := \bar{S}_\varepsilon(P_{XY}, P_X \otimes P_Y), \quad (1)$$

where $\bar{S}_\varepsilon$ is a divergence defined as $\bar{S}_\varepsilon(P_{XY}, P_X \otimes P_Y) := S_\varepsilon(P_{XY}, P_X \otimes P_Y) - \frac{1}{2}S_\varepsilon(P_{XY}, P_{XY}) - \frac{1}{2}S_\varepsilon(P_X \otimes P_Y, P_X \otimes P_Y)$. Here $S_\varepsilon(P_{XY}, P_X \otimes P_Y)$ is the entropy regularized optimal transport cost between P_{XY} and $P_X \otimes P_Y$, i.e.,

$$\inf_\gamma \left[\int c d\gamma + \varepsilon \, \mathrm{KL}(\gamma \| P_{XY} \otimes (P_X \otimes P_Y)) \right], \quad (2)$$

where the infimum is over $\Pi(P_{XY}, P_X \otimes P_Y)$ which is the set of couplings (or joint distributions) on $\mathbb{R}^{d \times d}$ with marginals P_{XY} and $P_X \otimes P_Y$, $\varepsilon > 0$ is the regularization parameter, and KL is the Kullback-Leibler divergence. The other two terms are defined similarly and are omitted for the sake of space.

As we will show later, it is computationally convenient to work with additive cost functions, i.e., $c((x,y),(x',y')) = c_1(x,x') + c_2(y,y')$. For this type of cost functions, we prove that the resulting ETIC is a valid independence criterion as long as the induced *Gibbs kernels*

$$k_1(x,x') = e^{-\frac{c_1(x,x')}{\varepsilon}} \text{ and } k_2(y,y') = e^{-\frac{c_2(y,y')}{\varepsilon}} \quad (3)$$

are positive universal; see Appendix A for the proof.

Proposition 1. *Let $\mathcal{X} \subset \mathbb{R}^{d_1}$ and $\mathcal{Y} \subset \mathbb{R}^{d_2}$ be compact subsets equipped with Lipschitz costs c_1 and c_2, respectively. Assume that the Gibbs kernels defined in (3) are positive universal. Then, the ETIC is a valid dependency measure on $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, i.e., $T(X,Y) \ge 0$ and*

$$T(X,Y) = 0 \text{ iff } P_{XY} = P_X \otimes P_Y. \quad (4)$$

Moreover, the claim holds true for measures with a bounded support on $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d$ with the costs $c_1(x,x') = \|x - x'\|^p / \lambda_1$ and $c_2(y,y') = \|y - y'\|^p / \lambda_2$ for $p \in \{1,2\}$ and for all $\lambda_1, \lambda_2 > 0$.

A running example we consider in this paper is the weighted quadratic cost.

Example 1 (Weighted quadratic cost)**.** *Let $\lambda_1, \lambda_2 \in (0,\infty)$. Consider the cost function*

$$c((x,y),(x',y')) = \frac{1}{\lambda_1} \|x - x'\|^2 + \frac{1}{\lambda_2} \|y - y'\|^2. \quad (5)$$

This cost induces two universal kernels $k_1(x,x') = e^{-\|x-x'\|^2/(\varepsilon\lambda_1)}$ and $k_2(y,y') = e^{-\|y-y'\|^2/(\varepsilon\lambda_2)}$. They play a similar role as the two kernels used in HSIC, and $\varepsilon\lambda_1$ and $\varepsilon\lambda_2$ serve as two kernel parameters.

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:** U and V.

ETIC-Based Independence Test. Given an i.i.d. sample $\{(X_i, Y_i)\}_{i=1}^n$ from P_{XY}, we are interested in determining whether X is independent of Y, which can be formalized as the following hypothesis testing problem:

$$\mathbf{H}_0 : P_{XY} = P_X \otimes P_Y \leftrightarrow \mathbf{H}_1 : P_{XY} \neq P_X \otimes P_Y. \quad (6)$$

For this purpose, we use the empirical estimate of $T(X,Y)$ as the test statistic, that is,

$$T_n(X,Y) := T_{n,\varepsilon}(X,Y) := \bar{S}_\varepsilon(\hat{P}_{XY}, \hat{P}_X \otimes \hat{P}_Y), \quad (7)$$

where $\hat{P}_{XY} := \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Y_i)}$ is the empirical measure of the pairs, and $\hat{P}_X := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{P}_Y := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ are the empirical measures of the two samples, respectively. Note that this is different from the standard plug-in estimator since the product measure $P_X \otimes P_Y$ is estimated by n^2 *dependent* (rather than independent) pairs $\{(X_i, Y_j)\}_{i,j=1}^n$. It raises challenges in the analysis of its statistical behavior as elaborated in Section 3. The statistical test (or decision rule) is then defined as

$$\psi(\alpha) := \mathbb{1}\{T_n(X,Y) > H_n(\alpha)\}, \quad (8)$$

where α is a prescribed significance level, e.g., $\alpha = 0.05$, and $H_n(\alpha)$ is a threshold chosen such that the *type I error rate* $\mathbb{P}(\psi(\alpha) = 1 \mid \mathbf{H}_0)$ is bounded by α. Here $\{\psi(\alpha) = 1\}$ indicates the rejection the null hypothesis. The *(statistical) power* of the test is defined as $\mathbb{P}(\psi(\alpha) = 1 \mid \mathbf{H}_1)$.

To avoid tuning the regularization parameter ε, we also consider an adaptive version of the test:

$$\psi_a(\alpha) := \mathbb{1}\left\{ \max_{\varepsilon \in \mathcal{E}} \bar{T}_{n,\varepsilon}(X,Y) > H_{n,\mathcal{E}}(\alpha) \right\}, \quad (9)$$

where $\mathcal{E}$ is a finite set of positive numbers selected by the user and $\bar{T}_{n,\varepsilon}(X,Y) := [T_{n,\varepsilon}(X,Y) - \mathbb{E}[T_{n,\varepsilon}(X,Y)]]/\mathrm{Sd}(T_{n,\varepsilon}(X,Y))$ is the studentized version of $T_{n,\varepsilon}(X,Y)$. In practice, $\mathbb{E}[T_{n,\varepsilon}(X,Y)]$ and $\mathrm{Sd}(T_{n,\varepsilon}(X,Y))$ can be estimated via resampling.

Tensor Sinkhorn Algorithm. We then derive an efficient algorithm to compute the test statistic. When P_{XY} admits a density, $\hat{P}_X \otimes \hat{P}_Y$ is supported on n^2 items $\{x_i\}_{i=1}^n \times \{y_i\}_{i=1}^n$ almost surely. If we compute

Table 1: Comparison of complexities, in time and in space, of Sinkhorn and Tensor Sinkhorn algorithms, Exact or Random Features approx.

	Sinkhorn		**Tensor Sinkhorn**	
	Exact	**RF**	**Exact**	**RF**
Time	$O(n^4)$	$\boldsymbol{O(n^2)}$	$O(n^3)$	$\boldsymbol{O(n^2)}$
Space	$O(n^4)$	$O(n^4)$	$\boldsymbol{O(n^2)}$	$\boldsymbol{O(n^2)}$

the ETIC statistic naively using the Sinkhorn algorithm (Cuturi, 2013), each iteration costs $O(n^4)$ time and space due to the matrix-vector product of sizes $n^2 \times n^2$ and $n^2 \times 1$. To speed up its computation, we develop a variant of the Sinkhorn algorithm to solve the EOT problem between two measures supported on the Cartesian product $\{x_i\}_{i=1}^n \times \{y_i\}_{i=1}^n$.

Let A and B be two probability measures on $\{x_i\}_{i=1}^n \times \{y_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^{d_1}$ and $y_j \in \mathbb{R}^{d_2}$. For convenience, both A and B are represented as a matrix, i.e., $A_{ij} = A(x_i, y_j)$. For instance, if we choose $A = \hat{P}_{XY}$ and $B = \hat{P}_X \otimes \hat{P}_Y$, then, in its matrix form, $A = I_n/n$ and $B = \mathbf{1}_{n\times n}/n^2$. Consider an additive cost function c, e.g., the weighted quadratic cost, such that $c((x,y),(x',y')) = c_1(x,x') + c_2(y,y')$ for $x, x' \in \{x_i\}_{i=1}^n$ and $y, y' \in \{y_j\}_{j=1}^n$. Let C_1 and C_2 be the cost matrices of $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^n$, respectively. Define Gibbs matrices $K_1 := e^{-C_1/\varepsilon}$ and $K_2 := e^{-C_2/\varepsilon}$, where the exponential function is element-wise. We show in Proposition 8 in Appendix A that Algorithm 1 can be used to compute $S_\varepsilon(A,B)$, where $\oslash$ represents element-wise division. We refer to it as the *Tensor Sinkhorn* algorithm. Each iteration in the Tensor Sinkhorn algorithm takes $O(n^3)$ time and $O(n^2)$ space, thanks to the additive cost function being used. This algorithm can be generalized to measures supported on the Cartesian product of $p > 2$ sets, which is also noted in (Peyré and Cuturi, 2019, Remark 4.17).

ETIC with Random Features. To further speed up the computation, we apply the random feature technique introduced by Scetbon and Cuturi (2020). On a high level, we approximate the gram matrix K_i by its low-rank approximation $\xi_i \xi_i^\top$ for $i \in \{1,2\}$, where $\xi_i \in \mathbb{R}^{n\times p}$ is the matrix of random features. Concretely, let ρ_1 and ρ_2 be two probability measures on measurable spaces $\mathcal{U}$ and $\mathcal{V}$, respectively. Following Scetbon and Cuturi (2020, Section 3), we focus on Gibbs kernels of the form

$$k_1(x,x') = \int \varphi(x,u)^\top \varphi(x',u) d\rho_1(u)$$
$$k_2(y,y') = \int \psi(y,v)^\top \psi(y',v) d\rho_2(v),$$

where $\varphi : \{x_i\}_{i=1}^n \times \mathcal{U} \to \mathbb{R}_+^{q_1}$ and $\psi : \{y_i\}_{i=1}^n \times \mathcal{V} \to \mathbb{R}_+^{q_2}$. Note that the Gibbs kernels induced by the weighted quadratic cost admit this expression. For $p \in \mathbb{N}_+$, we obtain two i.i.d. samples $\{u_i\}_{i=1}^p$ and $\{v_i\}_{i=1}^p$ from ρ_1 and ρ_2, respectively. We denote $\boldsymbol{u} := (u_1, \dots, u_p)$ and approximate $k_1(x,x')$ by

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

This new kernel induces the cost $c_{1,\boldsymbol{u}}(x,x') := -\varepsilon \log k_{1,\boldsymbol{u}}(x,x')$. Similarly, we define $\boldsymbol{v}$, $k_{2,\boldsymbol{v}}(y,y')$, and $c_{2,\boldsymbol{v}}(y,y')$. It is clear that Algorithm 1 with inputs A, B, $K_{1,\boldsymbol{u}}$, and $K_{2,\boldsymbol{v}}$ solves the entropy-regularized optimal transport problem with cost $c_{\boldsymbol{u},\boldsymbol{v}}((x,y),(x',y')) := c_{1,\boldsymbol{u}}(x,x') + c_{2,\boldsymbol{v}}(y,y')$. Let $S_{\varepsilon,c_{\boldsymbol{u},\boldsymbol{v}}}(A,B)$ be the entropic cost. The next proposition provides a high-probability guarantee for this random feature approximation.

Assumption 1. *There exists a constant $C > 0$ such that, for all $x, x' \in \{x_i\}_{i=1}^n$, $y, y' \in \{y_j\}_{j=1}^n$, $u \in \mathcal{U}$, and $v \in \mathcal{V}$,*

$$\varphi(x,u)^\top \varphi(x',u)/k_1(x,x') \le C$$
$$\psi(y,v)^\top \psi(y',v)/k_2(y,y') \le C.$$

Proposition 2. *Let $\delta > 0$, $\tau > 0$, and $p = \Omega\left(\frac{C^2}{\tau^2}\log\frac{n}{\delta}\right)$. Under Assumption 1, with probability at least $1-\delta$, it holds that*

$$\left|S_{\varepsilon,c_{\boldsymbol{u},\boldsymbol{v}}}(A,B) - S_{\varepsilon,c}(A,B)\right| \le \tau.$$

Replacing K_1 and K_2 by their random feature approximations $K_{1,\boldsymbol{u}}$ and $K_{2,\boldsymbol{v}}$ in Algorithm 1 leads to an algorithm with $O(pn^2)$ time complexity and $O(n^2)$ space complexity in each iteration. We note that if one applies the random feature approximation directly to the original Sinkhorn algorithm, then the resulting algorithm would have the same time complexity $O(pn^2)$ but $O(n^4)$ space complexity; see Table 1 for a comparison.

Dual Representation. The entropy regularized formulation of OT (2) is known as the *Schrödinger bridge problem* (Föllmer, 1988; Léonard, 2012, 2014) in continuum and the *Sinkhorn distance* (Cuturi, 2013; Ferradans et al., 2014) in the discrete case. It admits a dual representation (Genevay et al., 2016):

$$\sup_{f,g\in\mathcal{C}(\mathbb{R}^{d_1}\times\mathbb{R}^{d_2})}\Bigg[\int f dP_{XY} + \int g d(P_X\otimes P_Y) + \varepsilon - \varepsilon\int e^{-D_\varepsilon(x,y,x',y')}dP_{XY}(x,y)dP_X(x')dP_Y(y')\Bigg],$$

where $\mathcal{C}$ is the set of real-valued continuous functions and $D_\varepsilon(x,y,x',y') = \frac{1}{\varepsilon}[c_1(x,x') + c_2(y,y') - f(x,y) - g(x',y')]$. Due to (Csiszar, 1975; Rüschendorf and Thomsen, 1993), the optimal potentials $(f_\varepsilon, g_\varepsilon)$, to be called the *Schrödinger potentials*, satisfy

$$\begin{aligned}\int e^{-D_\varepsilon(x,y,x',y')}dP_X(x')dP_Y(y') &\overset{\text{a.s.}}{=} 1\\ \int e^{-D_\varepsilon(x,y,x',y')}dP_{XY}(x,y) &\overset{\text{a.s.}}{=} 1.\end{aligned} \tag{10}$$

Gradient Backpropagation through ETIC. We describe here how ETIC can fit into a differentiable programming framework, i.e., how one can run the reverse mode automatic differentiation through statistical quantities based on ETIC. Recently, Li et al. (2021) proposed a self-supervised learning approach using HSIC which we summarize below. Let (W, Y) be a pair of image and its identity. Given an i.i.d. sample $\{(W_i, Y_i)\}_{i=1}^n$, the goal is to learn a feature embedding model ϕ_θ such that the dependence between the image feature $X := \phi_\theta(W)$ and its identity Y is maximized, i.e., $\max_{\theta\in\Theta} \mathrm{HSIC}_n(\phi_\theta(W), Y)$. Similarly, one could also maximize the dependence measured by ETIC instead. This boils down to gradient backpropagating through $T_n(\phi_\theta(W), Y)$. We use the strategy in (Peyré and Cuturi, 2019, Section 9.1.3) and illustrate it on the entropy regularized OT $S_\varepsilon(\hat{P}_{XY}, \hat{P}_X\otimes\hat{P}_Y)$ in (2). For the forward pass, we construct the computational graph via the following steps. Firstly, we run Algorithm 1 (or its random feature variant) with $A = I_n/n$, $B = \mathbf{1}_{n\times n}/n^2$, $K_1 = \big(k_1(\phi_\theta(W_i), \phi_\theta(W_j))\big)_{n\times n}$, and $K_2 = \big(k_2(Y_i, Y_j)\big)_{n\times n}$ for L iterations to get $U^{(L)}$ and $V^{(L)}$. Secondly, we obtain the associated Schrödinger potentials $F^{(L)} := \varepsilon\log U^{(L)}$ and $G^{(L)} := \varepsilon\log V^{(L)}$. Thirdly, we approximate $S_\varepsilon(\hat{P}_{XY}, \hat{P}_X\otimes\hat{P}_Y)$ by $\hat{S}_\varepsilon(\theta) := \langle F^{(L)}, A\rangle_{\mathbf{F}} + \langle G^{(L)}, B\rangle_{\mathbf{F}}$ where $\langle\cdot,\cdot\rangle_{\mathbf{F}}$ is the Frobenius inner product. For the backward pass, we call the reverse mode automatic differentiation to evaluate $\nabla_\theta\hat{S}_\varepsilon(\theta)$. Since computing $\hat{S}_\varepsilon(\theta)$ only requires simple operations between matrices, the time complexity of the above procedure is of the same order as the one of Algorithm 1 for the computation of $\hat{S}_\varepsilon(\theta)$.

Connection to Previous Work. As shown in Appendix A, $T(X,Y)$ tends to the OT-based independence criterion $\mathrm{OT}(P_{XY}, P_X\otimes P_Y)$ as $\varepsilon\to 0$. If the cost c is chosen as the Euclidean distance to the power $p\ge 1$, it induces a distance (known as the Wasserstein-p distance) on the space of probability measures (Villani, 2016). As a result, $\mathrm{OT}(P_{XY}, P_X\otimes P_Y)$ is a valid independence criterion, i.e., $\mathrm{OT}(P_{XY}, P_X\otimes P_Y) = 0$ iff $P_{XY} = P_X\otimes P_Y$. The study of this independence criterion can be dated back to Gini; see (Cifarelli and Regazzini, 2017) for a discussion. Its normalized version—the so-called Wasserstein correlation coefficient—has recently gained attention in (Wiesel, 2021; Mordant and Segers, 2021; Nies et al., 2021). When $\varepsilon\to\infty$, $T(X,Y)$ tends to 0 if the cost is additive; if the cost is multiplicative, i.e., $c((x,y),(x',y')) = c_1(x,x')c_2(y,y')$, it recovers the negative of HSIC with kernels c_1 and c_2.

The quantity $\bar{S}_\varepsilon$ is known as the *Sinkhorn divergence*. It has been used in two-sample problems, where the goal is to quantify the distance of two distributions given i.i.d. samples from each of them. In particular, it is applied to two-sample testing (Ramdas et al., 2017) and generative modeling (Genevay et al., 2018). It is shown in (Feydy et al., 2019) that $\bar{S}_\varepsilon$ defines a semi-metric (metric without the triangle inequality) on the space of probability measures with bounded support if the Gibbs kernel induced by the cost is positive universal. The limiting behavior of the empirical estimator is to date not known in the literature, though non-asymptotic bounds are attainable using results in (Genevay et al., 2019; Mena and Weed, 2019). Our results also recover the two-sample case.

3 MAIN RESULTS

We give non-asymptotic bounds for the ETIC statistic with quadratic cost. We present the main results and their proof sketches here. We use C to denote a constant whose value may change from line to line, where subscripts are used to emphasize the dependency on other quantities. For instance, C_d represents a constant depending only on the dimension d. The detailed proofs are deferred to Appendices B and C.

Consistency. We first show that the ETIC statistic is a consistent estimator of its population counterpart under both the null and alternative.

Assumption 2. *We make the following assumptions:*

(i) c is chosen as the quadratic cost.

(ii) P_X and P_Y are subG(σ^2).

The quadratic cost is chosen for the sake of concision. We extend the results to weighted quadratic cost in Appendix B.

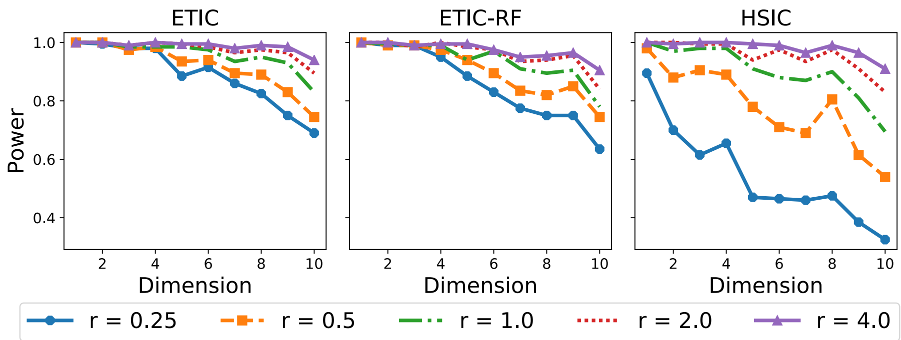

Figure 1: Power versus dimension in the linear dependency model (11).

Theorem 3. *Under Assumption 2, we have*

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

Remark 2. *According to Theorem 3, when $\varepsilon = \varepsilon_n$ is chosen such that $\varepsilon_n = \omega(n^{-1/(\lceil 5d/2\rceil+4)})$ and $\varepsilon_n = o(1)$, we have $T_n(X,Y)$ converges in $\mathbf{L}^1$ to $\mathrm{OT}(P_{XY}, P_X \otimes P_Y)$ as $n \to \infty$.*

We can upper bound the above $\mathbf{L}^1$ loss by the supremum of an empirical process and a U-process

$$\left\|\hat{P}_{XY} - P_{XY}\right\|_{\mathcal{F}^s}^2 \quad \text{and} \quad \left\|\hat{P}_X \otimes \hat{P}_Y - P_X \otimes P_Y\right\|_{\mathcal{F}^s}^2,$$

respectively, where $\mathcal{F}^s$ is the set of real-valued functions satisfying

$$|f(x,y)| \le C_{s,d}(1 + \|(x,y)\|^2)$$
$$|D^\alpha f(x,y)| \le C_{s,d}(1 + \|(x,y)\|^{|\alpha|}), \quad \forall 1 < |\alpha| \le s.$$

Mena and Weed (2019) used a similar strategy in their proofs. Empirical process theory has a long history in statistics and there are well-established tools to control them; see, e.g., (van der Vaart and Wellner, 1996). However, the theory of U-processes is much less well-developed. Moreover, many of the previous works focus on one-sample U-processes; see, e.g., (Peña and Giné, 1999). The second U-process here is a two-sample U-process on a paired sample, bringing about additional challenges in its analysis, compared to, e.g., Mena and Weed (2019). In order to control it, we develop the following results.

The first result is a metric entropy bound for *degenerate two-sample U-processes*. The main challenge comes from the dependence among the summands in $\sum_{i,j=1}^n f(X_i, Y_j)$. We get around that using the decoupling technique presented in (Peña and Giné, 1999).

Proposition 4. *Let $\mathcal{F}$ be a class of real-valued functions that are degenerate under $P_X \otimes P_Y$, i.e.,*

$$\mathbb{E}_{P_X\otimes P_Y}[f(X,Y) \mid X] \overset{a.s.}{=} \mathbb{E}_{P_X\otimes P_Y}[f(X,Y) \mid Y] \overset{a.s.}{=} 0$$

for any $f \in \mathcal{F}$. Under Assumption 2, we have

$$\mathbb{E}\left\|\hat{P}_X \otimes \hat{P}_Y - P_X \otimes P_Y\right\|_{\mathcal{F}}^2$$
$$\le \frac{C}{n}\mathbb{E}\left(\int_0^B \sqrt{\log N(\tau, \mathcal{F}, \mathbf{L}^2(\hat{P}_X \otimes \hat{P}_Y))}d\tau\right)^2,$$

where B is any measurable upper bound of $2\max_{f\in\mathcal{F}}\|f\|_{\mathbf{L}^2(\hat{P}_X\otimes\hat{P}_Y)}$.

Remark 3. *In classical two-sample U-statistics literature, it is usually assumed that the two samples are independent, i.e., X is independent of Y. However, Proposition 4 allows the sample to be paired since $(X,Y) \sim P_{XY}$.*

With Proposition 4 at hand, we can control the U-process $\|\hat{P}_X \otimes \hat{P}_Y - P_X \otimes P_Y\|_{\mathcal{F}^s}^2$ by upper bounding its covering number $N(\tau, \mathcal{F}^s, \mathbf{L}^2(\hat{P}_X \otimes \hat{P}_Y))$. The proof is inspired by (Mena and Weed, 2019) and relies on a result in (van der Vaart and Wellner, 1996, Chapter 2.7) to control the covering number of a class of smooth functions.

Proposition 5. *Under Assumption 2, there exists a random variable $L \ge 1$ depending on the samples $\{(X_i, Y_i)\}_{i=1}^n$ with $\mathbb{E}[L] \le 2$ such that, for any $s \ge 2$,*

$$\log N(\tau, \mathcal{F}^s, \mathbf{L}^2(\hat{P}_X \otimes \hat{P}_Y)) \le C_{s,d}\tau^{-d/s}L^{d/2s}(1 + \sigma^{2d})$$

and

$$\max_{f\in\mathcal{F}^s}\|f\|_{\mathbf{L}^2(\hat{P}_X\otimes\hat{P}_Y)}^2 \le C_{s,d}(1 + L\sigma^4).$$

In particular, when $s > d/2$, we have

$$\mathbb{E}\left\|\hat{P}_X \otimes \hat{P}_Y - P_X \otimes P_Y\right\|_{\mathcal{F}^s}^2 \le C_{s,d}(1 + \sigma^{2d+4})\frac{1}{n}.$$

Exponential Tail Bound. We also prove an exponential tail bound for the ETIC statistic. It follows from Theorem 3 and the McDiarmid inequality.

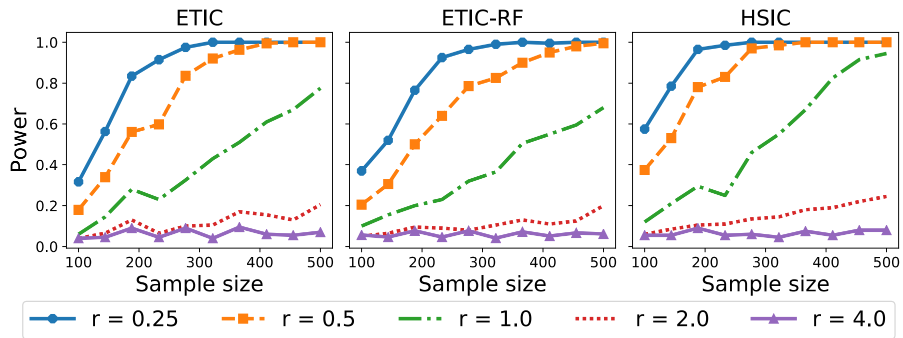

Figure 2: Power versus sample size in the Gaussian-sign model (12).

Theorem 6. *Let c be the quadratic cost. Assume that P_X and P_Y are supported on a bounded domain of radius D. Then we have, with probability at least $1-\delta$,*

$$|T_n(X,Y) - T(X,Y)| \leq C_d \left(1 + \frac{D^{5d+16}}{\varepsilon^{5d/2+8}} \sqrt{\log \frac{6}{\delta}}\right) \frac{\varepsilon}{\sqrt{n}}.$$

Under $\mathbf{H}_0$, we have $T(X,Y) = 0$, so Theorem 6 implies that

$$|T_n(X,Y)| > C_d \left(1 + \frac{D^{5d+16}}{\varepsilon^{5d/2+8}} \sqrt{\log \frac{6}{\delta}}\right) \frac{\varepsilon}{\sqrt{n}}$$

with probability at most δ. It gives an estimate of the tail behavior of $T_n(X,Y)$ which suggests that the critical value $H_n(\alpha)$ in (8) should be of order $O(n^{-1/2})$. Under $\mathbf{H}_1$, Theorem 6 implies that

$$T_n(X,Y) > T(X,Y) - C_d \left(1 + \frac{D^{5d+16}}{\varepsilon^{5d/2+8}} \sqrt{\log \frac{6}{\delta}}\right) \frac{\varepsilon}{\sqrt{n}}$$

with probability at least $1-\delta$. When $T(X,Y) > 0$, it is clear that the right hand side in the above inequality exceeds the threshold $H_n(\alpha)$ for large n. Hence, the ETIC test has power converging to 1 as $n \to \infty$.

4 EXPERIMENTS

We examine the empirical behavior of the proposed ETIC test for independence testing on both synthetic and real data. We consider synthetic benchmarks from (Gretton et al., 2007b; Jitkrittum et al., 2017; Zhang et al., 2018) and revisit an application from (Gretton et al., 2007b) with recent feature representations for text data. The performance of the adaptive ETIC test is also investigated but is deferred to Appendix E.2 due to the space limit. The code to reproduce the experiments is available online[1].

[1]https://github.com/langliu95/etic-experiments.

We focus on the weighted quadratic cost

$$c((x,y),(x',y')) = \frac{1}{\lambda_1} \|x - x'\|^2 + \frac{1}{\lambda_2} \|y - y'\|^2 .$$

For convenience, we absorb the regularization parameter ε into the weights and set $\varepsilon = 1$. It then induces two Gibbs kernels

$$k_1(x,x') = e^{-\frac{\|x-x'\|^2}{\lambda_1}} \text{ and } k_2(y,y') = e^{-\frac{\|y-y'\|^2}{\lambda_2}}$$

with λ_i being the parameter of kernel k_i for $i \in \{1,2\}$. To select the weights, we apply the median heuristic (Gretton et al., 2007b) widely used for HSIC, i.e.,

$$\lambda_1 = r_1 M_x \quad \text{and} \quad \lambda_2 = r_2 M_y$$

with r_1 and r_2 ranging from 0.25 to 4, where M_x and M_y are the medians of the quadratic costs $\{\|X_i - X_j\|^2\}$ and $\{\|Y_i - Y_j\|^2\}$, respectively. We also examine its variant ETIC-RF discussed in Section 2, where the number of random features is set to be 100 unless otherwise noted. We compare them with the HSIC statistic with kernels k_1 and k_2. For a fair comparison, we calibrate these tests by a Monte Carlo resampling technique (Feuerverger, 1993) with 200 permutations. For each of the experiment, we repeat the whole procedure 200 times and report the rejection frequency as either the type I error rate (when the null is true) or power (when the null is not true). Note that, even though we are using the same λ_1 and λ_2 in the cost and kernels, that does *not* mean we should compare ETIC and HSIC under the same hyperparameters. Our goal is to explore their performance over a range of values of the hyperparameters controlling the regularization penalties.

Our main findings are: 1) Both ETIC and ETIC-RF are consistent in power as the sample size approaches infinity. 2) In some scenarios, ETIC and ETIC-RF outperforms HSIC significantly; in the linear dependency model in particular, their power is much

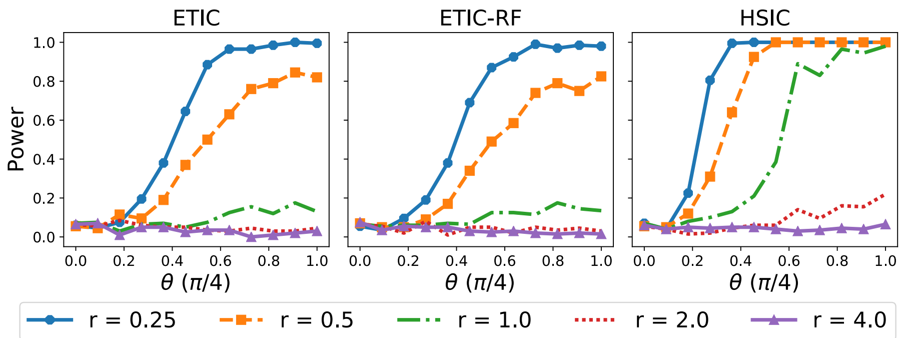

Figure 3: Power versus parameter in the subspaces dependency model.

more robust than HSIC to the value of the hyperparameters. 3) ETIC-RF performs reasonably good compared to ETIC with a moderate number (i.e., 100) of random features. 4) All three tests benefit from large hyperparameters in detecting simple linear dependency, but smaller values lead to higher power when the dependency is more complicated.

Hilbert-Schmidt Independence Criterion. Before we present our results, let us recall the definition of HSIC. Let $k : \mathbb{R}^{d_1} \times \mathbb{R}^{d_1} \to \mathbb{R}$ and $l : \mathbb{R}^{d_2} \times \mathbb{R}^{d_2} \to \mathbb{R}$ be two positive semi-definite kernels. The Hilbert-Schmidt independence criterion (HSIC) between X and Y, $\mathrm{HSIC}(X, Y)$, is defined as

$$\mathbb{E}[k(X, X')l(Y, Y')] + \mathbb{E}[k(X, X')]\,\mathbb{E}[l(Y, Y')] - \\ 2\,\mathbb{E}[\mathbb{E}[k(X, X') \mid X]\,\mathbb{E}[l(Y, Y') \mid Y]],$$

where (X', Y') is an independent copy of (X, Y). Given an i.i.d. sample $\{(X_i, Y_i)\}_{i=1}^n$ from P_{XY}, we can estimate $\mathrm{HSIC}(X, Y)$ by

$$\frac{1}{n^2}\sum_{i,j=1}^{n} k_{ij}l_{ij} + \frac{1}{n^4}\sum_{i,j,s,t=1}^{n} k_{ij}l_{st} - \frac{2}{n^3}\sum_{i,j,s=1}^{n} k_{ij}l_{is},$$

where $k_{ij} := k(X_i, X_j)$ and $l_{ij} := l(Y_i, Y_j)$. We refer to it as the HSIC statistic.

4.1 Synthetic Data

We first compare the performance of ETIC and ETIC-RF with HSIC on synthetic data. We consider synthetic benchmarks from (Zhang et al., 2018), (Jitkrittum et al., 2017), and (Gretton et al., 2007b). To facilitate the exploration, we set $r_1 = r_2 = r \in \{0.25, 0.5, 1, 2, 4\}$ in this section. Moreover, we examine the performance of two other independence tests discussed in (Gretton and Györfi, 2008) and summarize the results and findings in Appendix E.3.

Linear Dependency. We begin with a simple linear dependency model. Concretely,

$$X \sim \mathcal{N}_d(0, I_d) \quad \text{and} \quad Y = X_1 + Z, \tag{11}$$

where X_1 is the first coordinate of X, and $Z \sim \mathcal{N}(0, 1)$ is independent with X. We fix $n = 50$ and plot the power versus $d \in [1, 10]$ in Figure 1. All the tests have decaying power as the dimension increases. This is as expected since larger dimension results in weaker dependency between X and Y. It is clear that the power of both ETIC and HSIC increases as r *increases*, with the former more robust than the latter. While the performance of HSIC is similar to ETIC when r is large, it is much worse than ETIC when r is small. As for ETIC-RF, it has similar power curves as ETIC.

Gaussian Sign. We then consider a Gaussian sign model, i.e.,

$$X \sim \mathcal{N}_d(0, I_d) \quad \text{and} \quad Y = |Z| \prod_{i=1}^{d} \mathrm{sgn}(X_i), \tag{12}$$

where $\mathrm{sgn}(\cdot)$ is the sign function and $Z \sim \mathcal{N}(0, 1)$ is independent with X. This problem is challenging since Y is independent with any strict subset of $\{X_1, \ldots, X_d\}$. We fix $d = 3$ and plot the power versus $n \in [100, 500]$ in Figure 2. All the tests have improved power as the sample size increases. Additionally, they all benefit from a *small* regularization parameter, with HSIC performs the best and the other two perform similarly in this particular example.

Subspace Dependency. One important application of independence testing is independent component analysis (Gretton et al., 2005), which involves separating random variables from their linear mixtures. We construct our data by i) generating n i.i.d. copies of two random variables following independently $0.5\mathcal{N}(0.98, 0.04) + 0.5\mathcal{N}(-0.98, 0.04)$, ii) mixing the two random variables by a rotation matrix parameterized by $\theta \in [0, \pi/4]$ (larger θ leads to stronger

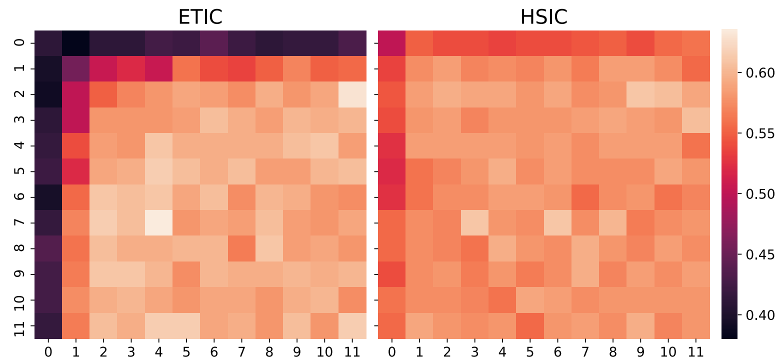

Figure 4: Heatmaps of power on the partially dependent sample of the bilingual data. The x-axis is for r_1 and y-axis is for r_2. The indices from 0 to 11 correspond to equally spaced values from 0.25 to 4. Lighter color indicates larger power.

dependency), iii) appending $\mathcal{N}_{d-1}(0, I_{d-1})$ to each of the two mixtures, and iv) multiplying each vector by an independent random d-dimensional orthogonal matrix. We refer to it as the *subspace dependency model*. We fix $n = 64$, $d = 2$, and plot the power versus $\theta \in [0, \pi/4]$ in Figure 3. As expected, the power of all three tests improves as θ becomes closer to $\pi/4$. Moreover, they all have improved power as r *decreases*. ETIC and ETIC-RF performs similarly, and they are outperformed by HSIC in this particular example.

4.2 Dependency between Bilingual Text

Inspired by Gretton et al. (2007b), we now investigate the performance of the proposed tests on bilingual data using recent developments in natural language processing. Our dataset is taken from the parallel European Parliament corpus (Koehn, 2005) which consists of a large number of documents of the same content in different languages. Note that it is also used in (Bounliphone et al., 2015) to test for relative dependency. For the hyperparameters, we consider different values of r_1 and r_2 ranging from 0.25 to 4.

To be more specific, we randomly select $n = 64$ English documents and a paragraph in each document from the corpus. We then 1) pair each paragraph with the corresponding paragraph in French to form the dependent sample, 2) pair each paragraph with a random paragraph in the same document in French to form the partially dependent sample, and 3) pair each paragraph with a random paragraph in French to form the independent sample.

Finally, we use LaBSE (Feng et al., 2020) to embed all the paragraphs into a common feature embedding space of dimension 768 and perform independence testing on these feature vectors. LaBSE is a state-of-the-art, language agnostic, sentence embedding model based on Bidirectional Encoder Representations from Transformers (BERT). This allows us to revisit the idea of Gretton et al. (2007b) yet with more modern feature embeddings.

Both ETIC and HSIC perform perfectly on the dependent sample (with power 1) and the independent sample (with low type I error) across all values of r_1 and r_2 considered. The results on the partially dependent sample is shown in Figure 4. ETIC performs better than HSIC when one of r_1 and r_2 is large; while HSIC has larger power when r_1 or r_2 is small. Overall ETIC appears to perform better than HSIC for large amounts of regularization parameters.

As for ETIC-RF, the high-dimensional natural of the feature embeddings imposes challenges on the random feature approximation. For its performance to be comparable, we first apply the principal component analysis (PCA) on the text embeddings to reduce the dimension, and then we perform ETIC-RF on the low-dimensional features. As presented in Appendix E.1, the random feature approximation equipped with PCA demonstrates similar performance as the exact ETIC with enough random features.

Conclusion. We introduced a new independence criterion ETIC based on entropy regularized optimal transport. The proposed criterion can be approximated using a random feature approximation. We established non-asymptotic bounds using U-process theory and optimal transport theory. The experimental results show that ETIC can exhibit stable behavior w.r.t. its hyperparameters. The extension of ETIC to multi-way dependence is an interesting venue for future work.

Acknowledgements

The authors would like to thank M. Scetbon for fruitful discussions. L. Liu is supported by NSF CCF-2019844 and NSF DMS-2023166. S. Pal is supported by NSF DMS-2052239 and a PIMS CRG (PIHOT). Z. Harchaoui is supported by NSF CCF-2019844, NSF DMS-2134012, NSF DMS-2023166, CIFAR-LMB, and faculty research awards. Part of this work was done while Z. Harchaoui was visiting the Simons Institute for the Theory of Computing. The authors would like to thank the Kantorovich Initiative of the Pacific Institute for the Mathematical Sciences (PIMS) for supporting this collaboration.

References

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser Basel, 1 edition, 2005.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3, 2002.

G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *NIPS*, 2011.

W. Bounliphone, A. Gretton, A. Tenenhaus, and M. Blaschko. A low variance consistent test of relative dependency. In *ICML*, 2015.

S. Chakraborty and X. Zhang. Distance metrics for measuring joint dependence with application to causal inference. *Journal of the American Statistical Association*, 114(528), 2019.

D. M. Cifarelli and E. Regazzini. On the centennial anniversary of Gini's theory of statistical relations. *Metron*, 2017.

G. M. Constantine and T. H. Savits. A multivariate Faà di Bruno formula with applications. *Transactions of the American Mathematical Society*, 348(2), 1996.

I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3, 1975.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, 2013.

N. Deb and B. Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, 2021.

R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1), 1969.

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *ICML*, 2018.

F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic BERT sentence embedding. *arXiv Preprint*, 2020.

S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3), 2014.

A. Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61(3), 1993.

J. Feydy, T. Séjourné, F. Vialard, S. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *AISTATS*, 2019.

H. Föllmer. Random fields and diffusion processes. In *École d'été de probabilités de Saint-Flour XV-XVII-1985-87*, volume 1362 of *Lecture Notes in Mathematics*. Springer, 1988.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162, 2015.

A. Genevay, M. Cuturi, G. Peyré, and F. R. Bach. Stochastic optimization for large-scale optimal transport. In *NIPS*, 2016.

A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *AISTATS*, 2018.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. In *AISTATS*, 2019.

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2015.

A. Gretton and L. Györfi. Nonparametric independence tests: Space partitioning and kernel approaches. In *ALT*, 2008.

A. Gretton, R. Herbrich, A. J. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6, 2005.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample problem. In *NIPS*, 2007a.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *NIPS*, 2007b.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13, 2012.

W. Hoeffding. A non-parametric test of independence. *The Annals of Mathematical Statistics*, 19(4), 1948.

W. Jitkrittum, Z. Szabó, and A. Gretton. An adaptive test of independence with analytic kernel embeddings. In *ICML*, 2017.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit*, 2005.

W. H. Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284), 1958.

E. L. Lehmann. Some concepts of dependence. *The Annals of Mathematical Statistics*, 37(5), 1966.

J. Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1), 2020.

C. Léonard. From the Schrödinger problem to the Monge-Kantorovich problem. *Journal of Functional Analysis*, 262(4), 2012.

C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4), 2014.

Y. Li, R. Pogodin, D. J. Sutherland, and A. Gretton. Self-supervised learning with kernel dependence maximization. *arXiv Preprint*, 2021.

R. Lyons. Distance covariance in metric spaces. *Annals of Probability*, 41(5), 2013.

G. Mena and J. Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *NeurIPS*, 2019.

G. Mordant and J. Segers. Measuring dependence between random vectors via optimal transport. *arXiv Preprint*, 2021.

T. G. Nies, T. Staudt, and A. Munk. Transport dependency: Optimal transport based dependency measures. *arXiv Preprint*, 2021.

Y. Nikitin. *Asymptotic Efficiency of Nonparametric Tests*. Cambridge University Press, 1995.

S. Ozair, C. Lynch, Y. Bengio, A. van den Oord, S. Levine, and P. Sermanet. Wasserstein dependency measure for representation learning. In *NeurIPS*, 2019.

V. d. l. Peña and E. Giné. *Decoupling: From Dependence to Independence*. Springer, 1 edition, 1999.

G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6), 2019.

N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1), 2018.

A. Ramdas, N. García Trillos, and M. Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2), 2017.

P. Rigollet and J. Weed. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11), 2018.

L. Rüschendorf and W. Thomsen. Note on the Schrödinger equation and I-projections. *Statistics & Probability Letters*, 17, 1993.

T. Salimans, H. Zhang, A. Radford, and D. N. Metaxas. Improving GANs using optimal transport. In *ICLR*, 2018.

M. Sanjabi, J. Ba, M. Razaviyayn, and J. D. Lee. On the convergence and robustness of training GANs with regularized optimal transport. In *NeurIPS*, 2018.

F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Basel, 2015.

M. Scetbon and M. Cuturi. Linear time Sinkhorn divergences using positive features. In *NeurIPS*, 2020.

B. Schweizer and E. F. Wolff. On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9(4), 1981.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5), 2013.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980.

H. Shi, M. Drton, and F. Han. Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association*, 2020.

G. J. Székely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6), 2007.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1 edition, 1996.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2016.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A), 2019.

J. Wiesel. Measuring association with Wasserstein distances. *arXiv Preprint*, 2021.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28, 2018.

Supplementary Material: Entropy Regularized Optimal Transport Independence Criterion

A PROPERTIES OF ETIC

In this section, we prove the properties of ETIC discussed in Section 2. For the sake of generality, we state the problem for general notations P and Q while keeping in mind that $P, Q \in \{P_{XY}, P_X \otimes P_Y\}$ in our case. Let $P \in \mathcal{M}_1(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2})$ and P_X and P_Y be the marginals on $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Define Q, Q_X, and Q_Y similarly. We are interested in the EOT cost between P and Q under the cost function c:

$$S_\varepsilon(P, Q) := \inf_{\gamma \in \Pi(P,Q)} \left[\int c d\gamma + \varepsilon \operatorname{KL}(\gamma \| P \otimes Q) \right]. \tag{13}$$

When $\varepsilon = 0$, $S_0(P, Q)$ is the optimal transport cost between P and Q. When $\varepsilon > 0$, it admits a dual representation:

$$S_\varepsilon(P, Q) := \sup_{f, g \in \mathcal{C}(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2})} \left[\int f dP + \int g dQ + \varepsilon - \varepsilon \int e^{\frac{1}{\varepsilon}[f(z) + g(z') - c(z,z')]} dP(z) dQ(z') \right]. \tag{14}$$

The Schrödinger bridge potentials $(f_\varepsilon, g_\varepsilon)$ satisfy the optimality conditions:

$$\begin{aligned} \int e^{\frac{1}{\varepsilon}[f_\varepsilon(z) + g_\varepsilon(z) - c(z,z')]} dQ(z') &\overset{\text{a.s.}}{=} 1 \\ \int e^{\frac{1}{\varepsilon}[f_\varepsilon(z) + g_\varepsilon(z) - c(z,z')]} dP(z) &\overset{\text{a.s.}}{=} 1. \end{aligned} \tag{15}$$

We first derive the limit ETIC as $\varepsilon \to 0$ and $\varepsilon \to \infty$.

Proposition 7. *Let c be a continuous cost function. If either c is bounded or P and Q have compact support, it holds that*

$$T_\varepsilon(X, Y) \to \begin{cases} 0 & \text{if } c = c_1 \oplus c_2 \\ -\frac{1}{2} \operatorname{HSIC}_{c_1, c_2}(X, Y) & \text{if } c = c_1 \otimes c_2, \end{cases} \quad \text{as } \varepsilon \to \infty. \tag{16}$$

Moreover, if both P and Q are densities (or discrete measures), then

$$T_\varepsilon(X, Y) \to S_0(P_{XY}, P_X \otimes P_Y), \quad \text{as } \varepsilon \to 0. \tag{17}$$

Proof. To show (16), we claim that, for all $P, Q \in \mathcal{M}_1(\mathbb{R}^d)$,

$$S_0(P, Q) \le S_\varepsilon(P, Q) \le (P \otimes Q)[c], \tag{18}$$

and

$$\lim_{\varepsilon \to \infty} S_\varepsilon(P, Q) = (P \otimes Q)[c]. \tag{19}$$

In fact, for any $\varepsilon_1 < \varepsilon_2$, we have

$$\int c d\gamma + \varepsilon_1 \operatorname{KL}(\gamma \| P \otimes Q) \le \int c d\gamma + \varepsilon_2 \operatorname{KL}(\gamma \| P \otimes Q), \quad \text{for all } \gamma \in \Pi(P, Q).$$

This yields that

$$S_{\varepsilon_1}(P, Q) \le S_{\varepsilon_2}(P, Q), \quad \text{for all } \varepsilon_1 \le \varepsilon_2,$$

and thus (18) follows.

We then study the limit of S_ε as $\varepsilon \to \infty$. By the assumption that c is bounded or P and Q have compact support, there exists $M > 0$ such that $\sup_{\gamma\in\Pi(P,Q)} \int c d\gamma \le M < \infty$. As a result,

$$\sup_{\gamma\in\Pi(P,Q)} \left| \frac{1}{\varepsilon} \int c d\gamma + \mathrm{KL}(\gamma \| P\otimes Q) - \mathrm{KL}(\gamma \| P\otimes Q) \right| \le \frac{M}{\varepsilon},$$

which implies that

$$\inf_{\gamma\in\Pi(P,Q)} \left[\frac{1}{\varepsilon} \int c d\gamma + \mathrm{KL}(\gamma \| P\otimes Q) \right] \to \inf_{\gamma\in\Pi(P,Q)} \mathrm{KL}(\gamma \| P\otimes Q) = 0, \quad \text{as } \varepsilon \to \infty.$$

By the strict convexity of KL, the problem on the LHS has a unique minimizer γ_ε and the problem on the RHS has a unique minimizer $\gamma_* = P \otimes Q$. Now, by the tightness of $\Pi(P, Q)$ (e.g., (Santambrogio, 2015, Theorem. 1.7)), every sequence of $\{\gamma_\varepsilon\}$ has a weakly converging subsequence whose limit must be γ_*. Therefore, the claim (19) holds true.

Let $c = c_1 \oplus c_2$. According to (19), we have

$$\lim_{\varepsilon\to\infty} S_\varepsilon(P_{XY}, P_X \otimes P_Y) = (P_{XY} \otimes P_X \otimes P_Y)[c] = (P_X \otimes P_X)[c_1] + (P_Y \otimes P_Y)[c_2].$$

Similarly, it holds that

$$\lim_{\varepsilon\to\infty} S_\varepsilon(P_{XY}, P_{XY}) = (P_X \otimes P_X)[c_1] + (P_Y \otimes P_Y)[c_2]$$
$$\lim_{\varepsilon\to\infty} S_\varepsilon(P_X \otimes P_Y, P_X \otimes P_Y) = (P_X \otimes P_X)[c_1] + (P_Y \otimes P_Y)[c_2].$$

Consequently, $\lim_{\varepsilon\to\infty} T_\varepsilon(X, Y) = 0$. An analogous argument implies that, when $c = c_1 \otimes c_2$

$$\lim_{\varepsilon\to\infty} T_\varepsilon(X, Y) = \mathbb{E}_{P_{XY}} \left[\mathbb{E}_{P_X}[c_1(X, X') \mid X]\, \mathbb{E}_{P_Y}[c_2(Y, Y') \mid Y]\right]$$
$$- \frac{1}{2} \mathbb{E}_{P_{XY}^2}[c_1(X, X')c_2(Y, Y')] - \frac{1}{2} \mathbb{E}_{(P_X\otimes P_Y)^2}[c_1(X, X')c_2(Y, Y')] = -\frac{1}{2} \mathrm{HSIC}_{c_1,c_2}(X, Y).$$

Note that

$$\lim_{\varepsilon\to 0} S_\varepsilon(P, Q) = S_0(P, Q)$$

when both P and Q are densities (Léonard, 2012) and when both of them are discrete measures (Peyré and Cuturi, 2019, Proposition 4.1). The statement (17) follows immediately from the fact that $S_0(P, P) = 0$ for all P. □

We then prove the validity of ETIC as a dependence measure as stated in Proposition 1.

Proof of Proposition 1. Due to Blanchard et al. (2011, Lemma 5.2), the Gibbs kernel

$$k_\varepsilon(z, z') := e^{-c(z,z')/\varepsilon} = k_1(x, x')k_2(y, y')$$

is universal since both k_x and k_y are. It is also clear that k_ε is positive since both k_x and k_y are. Consequently, the Sinkhorn divergence $\bar{S}_\varepsilon$ defines a semi-metric on $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ according to Feydy et al. (2019, Theorem 1). Hence, if $P_{XY}, P_X \otimes P_Y \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, then $T_\varepsilon(X, Y) := \bar{S}_\varepsilon(P_{XY}, P_X \otimes P_Y) = 0$ iff $P_{XY} = P_X \otimes P_Y$. □

Finally, we analyze the computational complexity of the Tensor Sinkhorn algorithm for additive cost functions, i.e.,

$$c(z, z') := c_1(x, x') + c_2(y, y'), \tag{20}$$

where $z = (x, y)$ and $z' = (x', y')$.

Let $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^n$ be two sets of atoms. Note that the two sets are assumed to be of the same size for convenience. Let A and B be two probability measures on $\{x_i\}_{i=1}^n \times \{y_j\}_{j=1}^n$. For convenience, both A and B are represented as a matrix, i.e., $A_{ij} = A(x_i, y_j)$. For instance, if we choose $A = \hat{P}_{XY}$ and $B = \hat{P}_X \otimes \hat{P}_Y$, then, in its matrix form, $A = I_n/n$ and $B = \mathbf{1}_{n\times n}/n^2$. Denote C_1 and C_2 as the cost matrices of $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^n$, respectively. Define Gibbs matrices $K_1 := e^{-C_1/\varepsilon}$ and $K_2 := e^{-C_2/\varepsilon}$, where the exponential function is applied element-wisely. Let $K := K_2 \otimes K_1 \in \mathbb{R}^{n^2 \times n^2}$ be the Gibbs matrix associated with the cost matrix on the pairs $\{(x_1, y_1), (x_2, y_1), \dots, (x_n, y_n)\}$, where $\otimes$ is the Kronecker product.

Proposition 8. *The Tensor Sinkhorn algorithm outputs an δ-accurate estimate of the entropic cost $S(A, B)$ in $O\left(n^3 \log(\kappa_1\kappa_2\kappa_3)/\delta\right)$ arithmetic operations, where $\kappa_1 := \max_{i,i'} k_1^{-1}(x_i, x_{i'})$, $\kappa_2 := \max_{j,j'} k_2^{-1}(y_j, y_{j'})$, and $\kappa_3 := \max_{i,j}\{a_{ij}^{-1}, b_{ij}^{-1}\}$.*

Proof. Let $a := \mathrm{Vec}(A) \in \mathbb{R}^{n^2}$ and $b := \mathrm{Vec}(B) \in \mathbb{R}^{n^2}$ be the probability vectors corresponding to A and B, respectively. Denote $u := \mathrm{Vec}(U) \in \mathbb{R}^{n^2}$ and $v := \mathrm{Vec}(V) \in \mathbb{R}^{n^2}$. The Sinkhorn algorithm to solve $S_\varepsilon(a, b)$ has the following two update steps:

$$u = a \oslash Kv \quad \text{and} \quad v = b \oslash K^\top u.$$

By the identity $\mathrm{Vec}(MNL) = (L^\top \otimes M)\,\mathrm{Vec}(N)$ for matrices M, N, and L of compatible dimensions, we obtain

$$\mathrm{Vec}(K_1 V K_2^\top) = (K_2 \otimes K_1)\,\mathrm{Vec}(V) = Kv.$$

Thus, the update $U = A \oslash (K_1 V K_2^\top)$ is equivalent to $u = a \oslash Kv$. Similarly, the updated $V = B \oslash (K_1^\top U K_2)$ is equivalent to $v = b \oslash K^\top u$. Due to Dvurechensky et al. (2018, Theorem 1), the Tensor Sinkhorn algorithm therefore outputs an δ-accurate estimate in $O(\log(\kappa_1\kappa_2\kappa_3)/\delta)$ iterations. Since each iteration costs $O(n^3)$ time, it has overall time complexity $O(n^3 \log(\kappa_1\kappa_2\kappa_3)/\delta)$. □

Remark 4. *A direct application of the Sinkhorn algorithm leads to $O(n^4 \log(\kappa_1\kappa_2\kappa_3)/\delta)$ time complexity, which is n times slower than the Tensor Sinkhorn algorithm.*

We then characterize the convergence of the Tensor Sinkhorn algorithm with the random feature approximation as presented in Proposition 2.

Proof of Proposition 2. The proof is heavily inspired by Scetbon and Cuturi (2020, Proof of Theorem 3.1). In consideration of the space, we only present the part that is significantly different from theirs, i.e., a counterpart of Scetbon and Cuturi (2020, Proposition 3.1). This proposition gives a uniform tail bound for the ratio between the approximated kernel and the original kernel. In our case, we are approximating the kernel $K := K_2 \otimes K_1$ by $K_{\boldsymbol{u},\boldsymbol{v}} := K_{2,\boldsymbol{v}} \otimes K_{1,\boldsymbol{u}}$. Hence, it suffices to bound

$$\sup_{x,x' \in \{x_i\}_{i=1}^n, y,y' \in \{y_i\}_{i=1}^n} \left| \frac{k_{1,\boldsymbol{u}}(x,x') k_{2,\boldsymbol{v}}(y,y')}{k_1(x,x') k_2(y,y')} - 1 \right|.$$

Note that

$$\frac{k_{1,\boldsymbol{u}}(x,x')}{k_1(x,x')} = \frac{1}{p} \sum_{k=1}^p \frac{\varphi(x,u_k)^\top \varphi(x',u_k)}{k_1(x,x')}$$

is a sum of nonnegative i.i.d. random variables with mean 1. Due to Assumption 1, they are also bounded. It follows from the Hoeffding inequality that

$$\mathbb{P}\left(\left| \frac{k_{1,\boldsymbol{u}}(x,x')}{k_1(x,x')} - 1 \right| \ge t \right) \le 2\exp\left(-\frac{pt^2}{C^2} \right).$$

The same inequality holds for the ratio $k_{2,\boldsymbol{v}}(y,y')/k_2(y,y')$. Since

$$\left| \frac{k_{1,\boldsymbol{u}}(x,x') k_{2,\boldsymbol{v}}(y,y')}{k_1(x,x') k_2(y,y')} - 1 \right| \le \left| \frac{k_{1,\boldsymbol{u}}(x,x')}{k_1(x,x')} - 1 \right| \left| \frac{k_{2,\boldsymbol{v}}(y,y')}{k_2(y,y')} - 1 \right| + \left| \frac{k_{1,\boldsymbol{u}}(x,x')}{k_1(x,x')} - 1 \right| + \left| \frac{k_{2,\boldsymbol{v}}(y,y')}{k_2(y,y')} - 1 \right|,$$

it follows that

$$\mathbb{P}\left(\left|\frac{k_{1,\boldsymbol{u}}(x,x')k_{2,\boldsymbol{v}}(y,y')}{k_1(x,x')k_2(y,y')}-1\right|\le t^2+2t\right)\ge\mathbb{P}\left(\left\{\left|\frac{k_{1,\boldsymbol{u}}(x,x')}{k_1(x,x')}-1\right|\le t\right\}\bigcap\left\{\left|\frac{k_{2,\boldsymbol{v}}(y,y')}{k_2(y,y')}-1\right|\le t\right\}\right)$$
$$=\mathbb{P}\left(\left|\frac{k_{1,\boldsymbol{u}}(x,x')}{k_1(x,x')}-1\right|\le t\right)\mathbb{P}\left(\left|\frac{k_{2,\boldsymbol{v}}(y,y')}{k_2(y,y')}-1\right|\le t\right)$$
$$\ge 1-4\exp\left(-\frac{pt^2}{C^2}\right).$$

Equivalently,

$$\mathbb{P}\left(\left|\frac{k_{1,\boldsymbol{u}}(x,x')k_{2,\boldsymbol{v}}(y,y')}{k_1(x,x')k_2(y,y')}-1\right|\ge t\right)\le 4\exp\left(-\frac{p(\sqrt{t+1}-1)^2}{C^2}\right).$$

A uniform bound yields

$$\mathbb{P}\left(\sup_{x,x'\in\{x_i\}_{i=1}^n,y,y'\in\{y_i\}_{i=1}^n}\left|\frac{k_{1,\boldsymbol{u}}(x,x')k_{2,\boldsymbol{v}}(y,y')}{k_1(x,x')k_2(y,y')}-1\right|\ge t\right)\le 4n^4\exp\left(-\frac{p(\sqrt{t+1}-1)^2}{C^2}\right).$$

□

Remark 5. *Let $\hat{S}_{\varepsilon,c_{\boldsymbol{u},\boldsymbol{v}}}(A,B)$ be the cost computed from Algorithm 1. Following Dvurechensky et al. (2018, Theorem 1), we can get that*

$$\left|\hat{S}_{\varepsilon,c_{\boldsymbol{u},\boldsymbol{v}}}(A,B)-S_{\varepsilon,c_{\boldsymbol{u},\boldsymbol{v}}}(A,B)\right|\le\tau$$

in $O\left(pn^2\log(\kappa_1\kappa_2\kappa_3)/\tau\right)$ arithmetic operations, where $\kappa_1:=\max_{i,i'}k_{1,\boldsymbol{u}}^{-1}(x_i,x_{i'})$, $\kappa_2:=\max_{j,j'}k_{2,\boldsymbol{v}}^{-1}(y_j,y_{j'})$, and $\kappa_3:=\max_{i,j}\{a_{ij}^{-1},b_{ij}^{-1}\}$.

B CONSISTENCY OF THE TEST STATISTIC

In this section, we prove the main results in Section 3. For the sake of generality, we start by considering the formulation in (13). We focus on the weighted quadratic cost function

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

where $z=(x,y)$, $z'=(x',y')$ and $w_1,w_2\in\mathbb{R}_+$. Denote $w:=\max\{w_1,w_2\}$. Due to Lemma 24, we assume, w.l.o.g., that $\varepsilon=1$ and write $S(P,Q):=S_1(P,Q)$.

B.1 Smoothness Properties of the Schrödinger Potentials

We start by deriving some smoothness properties of the Schrödinger potentials. Our proofs are deeply inspired by Mena and Weed (2019). Our results generalize theirs to weighted quadratic cost functions.

Assumption 3. *We assume that P_X, P_Y, Q_X, and Q_Y are all subG(σ^2).*

Proposition 9. *Under Assumption 3. there exist smooth Schrödinger potentials (f,g) for $S(P,Q)$ such that the optimality conditions (15) hold for all $z,z'\in\mathbb{R}^d$. Moreover, we have*

$$f(z)\ge -d\sigma^2\left[2w_1+2w_2+4w_1^2(\sqrt{2d_1}\sigma+\|x\|)^2+4w_2^2(\sqrt{2d_2}\sigma+\|y\|)^2\right]-1$$
$$f(z)\le w_1(\|x\|+\sqrt{2d_1}\sigma)^2+w_2(\|y\|+\sqrt{2d_2}\sigma)^2,$$

and for g similarly.

Proof. Let (f_0,g_0) be a pair of Schrödinger potentials. Since (f_0+C,g_0-C) is also a pair of Schrödinger potentials for any constant $C\in\mathbb{R}$, we assume, w.l.o.g., that $P[f_0]=Q[g_0]=\frac{1}{2}S(P,Q)\ge 0$. Define

$$f(z):=-\log\int e^{g_0(z')-c(z,z')}dQ(z')\quad\text{and}\quad g(z'):=-\log\int e^{f(z)-c(z,z')}dP(z).\tag{21}$$

We claim that the pair (f, g) satisfies the requirements.

Since (f_0, g_0) is a pair of Schrödinger potentials, it holds that

$$g_0(z') \overset{\text{a.s.}}{=} -\log \int e^{f_0(z)-c(z,z')} dP(z) \le -P[f_0] + w_1 \, \mathbb{E}_{P_X}[\|X - x'\|^2] + w_2 \, \mathbb{E}_{P_Y}[\|Y - y'\|^2],$$

by Jensen's inequality. Note that $P[f_0] \ge 0$ and, by Lemma 18, $\mathbb{E}_{P_X}[\|X\|^2] \le 2d_1\sigma^2$. It follows that

$$g_0(z') - c(z, z') \le w_1 \left[2d_1\sigma^2 + 2\,\|x'\|\,(\sqrt{2d_1}\sigma + \|x\|)\right] + w_2 \left[2d_2\sigma^2 + 2\,\|y'\|\,(\sqrt{2d_2}\sigma + \|y\|)\right],$$

and thus

$$\begin{aligned}\int e^{g_0(z')-c(z,z')} dQ(z') &\le e^{2(w_1 d_1 + w_2 d_2)\sigma^2} \left[\int e^{4w_1 \|x'\|(\sqrt{2d_1}\sigma + \|x\|)} dQ_X(x') \int e^{4w_2 \|y'\|(\sqrt{2d_2}\sigma + \|y\|)} dQ_Y(y')\right]^{1/2} \\ &\le 2e^{2(w_1 d_1 + w_2 d_2)\sigma^2} e^{4d_1\sigma^2 w_1^2(\sqrt{2d_1}\sigma + \|x\|)^2 + 4d_2\sigma^2 w_2^2(\sqrt{2d_2}\sigma + \|y\|)^2} < \infty, \quad \text{by Lemma 18.}\end{aligned}$$

Hence, $f(z)$ is well-defined for all $z \in \mathbb{R}^d$. Moreover, we have the lower bound

$$\begin{aligned} f(z) &\ge -d_1\sigma^2 \left[2w_1 + 4w_1^2(\sqrt{2d_1}\sigma + \|x\|)^2\right] - d_2\sigma^2 \left[2w_2 + 4w_2^2(\sqrt{2d_2}\sigma + \|y\|)^2\right] - 1 \\ &\ge -d\sigma^2 \left[4w + 4w_1^2(\sqrt{2d_1}\sigma + \|x\|)^2 + 4w_2^2(\sqrt{2d_2}\sigma + \|y\|)^2\right] - 1 \end{aligned}$$

For the upper bound, by Jensen's inequality, it holds that

$$\begin{aligned} f(z) &\le -Q[g_0] + w_1 \, \mathbb{E}_{Q_X} \|x - X'\|^2 + w_2 \, \mathbb{E}_{Q_Y} \|y - Y'\|^2 \\ &\le w_1(\|x\| + \sqrt{2d_1}\sigma)^2 + w_2(\|y\| + \sqrt{2d_2}\sigma)^2. \end{aligned}$$

Similar arguments prove the claim for g. Now, it remains to show that (f, g) satisfies the optimality conditions (15) for all $z, z \in \mathbb{R}^d$. By definition, it is clear that

$$\int e^{f(z)+g(z')-c(z,z')} dP(z) = 1 \quad \text{and} \quad \int e^{f(z)+g_0(z')-c(z,z')} dQ(z') = 1, \quad \forall z, z' \in \mathbb{R}^d.$$

Since (f_0, g_0) is a pair of Schrödinger potentials, we also have

$$\int e^{f_0(z)+g_0(z')-c(z,z')} dP(z) dQ(z') = 1.$$

Consequently, by Jensen's inequality

$$\begin{aligned} &\int (f - f_0) dP + \int (g - g_0) dQ \\ &\ge -\log \int e^{f_0 - f} dP - \log \int e^{g_0 - g} dQ \\ &= -\log \int e^{f_0(z)+g_0(z')-c(z,z')} dP(z) dQ(z') - \log \int e^{f(z)+g_0(z')-c(z,z')} dP(z) dQ(z') \\ &= 0. \end{aligned}$$

Since both (f_0, g_0) and (f, g) are Schrödinger potentials, the above equality holds true. This implies that $\int (g_0 - g) dQ = \log \int e^{g_0 - g} dQ$, and thus $g = g_0 + C$ Q-almost surely by the strict concavity of log. Therefore, we have

$$\int e^{f(z)+g(z')-c(z,z')} dQ(z') = e^C \int e^{f(z)+g_0(z')-c(z,z')} dQ(z') = e^C, \quad \forall z, z' \in \mathbb{R}^d.$$

Taking integrals with respect to P implies that $C = 0$, which completes the proof. □

The next proposition shows that there exist Schrödinger potentials satisfying Hölder-type conditions.

Definition 4. *For any $\sigma \in \mathbb{R}_+$, $d \in \mathbb{N}_+$, and $w = (w_1, w_2) \in \mathbb{R}_+^2$, let $\mathcal{F}_\sigma := \mathcal{F}_{\sigma,d,w}$ be the set of smooth functions such that, for any $k \in \mathbb{N}_+$ and any multi-index α with $|\alpha| = k$,*

$$\left|D^\alpha \left(f(x,y) - w_1 \|x\|^2 - w_2 \|y\|^2\right)\right| \leq C_{k,d,w} \begin{cases} (1+\sigma^4) & \text{if } k = 0 \\ \sigma^k (1+\sigma)^k & \text{otherwise,} \end{cases} \tag{22}$$

if $\|z\| \leq \sqrt{d}\sigma$, and

$$\left|D^\alpha \left(f(x,y) - w_1 \|x\|^2 - w_2 \|y\|^2\right)\right| \leq C_{k,d,w} \begin{cases} [1 + (1+\sigma^2)\|x\|^2] & \text{if } k = 0 \\ \sigma^k (\sqrt{\sigma \|x\|} + \sigma \|x\|)^k & \text{otherwise,} \end{cases} \tag{23}$$

if $\|z\| > \sqrt{d}\sigma$, where $C_{k,d,w}$ is a constant depending on k, d, and w.

Proposition 10. *Under Assumption 3, there exist Schrödinger potentials (f, g) such that the optimality conditions (15) hold for all $z, z' \in \mathbb{R}^d$ and $f, g \in \mathcal{F}_\sigma$.*

Proof. Let (f, g) be a pair of Schrödinger potentials satisfying the requirements in Proposition 9. Denote $\bar{f}(x,y) := f(x,y) - w_1 \|x\|^2 - w_2 \|y\|^2$. Note that

$$\begin{aligned} \bar{f}(z) &= -\log e^{-\bar{f}(x,y)} = -\log \int e^{w_1 \|x\|^2 + w_2 \|y\|^2 + g(z') - c(z,z')} dQ(z') \\ &= -\log \int e^{g(z') - w_1 \|x'\|^2 - w_2 \|y'\|^2 + 2w_1 \langle x, x' \rangle + 2w_2 \langle y, y' \rangle} dQ(z'). \end{aligned}$$

The desired inequalities for $k = 0$ follow directly from Proposition 9. We focus on $k > 0$. According to the multivariate Faá di Bruno formula (Constantine and Savits, 1996), we have

$$D^\alpha \bar{f}(z) = \sum_{\lambda_1 + \cdots + \lambda_k = \alpha} C_{\alpha, \lambda_1, \ldots, \lambda_k} \prod_{i=1}^k M_{\lambda_i},$$

where

$$M_\lambda = \frac{\int (\tilde{z}')^\lambda \exp\left\{g(z') - w_1 \|x'\|^2 - w_2 \|y'\|^2 + 2w_1 \langle x, x' \rangle + 2w_2 \langle y, y' \rangle\right\} dQ(z')}{\int \exp\left\{g(z') - w_1 \|x'\|^2 - w_2 \|y'\|^2 + 2w_1 \langle x, x' \rangle + 2w_2 \langle y, y' \rangle\right\} dQ(z')}. \tag{24}$$

Here $\tilde{z}' = (2w_1 x'; 2w_2 y')$ and $z^\lambda = \prod_{i=1}^d z_i^{\lambda_i}$. By Lemma 11 below, it holds that

$$\left|D^\alpha \bar{f}(z)\right| \leq C_{k,d,w} \begin{cases} \sigma^k (1 + \sigma^k) & \text{if } \|z\| \leq \sqrt{d}\sigma \\ \sigma^k (\sigma \|z\| + \sqrt{\sigma \|z\|})^k & \text{if } \|z\| > \sqrt{d}\sigma, \end{cases}$$

which proves the claim. □

Lemma 11. *Recall M_λ in (24). Under Assumption 3, for $|\lambda| > 0$, we have*

$$|M_\lambda| \leq C_{|\lambda|,d,w} \begin{cases} \sigma^{|\lambda|} (\sigma + \sigma^2)^{|\lambda|} & \text{if } \|z\| \leq \sqrt{d}\sigma \\ \sigma^{|\lambda|} (\sigma \|z\| + \sqrt{\sigma \|z\|})^{|\lambda|} & \text{if } \|z\| > \sqrt{d}\sigma \end{cases}.$$

Proof. We first bound the denominator. By the optimality conditions (15), it holds that

$$\begin{aligned} &\left(\int \exp\left\{g(z') - w_1 \|x'\|^2 - w_2 \|y'\|^2 + 2w_1 \langle x, x' \rangle + 2w_2 \langle y, y' \rangle\right\} dQ(z')\right)^{-1} \\ &= e^{f(x,y) - w_1 \|x\|^2 - w_2 \|y\|^2} \leq e^{w_1 (2d_1 \sigma^2 + 2\sqrt{2d_1}\sigma \|x\|) + w_2 (2d_2 \sigma^2 + 2\sqrt{2d_2}\sigma \|y\|)}, \end{aligned}$$

where the last inequality follows from Proposition 9. To bound the numerator, we use the truncation technique. Let $A := \{(x', y') : \|2w_1 x'\| \le K, \|2w_2 y'\| \le K\}$ for some constant K to be determined later. On the set A, it is clear that $(\tilde{z}')^\lambda \le \|\tilde{z}'\|^{|\lambda|} \le K^{|\lambda|}$, and thus

$$\frac{\int_A (\tilde{z}')^\lambda \exp\left\{g(z') - w_1 \|x'\|^2 - w_2 \|y'\|^2 + 2w_1\langle x, x'\rangle + 2w_2\langle y, y'\rangle\right\} dQ(z')}{\int \exp\left\{g(z') - w_1 \|x'\|^2 - w_2 \|y'\|^2 + 2w_1\langle x, x'\rangle + 2w_2\langle y, y'\rangle\right\} dQ(z')} \le K^{|\lambda|}.$$

On the set A^c, we proceed as follows. According to Proposition 9, we have

$$e^{g(x',y') - w_1\|x'\|^2 - w_2\|y'\|^2} \le e^{w_1(2d_1\sigma^2 + 2\sqrt{2d_1}\sigma\|x'\|) + w_2(2d_2\sigma^2 + 2\sqrt{2d_2}\sigma\|y'\|)},$$

which yields

$$\begin{aligned}
&\int_{A^c} (\tilde{z}')^\lambda \exp\left\{g(z') - w_1 \|x'\|^2 - w_2 \|y'\|^2 + 2w_1\langle x, x'\rangle + 2w_2\langle y, y'\rangle\right\} dQ(z') \\
&\le e^{2(w_1 d_1 + w_2 d_2)\sigma^2} \left[\int_{A^c} (\tilde{z}')^{2\lambda} dQ(z') \int_{A^c} e^{2w_1\|x'\|(\|x\| + \sqrt{2d_1}\sigma) + 2w_2\|y'\|(\|y\| + \sqrt{2d_2}\sigma)} dQ(z')\right]^{1/2}.
\end{aligned}$$

For any $z' \in A^c$, we have either $\|2w_1 x'\| > K$ or $\|2w_2 y'\| > K$. If the former is true, then

$$\int_{A^c} (\tilde{z}')^{2\lambda} dQ(z') \le \int_{A^c} e^{-\frac{K^2}{16w_1^2 d_1\sigma^2}} e^{\frac{\|2w_1 x'\|^2}{16 w_1^2 d_1 \sigma^2}} (\tilde{z}')^{2\lambda} dQ(z') \le C_{|\lambda|,d,w} e^{-\frac{K^2}{16w^2 d\sigma^2}} \sigma^{2|\lambda|},$$

where $w = \max\{w_1, w_2\}$. The same bound holds if the latter is true. Furthermore, by the Cauchy-Schwartz inequality and Lemma 18 in Appendix D, we have

$$\int_{A^c} e^{2w_1\|x'\|(\|x\| + \sqrt{2d_1}\sigma) + 2w_2\|y'\|(\|y\| + \sqrt{2d_2}\sigma)} dQ(z') \le e^{4w_1^2 d_1\sigma^2(\|x\| + \sqrt{2d_1}\sigma)^2 + 4w_2^2 d_2\sigma^2(\|y\| + \sqrt{2d_2}\sigma)^2}$$

Putting all together, we get

$$\begin{aligned}
&\frac{\int_{A^c} (\tilde{z}')^\lambda \exp\left\{g(z') - w_1 \|x'\|^2 - w_2 \|y'\|^2 + 2w_1\langle x, x'\rangle + 2w_2\langle y, y'\rangle\right\} dQ(z')}{\int \exp\left\{g(z') - w_1 \|x'\|^2 - w_2 \|y'\|^2 + 2w_1\langle x, x'\rangle + 2w_2\langle y, y'\rangle\right\} dQ(z')} \\
&\le C_{|\lambda|,d,w} e^{-\frac{K^2}{32w^2 d\sigma^2}} e^{2w_1^2 d_1\sigma^2(\|x\| + \sqrt{2d_1}\sigma)^2 + 2w_2^2 d_2\sigma^2(\|y\| + \sqrt{2d_2}\sigma)^2} \sigma^{|\lambda|} \\
&\le C_{|\lambda|,d,w} e^{-\frac{K^2}{32w^2 d\sigma^2}} e^{2w^2 d\sigma^2[(\|x\| + \sqrt{2d}\sigma)^2 + (\|y\| + \sqrt{2d}\sigma)^2]} \sigma^{|\lambda|}
\end{aligned}$$

When $\|z\| \le \sqrt{d}\sigma$, it holds that $\|x\| \le \sqrt{2d}\sigma$ and $\|y\| \le \sqrt{2d}\sigma$. Hence, if we choose $K^2 = C_{|\lambda|,d,w}(\sigma^4 + \sigma^6)$ for some sufficiently large constant $C_{|\lambda|,d,w}$, then we have

$$|M_\lambda| \le C_{|\lambda|,d,w}\sigma^{|\lambda|}(\sigma + \sigma^2)^{|\lambda|}.$$

When $\|z\| > \sqrt{d}\sigma$, if we choose $K^2 = C_{|\lambda|,d,w}(\sigma^4\|z\|^2 + \sigma^3\|z\|)$, then we have

$$|M_\lambda| \le C_{|\lambda|,d,w}\sigma^{|\lambda|}\left(\sigma\|z\| + \sqrt{\sigma\|z\|}\right)^{|\lambda|}.$$

□

When P and Q have bounded support, we can further show that the Schrödinger potentials can be chosen to be bounded.

Proposition 12. *Assume that P and Q are supported on a bounded domain of radius D. Then there exist Schrödinger potentials (f, g) such that 1) the optimality conditions (15) hold for all $x, y \in \mathbb{R}^d$ and 2) $\|f\|_\infty \le 8wD^2$ and $\|g\|_\infty \le 8wD^2$.*

Proof. Let (f, g) the Schrödinger potentials defined in (21). By the proof of Proposition 9, they satisfy (15) everywhere. Moreover, we have

$$f(z) \le w_1 \mathbb{E}_{Q_X}\|x - X'\|^2 + w_2 \mathbb{E}_{Q_Y}\|y - Y'\|^2 \le 8wD^2$$

and g similarly. □

B.2 Controlling the Empirical Process and the U-Process

We then upper bound the $\mathbf{L}^1$ loss $\mathbb{E}\left|T_n(X,Y)-T(X,Y)\right|$ by empirical processes and U-processes.

Proposition 13 (Corollary 2 (Mena and Weed, 2019))**.** *Let $P,Q,P',Q'\in\mathcal{M}_1(\mathbb{R}^d)$ be subG(σ^2). Then we have*

$$|S(P',Q')-S(P,Q)|\le\sup_{f\in\mathcal{F}_\sigma}\left|\int f(dP'-dP)\right|+\sup_{g\in\mathcal{F}_\sigma}\left|\int g(dQ'-dQ)\right|,$$

where $\mathcal{F}_\sigma$ is defined in Definition 4.

To simply the function class $\mathcal{F}_\sigma$, we show in Lemma 22 in Appendix D that $(1+\sigma^{3s})^{-1}\mathcal{F}_\sigma\subset\mathcal{F}^s$ for $\mathcal{F}^s$ defined below. Consequently, we can separate the sub-Gaussian parameter σ from the function class $\mathcal{F}_\sigma$.

Definition 5. *For any $s\ge 2$, $d\in\mathbb{N}_+$, and $w=(w_1,w_2)\in\mathbb{R}^2_+$, let $\mathcal{F}^s:=\mathcal{F}^{s,d,w}$ be the set of functions satisfying*

$$\begin{aligned}|f(z)|&\le C_{s,d,w}(1+\|z\|^2)\\ |D^\alpha f(z)|&\le C_{s,d,w}(1+\|z\|^{|\alpha|}),\quad\forall 1\le|\alpha|\le s,\end{aligned}$$

where $C_{s,d,w}$ is a constant depending on s, d, and w.

In order to handle the U-process, we also need a variant function class of $\mathcal{F}^s$ which we also define below.

Definition 6. *For any $\sigma\in\mathbb{R}_+$, $s\ge 2$, $d\in\mathbb{N}_+$, and $w=(w_1,w_2)\in\mathbb{R}^2_+$, let $\mathcal{F}^s_\sigma:=\mathcal{F}^{s,d,w}_\sigma$ be the set of functions satisfying*

$$\begin{aligned}|f(z)|&\le C_{s,d,w}(1+\max\{\|z\|^2,\sigma^2\})\\ |D^\alpha f(z)|&\le C_{s,d,w}(1+\max\{\|z\|^{|\alpha|},\sigma^{|\alpha|}\}),\quad\forall 1\le|\alpha|\le s,\end{aligned}$$

where $C_{s,d,w}$ is a constant depending on s, d, and w.

Let us control the complexity of $\mathcal{F}^s$ and $\mathcal{F}^s_\sigma$, which is achieved by the following covering number bound.

Proposition 14. *Let $P\in\mathcal{M}_1(\mathbb{R}^d)$ be subG(σ^2). Let $\{Z_i\}_{i=1}^n\overset{i.i.d.}{\sim}P$ and $\hat{P}_n$ be the empirical measure. There exists a random variable $L\ge 1$ depending on the sample $\{Z_i\}_{i=1}^n$ with $\mathbb{E}[L]\le 2$ such that*

$$\log N(\tau,\mathcal{F}^s,\mathbf{L}^2(\hat{P}_n))\le C_{s,d,w}\tau^{-d/s}L^{d/2s}(1+\sigma^{2d})\quad\text{and}\quad\max_{f\in\mathcal{F}^s}\|f\|^2_{\mathbf{L}^2(\hat{P}_n)}\le C_{s,d,w}(1+L\sigma^4).$$

Moreover, the same bounds hold for $\mathcal{F}^s_\sigma$.

Proof of Proposition 14. Define $L:=\frac{1}{n}\sum_{i=1}^n e^{\|Z_i\|^2/2d\sigma^2}\ge 1$. By the sub-Gaussianity of P, we have $\mathbb{E}[L]\le 2$. In order to apply (van der Vaart and Wellner, 1996, Corollary 2.7.4), we partition $\mathbb{R}^d$ into $\cup_{j\ge1}B_j$ where $B_1:=[-\sigma,\sigma]^d$ and $B_j:=[-j\sigma,j\sigma]^d\backslash[-(j-1)\sigma,(j-1)\sigma]^d$ for $j\ge 2$. Since B_j is *not convex* for $j\ge 2$, we further partition it into disjoint hypercubes $\{B_{j,k}\}_{k=1}^{2d}$, e.g.,

$$B_{j,1}=[(j-1)\sigma,j\sigma]\times[-j\sigma,j\sigma]^{d-1}.$$

Take any $j\ge 2$ and $k\in[2d]$. Firstly, it holds that

$$\lambda\{x:d(x,B_{j,k})\le 1\}\le(\sigma+2)(2j\sigma+2)^{d-1}\le C_d(1+j^d\sigma^d),$$

where λ is the Lebesgue measure. Secondly, the mass that $\hat{P}_n$ assigns to $B_{j,k}$ can be bounded as follows:

$$\hat{P}_n(Z\in B_{j,k})\le\hat{P}_n\left(\|Z\|^2>d\sigma^2(j-1)^2\right)\le\hat{P}_n\left[e^{\|Z\|^2/2d\sigma^2}\right]e^{-(j-1)^2/2}=Le^{-(j-1)^2/2}.\tag{25}$$

Finally, we prove that $\mathcal{F}^s\subset\mathcal{C}^s_M(B_{j,k})$ with $M=C_{s,d,w}(1+j^s\sigma^s)$, where $\mathcal{C}^s_M(B_{j,k})$ is the set of continuous functions satisfying

$$\|f\|_s:=\max_{|\alpha|\le s}\sup_{z\in B_{j,k}}|D^\alpha f(z)|+\max_{|\alpha|=s}\sup_{z,w\in B_{j,k}}|D^\alpha f(z)-D^\alpha f(w)|\le M.$$

In fact, for any $f \in \mathcal{F}^s$, we have

$$\max_{|\alpha|\le s} \sup_{z\in B_{j,k}} |D^\alpha f(z)| \le C_{s,d,w} \sup_{z\in B_{j,k}} (1+\|z\|^s) \le C_{s,d,w}(1+j^s\sigma^s),$$

and

$$\max_{|\alpha|=s} \sup_{z,w\in B_{j,k}} |D^\alpha f(z) - D^\alpha f(w)| \le 2 \max_{|\alpha|=s} \sup_{z\in B_{j,k}} |D^\alpha f(z)| \le C_{s,d}(1+j^s\sigma^s).$$

Note that the same argument holds for any $f \in \mathcal{F}^s_\sigma$ since we can simply replace $1+\|z\|^s$ by $1+\max\{\|z\|^s, \sigma^s\}$. Now, applying (van der Vaart and Wellner, 1996, Corollary 2.7.4) with $r = 2$ and $V = d/s$ leads to

$$\begin{aligned}\log N(\tau, \mathcal{F}^s, \mathbf{L}^2(\hat{P}_n)) &\le C_{s,d,w}\tau^{-d/s}L^{d/2s}\left(1+\sum_{j=2}^\infty\sum_{k=1}^{2d}(1+j^d\sigma^d)^{\frac{2s}{d+2s}}(1+j^s\sigma^s)^{\frac{2d}{d+2s}}e^{-\frac{d(j-1)^2}{d+2s}}\right)^{\frac{d+2s}{2s}} \\ &\le C_{s,d,w}\tau^{-d/s}L^{d/2s}(1+\sigma^{2d})\left(2d\sum_{j=1}^\infty j^{\frac{4ds}{d+2s}}e^{-\frac{d(j-1)^2}{d+2s}}\right)^{\frac{d+2s}{2s}} \\ &\le C_{s,d,w}\tau^{-d/s}L^{d/2s}(1+\sigma^{2d}), \quad \text{by the summability.}\end{aligned}$$

To verify the second inequality, we obtain

$$\max_{f\in\mathcal{F}^s} \|f\|^2_{\mathbf{L}^2(\hat{P}_n)} = \max_{f\in\mathcal{F}^s} \hat{P}_n[|f(Z)|^2] \le C_{s,d,w}\hat{P}_n[(1+\|Z\|^4)]. \tag{26}$$

Note that $\|Z\|^4 \le C_d e^{\|Z\|^2/2d\sigma^2}\sigma^4$. It follows that $\hat{P}_n[\|Z\|^4] \le C_d L\sigma^4$, and thus

$$\max_{f\in\mathcal{F}^s} \|f\|^2_{\mathbf{L}^2(\hat{P}_n)} \le C_{s,d,w}(1+L\sigma^4).$$

Again, the same argument hold for $\mathcal{F}^s_\sigma$ by replacing $\|Z\|^4$ with $\max\{\|Z\|^4, \sigma^4\}$. □

With this covering number bound at hand, we can control the empirical process by the metric entropy.

Proposition 15. *Let $P \in \mathcal{M}_1(\mathbb{R}^d)$ be subG(σ^2). Let $\{Z_i\}_{i=1}^n \overset{i.i.d.}{\sim} P$ and $\hat{P}_n$ be the empirical measure. Then,*

$$\mathbb{E}\left\|\hat{P}_n - P\right\|^2_{\mathcal{F}^s} \le C_{s,d,w}(1+\sigma^{2d+4})\frac{1}{n}, \quad \textit{for all } s > d/2.$$

Moreover, the same bound holds for $\mathcal{F}^s_\sigma$.

Proof. Define the symmetrized version of $\left\|\hat{P}_n - P\right\|_{\mathcal{F}^s}$ by

$$\left\|\hat{\mathbb{S}}_n\right\|_{\mathcal{F}^s} := \sup_{f\in\mathcal{F}^s}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Z_i)\right|, \tag{27}$$

where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. Rademacher random variables that are independent with $\{Z_i\}_{i=1}^n$. According to (Wainwright, 2019, Proposition 4.11), it holds that

$$\mathbb{E}\left\|\hat{P}_n - P\right\|^2_{\mathcal{F}^s} \le 4\,\mathbb{E}\left\|\hat{\mathbb{S}}_n\right\|^2_{\mathcal{F}^s}.$$

Conditioning on $\{Z_i\}_{i=1}^n$, the random variable $Z(f) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(Z_i)$ is a linear combination of independent Rademacher random variables. Hence, $Z(f)$ is a sub-Gaussian process (see Definition 7) with respect to

$$\|f-g\|_{\mathbf{L}^2(\hat{P}_n)} = \sqrt{\frac{1}{n}\sum_{i=1}^n [f(Z_i)-g(Z_i)]^2}.$$

It then follows from Proposition 16 below that

$$\begin{aligned}
\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}^s} |Z(f)|^2 &\le C\left(\int_0^{2\max_{f\in\mathcal{F}^s}\|f\|_{\mathbf{L}^2(\hat{P}_n)}} \sqrt{\log N(\tau,\mathcal{F}^s,\mathbf{L}^2(\hat{P}_n))}d\tau\right)^2 \\
&\le C_{s,d,w}\left(\int_0^{C_{s,d}\sqrt{1+L\sigma^4}} \tau^{-d/2s}L^{d/4s}\sqrt{1+\sigma^{2d}}d\tau\right)^2, \quad \text{by Proposition 14} \\
&= C_{s,d,w}(1+\sigma^{2d})L^{d/2s}(1+L\sigma^4)^{1-d/2s}, \quad \text{by } s>d/2 \\
&\le C_{s,d,w}(1+\sigma^{2d+4})L, \quad \text{by } L\ge 1.
\end{aligned}$$

Note that $\mathbb{E}\left\|\hat{\mathbb{S}}_n\right\|^2_{\mathcal{F}^s} = \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}^s}|Z(f)|^2$. Consequently, we have

$$\mathbb{E}\left\|\hat{P}_n - P\right\|^2_{\mathcal{F}^s} \le C_{s,d,w}(1+\sigma^{2d+4})\frac{1}{n}. \tag{28}$$

The same argument holds for $\mathcal{F}^s_\sigma$ since Proposition 14 holds true for $\mathcal{F}^s_\sigma$. ☐

The following proposition controls the $\mathbf{L}^2$ norm of the supremum of a sub-Gaussian process. It can be obtained from Giné and Nickl (2015, Exercise 2.3.1). We give its proof here for self-completeness.

Proposition 16. *Let $\{Z(\theta)\}_{\theta\in\Theta}$ be a sub-Gaussian process with respect to a metric ρ in Θ such that $\int_0^\infty \sqrt{\log N(\tau,\Theta,\rho)}d\tau < \infty$. Then it holds that, for any separable version of Z,*

$$\left\|\sup_{\theta\in\Theta}|Z(\theta)|\right\|_{\mathbf{L}^2} \le \|Z(\theta_0)\|_{\mathbf{L}^2} + C\int_0^D \sqrt{\log N(\tau,\Theta,\rho)}d\tau, \tag{29}$$

where $\theta_0\in\Theta$ is arbitrary and D is the ρ-diameter of Θ.

Proof. Due to the separability, it suffices to prove

$$\left\|\sup_{\theta\in\Theta'}|Z(\theta)|\right\|_{\mathbf{L}^2} \le \|Z(\theta_0)\|_{\mathbf{L}^2} + C\int_0^D \sqrt{\log N(\tau,\Theta,\rho)}d\tau \tag{30}$$

for any finite $\Theta'\subset\Theta$. When the diameter $D=0$, the claim holds trivially and thus we only need to focus on the case when $|\Theta'|\ge 2$. By considering $(Z(\theta)-Z(\theta_0))/(1+\delta)D$ and $\rho/(1+\delta)D$ instead of $Z(\theta)$ and ρ for some any small $\delta>0$, we may assume that $Z(\theta_0)=0$ and $D\in(1/2,1)$. Our proof relies on the classical chaining argument.

Step 1. Construct a chain of projections. Let $r_1\in\mathbb{N}$ be such that, for any $\theta\in\Theta$, the ball $B(\theta,2^{-r_1})$ centered at θ of radius 2^{-r_1} contains at most 1 element in Θ'. Denote $\Theta_{r_1} := \Theta'$ and $\Theta_0 := \{\theta_0\}$. For each $1\le r<r_1$, we take a 2^{-r} covering of Θ and let Θ_r be the collection of these centers. By definition, we get $|\Theta_r|\le N(2^{-r},\Theta,\rho)$ for all $0\le r\le r_1$. For each $\theta\in\Theta'$, we construct a chain $(\pi_{r_1}(\theta),\pi_{r_1-1}(\theta),\dots,\pi_0(\theta))$ such that $\pi_r(\theta)\in\Theta_r$ as follows. For $r=r_1$, we let $\pi_r(\theta)=\theta$. For any $0\le r<r_1$, we define $\pi_r(\theta)$ to be a point in Θ_r for which the ball $B(\pi_r(\theta),2^{-r})$ contains $\pi_{r+1}(\theta)$. Note that there may be multiple points satisfying this requirement, but we select the same one for θ and θ' as long as $\pi_{r+1}(\theta)=\pi_{r+1}(\theta')$.

Step 2. Telescoping. By the triangle inequality, we have

$$\left\|\max_{\theta\in\Theta'}|Z(\theta)|\right\|_{\mathbf{L}^2} = \left\|\max_{\theta\in\Theta'}|Z(\pi_{r_1}(\theta))-Z(\pi_0(\theta))|\right\|_{\mathbf{L}^2} \le \sum_{r=1}^{r_1}\left\|\max_{\theta\in\Theta'}|Z(\pi_r(\theta))-Z(\pi_{r-1}(\theta))|\right\|_{\mathbf{L}^2}.$$

Note that

$$|\{(\pi_r(\theta),\pi_{r-1}(\theta)):\theta\in\Theta'\}| = |\{\pi_r(\theta):\theta\in\Theta'\}| \le |\Theta_r| \le N(2^{-r},\Theta,\rho).$$

According to (Giné and Nickl, 2015, Lemma 2.3.3), we obtain

$$\begin{aligned}
\left\|\max_{\theta\in\Theta'}|Z(\pi_r(\theta))-Z(\pi_{r-1}(\theta))|\right\|_{\mathbf{L}^2} &\le C\sqrt{\log N(2^{-r},\Theta,\rho)}\max_{\theta\in\Theta'}\|Z(\pi_r(\theta))-Z(\pi_{r-1}(\theta))\| \\
&\le C2^{-r+1}\sqrt{\log N(2^{-r},\Theta,\rho)}.
\end{aligned}$$

Consequently, it holds that

$$\left\|\max_{\theta\in\Theta'}|Z(\theta)|\right\|_{\mathbf{L}^2} \le C\sum_{r=1}^{r_1} 2^{-r+1}\sqrt{\log N(2^{-r},\Theta,\rho)} \le C\int_0^1 \sqrt{\log N(\tau,\Theta,\rho)}d\tau,$$

which completes the proof. □

B.3 Proofs of Main Results

We now prove the main consistency results in Section 3. For simplicity of the notation, we focus on the quadratic cost function, i.e., $w_1 = w_2 = 1$, and drop the dependency on w (e.g., we write $C_{s,d} = C_{s,d,w}$. The proofs can be adapted to weighted quadratic costs with minor modifications. Let $P_X \in \mathcal{M}_1(\mathbb{R}^{d_1})$ and $P_Y \in \mathcal{M}_1(\mathbb{R}^{d_2})$ with $d := d_1 + d_2$. Suppose that $\{(X_i, Y_i)\}_{i=1}^n$ is an i.i.d. sample from some joint distribution P_{XY} with marginals P_X and P_Y, where P_{XY} may or may not equal $P_X \otimes P_Y$. Let $\hat{P}_n$ and $\hat{Q}_n$ be the empirical measures of $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, respectively.

Proof of Proposition 4. Step 1. Decoupling. Due to the degeneracy, it suffices to bound

$$\mathbb{E}\left\|\hat{P}_X \otimes \hat{P}_Y\right\|_{\mathcal{F}}^2 = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n f(X_i,Y_j)\right|^2\right]. \tag{31}$$

We prove in the following that it boils down to control (31) under the product measure $P_X \otimes P_Y$. When $P_{XY} = P_X \otimes P_Y$, the claim holds trivially. When $P_{XY} \neq P_X \otimes P_Y$, we use the decoupling technique (Peña and Giné, 1999). Note that, by the Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n f(X_i,Y_j)\right|^2\right] \le C\,\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i\neq j} f(X_i,Y_j)\right|^2 + \sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i=1}^n f(X_i,Y_i)\right|^2\right].$$

Note that the second term on the RHS is a lower order term and can be taken care of by Proposition 15. Hence, it suffices to upper bound the first term. Let $\{\varepsilon_i\}_{i=1}^n$ be i.i.d. Rademacher random variables and $\{(X_i', Y_i')\}_{i=1}^n$ be an independent copy of $\{(X_i, Y_i)\}_{i=1}^n$. Define

$$A_i := \begin{cases} X_i & \text{if } \varepsilon_i = 1 \\ X_i' & \text{if } \varepsilon_i = -1 \end{cases} \quad \text{and} \quad B_i := \begin{cases} Y_i' & \text{if } \varepsilon_i = 1 \\ Y_i & \text{if } \varepsilon_i = -1 \end{cases}.$$

For any functional $F : \mathcal{F} \to \mathbb{R}_+$, let $\Phi(F) := \sup_{f\in\mathcal{F}} F(f)^2$. For instance, we define $U_{X,Y}(f) := \frac{1}{n^2}\left|\sum_{i\neq j} f(X_i,Y_j)\right|$. It is clear that Φ is convex and increasing, and the target reads

$$\mathbb{E}\left[\Phi(U_{X,Y})\right] = \mathbb{E}\left[\Phi\left(\left|\frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\left[f(X_i,Y_j) + f(X_i',Y_j) + f(X_i,Y_j') + f(X_i',Y_j') \mid \mathcal{Z}\right]\right|\right)\right],$$

where $\mathcal{Z} := \{(X_i, Y_i)\}_{i=1}^n$. Since, for any $i \neq j$,

$$f(X_i,Y_j) + f(X_i',Y_j) + f(X_i,Y_j') + f(X_i',Y_j') = 4\,\mathbb{E}\left[f(A_i,B_j) \mid \mathcal{Z},\mathcal{Z}'\right],$$

it follows from the convexity and the monotonicity of Φ that

$$\mathbb{E}\left[\Phi(U_{X,Y})\right] \le \mathbb{E}\left[\Phi(4U_{A,B})\right].$$

Finally, the joint distribution of $(X_1,\ldots,X_n,Y_1',\ldots,Y_n')$ is the same as $(A_1,\ldots,A_n,B_1,\ldots,B_n)$, so we have

$$\mathbb{E}\left[\Phi(U_{X,Y})\right] \le \mathbb{E}\left[\Phi(4U_{X,Y'})\right].$$

Adding back the diagonal terms proves the claim since $(X_i, Y_i') \sim P_X \otimes P_Y$.

Step 2. Randomization. We work under the measure $P_{XY} = P_X \otimes P_Y$. Note that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n f(X_i,Y_j)\right|^2\right] \\
&= \mathbb{E}_Y\,\mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i=1}^n\left[\sum_{j=1}^n f(X_i,Y_j) - \mathbb{E}_{X'}\left[\sum_{j=1}^n \bar{f}(X_i',Y_j)\right]\right]\right|^2\right], \quad \text{by (36)} \\
&\le \mathbb{E}_Y\,\mathbb{E}_{X,X'}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i=1}^n\left[\sum_{j=1}^n f(X_i,Y_j) - \sum_{j=1}^n \bar{f}(X_i',Y_j)\right]\right|^2\right], \quad \text{by Jensen's inequality} \\
&= \mathbb{E}_Y\,\mathbb{E}_{X,X',\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i=1}^n\varepsilon_i\left[\sum_{j=1}^n \bar{f}(X_i,Y_j) - \sum_{j=1}^n \bar{f}(X_i',Y_j)\right]\right|^2\right] \\
&\le C\,\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\varepsilon_i f(X_i,Y_j)\right|^2\right], \quad \text{by the Cauchy-Schwarz inequality.}
\end{aligned}$$

Repeating above arguments gives

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n f(X_i,Y_j)\right|^2\right] &\le C\,\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n \varepsilon_i\varepsilon_j' f(X_i,Y_j)\right|^2\right] \\
&\le C\,\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n \varepsilon_i\varepsilon_j' f(X_i,Y_j)\right|^2\right],
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality and Jensen's inequality. Hence, it suffices to bound

$$A := \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n \varepsilon_i\varepsilon_j' f(X_i,Y_j)\right|^2.$$

Step 3. Metric entropy. Define the process $Z(f) := \frac{1}{n^{3/2}}\sum_{i,j=1}^n \varepsilon_i\varepsilon_j' f(X_i,Y_j)$ for any $f\in\mathcal{F}$. We claim that it is a sub-Gaussian process with respect to

$$\|f-g\|_{\mathbf{L}^2(\hat{P}_n\otimes\hat{Q}_n)} = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^n [f(X_i,Y_j)-g(X_i,Y_j)]^2}. \tag{32}$$

To prove it, let us control the moment generating function of the increment $Z(f)-Z(g)$. Denote $a_i := \sum_{j=1}^n \varepsilon_j'[f(X_i,Y_j)-g(X_i,Y_j)]$. Conditioning on $\{X_i,Y_i,\varepsilon_i'\}_{i=1}^n$,

$$Z(f)-Z(g) = \frac{1}{n^{3/2}}\sum_{i=1}^n a_i\varepsilon_i$$

is a linear combination of independent Rademacher random variables. Consequently,

$$\mathbb{E}_\varepsilon \exp\{\lambda[Z(f)-Z(g)]\} \le \exp\left\{\frac{\lambda^2\sum_{i=1}^n a_i^2}{2n^3}\right\}. \tag{33}$$

Note that, by the Cauchy-Schwarz inequality,

$$a_i^2 \le \left[\sum_{j=1}^n (\varepsilon_j')^2\right] \left[\sum_{j=1}^n [f(X_i, Y_j) - g(X_i, Y_j)]^2\right] = n \left[\sum_{j=1}^n [f(X_i, Y_j) - g(X_i, Y_j)]^2\right].$$

This yields that

$$\mathbb{E}_\varepsilon \exp\{\lambda[Z(f) - Z(g)]\} \le \exp\left\{\frac{\lambda^2 \sum_{i,j=1}^n [f(X_i, Y_j) - g(X_i, Y_j)]^2}{2n^2}\right\} = \exp\left\{\frac{\lambda^2 \|f - g\|^2_{\mathbf{L}^2(\hat{P}_n \otimes \hat{Q}_n)}}{2}\right\}, \tag{34}$$

and thus the claim follows. Therefore, the conclusion in Proposition 4 holds true due to Proposition 16. □

Proof of Proposition 5. The proof of the first part is similar to Proposition 14. Define $L_1 := \hat{P}_X[e^{\|X\|^2/2d\sigma^2}] \ge 1$ and $L_2 := \hat{P}_Y[e^{\|Y\|^2/2d\sigma^2}] \ge 1$. By the sub-Gaussian assumption, it is clear that $\mathbb{E}[L_1] \le 2$ and $\mathbb{E}[L_2] \le 2$. There are two places in the proof of Proposition 14 where the measure is involved. The first place is (25), where we replace it by

$$\begin{aligned}(\hat{P}_X \otimes \hat{P}_Y)\{(X, Y) \in B_{j,k}\} &\le (\hat{P}_X \otimes \hat{P}_Y)\left\{\|X\|^2 + \|Y\|^2 > d\sigma^2 (j-1)^2\right\} \\ &\le (\hat{P}_X \otimes \hat{P}_Y)\left[\exp\left(\frac{\|X\|^2 + \|Y\|^2}{4d\sigma^2}\right)\right] e^{-(j-1)^2/4}, \quad \text{by Chernoff bound} \\ &= L_1 L_2 e^{-(j-1)^2/4}.\end{aligned}$$

The second place is (26), where we replace it by

$$\max_{f \in \mathcal{F}^s} \|f\|^2_{\mathbf{L}^2(\hat{P}_X \otimes \hat{P}_Y)} = \max_{f \in \mathcal{F}^s} (\hat{P}_X \otimes \hat{P}_Y)[|f(X, Y)|^2] \le C_{s,d}(\hat{P}_X \otimes \hat{P}_Y)[1 + \|X\|^4 + \|Y\|^4].$$

Note that $\|Z\|^4 \le C_d e^{\|Z\|^2/2d\sigma^2}\sigma^4$. It follows that $(\hat{P}_X \otimes \hat{P}_Y)[\|X\|^4 + \|Y\|^4] \le C_d(L_1 + L_2)\sigma^4$. Hence, the claim holds true for $L := (L_1 + L_2)/2$.

For the second part, we define $\theta_f := \mathbb{E}_{P_X \otimes P_Y}[f(X, Y)]$,

$$f_{1,0}(X) := \mathbb{E}_{P_X \otimes P_Y}[f(X, Y) \mid X] \quad \text{and} \quad f_{0,1}(Y) := \mathbb{E}_{P_X \otimes P_Y}[f(X, Y) \mid Y] \tag{35}$$

for each $f \in \mathcal{F}^s$. As a result, $\bar{f}(x, y) := f(x, y) - f_{1,0}(x) - f_{0,1}(y) + \theta_f$ satisfies

$$\mathbb{E}_{P_X \otimes P_Y}[\bar{f}(X, Y) \mid X] \overset{\text{a.s.}}{=} 0 \overset{\text{a.s.}}{=} \mathbb{E}_{P_X \otimes P_Y}[\bar{f}(X, Y) \mid Y]. \tag{36}$$

Note that

$$\begin{aligned}&\mathbb{E}\left\|\hat{P}_X \otimes \hat{P}_Y - P_X \otimes P_Y\right\|^2_{\mathcal{F}^s} \\ &= \mathbb{E}\left[\sup_{f \in \mathcal{F}^s}\left|\frac{1}{n^2}\sum_{i,j=1}^n \big(f(X_i, Y_j) - \theta_f\big)\right|^2\right] \\ &\le C\,\mathbb{E}\left[\sup_{f \in \mathcal{F}^s}\left|\frac{1}{n^2}\sum_{i,j=1}^n \bar{f}(X_i, Y_j)\right|^2 + \sup_{f \in \mathcal{F}^s}\left|\frac{1}{n}\sum_{i=1}^n f_{1,0}(X_i) - \theta_f\right|^2 + \sup_{f \in \mathcal{F}^s}\left|\frac{1}{n}\sum_{i=1}^n f_{0,1}(Y_i) - \theta_f\right|^2\right] \\ &\le C\,\mathbb{E}\left[\sup_{f \in \mathcal{F}^s}\left|\frac{1}{n^2}\sum_{i,j=1}^n \bar{f}(X_i, Y_j)\right|^2 + \left\|\hat{P}_X - P_X\right\|^2_{\mathcal{F}^s_\sigma} + \left\|\hat{P}_Y - P_Y\right\|^2_{\mathcal{F}^s_\sigma}\right], \quad \text{by Lemma 23.}\end{aligned} \tag{37}$$

Since the last two terms above can be controlled by Proposition 15, it remains to consider the first term. Analogous to the proof of Proposition 15, we obtain, by Proposition 4 and the first part, that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}^s}\left|\frac{1}{n^2}\sum_{i,j=1}^{n}\bar{f}(X_i,Y_j)\right|^2\right]\leq C_{s,d}(1+\sigma^{2d+4})\frac{1}{n}.$$

Therefore, by (37), we have

$$\mathbb{E}\left\|\hat{P}_n\otimes\hat{Q}_n-P\otimes Q\right\|_{\mathcal{F}^s}^2\leq C_{s,d}(1+\sigma^{2d+4})\frac{1}{n}.$$

□

Now we are ready to prove Theorem 3.

Proof of Theorem 3. We prove the statement for $\varepsilon=1$ and write $S:=S_1$. The result for general $\varepsilon>0$ follows immediately from Lemma 24. By the triangle inequality, it holds that

$$\begin{aligned}|T_n(X,Y)-T(X,Y)|\leq\left|S(\hat{P}_{XY},\hat{P}_X\otimes\hat{P}_Y)-S(P_{XY},P_X\otimes P_Y)\right|+\frac{1}{2}\left|S(\hat{P}_{XY},\hat{P}_{XY})-S(P_{XY},P_{XY})\right|\\+\frac{1}{2}\left|S(\hat{P}_X\otimes\hat{P}_Y,\hat{P}_X\otimes\hat{P}_Y)-S(P_X\otimes P_Y,P_X\otimes P_Y)\right|.\end{aligned}\tag{38}$$

We begin with deriving the bound for the first term

$$A:=\left|S(\hat{P}_{XY},\hat{P}_X\otimes\hat{P}_Y)-S(P_{XY},P_X\otimes P_Y)\right|.\tag{39}$$

Step 1. Upper bound via empirical processes. According to Lemma 19 and Lemma 20, the joint distribution P_{XY} is subG$(2\sigma^2)$, and thus there exist a zero-measure set $S_{P_{XY}}\subset\Omega$ and a random variable $\sigma^2_{P_{XY}}$ such that $\hat{P}_{XY}(\omega)$ and P_{XY} are subG$(\sigma^2_{P_{XY}}(\omega))$ for every $\omega\in S^c_{P_{XY}}$. Similarly, by Lemma 21, there exist a zero-measure set $S_{P_X,P_Y}\subset\Omega$ and a random variable $\sigma^2_{P_X,P_Y}$ such that $\hat{P}_X(\omega)\otimes\hat{P}_Y(\omega)$ and $P_X\otimes P_Y$ are subG$(\sigma^2_{P,Q}(\omega))$ for every $\omega\in S^c_{P_X,P_Y}$. Take $S:=S^c_{P_{XY}}\cap S^c_{P_X,P_Y}$ and $\bar{\sigma}^2:=\max\{\sigma^2_{P_{XY}},\sigma^2_{P_X,P_Y}\}$. It follows that $\hat{P}_{XY}(\omega)$, $\hat{P}_X(\omega)\otimes\hat{P}_Y(\omega)$, P_{XY}, and $P_X\otimes P_Y$ are subG$(\bar{\sigma}^2(\omega))$ for every $\omega\in S$. Now, by Proposition 13,

$$\begin{aligned}&\left|S(\hat{P}_{XY}(\omega),\hat{P}_X(\omega)\otimes\hat{P}_Y(\omega))-S(P_{XY},P_X\otimes P_Y)\right|\\&\leq\sup_{f\in\mathcal{F}_{\bar{\sigma}(\omega)}}\left|\int f(d\hat{P}_{XY}(\omega)-dP_{XY})\right|+\sup_{g\in\mathcal{F}_{\bar{\sigma}(\omega)}}\left|\int g(d\hat{P}_X(\omega)\otimes\hat{P}_Y(\omega)-dP_X\otimes P_Y)\right|,\quad\forall\omega\in S.\end{aligned}$$

Note that $\mathbb{P}(S)=\mathbb{P}(S^c_{P_{XY}}\cap S^c_{P_X,P_Y})=1$. This implies, almost surely,

$$A\leq\sup_{f\in\mathcal{F}_{\bar{\sigma}}}\left|\int f(d\hat{P}_{XY}-dP_{XY})\right|+\sup_{g\in\mathcal{F}_{\bar{\sigma}}}\left|\int g(d\hat{P}_X\otimes\hat{P}_Y-dP_X\otimes P_Y)\right|.\tag{40}$$

According to Lemma 22, we have

$$\begin{aligned}\mathbb{E}[A]&\leq\mathbb{E}\left[(1+\bar{\sigma}^{3s})\left\|\hat{P}_{XY}-P_{XY}\right\|_{\mathcal{F}^s}\right]+\mathbb{E}\left[(1+\bar{\sigma}^{3s})\left\|\hat{P}_X\otimes\hat{P}_Y-P_X\otimes P_Y\right\|_{\mathcal{F}^s}\right]\\&\leq\sqrt{\mathbb{E}[(1+\bar{\sigma}^{3s})^2]}\left[\sqrt{\mathbb{E}\left\|\hat{P}_{XY}-P_{XY}\right\|^2_{\mathcal{F}^s}}+\sqrt{\mathbb{E}\left\|\hat{P}_X\otimes\hat{P}_Y-P_X\otimes P_Y\right\|^2_{\mathcal{F}^s}}\right].\end{aligned}$$

Step 2. Control empirical processes via metric entropy. Let $s=\lceil d/2\rceil+1$. Since the joint probability P_{XY} is subG$(2\sigma^2)$, it follows from Proposition 15 that

$$\sqrt{\mathbb{E}\left\|\hat{P}_{XY}-P_{XY}\right\|^2_{\mathcal{F}^s}}\leq C_d(1+\sigma^{d+2})\frac{1}{\sqrt{n}}.\tag{41}$$

The same bound holds for $\sqrt{\mathbb{E}\left\|\hat{P}_X \otimes \hat{P}_Y - P_X \otimes P_Y\right\|_{\mathcal{F}^s}^2}$ by Proposition 4. Note that

$$\mathbb{E}[(1+\tilde{\sigma}^{3s})^2] \le C(1+\mathbb{E}\,\tilde{\sigma}^{6s}) \le C_s(1+\mathbb{E}\,\sigma_{P_{XY}}^{6s} + \mathbb{E}\,\sigma_{P_X,P_Y}^{6s}) \le C_s(1+\sigma^{6s}),$$

where the last inequality follows from Lemma 19 and Lemma 21. Recall that we have chosen $s = \lceil d/2 \rceil + 1$. As a result, $\mathbb{E}[A] \le C_d(1+\sigma^{\lceil 5d/2\rceil+6})n^{-1/2}$. A similar argument shows that the same bound hold for the second and third term in (38). Hence,

$$\mathbb{E}\,|T_n(X,Y)| \le C_d(1+\sigma^{\lceil 5d/2\rceil+6})\frac{1}{\sqrt{n}}. \tag{42}$$

□

C EXPONENTIAL TAIL BOUNDS

We now prove the exponential tail bound in Section 3. For simplicity of the notation, we focus on the quadratic cost function, i.e., $w_1 = w_2 = 1$, and drop the dependency on w (e.g., we write $C_{s,d} = C_{s,d,w}$. The proofs can be adapted to weighted quadratic costs with minor modifications. Let $P_X \in \mathcal{M}_1(\mathbb{R}^{d_1})$ and $P_Y \in \mathcal{M}_1(\mathbb{R}^{d_2})$ with $d := d_1 + d_2$. Suppose that $\{(X_i, Y_i)\}_{i=1}^n$ is an i.i.d. sample from some joint distribution P_{XY} with marginals P_X and P_Y, where P_{XY} may or may not equal $P_X \otimes P_Y$. Let $\hat{P}_n$ and $\hat{Q}_n$ be the empirical measures of $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, respectively.

Proposition 17. *For any b-uniformly bounded class of functions $\mathcal{F}$, we have*

$$\mathbb{P}\left\{\left\|\hat{P}_X \otimes \hat{P}_Y - P_X \otimes P_Y\right\|_{\mathcal{F}} - \mathbb{E}\left\|\hat{P}_X \otimes \hat{P}_Y - P_X \otimes P_Y\right\|_{\mathcal{F}} > t\right\} \le \exp\left(-\frac{nt^2}{8b^2}\right), \quad \textit{for any } t \ge 0.$$

Proof. For any function f defined on $\mathbb{R}^d$, we define $\bar{f}(x,y) = f(x,y) - (P_X \otimes P_Y)[f]$. As a results, we have $\left\|\hat{P}_X \otimes \hat{P}_Y - P_X \otimes P_Y\right\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n \bar{f}(X_i, Y_j)\right|$. Consider the function

$$F(z_1, \dots, z_n) := \sup_{f\in\mathcal{F}}\left|\frac{1}{n^2}\sum_{i,j=1}^n \bar{f}(x_i, y_j)\right|, \tag{43}$$

where $z_i = (x_i, y_i) \in \mathbb{R}^d$. We claim that F satisfies the bounded difference property required in the McDiarmid inequality. Since F is permutation invariant, it suffices to verify the property for the first coordinate. Let $z_1' \ne z_1$ and $z_i' = z_i$ for all $i \ne 1$. It holds that

$$\begin{aligned}\left|\frac{1}{n^2}\sum_{i,j=1}^n \bar{f}(x_i, y_j)\right| - F(z_1', \dots, z_n') &\le \left|\frac{1}{n^2}\sum_{i,j=1}^n \bar{f}(x_i, y_j)\right| - \left|\frac{1}{n^2}\sum_{i,j=1}^n \bar{f}(x_i', y_j')\right| \\ &\le \frac{1}{n^2}\sum_{i=1 \text{ or } j=1}\left|\bar{f}(x_i, y_j) - \bar{f}(x_i', y_j')\right| \le \frac{4b}{n},\end{aligned}$$

where the last inequality uses the boundedness of f. Taking the supremum over $\mathcal{F}$ yields that $F(z_1, \dots, z_n) - F(z_1', \dots, z_n') \le 4b/n$. By symmetry, it follows that $|F(z_1, \dots, z_n) - F(z_1', \dots, z_n')| \le 4b/n$. Note that $\{Z_i := (X_i, Y_i)\}_{i=1}^n$ is an i.i.d. sample. According to the McDiarmid inequality, it holds that

$$\mathbb{P}\left\{\left\|\hat{P}_X \otimes \hat{P}_Y - P_X \otimes P_Y\right\|_{\mathcal{F}} - \mathbb{E}\left\|\hat{P}_X \otimes \hat{P}_Y - P_X \otimes P_Y\right\|_{\mathcal{F}} > t\right\} \le \exp\left(-\frac{nt^2}{8b^2}\right), \quad \text{for any } t \ge 0.$$

□

Proof of Theorem 6. We prove the statement for $\varepsilon = 1$ and write $S := S_1$. The result for general $\varepsilon > 0$ follows immediately from Lemma 24. By the bounded support assumption, it holds that P_X and P_Y are both $\mathrm{subG}(D^2/d)$. According to the proof of Lemma 19, we have $\{\hat{P}_X\}_{n\ge1}$, $\{\hat{P}_Y\}_{n\ge1}$, P_X, and P_Y are uniformly subG(τ^2) for $\tau^2 := D^2 e^{1/2}/d \le 2D^2/d$. Moreover, it follows from Lemma 20 that $\{\hat{P}_{XY}\}_{n\ge1}$ and P_{XY} are uniformly subG($2\tau^2$). As a result, we obtain, by Proposition 13,

$$\begin{aligned} A &:= \left|S(\hat{P}_{XY}, \hat{P}_X \otimes \hat{P}_Y) - S(P_{XY}, P_X \otimes P_Y)\right| \\ &\le \sup_{f\in\mathcal{F}_{2\tau}} \left|\int f(d\hat{P}_{XY} - dP_{XY})\right| + \sup_{g\in\mathcal{F}_{2\tau}} \left|\int g(d\hat{P}_X \otimes \hat{P}_Y - dP_X \otimes P_Y)\right|. \end{aligned}$$

Fix $s = \lceil d/2 \rceil + 1$. According to Lemma 22, we have

$$A \le C_d(1 + D^{3d+12}) \left[\left\|\hat{P}_{XY} - P_{XY}\right\|_{\mathcal{F}^s} + \left\|\hat{P}_X \otimes \hat{P}_Y - P_X \otimes P_Y\right\|_{\mathcal{F}^s}\right], \tag{44}$$

where we have used $\tau^{3s} \le C_d D^{3d+12}$. Proposition 12 shows that we can further constraint the function class $\mathcal{F}^s$ to $\mathcal{F}^s_b := \{f \in \mathcal{F}^s : \|f\|_\infty \le b\}$ for $b = 2D^2$. Hence, by (Wainwright, 2019, Theorem 4.10), it holds that

$$\mathbb{P}\left\{\left\|\hat{P}_{XY} - P_{XY}\right\|_{\mathcal{F}^s_b} - \mathbb{E}\left\|\hat{P}_{XY} - P_{XY}\right\|_{\mathcal{F}^s_b} > t\right\} \le \exp\left(-\frac{nt^2}{2b^2}\right), \quad \text{for any } t \ge 0.$$

It is clear from Proposition 15 that

$$\mathbb{E}\left\|\hat{P}_{XY} - P_{XY}\right\|_{\mathcal{F}^s_b} \le \mathbb{E}\left\|\hat{P}_{XY} - P_{XY}\right\|_{\mathcal{F}^s} \le C_d(1 + D^{2d+4})\frac{1}{\sqrt{n}}.$$

Consequently, we get

$$\mathbb{P}\left\{\left\|\hat{P}_{XY} - P_{XY}\right\|_{\mathcal{F}^s_b} > t + C_d(1 + D^{2d+4})\frac{1}{\sqrt{n}}\right\} \le \exp\left(-\frac{nt^2}{2b^2}\right), \quad \text{for any } t \ge 0.$$

Similarly, using Proposition 4 and Proposition 17, we obtain

$$\mathbb{P}\left\{\left\|\hat{P}_{XY} - P_{XY}\right\|_{\mathcal{F}^s_b} > t + C_d(1 + D^{2d+4})\frac{1}{\sqrt{n}}\right\} \le \exp\left(-\frac{nt^2}{8b^2}\right), \quad \text{for any } t \ge 0.$$

Now it follows from (44) that

$$\mathbb{P}\left\{A \ge C_d(1 + D^{3d+12})\left[t + (1 + D^{2d+4})\frac{1}{\sqrt{n}}\right]\right\} \le 2\exp\left(-\frac{nt^2}{8b^2}\right), \quad \text{for any } t \ge 0.$$

Analogously, we have, for any $t \ge 0$

$$\begin{aligned} \mathbb{P}\left\{B \ge C_d(1 + D^{3d+12})\left[t + (1 + D^{2d+4})\frac{1}{\sqrt{n}}\right]\right\} &\le 2\exp\left(-\frac{nt^2}{8b^2}\right) \\ \mathbb{P}\left\{B' \ge C_d(1 + D^{3d+12})\left[t + (1 + D^{2d+4})\frac{1}{\sqrt{n}}\right]\right\} &\le 2\exp\left(-\frac{nt^2}{8b^2}\right), \end{aligned}$$

where $B := \left|S(\hat{P}_{XY}, \hat{P}_{XY}) - S(P_{XY}, P_{XY})\right|$ and $B' := \left|S(\hat{P}_X \otimes \hat{P}_Y, \hat{P}_X \otimes \hat{P}_Y) - S(P_X \otimes P_Y, P_X \otimes P_Y)\right|$. Since $|T_n(X,Y) - T(X,Y)| \le A + \frac{B}{2} + \frac{B'}{2}$, it holds that

$$\mathbb{P}\left\{|T_n(X,Y) - T(X,Y)| \ge C_d(1 + D^{3d+12})\left[t + (1 + D^{2d+4})\frac{1}{\sqrt{n}}\right]\right\} \le 6\exp\left(-\frac{nt^2}{8b^2}\right). \tag{45}$$

Therefore, we have, with probability at least $1 - \delta$,

$$|T_n(X,Y) - T(X,Y)| \le C_d\left(1 + D^{2d+2}\sqrt{\log\frac{6}{\delta}}\right)\frac{D^{3d+14}}{\sqrt{n}}.$$

□

D TECHNICAL LEMMAS

In this section, we give several technical lemmas used to prove the main results. We use C to denote a constant whose value may change from line to line.

Lemma 18. *If $P \in \mathcal{M}_1(\mathbb{R}^d)$ is subG(σ^2), then, for any $k \in \mathbb{N}_+$,*

$$\mathbb{E}_P[\|Z\|^{2k}] \le (2d\sigma^2)^k k!.$$

Moreover, for any $v \in \mathbb{R}^d$, it holds that

$$\mathbb{E}_P\, e^{\langle v, Z\rangle} \le \mathbb{E}_P\, e^{\|v\|\|Z\|} \le 2e^{d\sigma^2\|v\|^2/2}. \tag{46}$$

Proof. By Taylor's expansion, we have

$$e^{\|Z\|^2/2d\sigma^2} - 1 \ge \frac{\|Z\|^{2k}}{(2d\sigma^2)^k k!}.$$

Taking the expectation on both sides gives

$$\mathbb{E}_P[\|Z\|^{2k}] \le (2d\sigma^2)^k k!.$$

The inequalities (46) follows from the Cauchy-Schwarz inequality and the sub-gaussianity of P. □

Lemma 19. *Let $P \in \mathcal{M}_1(\mathbb{R}^d)$ be subG(σ^2) and $\hat{P}_n$ be the empirical measure. There exist a zero-measure set $S_P \subset \Omega$ and a random variable σ_P^2 depending on the sample $\{Z_i\}_{i=1}^n$ such that $\hat{P}_n(\omega)$ and P are subG$(\sigma_P^2(\omega))$ for any $\omega \in S_P^c$, and, for any $k \in \mathbb{N}_+$,*

$$\mathbb{E}\, \sigma_P^{2k} \le 2k^k\sigma^{2k}.$$

Proof. By the strong law of large numbers, there exists a zero-measure set $S_P \subset \Omega$ such that, for all $\omega \in S_P$,

$$\hat{P}_n(\omega)\left[e^{\|Z\|^2/2d\sigma^2}\right] \to P\left[e^{\|Z\|^2/2d\sigma^2}\right] \le 2, \quad \text{as } n \to \infty. \tag{47}$$

Let $\tau^2 := \sup_n \hat{P}_n\left[e^{\|Z\|^2/2d\sigma^2}\right]$. It follows from (47) that $\tau^2(\omega)$ is finite for all $\omega \in S_P$. Since $\tau^2(\omega) \ge 1$, by Jensen's inequality, we obtain, for all $\omega \in S_P$

$$\hat{P}_n(\omega)\left[e^{\|Z\|^2/2d\sigma^2\tau^2(\omega)}\right] \le \left(\hat{P}_n(\omega)\left[e^{\|Z\|^2/2d\sigma^2}\right]\right)^{1/\tau^2(\omega)} = \left(\tau^2(\omega)\right)^{1/\tau^2(\omega)} < 2.$$

As a result, $\hat{P}_n(\omega)$ is subG$(\sigma^2\tau^2(\omega))$. Moreover, P is also subG$(\sigma^2\tau^2(\omega))$ since $\tau^2(\omega) \ge 1$. Applying the same argument to $\tau_k^2 := \sup_n \hat{P}_n\left[e^{\|Z\|^2/2kd\sigma^2}\right]$ implies that $\hat{P}_n(\omega)$ and P are both subG$(k\sigma^2\tau_k^2(\omega))$. Define $\sigma_P^2 := \min_{k\ge 1} k\sigma^2\tau_k^2$. Then we have, for each $k \ge 1$,

$$\mathbb{E}_P[\sigma_P^{2k}] \le \mathbb{E}_P\left[\hat{P}_n\left[k^k\sigma^{2k}e^{\|Z\|^2/2d\sigma^2}\right]\right] = k^k\sigma^{2k}\,\mathbb{E}_P[e^{\|Z\|^2/2d\sigma^2}] \le 2k^k\sigma^{2k}.$$

□

The sub-Gaussianity of two marginals implies the sub-Gaussianity of the joint.

Lemma 20. *If P_X and P_Y are subG(σ^2), then P_{XY} is subG$(2\sigma^2)$ for any $P_{XY} \in \Pi(P_X, P_Y)$.*

Proof. By the Cauchy-Schwarz inequality,

$$\mathbb{E}_{P_{XY}}\, e^{\|Z\|^2/4d\sigma^2} = \mathbb{E}_{P_{XY}}[e^{\|X\|^2/4d\sigma^2}e^{\|Y\|^2/4d\sigma^2}] \le \sqrt{\mathbb{E}_{P_X}[e^{\|X\|^2/2d\sigma^2}]\,\mathbb{E}_{P_Y}[e^{\|Y\|^2/2d\sigma^2}]}.$$

Since P_X and P_Y are subG(σ^2), it follows that $\mathbb{E}_{P_{XY}}\, e^{\|Z\|^2/4d\sigma^2} \le 2$ and thus P_{XY} is subG$(2\sigma^2)$. □

The next result is for the uniform sub-Gaussianity of the product of two empirical measures.

Lemma 21. *If P_X and P_Y are subG(σ^2), then there exist a zero-measure set $S_{P_X,P_Y} \subset \Omega$ and a random variable $\sigma^2_{P_X,P_Y}$ depending on the sample $\{(X_i, Y_i)\}_{i=1}^n$ such that $\hat{P}_X(\omega) \otimes \hat{P}_Y(\omega)$ and $P_X \otimes P_Y$ are subG$(\sigma^2_{P_X,P_Y}(\omega))$ for any $\omega \in S^c_{P_X,P_Y}$, and, for any $k \in \mathbb{N}_+$,*

$$\mathbb{E}\,\sigma^{2k}_{P_X,P_Y} \le 2^{k+1} k^k \sigma^{2k}.$$

Proof. Similar to Lemma 19. □

The sub-Gaussian processes play an central role in our analysis. We give its definition here; see, e.g., (Wainwright, 2019, Section 5.3).

Definition 7 (Sub-Gaussian process)**.** *Let $\{Z(\theta) : \theta \in \Theta\}$ be a collection of mean-zero random variables. We call it a sub-Gaussian process with respect to a metric ρ in Θ if*

$$\mathbb{E}[e^{\lambda(Z(\theta)-Z(\theta'))}] \le \exp\left[\lambda^2\rho^2(\theta,\theta')/2\right].$$

To facilitate the analysis of $\mathcal{F}_\sigma$ defined in Definition 4, it is convenient to separate the sub-Gaussian parameter from the function class by the following lemma. Note that this result is used in (Mena and Weed, 2019) without proof.

Lemma 22. *For any $\sigma > 0$ and $s \ge 2$. we have $\frac{1}{1+\sigma^{3s}}\mathcal{F}_\sigma \subset \mathcal{F}^s$, where $\mathcal{F}^s := \mathcal{F}^{s,d,w}$ is defined in Definition 5.*

Proof. Take any $f \in \mathcal{F}_\sigma$, it suffices to show $f/(1+\sigma^{3s}) \in \mathcal{F}^s$. According to Proposition 10, it holds that

$$|f(z)| - w_1\|x\|^2 - w_2\|y\|^2 \le \left|f(z) - w_1\|x\|^2 - w_2\|y\|^2\right| \le C_{k,d,w}\begin{cases}(1+\sigma^4) & \text{if } \|z\| \le \sqrt{d}\sigma\\ [1+(1+\sigma^2)\|z\|^2] & \text{if } \|z\| > \sqrt{d}\sigma.\end{cases}$$

Consequently,

$$\left|\frac{f(z)}{1+\sigma^{3s}}\right| \le C_{k,d,w}\begin{cases}\frac{1+\sigma^4}{1+\sigma^{3s}} & \text{if } \|z\| \le \sqrt{d}\sigma\\ \frac{1+(1+\sigma^2)\|z\|^2}{1+\sigma^{3s}} & \text{if } \|z\| > \sqrt{d}\sigma.\end{cases}$$

Since $s \ge 2$, it is clear that $\frac{1+\sigma^4}{1+\sigma^{3s}} \le C$ and $\frac{1+\sigma^2}{1+\sigma^{3s}} \le C$, and thus

$$\left|\frac{f(z)}{1+\sigma^{3s}}\right| \le C_{k,d,w}(1+\|z\|^2).$$

The other inequality can be proved analogously. □

Lemma 23. *Let $P \in \mathcal{M}_1(\mathbb{R}^{d_1})$ and $Q \in \mathcal{M}_1(\mathbb{R}^{d_2})$ be subG(σ^2). Denote $d := d_1 + d_2$. For any $s \ge 1$ and $f \in \mathcal{F}^s$, there exist constants $C_{s,d,w}$ such that $f_{1,0} \in \mathcal{F}^s_\sigma$ and $f_{0,1} \in \mathcal{F}^s_\sigma$, where $\mathcal{F}^s_\sigma$ is defined in Definition 6,*

$$f_{1,0}(x) := \int f(x,y)dQ(y) \quad \text{and} \quad f_{0,1}(y) := \int f(x,y)dP(x).$$

Proof. We only prove it for $f_{1,0}$. By Jensen's inequality, it holds that

$$|f_{1,0}(x)| \le \int |f(x,y)|\,dQ(y) \le C_{s,d,w}\left(1+\|x\|^2 + \int \|y\|^2\,dQ(y)\right) \le C_{s,d,w}(1+\max\{\|x\|^2,\sigma^2\}),$$

where the last inequality follows from Lemma 18. The inequality for $|D^\alpha f_{1,0}(x)|$ can be verified similarly. □

The next lemma suggests that it is enough to consider the case $\varepsilon = 1$ for S_ε.

Lemma 24. *Let $\varepsilon > 0$. For any $P, Q \in \mathcal{M}_1(\mathbb{R}^d)$, it holds that*

$$S_\varepsilon(P,Q) = \varepsilon S(P^\varepsilon, Q^\varepsilon),$$

where P^ε and Q^ε are the pushforwards of P and Q under the map $x \mapsto \varepsilon^{-1/2}x$, respectively.

Proof. By a change of variable argument. □

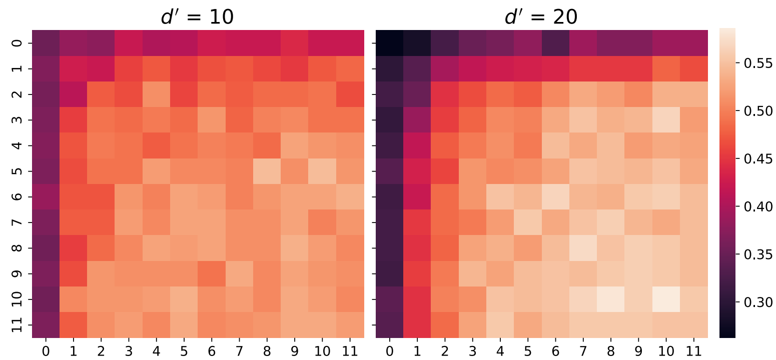

Figure 5: Heatmaps of power for ETIC-RF with $p = 700$ random features and d' PCs on the partially dependent sample of the bilingual data. The x-axis is for r_1 and y-axis is for r_2. The indices from 0 to 11 correspond to equally spaced values from 0.25 to 4. Lighter color indicates larger power.

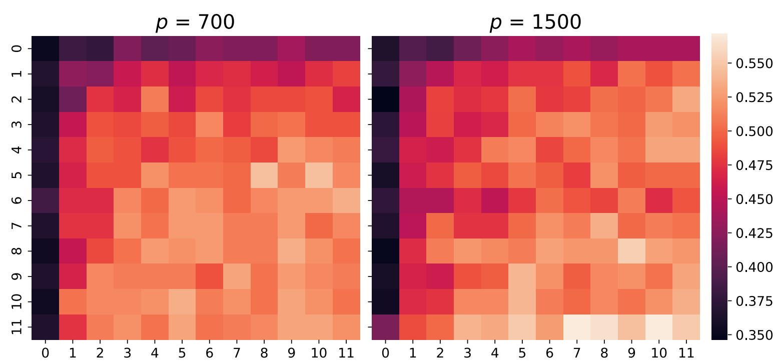

Figure 6: Heatmaps of power for ETIC-RF with $d' = 10$ PCs and p random features on the partially dependent sample of the bilingual data. The x-axis is for r_1 and y-axis is for r_2. The indices from 0 to 11 correspond to equally spaced values from 0.25 to 4. Lighter color indicates larger power.

E ADDITIONAL EXPERIMENTAL RESULTS

E.1 ETIC-RF on Bilingual Text

We examine the ETIC-RF test on bilingual data discussed in Section 4.2. The feature embeddings are of high dimension (i.e., 768), which imposes challenges on the random feature approximation. Hence, we first use dimension reduction (PCA) on the English embeddings and French embeddings separately to reduce the dimension to $d' \ll 768$, and then perform ETIC-RF on the low-dimensional embeddings. Since the dimension reduction step does not utilize information about the joint distribution P_{XY}, it will not violate the level consistency of the test. This is also validated in our experimental results, i.e., all the tests have type I error rate close to 0.05 as expected.

As shown in Figure 5, The number of PCs d' has an interesting effect on the power. Intuitively, the larger d' is the less information we lose, and thus the larger power the test has. This can be seen at the lower right corner where both r_1 and r_2 are large. However, larger d' also means the random feature approximation is harder, especially when r_1 and r_2 are small. This is reflected at the upper left corner where the power decreases as d' increases. We then investigate the effect of p—the number of random features. As shown in Figure 6, the power increases with the number of random features. Overall, the random feature approximation demonstrates similar performance as the exact ETIC with enough random features.

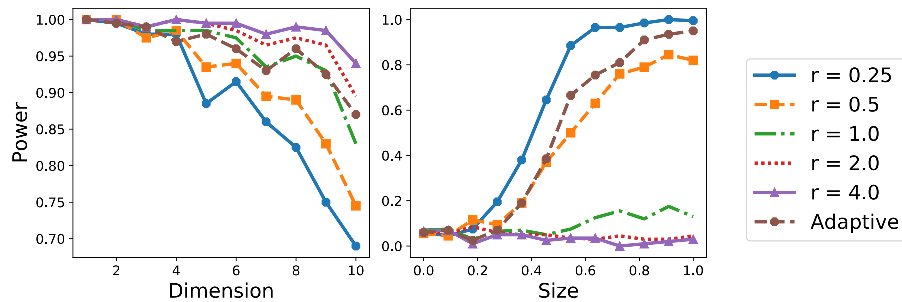

Figure 7: Power curves in the linear dependency model (left) and subspaces dependency model (right).

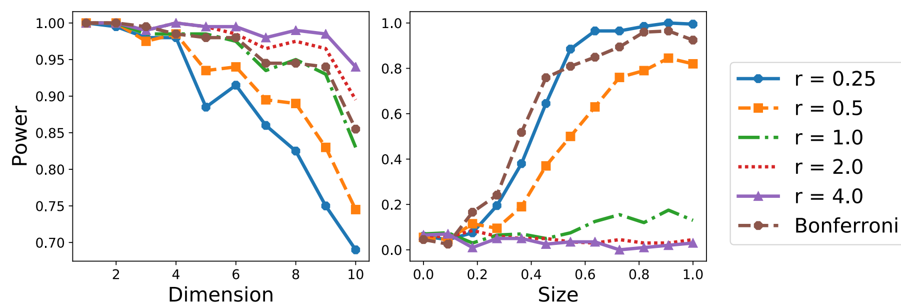

Figure 8: Power curves in the linear dependency model (left) and subspaces dependency model (right).

E.2 Adaptive ETIC Test

To ensure that $(T_{n,\varepsilon}(X,Y))_\varepsilon$ are roughly on the same scale for different values of ε, we use

$$\psi_a(\alpha) := \mathbb{1}\left\{\max_{\varepsilon\in\mathcal{E}} \bar{T}_{n,\varepsilon}(X,Y) > H_{n,\mathcal{E}}(\alpha)\right\}.$$

where each $T_{n,\varepsilon}(X,Y)$ is studentized via resampling (20 permutations) under the null to yield $\bar{T}_{n,\varepsilon}(X,Y)$.

Following the trick in Section 4, we select the cost function to be the weighted quadratic cost with weights given by the median heuristic. We set $\mathcal{E} = \{0.25, 1, 4\}$ and perform the adaptive ETIC test on the linear dependency model and the subspace dependency model. As shown in Figure 7, it is slightly worse than the best ETIC test in both models. We also run it on the bilingual text data. The power and type I error rate of adaptive ETIC are 1 and 0.07 on the dependent sample and the independent sample, respectively. The power achieved is 0.535 on the partially dependent sample; whereas the worst and best power of ETIC are 0.38 and 0.635, respectively.

Finally, we consider a Bonferroni-type ETIC test which is adaptive to both the regularization parameter and the weights in the cost function. Following the formulation in Section 4, we let $\psi_{r_1,r_2}(\alpha)$ be the decision rule of ETIC with hyper-parameters r_1 and r_2. Consider the following Bonferroni-type ETIC test

$$\psi(\alpha) := \max_{r_1,r_2\in\mathcal{R}} \psi_{r_1,r_2}(\alpha/\left|\mathcal{R}^2\right|).$$

We perform this Bonferroni-type ETIC test on the linear dependency model and the subspace dependency model for $\mathcal{R} = \{0.25, 4\}$. As shown in Figure 8, it is slightly worse than the best ETIC test in both models. Compared to the adaptive ETIC test, it performs similar in the linear dependency model and slightly better in the subspace dependency model. We also run it on the bilingual text data. The power and type I error rate of adaptive ETIC are 1 and 0.045 on the dependent sample and the independent sample, respectively. The power achieved is 0.5 on the partially dependent sample, which is smaller than the power of the adaptive ETIC.

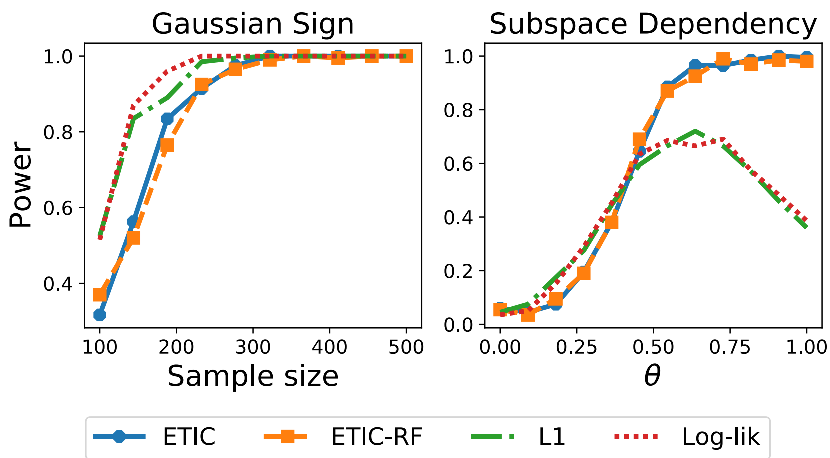

Figure 9: Power curves in the linear dependency model (left) and subspaces dependency model (right).

E.3 Comparison with Other Independence Tests

We implemented another two independence tests considered in (Gretton and Györfi, 2008): the L_1 test and the Log-likelihood test. We apply them to the Gaussian sign model and the subspace dependency model and compare them with ETIC and ETIC-RF with $r = 0.25$. As shown in Figure 9, they slightly outperform ETIC in the Gaussian sign model. In the subspace dependency model, they perform similarly to ETIC for small θ and significantly worse for large θ.