Journal of Machine Learning Research 24 (2023) 1-65 Submitted 6/22; Revised 4/23; Published 5/23

Multivariate Soft Rank via Entropy-Regularized Optimal Transport: Sample Efficiency and Generative Modeling

Shoaib Bin Masud* SHOAIB_BIN.MASUD@TUFTS.EDU
Department of Electrical and Computer Engineering
Tufts University
Medford, MA 02155, USA

Matthew Werenski* MATTHEW.WERENSKI@TUFTS.EDU
Department of Computer Science
Tufts University
Medford, MA 02155, USA

James M. Murphy JM.MURPHY@TUFTS.EDU
Department of Mathematics
Tufts University
Medford, MA 02155, USA

Shuchin Aeron SHUCHIN@ECE.TUFTS.EDU
Department of Electrical and Computer Engineering
Tufts University
Medford, MA 02155, USA

Editor: Maxim Raginsky

Abstract

The framework of optimal transport has been leveraged to extend the notion of rank to the multivariate setting as corresponding to an optimal transport map, while preserving desirable properties of the resulting goodness-of-fit (GoF) statistics. In particular, the rank energy (RE) and rank maximum mean discrepancy (RMMD) are distribution-free under the null, exhibit high power in statistical testing, and are robust to outliers. In this paper, we point to and alleviate some of the shortcomings of these GoF statistics that are of practical significance, namely high computational cost, curse of dimensionality in statistical sample complexity, and lack of differentiability with respect to the data. We show that all these issues are addressed by defining multivariate rank as an entropic transport map derived from the entropic regularization of the optimal transport problem, which we refer to as the *soft rank*. We consequently propose two new statistics, the *soft rank energy (sRE)* and *soft rank maximum mean discrepancy (sRMMD)*. Given n sample data points, we provide non-asymptotic convergence rates for the sample estimate of the entropic transport map to its population version that are essentially of the order $n^{-1/2}$ when the source measure is subgaussian and the target measure has compact support. This result is novel compared to existing results which achieve a rate of n^{-1} but crucially rely on both measures having compact support. In contrast, the corresponding convergence rate of estimating an optimal transport map, and hence the rank map, is exponential in the data dimension. We leverage these fast convergence rates to show that the sample estimates of sRE and sRMMD converge rapidly to their population versions. Combined with the

∗. Equal contribution

computational efficiency of methods in solving the entropy-regularized optimal transport problem, these results enable efficient rank-based GoF statistical computation, even in high dimensions. Furthermore, the sample estimates of sRE and sRMMD are differentiable with respect to the data and amenable to popular machine learning frameworks that rely on gradient methods. We leverage these properties towards showcasing their utility for generative modeling on two important problems: image generation and generating valid knockoffs for controlled feature selection.

Keywords: optimal transport, multivariate rank, high-dimensional statistics, goodness-of-fit testing, generative modeling, knockoff filtering

1. Introduction

It is well-known that in one dimension, the notions of rank and quantile with respect to the distribution of the data are naturally defined via the cumulative distribution function (cdf) and its generalized inverse, respectively. This is because the set of real numbers has a canonical ordering, which is naturally captured by the cdf. Based on these notions, a number of statistical tests for independence and goodness-of-fit (GoF) testing have been proposed in the literature, such as the two-sample Kolmogorov-Smirnov test (Smirnov, 1939), Wilcoxon signed-rank test (Wilcoxon, 1947), Wald-Wolfowitz runs test (Wald and Wolfowitz, 1940), and Hoeffding's D test (Hoeffding, 1994). These statistics possess several desirable properties such as being computationally feasible, non-parametric, and distribution-free under the null.

Recently, meaningful multivariate extensions of the notions of rank and quantile maps were proposed in the pioneering works (Hallin, 2017; Chernozhukov et al., 2017; Hallin et al., 2021), and more recently in (Deb and Sen, 2023) based on the theory of optimal transportation (Villani, 2009; Santambrogio, 2015). For a detailed discussion, we refer the reader to a recent survey on the topic (Hallin, 2021) and references therein. Essentially, these ideas leverage the geometry of the optimal transport (OT) problem with the squared Euclidean metric as the ground cost, where under some mild conditions the optimal maps are gradients of convex functions (Brenier, 1991; McCann, 1995). These extensions and the corresponding high-dimensional analogues of the rank-based GoF statistics based on these extensions retain some of the useful properties of their one-dimensional counterparts, namely they are computationally feasible for small sample sizes and are distribution-free under the null.

In this paper we focus on the multivariate rank-based GoF statistics proposed in (Deb and Sen, 2023), namely the rank energy (RE) and rank maximum mean discrepancy (RMMD), based on a particular choice of the reference measure when defining the multivariate ranks via optimal transport maps. These statistics are shown to be distribution-free in finite samples (under the null), consistent against alternatives, exhibit high power in statistical testing for heavy-tailed distributions, and are robust to outliers.

However, as we discuss in detail in Section 2.2.2, practical use of the RE and RMMD suffer from the well-known curse of dimensionality associated with the estimation of OT maps as well as high computational costs for large sample sizes. Furthermore, RE and RMMD suffer from lack of differentiability with respect to data. This limits a direct use of iterative gradient-based optimization methods and therefore inhibits their potential utility in learning a generative model when using these GoF statistics as a loss function, as has been

successfully done with maximum mean discrepancy (MMD) and the Wasserstein distances (Li et al., 2017; Arjovsky et al., 2017).

In this context, our paper makes the following main contributions.

(C1) We introduce the notion of *soft rank* that utilizes the recent developments in computational optimal transport, namely entropic regularization (Peyré et al., 2019). In particular, we define soft rank as the entropic map (Pooladian and Niles-Weed, 2021) derived from entropy-regularized optimal transport. Based on this notion, we then propose the *soft rank energy (sRE)*, a new multivariate GoF statistic, and the related *soft rank maximum mean discrepancy (sRMMD)*. In Proposition 15 and Proposition 16, we establish the properties of sRE and its convergence to RE, which together justify its utility as a GoF statistic for two-sample testing.

(C2) We provide a new result (Theorem 13) on the convergence rate of a sample-driven estimate for general entropic maps which enjoys a fast convergence rate of $n^{-1/2}$ to the population entropic map, even in high dimensions. We note that the subgaussian assumption in Theorem 13 on the source measure is a significant weakening of the assumptions compared to a recent result of (Rigollet and Stromme, 2022) that assumes compactness of both measures albeit providing a faster rate of n^{-1}. This result is then used in Theorem 17 and Theorem 18 to establish the statistical convergence of sample sRE and sample sRMMD, respectively, to their population versions with rate $n^{-1/2}$. Our analysis also clarifies the impact of key problem parameters, such as the data dimension, entropic regularization strength, support, and subgaussian constants of the distributions.

(C3) We show the practical utility of the proposed statistics on several real generative modeling problems. First, we use sRE and sRMMD as the loss functions in a simple generative model architecture to produce MNIST-digits. Under an appropriate choice of the entropic regularization parameter, we show that using sRE and sRMMD as the loss functions can generate all of the digits successfully and does not suffer from mode collapse. We then utilize the sRMMD in a deep generative model in order to produce valid knockoffs (Barber and Candès, 2015). We showcase improved tradeoffs between detection power versus false discovery rate (FDR) compared to other benchmarks of knockoff generation techniques on different Gaussian and non-Gaussian distributional settings. We also test our approach for provable biomarker selection in metabolomics.[1]

Paper outline: The paper is organized as follows. In Section 2, we provide the required background on optimal transport theory and its entropy-regularized variant as well as discuss the multivariate RE. In Section 3, we introduce the sRE and the sRMMD, and sample versions thereof. In Section 4, we state our main theorems which establish the properties of the sRE and prove finite sample convergence rates for the sample sRE and sRMMD to their population versions. In Section 5, we provide extensive simulations to establish the efficacy of sRE and sRMMD as loss functions for learning generative models.

1. Codes are available at `https://github.com/ShoaibBinMasud/soft-rank-energy-and-applications`

Notation: We will let X, Y denote random vectors in $\mathbb{R}^d$ and we will use superscripts to denote their entries $X = (X^1, \cdots, X^d)$. A bold $\mathbf{X}$ will denote a matrix. $\|\cdot\|$ will denote the Euclidean norm in $\mathbb{R}^d$. $\overset{d}{=}$ will denote equality in distribution. We let $\mathcal{P}(\mathbb{R}^d)$ denote the space of Borel probability measures on $\mathbb{R}^d$ and $\mathcal{P}_{ac}(\mathbb{R}^d)$ the space of absolutely continuous measures (with respect to the Lebesgue measure) on $\mathbb{R}^d$.

Throughout we consider measures of two types, one being measures with bounded support and the another being those with subgaussian concentration. A measure P is said to be σ^2-subgaussian if the random vector X with law P satisfies $\mathbb{E}\left[\exp\left(\frac{\|X\|^2}{2d\sigma^2}\right)\right] \leq 2$; see Appendix B for details.

We write $a \lesssim b$ if there exists a constant C such that $a \leq Cb$. The rest of the notation is standard and clear from the context. We also include in Table 1 a list of notation introduced later for easy reference.

Symbol	Meaning
$B_2^d(0, r)$	Euclidean ball of radius r in $\mathbb{R}^d$
$P_\lambda = \lambda P_X + (1-\lambda)P_Y$	Mixture distribution, $\lambda \in (0, 1)$.
T	Optimal transport map
T_ε	Entropic map
$T_\varepsilon^{n,n}$	Two-sample entropic map
$\mathtt{R}_\lambda$	Rank map. T from P_λ to Unif$([0,1]^d)$
$\mathtt{R}_{\lambda,\varepsilon}$	Soft rank map. T_ε from P_λ to Unif$([0,1]^d)$
$\mathtt{R}_{\lambda,\varepsilon}^{m+n}$	Sample soft rank map
$\mathtt{RE}_\lambda$	Rank energy
$\mathtt{RE}_{m,n}$	Sample rank energy
$\mathtt{sRE}_{\lambda,\varepsilon}$	Soft rank energy
$\mathtt{sRE}_{\lambda,\varepsilon}^{m,n}$	Sample soft rank energy
$\mathtt{RMMD}_\lambda$	Rank maximum mean discrepancy
$\mathtt{RMMD}_{m,n}$	Sample rank maximum mean discrepancy
$\mathtt{sRMMD}_{\lambda,\varepsilon}$	Soft rank maximum mean discrepancy
$\mathtt{sRMMD}_{\lambda,\varepsilon}^{m,n}$	Sample rank maximum mean discrepancy
σ^2	Subgaussian constant
$\|F\|_{L^2(P)}^2$	$\int \|F(x)\|_2^2 dP(x)$ for $F : \mathbb{R}^n \to \mathbb{R}^n$
$\|X\|_{L^2}$	$\mathbb{E}[X^2]^{1/2}$ for a scalar random variable X

Table 1: Frequently used notation throughout the paper.

2. Background on Optimal Transport and Rank Energy

2.1 Optimal Transport

Given two distributions $P, Q \in \mathcal{P}(\mathbb{R}^d)$, the *Monge problem* (Monge, 1781) seeks a measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$ that pushes P to Q with a minimal cost. Precisely, it solves

$$\inf_T \int \frac{1}{2}\|x - T(x)\|^2 dP(x), \quad \text{subject to } T_\# P = Q, \tag{1}$$

where $T_{\#}P$ denotes the *push-forward measure*, which satisfies $(T_{\#}P)[A] = P[T^{-1}(A)]$ for all measurable sets A.

Throughout we will make heavy use of the optimal map T which minimizes (1). It is therefore important to establish the existence, uniqueness, and important properties of T, which are established by the following celebrated theorem.

Theorem 1 (Brenier-McCann (Brenier, 1991; McCann, 1995))**.** *Let $P \in \mathcal{P}_{ac}(\mathbb{R}^d)$ and $Q \in \mathcal{P}(\mathbb{R}^d)$. Then there exists a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ whose gradient $\nabla\phi : \mathbb{R}^d \to \mathbb{R}^d$ pushes P forward to Q. Moreover, if P and Q have finite second moments, then $\nabla\phi$ is the unique (up to sets of measure 0) solution to the Monge problem* (1)*.*

In the rest of the paper we will assume that all measures have finite second moments.

2.1.1 Entropy-regularized Optimal Transport

Towards developing the notion of sRE, we first state a relaxation of the *Monge problem*, where instead of a map, one seeks an optimal "coupling" π between a source distribution P and a target distribution Q. The *Kantorovich relaxation* (Kantorovich, 1942; Santambrogio, 2015) solves

$$\min_{\pi \in \Pi(P,Q)} \int \frac{1}{2}\|x - y\|^2 d\pi(x, y), \tag{2}$$

where $\Pi(P, Q)$ is the set of joint probability measures with marginals P and Q. When a solution to the Monge problem (1) exists, then the solution to Kantorovich relaxation coincides with it in the sense that the optimal plan is concentrated on $\{(x, T(x)) : x \in \text{supp}(P)\}$ (Santambrogio, 2015).

The statistic we propose relies on an *entropy-regularized* version of (2). For $\varepsilon > 0$, the primal formulation of the entropy-regularized optimal transport is given by:

$$\min_{\pi \in \Pi(P,Q)} \int \frac{1}{2}\|x - y\|^2 \mathrm{d}\pi(x, y) + \varepsilon \mathrm{KL}(\pi \mid\mid P \otimes Q), \tag{3}$$

where

$$\mathrm{KL}(\pi | P \otimes Q) \triangleq \int \log\left(\frac{d\pi(x, y)}{dP(x)dQ(y)}\right) d\pi(x, y).$$

This problem has been extensively studied both for its theoretical properties, as well as for the efficient algorithms that are used to solve it (see (Cuturi, 2013; Genevay et al., 2016; Peyré et al., 2019) and references therein). Importantly, (3) admits the following dual formulation, a derivation of which may be found in (Genevay, 2019; Peyré et al., 2019):

$$\max_{f,g} \int f(x)dP(x) + \int g(y)dQ(y) - \varepsilon \iint \exp\left[\frac{1}{\varepsilon}\left(f(x) + g(y) - \frac{1}{2}\|x - y\|^2\right)\right] dP(x)dQ(y) + \varepsilon, \tag{4}$$

where the maximization is over the pairs $f \in L^1(P), g \in L^1(Q)$. The optimal entropic potentials for ε are the pair of functions $(f_\varepsilon, g_\varepsilon)$ which achieve the maximum in (4). Furthermore, there is an optimality relation between (3) and (4) given by:

$$d\pi_\varepsilon(x, y) = \exp\left[\frac{1}{\varepsilon}\left(f_\varepsilon(x) + g_\varepsilon(y) - \frac{1}{2}\|x - y\|^2\right)\right] dP(x)dQ(y), \tag{5}$$

where π_ε denotes the solution to (3). We emphasize the fact that π_ε is not a map, but a diffused coupling, and that the degree of diffusion depends on the entropic regularization parameter ε (Peyré et al., 2019). In the finite sample setting, entropic regularization of optimal transport significantly reduces computational complexity (Cuturi, 2013) and also yields a differentiable loss function (Schmitz et al., 2018).

2.2 Rank and Quantile Maps

Let $P \in \mathcal{P}(\mathbb{R}^d)$ and consider a random variable $X \sim P$. When $d = 1$, the rank map, or cumulative distribution function, is $F_X(t) = \mathbb{P}\{X \leq t\}$, and the quantile map is its generalized inverse, $F_X^{-1}(p) = \inf\{x \in \mathbb{R} : p \leq F_X(x)\}$. The rank and quantile maps are always monotonic increasing and are continuous when P has a density. When F_X is continuous, one can show that the random variable $F_X(X)$ is distributed according to $\text{Unif}([0,1])$. Similarly, when the quantile map is continuous, $F_X^{-1}(U)$ is distributed according to X where $U \sim \text{Unif}([0,1])$.

The key insight in using the theory of optimal transport to define multivariate rank and quantile maps comes from noticing that in one dimension, the optimal map between two measures, P and Q, is given by $T = F_Q^{-1} \circ F_P$, where F_P is the rank map of P and F_Q^{-1} is the quantile map of Q (Chernozhukov et al., 2017). When $Q = \text{Unif}([0,1])$, one has $F_Q^{-1} = \text{Id}$ and therefore $F_Q^{-1} \circ F_P = \text{Id} \circ F_P = F_P$. By the push-forward constraint we know that $F_P \# P = \text{Unif}([0,1])$ which is just one way of observing that $F_P(X) \sim \text{Unif}([0,1])$. The main intuition is that the rank map is exactly the optimal map from P to $\text{Unif}([0,1])$. Analogously, with Theorem 1 ensuring the existence of a unique optimal T that is monotone (being a gradient of a convex function), (Deb and Sen, 2023) generalizes the notion of rank and quantile maps to dimensions $d \geq 2$ as optimal transport maps to and from $\text{Unif}([0,1]^d)$.

Definition 2 (Definition 2.1 (Deb and Sen, 2023)). *Let $P \in \mathcal{P}_{ac}(\mathbb{R}^d)$ and let $Q = Unif([0,1]^d)$. The multivariate rank and quantile maps for P are defined as $\mathtt{R} = \nabla\phi$ and $\mathtt{Q} = \nabla\phi^*$, respectively, where ϕ is the strictly convex function as in Theorem 1 such that $\nabla\phi$ optimally transports P to Q. Here ϕ^* denotes the standard convex conjugate[2] of the convex function ϕ.*

With the high-dimensional analogue of the rank defined, we can now state the candidate GoF statistics proposed in (Deb and Sen, 2023).

Definition 3 (Definition 3.3 (Deb and Sen, 2023)). *Let $P_X, P_Y \in \mathcal{P}_{ac}(\mathbb{R}^d)$ and let $X, X' \overset{i.i.d.}{\sim} P_X$ and $Y, Y' \overset{i.i.d.}{\sim} P_Y$. Let $P_\lambda = \lambda P_X + (1-\lambda)P_Y$ denote the mixture distribution for any $\lambda \in (0,1)$ and let $\mathtt{R}_\lambda$ be the multivariate rank map of P_λ as in Definition 2. The (population) rank energy (RE) is defined as:*

$$\begin{aligned}\mathtt{RE}_\lambda(P_X, P_Y)^2 &\triangleq C_d \int_{\mathcal{S}^{d-1}} \int_{\mathbb{R}} \Big(\mathbb{P}\big(a^\top \mathtt{R}_\lambda(X) \leq t\big) - \mathbb{P}\big(a^\top \mathtt{R}_\lambda(Y) \leq t\big)\Big)^2 \, dt d\kappa(a) \\ &= 2\mathbb{E}||\mathtt{R}_\lambda(X) - \mathtt{R}_\lambda(Y)|| - \mathbb{E}||\mathtt{R}_\lambda(X) - \mathtt{R}_\lambda(X')|| - \mathbb{E}||\mathtt{R}_\lambda(Y) - \mathtt{R}_\lambda(Y')||, \quad (6)\end{aligned}$$

2. For any proper function $f : \mathbb{R}^d \to \mathbb{R}$, the convex conjugate $f^* : \mathbb{R}^d \to \mathbb{R}$ is defined as: $f^*(y) \triangleq \sup_y \langle x, y\rangle - f(y)$ for all $y \in \mathbb{R}^d$.

where $\mathcal{S}^{d-1} \triangleq \{x \in \mathbb{R}^d : \|x\| = 1\}$ *is the unit sphere in* $\mathbb{R}^d$, $\kappa(\cdot)$ *is the uniform measure on* $\mathcal{S}^{d-1}$, *and* $C_d = (2\Gamma(d/2))^{-1}\sqrt{\pi}(d-1)\Gamma\big((d-1)/2\big)$ *is an appropriate normalizing constant.*

Note that $\texttt{RE}^2_\lambda$ closely resembles the definition of energy distance (Székely and Rizzo, 2013),

$$\mathrm{En}(P_X, P_Y)^2 \triangleq C_d \int_{\mathcal{S}^{d-1}} \int_{\mathbb{R}} \Big(\mathbb{P}\big(a^\top X \le t\big) - \mathbb{P}\big(a^\top Y \le t\big)\Big)^2 dt d\kappa(a),$$

which is a widely used GoF statistic for two-sample testing. This definition is motivated by the continuity and uniqueness of characteristic functions. Namely, $P_X = P_Y$ if and only if $a^\top X \overset{d}{=} a^\top Y$ (that is, equality in distribution) for κ-almost everywhere a; indeed, the integration over the sphere aggregates the discrepancy in characteristic functions in every direction. For a discussion showing the equivalence of the formulation in (6) using integrals over $\mathcal{S}^{d-1}$ and the formulation in terms of expectations of $\texttt{R}_\lambda$, see (Baringhaus and Franz, 2004). One advantage of $\texttt{RE}^2_\lambda$ over the energy distance is that $\texttt{RE}^2_\lambda$ is distribution-free under the null for all sample sizes (Deb and Sen, 2023).

For the RE, the natural choice for λ is $1/2$. This extra parameter is made use of when one only has access to a finite set of samples as is discussed in the next section.

One can generalize the RE by replacing the pairwise distance with a kernel function (Phillips and Venkatasubramanian, 2011).

Definition 4. *Let* $P_X, P_Y \in \mathcal{P}_{ac}(\mathbb{R}^d)$ *and let* $X, X' \overset{i.i.d.}{\sim} P_X$ *and* $Y, Y' \overset{i.i.d.}{\sim} P_Y$. *Let* $P_\lambda = \lambda P_X + (1-\lambda)P_Y$ *denote the mixture distribution for any* $\lambda \in (0,1)$ *and let* $\texttt{R}_\lambda$ *be the multivariate rank map of* P_λ *as in Definition 2. Let* $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *a characteristic kernel*[3]. *The (population) rank maximum mean discrepancy (RMMD) is defined as:*

$$\texttt{RMMD}_\lambda(P_X, P_Y)^2 \triangleq \mathbb{E}[k(\texttt{R}_\lambda(X), \texttt{R}_\lambda(X'))] + \mathbb{E}[k(\texttt{R}_\lambda(Y), \texttt{R}_\lambda(Y'))] - 2\mathbb{E}[k(\texttt{R}_\lambda(X), \texttt{R}_\lambda(Y))]. \quad (7)$$

Note that $\texttt{RMMD}_\lambda(P_X, P_Y)^2$ closely follows the definition of *maximum mean discrepancy (MMD)* (Gretton et al., 2012) which is a widely used statistic in the framework of two-sample testing: $\texttt{MMD}(P_X, P_Y)^2 \triangleq \mathbb{E}[k(X, X')] + \mathbb{E}[k(Y, Y')] - 2\mathbb{E}[k(X, Y)]$.

One can view $\texttt{RE}^2_\lambda$ and $\texttt{RMMD}^2_\lambda$ as the "rank-transformed" energy distance and MMD, respectively, in the sense that they are the energy distance and MMD of the samples after being transformed by the rank map.

In practice, the $\texttt{RE}^2_\lambda$ and $\texttt{RMMD}^2_\lambda$ must be estimated from samples, a procedure outlined in the next subsection.

2.2.1 Sample Rank Map and RE

Given i.i.d. samples $X_1, \ldots, X_m \sim P$, the empirical measure is defined as $P^m = \frac{1}{m}\sum_{i=1}^m \delta_{X_i}$ where δ_{X_i} is a Dirac distribution placed at X_i. Given the empirical measures P^m and Q^m we can can define the optimal transport map between them as follows:

Definition 5. *Let* $P^m = \frac{1}{m}\sum_{i=1}^m \delta_{X_i}$, $Q^m = \sum_{j=1}^m \delta_{Y_j}$ *be empirical measures of* P *and* Q *respectively. The plug-in estimate* T^m *of the transport map* T *between* P *and* Q *is the*

3. A kernel k is said to be characteristic if the map $P \to \int_{\mathcal{X}} k(\cdot, x)dP(x)$ is injective, where P is a measure defined on the topological space $\mathcal{X}$.

solution to

$$T^m \triangleq \arg\min_T \int \frac{1}{2} \|x - T(x)\|^2 \, dP^m(x), \quad \textit{subject to } T_{\#}P^m = Q^m. \tag{8}$$

The minimization problem defining T^m can be converted to a standard linear program and solved either by tailored methods or general linear program solvers (Peyré et al., 2019). The plug-in estimate of the transport map can be specialized to obtain the sample rank map.

Definition 6. *The sample rank map $\mathtt{R}^m$ for a measure P is the plug-in estimate of the transport map from P to $Q = Unif([0,1]^d)$.*

Even though we have exact knowledge of the target measure $Q = \text{Unif}([0,1]^d)$, it is still desirable to approximate Q using precisely m samples. This is because it ensures that there is a proper map $\mathtt{R}^m$ defined on the sample points $X_1, ..., X_m$. If instead we consider Q or Q^n with $n \neq m$ there may not exist a map transporting P^m to Q. It will still be possible to come up with an optimal coupling in these cases, however further considerations will be required to convert that coupling into a statistic.

In practice, one may generate the samples from $\text{Unif}([0,1]^d)$ using a pseudo-random sequence of points. In (Deb and Sen, 2023), Halton sequences (Hofer, 2009) are used and we do the same in our experiments. Nevertheless, any sequence which weakly converges to $\text{Unif}([0,1]^d)$ can be used.

We can now define the sample RE and sample RMMD.

Definition 7. *Let $X_1, \ldots, X_m \overset{i.i.d.}{\sim} P_X$ and $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} P_Y$, define P^{m+n} as*

$$P^{m+n} = \frac{1}{m+n}\left(\sum_{i=1}^m \delta_{X_i} + \sum_{j=1}^n \delta_{Y_j}\right),$$

the empirical mixture of the two sets of samples. Let $Q^{m+n} = \frac{1}{m+n}\sum_{i=1}^{n+m} \delta_{U_i}$ where $U_i \sim Unif([0,1]^d)$. Let $\mathtt{R}^{m+n}$ be the sample rank map obtained from the two empirical measures. The sample RE is given by:

$$\begin{aligned}\mathtt{RE}_{m,n}(P_X, P_Y)^2 \triangleq \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n \|\mathtt{R}^{m+n}(X_i) - \mathtt{R}^{m+n}(Y_j)\| - \frac{1}{m^2}\sum_{i,j=1}^m \|\mathtt{R}^{m+n}(X_i) - \mathtt{R}^{m+n}(X_j)\| \\ - \frac{1}{n^2}\sum_{i,j=1}^n \|\mathtt{R}^{m+n}(Y_i) - \mathtt{R}^{m+n}(Y_j)\|. \end{aligned} \tag{9}$$

Definition 8. *Consider the same setting as Definition 7, and let k be a characteristic kernel. The sample RMMD is given by:*

$$\begin{aligned}\mathtt{RMMD}_{m,n}(P_X, P_Y)^2 \triangleq \frac{1}{m^2}\sum_{i,j=1}^m k(\mathtt{R}^{m+n}(X_i), \mathtt{R}^{m+n}(X_j)) + \frac{1}{n^2}\sum_{i,j=1}^n k(\mathtt{R}^{m+n}(Y_i), \mathtt{R}^{m+n}(Y_j)) \\ - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k(\mathtt{R}^{m+n}(X_i), \mathtt{R}^{m+n}(Y_j)). \end{aligned} \tag{10}$$

Since the sample RE and sample RMMD are transporting from a mixture with $m/(m+n)$ mass on samples from P_X and $n/(m+n)$ mass on samples from P_Y, in the limit as $m+n \to \infty$, and $m/(m+n) \to \lambda$ one will have $\texttt{RE}_{m,n}(P_X, P_Y)^2 \to \texttt{RE}_\lambda(P_X, P_Y)^2$ (Deb and Sen, 2023). This is the reason the extra parameter λ is included in the definition of the population RE (and similarly for the population RMMD).

It is known that both $\texttt{RE}^2_{m,n}$ and $\texttt{RMMD}^2_{m,n}$ are distribution-free when $P_X = P_Y$ for any fixed sample size (Deb and Sen, 2023). This follows from the fact that when $P_X = P_Y$ one has

$$\left(\texttt{R}^{m+n}(X_1), ..., \texttt{R}^{m+n}(X_m), \texttt{R}^{m+n}(Y_1), ..., \texttt{R}^{m+n}(Y_n)\right) \sim \text{Unif}([0,1]^d)^{m+n},$$

i.e., the image of the samples has the same distribution as $m+n$ independent samples from the uniform distribution on the hypercube (this is a consequence of Proposition 2.2 in (Deb and Sen, 2023)). As a result, when $P_X = P_Y$ the distributions of $\texttt{RE}_{m,n}$ and $\texttt{RMMD}_{m,n}$ are precisely the same as the energy distance and MMD, respectively, when drawing m and n samples from $\text{Unif}([0,1]^d)$. In particular, this implies $\texttt{RE}_{m,n}$ and $\texttt{RMMD}_{m,n}$ are distribution-free under the null $P_X = P_Y$.

We note that other simpler OT-based statistics are not distribution-free when $P_X = P_Y$. These include the OT distance between P_X^m and P_Y^m as well as the L^2 difference between the maps from a fixed measure Q^m to P_X^m and P_Y^m (Ramdas et al., 2017).

2.2.2 Practical Issues with RE

While the RE enjoys certain desirable properties, it also suffers from important drawbacks. We focus on two below.

Complexity, sample and computational: In practice, to compute RE and RMMD one needs to solve the discrete version of the Monge problem. Given n samples, solving this problem exactly using typical methods requires $O(n^3 \log n)$ computations (Peyré et al., 2019). To make matters worse, when the samples are in $\mathbb{R}^d$, approximating $\texttt{R}_\lambda$ has a large sample complexity in the sense that when $\lambda = m/(m+n)$ and in the absence of further assumptions, one only has:

$$\frac{1}{m+n}\mathbb{E}\left[\sum_{i=1}^m \|\texttt{R}^{m+n}(X_i) - \texttt{R}_\lambda(X_i)\| + \sum_{j=1}^n \|\texttt{R}^{m+n}(Y_j) - \texttt{R}_\lambda(Y_j)\|\right] \lesssim (m+n)^{-1/d}.$$

Unfortunately, this rate which depends exponentially on d can be tight (Dudley, 1969), implying the need for extremely large sample sizes when working in high dimensions to get a faithful estimate of the map $\texttt{R}_\lambda$ which is required to estimate RE and RMMD. Together, these observations mean that RE and RMMD suffer from a statistical-computational bottleneck that precludes their use in high dimension: one needs n large to get a good estimate, but computing the estimate for large n is computationally infeasible.

Gradient issues: In a similar vein as the MMD (Li et al., 2015), energy distance (Bellemare et al., 2017), and Wasserstein-1 distances (Arjovsky and Bottou, 2017), we are interested in utilizing RE or RMMD as a loss function for learning generative models. For example, if $X_1, ..., X_m$ are real samples and $Y_1, ..., Y_n$ are from a standard model (e.g.,

$Y_i \sim N(0, I_d)$) one may seek to solve[4]

$$\min_{\theta} \ \texttt{RE}_{m,n}(\{X_i\}_{i=1}^m, \{T_\theta(Y_j)\}_{j=1}^n)^2, \tag{11}$$

where $\texttt{RE}_{m,n}^2$ is the (squared) sample RE (defined in (9)) and $T_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is a learnable transformation parameterized by θ. In this setting, a small statistic indicates that T_θ is successfully transforming the samples $\{Y_j\}_{j=1}^n$ in such a way that they are difficult to distinguish from the $\{X_i\}_{i=1}^m$. To obtain θ^*, a (near) minimizer of (11), many standard procedures start at a random θ_0 and then use a gradient based procedure which requires access to $\nabla_\theta \texttt{RE}_{m,n}(\{X_i\}_{i=1}^m, \{T_\theta(Y_i)\}_{i=1}^n)^2$ to obtain a sequence of improving θ_i (for example, $\theta_{i+1} = \theta_i - \eta \nabla_{\theta_i} \texttt{RE}_{m,n}(\{X_i\}_{i=1}^m, \{T_\theta(Y_j)\}_{j=1}^n)^2$ for some learning rate $\eta > 0$) (Bottou, 2012; Boyd and Vandenberghe, 2004).

This gradient descent approach will not work for the RE. First, one needs to rely on special methods to back-propagate through the construction of the rank map which is the argmin of a convex optimization problem (one can rely on so-called convex optimization layers (Agrawal et al., 2019), for example). Second, when $n = m$, the gradient will either be undefined or zero (Blondel et al., 2020). This is because either the optimal transport map does not change, and as a result the RE does not change, or it experiences a jump to a new transport map which is not differentiable; see Appendix A for a precise description. In the absence of a non-zero gradient, the methods above will not be able to incrementally improve the setting of θ, or even choose a direction in the parameter space to search along. In practice the first-order methods above are by far the most popular and are often the only feasible methods when θ is extremely high-dimensional (e.g., represents the weights of a deep neural network). The absence of a proper gradient severely restricts the utility of the RE in broader contexts.

These two drawbacks are not isolated to the RE but are present whenever an empirical rank map is involved. This also limits the utility of the RMMD. To alleviate these drawbacks and to enable the use of these rank-based GoF statistic for learning generative models, in the next section we propose *soft rank energy (sRE)* and *soft rank maximum mean discrepancy (sRMMD)*.

3. Soft Rank and sRE

Our proposed statistics rely on the notion of entropic map derived from entropic regularization of the optimal transport problem. Based on the primal-dual relationship in equation (5), we note the following definition of the entropic map.

Definition 9 (Entropic map (Pooladian and Niles-Weed, 2021)). *Given an optimal entropic plan π_ε or the optimal entropic potentials $(f_\varepsilon, g_\varepsilon)$ between P and Q, the* entropic map *is defined by:*

$$T_\varepsilon(x) \triangleq \int \pi_\varepsilon(y|x) = \mathbb{E}_{\pi_\varepsilon}[Y|X=x] = \frac{\displaystyle\int y \exp\left(\frac{1}{\varepsilon}\left(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2\right)\right) dQ(y)}{\displaystyle\int \exp\left(\frac{1}{\varepsilon}\left(g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2\right)\right) dQ(y)}, \tag{12}$$

4. Technically this should be written $\texttt{RE}_{m,n}(P_X, (T_\theta)_\# P_Y)^2$, but for clarity we avoid this notation here.

where $\pi_\varepsilon(y|x)$ denotes the conditional distribution of π_ε.

In (Pooladian and Niles-Weed, 2021), the latter form is used to define a sample version of the entropic map. Given samples $X_1, ..., X_n \sim P$ and $Y_1, ..., Y_n \sim Q$ and optimal entropic potentials $(f_\varepsilon^n, g_\varepsilon^n)$ solving (4) between P^n, Q^n, the sample entropic map is defined as:

$$T_\varepsilon^{n,n}(x) \triangleq \frac{\sum_{i=1}^n Y_i \exp\left(\frac{1}{\varepsilon}\left(g_\varepsilon^n(Y_i) - \frac{1}{2}\|x - Y_i\|^2\right)\right)}{\sum_{i=1}^n \exp\left(\frac{1}{\varepsilon}\left(g_\varepsilon^n(Y_i) - \frac{1}{2}\|x - Y_i\|^2\right)\right)}. \tag{13}$$

We adopt these definitions in constructing the soft rank maps.

Definition 10 (Soft Rank Map)**.** *The soft rank map $\mathtt{R}_\varepsilon$ for a measure P is the entropic map from P to $Q = Unif([0,1]^d)$ as in (12). The sample soft rank map $\mathtt{R}_\varepsilon^n$ is defined as the sample entropic map from P^n to Q^n as in (13).*

With these alternatives to the original rank map, the definition of the sRE is as follows.

Definition 11 (Soft Rank Energy)**.** *Let $P_X, P_Y \in \mathcal{P}(\mathbb{R}^d)$ and let $X, X' \overset{i.i.d.}{\sim} P_X, Y, Y' \overset{i.i.d.}{\sim} P_Y$. Let $P_\lambda = \lambda P_X + (1-\lambda)P_Y$ for $\lambda \in (0,1)$ and let $\mathtt{R}_{\lambda,\varepsilon}$ be the soft rank map of P_λ.*

(a) soft rank energy (sRE) is defined as:

$$\begin{aligned}\mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2 &\triangleq C_d \int_{\mathcal{S}^{d-1}} \int_{\mathbb{R}} \Big(\mathbb{P}\big(a^\top \mathtt{R}_{\lambda,\varepsilon}(X) \le t\big) - \mathbb{P}\big(a^\top \mathtt{R}_{\lambda,\varepsilon}(Y) \le t\big)\Big)^2 dt d\kappa(a) \\ &= 2\mathbb{E}\big\|\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(Y)\big\| - \mathbb{E}\big\|\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(X')\big\| - \mathbb{E}\big\|\mathtt{R}_{\lambda,\varepsilon}(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y')\big\|, \end{aligned} \tag{14}$$

where C_d and κ are the same as in Definition 3.

(b) Let $X_1, ..., X_m \overset{i.i.d.}{\sim} P_X, Y_1, ..., Y_n \overset{i.i.d.}{\sim} P_Y$ and let $\mathtt{R}_{\lambda,\varepsilon}^{m+n}$ be an independently estimated soft rank map using $m+n$ samples from P_λ. The sample sRE is defined as:

$$\begin{aligned}\mathtt{sRE}_{\lambda,\varepsilon}^{m,n}(P_X, P_Y)^2 \triangleq \frac{2}{mn}\sum_{i=1}^m \sum_{j=1}^n \|\mathtt{R}_{\lambda,\varepsilon}^{m+n}(X_i) - \mathtt{R}_{\lambda,\varepsilon}^{m+n}(Y_j)\| \\ - \frac{1}{m^2}\sum_{i,j=1}^m \|\mathtt{R}_{\lambda,\varepsilon}^{m+n}(X_i) - \mathtt{R}_{\lambda,\varepsilon}^{m+n}(X_j)\| - \frac{1}{n^2}\sum_{i,j=1}^n \|\mathtt{R}_{\lambda,\varepsilon}^{m+n}(Y_i) - \mathtt{R}_{\lambda,\varepsilon}^{m+n}(Y_j)\|. \end{aligned} \tag{15}$$

The equivalence of the formulation in terms of integrals over $\mathcal{S}^{d-1}$ and the formulation in terms of expectations of $\mathtt{R}_{\lambda,\varepsilon}$ is shown in Proposition 15. Comparing the definition of the RE (Definition 3) to the sRE, the only change is the use of the soft rank map instead of the rank map. In the subsequent sections we establish several properties of the soft rank map and sRE which motivate this choice both practically and theoretically.

We remark that the reason for using separate batches of samples is a technical artifact and is required only because our analysis requires independence between the samples used to compute the estimate of the sRE and those used to estimate the map. Imposing this

condition only requires a doubling of the number of samples, and in fact the choice to use $m+n$ samples to estimate the map is somewhat arbitrary, and we use this convention to make the notation and statement of the results more compact. In practice, one may even choose not to use separate batches of samples for map estimation and calculating the statistic at all, and we adopt this strategy in our experiments in Section 5.

Using a similar approach to the one taken in (7), one can also generalize sRE by using a kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ instead of pairwise Euclidean distances.

Definition 12. *Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a characteristic kernel. Let P_X and P_Y be two probability measures and let $X, X' \overset{i.i.d.}{\sim} P_X, Y, Y' \overset{i.i.d.}{\sim} P_Y$. Let $\mathtt{R}_{\lambda,\varepsilon}$ denote the soft rank map of P_λ for $\lambda \in (0,1)$.*

(a) The soft rank maximum mean discrepancy (sRMMD) is defined as:

$$\begin{aligned}\mathtt{sRMMD}_{\lambda,\varepsilon}(P_X, P_Y)^2 \triangleq \mathbb{E}[k(\mathtt{R}_{\lambda,\varepsilon}(X), \mathtt{R}_{\lambda,\varepsilon}(X'))] + \mathbb{E}[k(\mathtt{R}_{\lambda,\varepsilon}(Y), \mathtt{R}_{\lambda,\varepsilon}(Y'))] \\ - 2\mathbb{E}[k(\mathtt{R}_{\lambda,\varepsilon}(X), \mathtt{R}_{\lambda,\varepsilon}(Y))].\end{aligned}$$

(b) Let $X_1, ..., X_m \overset{i.i.d.}{\sim} P_X, Y_1, ..., Y_n \overset{i.i.d.}{\sim} P_Y$ and let $\mathtt{R}^{m+n}_{\lambda,\varepsilon}$ be an independently estimated rank map using $m+n$ samples from P_λ. The sample sRMMD is defined as:

$$\begin{aligned}\mathtt{sRMMD}^{m,n}_{\lambda,\varepsilon}(P_X, P_Y)^2 \triangleq \frac{1}{m^2}\sum_{i,j=1}^{m} k(\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i), \mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_j)) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(\mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_i), \mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_j)) \\ - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i), \mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_j)).\end{aligned}$$

4. Properties of the sRE

4.1 Estimation of the Entropic Map

As a first step for many of our results, we give a convergence rate of the sample entropic map to the population entropic map. In this work, we consider results in two regimes. The first regime is when the measure P is subgaussian and Q has bounded support. In this case we have the following:

Theorem 13. *Suppose that $Q \in \mathcal{P}(B_2^d(0,r))$ and that P is σ^2-subgaussian. Let T_ε be the entropic map from P to Q and let $T_\varepsilon^{n,n}$ be the estimated entropic map from P^n to Q^n. Then:*

$$\mathbb{E}||T_\varepsilon^{n,n} - T_\varepsilon||^2_{L^2(P)} \leq b_1(r, d, \sigma^2, \varepsilon)n^{-1/2},$$

for some function b_1 independent of n.

An exact expression for b_1 can be found in (26) in the Appendix. The factor b_1 grows exponentially with r, σ^2, d, but can be controlled by taking ε sufficiently large. The proof of Theorem 13 is deferred to Appendix B. The proof is done in essentially three steps. First, introduce a "one-sample" T_ε^n which is the entropic map from P to Q^n and use

$$\mathbb{E}||T_\varepsilon^{n,n} - T_\varepsilon||^2_{L^2(P)} \leq 2\mathbb{E}||T_\varepsilon^{n,n} - T_\varepsilon^n||^2_{L^2(P)} + 2\mathbb{E}||T_\varepsilon^n - T_\varepsilon||^2_{L^2(P)}.$$

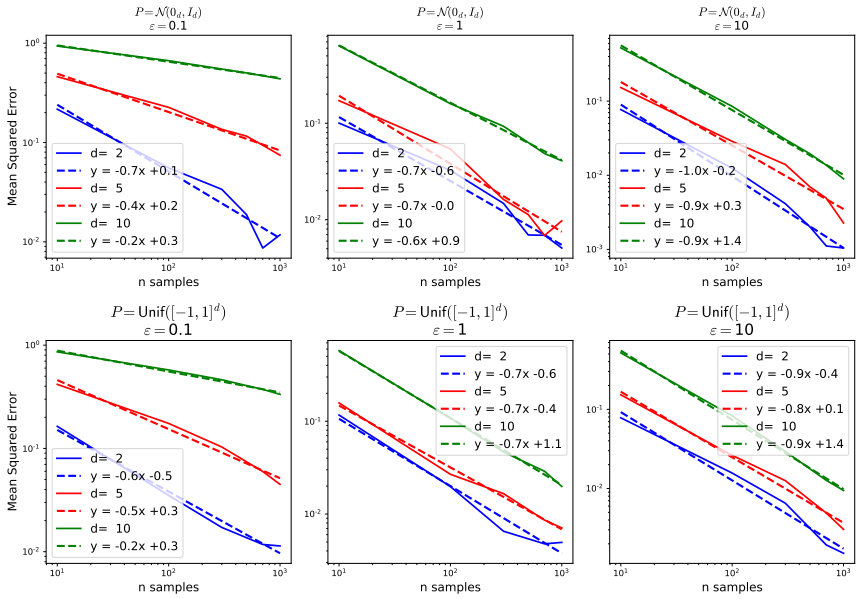


Figure 1: Convergence rate of the entropic map for different source measures and levels of regularization. The solid lines in the graph depict the mean squared error between the true map T_ε (since there is no closed form available for T_ε, we have used $n = 10000$ samples to approximate the true entropic map) and the estimated map $T_\varepsilon^{n,n}$ with respect to the sample numbers, whereas the dashed lines represent the best fit and indicate the slopes. For all cases, $Q = \text{Unif}([0,1]^d)$.

The second term is the more difficult of the two to control. The second step of the proof is to build on tools developed in (Pooladian and Niles-Weed, 2021) which allow $||T_\varepsilon^n - T_\varepsilon||^2_{L^2(P)}$ to be controlled by a supremum over a suitable class of test functions. The final step is to use tools from empirical process theory in order to control the supremum. A similar but much simpler strategy is possible for the other term.

The second regime is when both P and Q have bounded support. An important result in this second regime is Theorem 14, which was shown in (Rigollet and Stromme, 2022) and provides a useful contrast with Theorem 13.

Theorem 14 ((Rigollet and Stromme, 2022), Theorem 4, adapted)**.** *Let $P, Q \in \mathcal{P}(B_2^d(0,r))$, let T_ε be the entropic map from P to Q, and let $T_\varepsilon^{n,n}$ be the estimated entropic map from P^n to Q^n. Then:*

$$\mathbb{E}||T_\varepsilon^{n,n} - T_\varepsilon||^2_{L^2(P)} \leq b_2(r,d,\varepsilon)n^{-1},$$

for some function b_2 independent of n.

The analysis leading to this result requires a strong concavity property of the entropic dual problem which is only present when both P and Q have bounded support. As such, this result does not appear to be directly extendable to the case where P has unbounded support.

These results stand out against results for the non-regularized transport map in that the rate is either n^{-1} or $n^{-1/2}$ and only requires the source P have bounded support or be subgaussian and the target Q to have bounded support. In contrast in (Hütter and Rigollet, 2021) the authors showed that under some technical conditions the minimax optimal rate for the unregularized transport map (up to log factors) is $n^{-\frac{2\alpha}{2\alpha-2+d}}$ where α is an assumed smoothness parameter of the optimal map. These rates can be incredibly slow when the dimension d dominates the smoothness α. This suggests that entropic regularization helps break the curse of dimensionality for OT map estimation.

In Figure 1, the rate of convergence of the entropic map is observed empirically when the source measure has bounded and unbounded supports. We observe that when $\varepsilon = 1, 10$ that the convergence rate is always at least as fast as $n^{-1/2}$, which can be seen by the slopes of the lines of best fit being -0.5 or less. When $\varepsilon = 0.1$, we see a dimension dependent rate for $d = 10$, however this does not contradict Theorems 13 and 14 because we are in the small sample regime and the constants in b_1, b_2 may still provide valid upper bounds. Indeed Theorems 13 and 14 only ensure that for large enough n the convergence rates will be at least n^{-1} or $n^{-1/2}$. Obtaining matching lower bounds which are tight in the small sample setting is an important direction for future work.

4.2 Fast Convergence of $\mathtt{sRE}^{m,n}_{\lambda,\varepsilon}(P_X, P_Y)^2$

We first establish a few preliminary facts about $\mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2$ which will be useful when combined with Theorem 13 and also demonstrate some of the important properties of the sRE as a GoF statistic.

Proposition 15. *The sRE satisfies the following properties:*

(a) Let $X, X' \overset{i.i.d.}{\sim} P_X$ *and* $Y, Y' \overset{i.i.d.}{\sim} P_Y$. *Then the sRE can also be expressed as:*

$$\begin{aligned}\mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2 = 2\mathbb{E}\big\|\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(Y)\big\| - \mathbb{E}\big\|\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(X')\big\| \\ - \mathbb{E}\big\|\mathtt{R}_{\lambda,\varepsilon}(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y')\big\|.\end{aligned}$$

(b) $\mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2 = \mathtt{sRE}_{\lambda,\varepsilon}(P_Y, P_X)^2$.

(c) $\mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2 = 0$ *if* $P_X = P_Y$.

The proof is deferred to Appendix C.1. Property (a) is particularly useful because it allows one to use an easily computed formula instead of estimating the integrals and probabilities in (14). Properties (b) and (c) make $\mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2$ a good candidate for measuring the GoF between distributions.

One can also show that under some regularity conditions that the original RE is recovered by sending ε to 0, which further motivates the notion of the sRE.

Proposition 16. *Let* $P_X, P_Y \in \mathcal{P}_{ac}([0,1]^d)$ *and* $\lambda \in (0,1)$ *be such that for* $\mathtt{R}_\lambda = \nabla\phi$ *we have* $aI \preceq \nabla\phi \preceq bI$ *for some* $0 < a, b$, *and* $\mathtt{R}_\lambda^{-1} \in \mathcal{C}^\alpha$ *for some* $\alpha \geq 2$. *Then:*

$$\lim_{\varepsilon \to 0^+} \mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2 = \mathtt{RE}_\lambda(P_X, P_Y)^2.$$

The proof appears in Appendix C.2. Implicit in the proof of Proposition 16 is a result that also characterizes the degree of approximation of sRE with RE as a function of the entropy regularization ε.

An important consequence of Theorems 13 and 14 is a matching convergence rate for the sample sRE:

Theorem 17. *Let $\lambda \in (0,1)$. If P_X, P_Y are σ^2-subgaussian for then:*

$$||\mathtt{sRE}^{m,n}_{\lambda,\varepsilon}(P_X,P_Y)^2 - \mathtt{sRE}_{\lambda,\varepsilon}(P_X,P_Y)^2||^2_{L^2} \lesssim \frac{b_1(\sqrt{d},d,\sigma^2,\varepsilon)}{\min(\lambda,1-\lambda)}(m+n)^{-1/2} + \frac{d(m+n)}{mn}.$$

If $P_X, P_Y \in \mathcal{P}(B(0,r))$ then:

$$||\mathtt{sRE}^{m,n}_{\lambda,\varepsilon}(P_X,P_Y)^2 - \mathtt{sRE}_{\lambda,\varepsilon}(P_X,P_Y)^2||^2_{L^2} \lesssim \frac{b_2(\max(r,\sqrt{d}),d,\varepsilon)}{\min(\lambda,1-\lambda)}(m+n)^{-1} + \frac{d(m+n)}{mn}.$$

The proof, which is deferred to Appendix D.1, relies on four ingredients: Theorem 13 or Theorem 14, Proposition 15 (a), several applications of the triangle and reverse triangle inequalities, as well as the Efron-Stein inequality ((Boucheron et al., 2013) Theorem 3.1). The first term in the bound can be thought of as the amount of error incurred from estimating $\mathtt{R}_{\lambda,\varepsilon}$ using $m+n$ samples, while the second term bounds the error incurred from estimating an expectation by sampling.

The factor of $\frac{1}{\min(\lambda,1-\lambda)}$ arises from the use of the bounds like

$$a+b = \frac{\lambda a}{\lambda} + \frac{(1-\lambda)b}{(1-\lambda)} \leq \frac{1}{\min(\lambda,1-\lambda)}(\lambda a + (1-\lambda)b),$$

which arise when working with P_λ. This is a natural quantity to appear because in $\mathtt{sRE}^{m,n}_{\lambda,\varepsilon}(P_X,P_Y)^2$ there is an equal weight given to both $\{X_1,...,X_m\}$ and $\{Y_1,...,Y_n\}$, even if $m \ll n$ or vice versa. In contrast, the estimate of $\mathtt{R}^{m+n}_{\lambda,\varepsilon}$ weights P_X and P_Y proportionally to λ and $(1-\lambda)$, respectively. The intuition here is that if one pays little attention to constructing a good map on the support of P_X then one should expect a large amount of error for $\mathtt{R}_\lambda(X_i)$ which the weighting scheme only amplifies when computing the statistic.

In a similar vein, one can show that the sample version of sRMMD converges quickly in expectation to the population version. Again this is a consequence of Theorem 13.

Theorem 18. *Let k be a characteristic kernel such that for all x, the function $k(x,\cdot)$ is l-Lipschitz with respect to the Euclidean norm. Let $\lambda \in (0,1)$. If P_X, P_Y are σ^2-subgaussian then:*

$$||\mathtt{sRMMD}^{m,n}_{\lambda,\varepsilon}(P_X,P_Y)^2 - \mathtt{sRMMD}_{\lambda,\varepsilon}(P_X,P_Y)^2||^2_{L^2} \lesssim \frac{l^2 b_1(\sqrt{d},d,\sigma^2,\varepsilon)}{\min(\lambda,1-\lambda)}(m+n)^{-1/2} + \frac{l^2 d(m+n)}{mn}.$$

If $P_X, P_Y \in \mathcal{P}(B(0,r))$ then:

$$||\mathtt{sRMMD}^{m,n}_{\lambda,\varepsilon}(P_X,P_Y)^2 - \mathtt{sRMMD}_{\lambda,\varepsilon}(P_X,P_Y)^2||^2_{L^2} \lesssim \frac{l^2 b_2(\max(r,\sqrt{d}),d,\varepsilon)}{\min(\lambda,1-\lambda)}(m+n)^{-1} + \frac{l^2 d(m+n)}{mn}.$$

The proof follows essentially the same arguments as the one used to prove Theorem 17, but uses the Lipschitz assumption in place of the reverse-triangle inequality. A discussion of this trick is given in Appendix D.2. We remark that many of the most popular kernels do satisfy a Lipschitz continuity condition, and practically all do when restricted to bounded domains, so this restriction is not stringent.

4.3 Utility as a Loss Function

There are many examples in practice where one will try to train a generative model to create high-dimensional data (for example images, video or, audio signals), which requires a measure of how closely an artificially generated dataset matches a real dataset. In high dimensions there is a need for test statistics that converge rapidly since otherwise one won't be able to distinguish model performance from statistical fluctuations. This is one of the main reasons (along with computational ease) that MMD (Li et al., 2015) and energy distances (Bellemare et al., 2017) have become prominent in generative modeling. These statistics both have $n^{-1/2}$ convergence rates to their population variants, which is also achieved by our proposed sRE and sRMMD. We believe this makes sRE and sRMMD strong contenders for high-dimensional generative modeling.

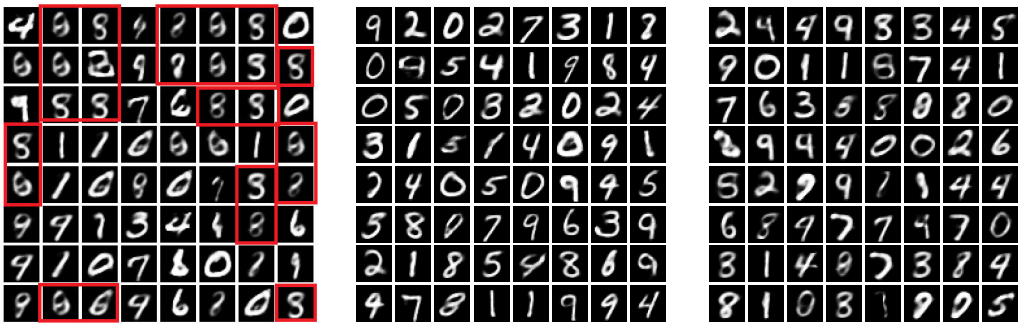
Another important note is that the computation of high fidelity estimates of the sRE and sRMMD can be easily done through the use of Sinkhorn's algorithm (Cuturi, 2013) which is based on fixed point iterations. One can perform a fixed number of iterations and then employ automatic differentiation (Paszke et al., 2017) to obtain a gradient through this method. This approach is not novel to this work and so-called "Sinkhorn layers" have become popular in the neural network literature (Adams and Zemel, 2011; Emami and Ranka, 2018; Feydy et al., 2019). This approach is implemented in our publicly available code, linked in Section 1.

5. Applications

5.1 sRE and sRMMD as the Loss Function in a Generative Model

Generative modeling is used to implicitly approximate a complex, high-dimensional distribution from a finite number of samples. When trained successfully, it allows one to draw new samples from the underlying distribution. A seminal framework to learn the generative model is the *generative adversarial network (GAN)* (Goodfellow et al., 2014) that optimizes a minimax program. A simpler approach known as a *generative moment matching network (GMMN)* (Li et al., 2015) instead minimizes the differentiable MMD (Gretton et al., 2012). In this section, we train a generative model to minimize the proposed sRE and sRMMD and illustrate their effectiveness compared to GMMN on the benchmark MNIST digits data set (LeCun et al., 1998). A brief description of the model architecture and the training procedure is given below.

Architecture: We use the same architecture used in (Li et al., 2015) which consists of (a) a generative network and (b) an auto-encoder (Kingma and Welling, 2013). The generative network consists of 3 intermediate ReLU nonlinear layers and one logistic sigmoid output layer. The auto-encoder has 4 layers, 2 for the encoder and 2 for the decoder with sigmoid nonlinearities. The auto-encoder is used to learn an almost lossless low-dimensional latent space for the high-dimensional, complex data. The generative network is used to generate samples in this low-dimensional space. The trained decoder then turns these samples into meaningful high-dimensional data. For further details, we refer the reader to (Li et al., 2015).



(a) MMD (b) sRE (c) sRMMD

Figure 2: Comparison of the MNIST digits generated via minimizing MMD, sRE, and sRMMD. Subsamples of digits generated by each method are totally random and are not handpicked, while the model architecture and training procedure are kept the same for all methods. Red boxes in (a) indicate the abundance of the same digit e.g., 8 when using MMD.

Training: We train the auto-encoder and generator network separately. First an auto-encoder is learned to produce a low-dimensional representation for the MNIST digits with latent dimension of 8 via minimizing the mean squared error. Then we train the generative network to learn to generate samples in this low-dimensional representation space via minimizing either sRE or sRMMD. Both sRE and sRMMD are computed with entropic regularizer $\varepsilon = 1$, where the soft rank maps are obtained via Sinkhorn's algorithm (Peyré et al., 2019) with a maximum of 5000 iterations. To compute sRMMD, we employ a Gaussian mixture kernel $k(x, x') = \frac{1}{6}\sum_{i=1}^{6} \exp(-\frac{\|x-x'\|^2}{2\sigma_i^2})$ with the bandwidth parameter $\sigma = (1, 2, 4, 8, 16, 32)$. Both the auto-encoder and the generator are trained on a minibatch size of 256 using the Adam optimizer (Kingma and Welling, 2019) with a learning rate of 0.001 over 100 epochs.

Method	GAN-train accuracy	GAN-test accuracy
MMD	74.5	88.7
sRE	86.6	95.6
sRMMD	84.2	94.5

Table 2: MNIST image experiments.

Result: It is apparent from Figure 2 that the generator minimizing MMD lacks in diversity (mostly producing 8's). Most of the digits are also barely recognizable which indicates that the generator performs poorly if it is trained to minimize MMD. In contrast, the generator minimizing either sRE or sRMMD produces a diverse set of digits and almost all of them are unambiguous.

To numerically evaluate the performance of the MMD, sRMMD, and sRE-based generators, we use the GAN-train and GAN-test scores (Shmelkov et al., 2018). The GAN-train score is the accuracy of a classifier trained on generated images but tested on real images whereas GAN-test is the accuracy of a classifier trained on real images but tested on generated images. We train a two-layer convnet classifier on generated images and real images and report the GAN-train and GAN-test score in Table 2. We observe a better GAN-test score for sRMMD and sRE compared to MMD. This happens because the MMD-based generator produces a lot of ambiguous digits. Also, due to the lack of diversity in generated MNIST digits using MMD, we see a poor GAN-train score, whereas both sRE and sRMMD perform well enough to capture the diversity in MNIST digits and show comparatively better GAN-train scores. This establishes both sRE and sRMMD as potentially better choices than MMD for generative modeling purposes.

Though in the current setup the generator minimizing sRMMD works well, we empirically observe that the performance (e.g., diversity, unambiguity) heavily depends on the choice of the bandwidth parameter σ and the entropic regularizer ε. In our case, we find it beneficial to use smaller σ when using a larger ε. Plots showing the dependency of the sRMMD generator on ε and σ can be found in Appendix F.1.

5.2 Generating Valid Knockoffs using sRMMD

In applications where the goal is to discover relevant features that can explain certain outcomes (e.g., metabolites or genes related to Crohn's disease (Lloyd-Price et al., 2019; Franke et al., 2010)), it is important that the set of selected features contains as few false discoveries as possible. One way to do that is to control the false discovery rate (FDR) at a prespecified level $q \in (0, 1)$. The classical setup to control FDR depends on assumptions on how the features and the outcomes are related (Benjamini and Hochberg, 1995; Gavrilov et al., 2009). A novel FDR control framework, namely Model-X knockoffs (Candès et al., 2016), provides an alternative to the traditional methods by assuming no knowledge about the association between the features and the outcomes. Given the set of explanatory random variables $X = (X^1, ..., X^d) \in \mathbb{R}^d$ and the outcome variable $Y \in \mathbb{R}$, the Model-X knockoff framework works in four steps to select relevant variables while controlling the FDR.

(a) Generate a synthetic set of features called knockoffs $\tilde{X} = (\tilde{X}^1, \ldots, \tilde{X}^d)$ which are independent of Y conditionally on X, and satisfy what is referred to as the *pairwise exchangeability condition*:

$$(X, \tilde{X})_{swap(B)} \stackrel{d}{=} (X, \tilde{X}), \forall B \subset \{1, \ldots, d\}, \tag{16}$$

where $swap(B)$ exchanges the positions of any variable X_j, $j \in B$, with its knockoff $\tilde{X}_j$.

(b) Produce a knockoff statistic $W_j = w_j([X, \tilde{X}], y)$ for $j \in \{1, \ldots, d\}$ to assess the importance of each feature. Here, $w_j(\cdot)$ is any function with the flip sign property[5].

5. A function f is said to have flip-sign property if $f(u, v) = -f(v, u)$.

(c) Find a data dependent threshold τ via,

$$\tau = \min_{t>0} \Big\{ t : \frac{1 + |\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}|} \leq q \Big\}. \tag{17}$$

(d) Select the set of variables: $\hat{\mathcal{S}} = \{j : W_j \geq \tau\}$.

Performance of the Model-X framework depends on the quality of the knockoffs, that is, to what extent they satisfy pairwise exchangeability. One way to achieve this is to approximate only the first two moments (mean and covariance) assuming that the joint distribution of X is a multivariate Gaussian. This is often called a second-order method (Candès et al., 2016). Other methods proposed in (Salimans et al., 2016; Liu and Zheng, 2018; Romano et al., 2020; Sudarshan et al., 2020) take a generative modeling approach to satisfy (18) and sample the knockoffs. A brief description of these methods is provided in Appendix E.1.

In this paper, we take a generative modeling approach where we propose to use sRMMD as the loss to satisfy the pairwise exchangeability condition (16).

5.2.1 An sRMMD-Based Knockoff Generator

We use a generative model similar to the one used in (Romano et al., 2020). The generative model has a deep neural network f_θ that takes $X \sim P_X \in \mathbb{R}^d$ and a noise vector $V \sim \mathcal{N}(0, I_d) \in \mathbb{R}^d$ as inputs and returns an approximate copy of knockoff $\tilde{X} = f_\theta(X, V) \in \mathbb{R}^d$. Here θ denotes the set of the parameters which is learned from the data. The network is fed with $\{X_i\}_{i=1}^n \in \mathbb{R}^d$ independent observations and generates $\tilde{X}_i = f_\theta(X_i, V_i)$ for $1 \leq i \leq n$. Let $\mathbf{X}, \tilde{\mathbf{X}} \in \mathbb{R}^{n \times d}$ be the matrices having these observations and their knockoffs as row vectors, respectively. To ensure that the knockoffs are of good quality (i.e., the joint distribution of $(X_i, \tilde{X}_i)$ satisfies (16) and X_i and $\tilde{X}_i$ are as different as possible, for $1 \leq i \leq n$), we minimize the following loss,

$$\ell(\mathbf{X}, \tilde{\mathbf{X}}) = \underbrace{\texttt{sRMMD}\Big[(\mathbf{X}', \tilde{\mathbf{X}}'), (\tilde{\mathbf{X}}'', \mathbf{X}'')\Big]}_{\text{full-swap}} + \underbrace{\texttt{sRMMD}\Big[(\mathbf{X}', \tilde{\mathbf{X}}'), (\mathbf{X}'', \tilde{\mathbf{X}}'')_{\text{swap}(B)}\Big]}_{\text{partial-swap}} + \gamma \ell_{Decor}(\mathbf{X}, \tilde{\mathbf{X}}), \tag{18}$$

where $\mathbf{X}', \mathbf{X}'' \in \mathbb{R}^{n/2 \times d}$ and $\tilde{\mathbf{X}}', \tilde{\mathbf{X}}'' \in \mathbb{R}^{n/2 \times d}$ are obtained by randomly splitting $\mathbf{X}$ and $\tilde{\mathbf{X}}$ in half and B is a chosen random subset of $\{1, \ldots, d\}$, such that $j \in B$ with probability $1/2$. We adapt the idea of splitting and swapping from (Romano et al., 2020). The first two terms in (18) help to achieve pairwise exchangeability. The last term in (18) trades off power versus FDR by decorrelating the variables with the knockoffs. We adapt this loss term from (Romano et al., 2020) with hyperparameter $\gamma > 0$, which is defined as:

$$\ell_{Decor}(\mathbf{X}, \tilde{\mathbf{X}}) = \|\texttt{diag}(\Sigma_{X\tilde{X}}) - 1 + s^*_{\texttt{SDP}}(\Sigma_{XX})\|_2^2.$$

Σ_{XX} and $\Sigma_{X\tilde{X}}$ are the empirical covariance matrix of X and the empirical cross covariance matrix between X and $\tilde{X}$, respectively, and $s^*_{\texttt{SDP}}(\Sigma_{XX}) = \arg\min_{s \in [0,1]^d} \sum_{j=1}^p |1 - s_j|$ such that $2\Sigma_{XX} \succeq \texttt{diag}(s) \succeq 0$. The loss (18) is differentiable and therefore any gradient descent method can be adopted to train the generative model. Generally training is done

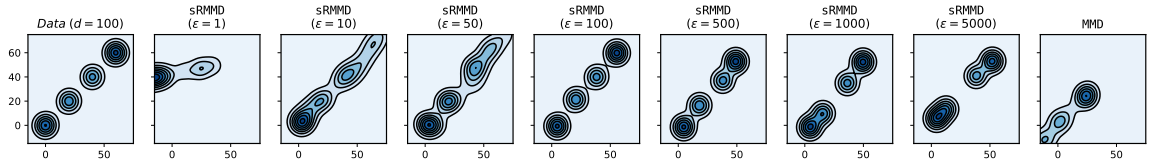


Figure 3: Visualizing two randomly selected dimensions of the original data and the generated knockoffs, where the original data are sampled from a $d = 100$ dimensional Gaussian mixture model distribution consisting of 4 different modes, $P_X = \sum_{k=1}^{4} \tau_k \mathcal{N}(\xi_k, \Sigma_k)$, with $\Sigma_k = \rho_k^{|i-j|}$, $(\rho_1, \rho_2, \rho_3, \rho_4) = (0.6, 0.4, 0.2, 0.1)$, $(\xi_1, \xi_2, \xi_3, \xi_4) = (0, 20, 40, 60)$, and $(\tau_1, \tau_2, \tau_3, \tau_4) = (0.27, 0.23, 0.23, 0.27)$. When ε is small, sRMMD fails to reconstruct the original data accurately. As ε increases, sRMMD knockoffs get better and capture all four modes eventually. However, when ε is extremely large, a performance degradation does occur. These results indicate the importance of selecting an appropriate value of ε for a successful sRMMD-generator. The last panel shows the performance of the MMD-based generator on the same dataset, where MMD fails to capture all four modes. These results demonstrate that sRMMD with an appropriate ε is a better choice than MMD for the generation of valid knockoffs.

in minibatches size of $m \ll n$. At each epoch of training, for each batch of size m, only one B is picked randomly to compute (18), which may prevent one to achieve pairwise exchangeability (16) to a great extent. That is why, we recommend generating $n_b > (n/m)$ batches by reshuffling the training set of size n several times at each epoch so that multiple sets of random B's are picked. Also, to generate valid knockoffs with sRMMD, choosing the right ε is crucial. As seen from Figure 3, a very small ε, or an extremely large ε fails to imitate the original data distribution which results in a poor-quality knockoff generator.

The details of the generative model and the training procedure for any fixed ε and γ are summarized in Appendix E.2 and Appendix E.3, respectively. In addition, we provide empirical justification for using sRMMD over sRE in (18) in Appendix F.2.

5.2.2 Knockoff Experiments on Synthetic Benchmarks

We compare the performance of sRMMD knockoffs with other benchmarks, namely second-order knockoffs (Candès et al., 2016), knockoffGAN (Jordon et al., 2018), deep knockoff (Romano et al., 2020), and deep direct likelihood knockoff (DDLK) (Sudarshan et al., 2020) on several synthetic (four Gaussian and non-Gaussian distributional settings adapted from (Romano et al., 2020)) and a real-world dataset. To be on an equal footing with deep knockoff, we remove the second-order term from the loss used in (Romano et al., 2020) and denote it as the MMD knockoff. For all comparisons, we use publicly available implementations of the code and used their recommended configurations and hyperparameters (if available).

For each distributional setting, we train the knockoff generator on a set of $n = 2000$ samples each of dimension $d = 100$. We compute sRMMD using a Gaussian mixture kernel

$k(x, x') = \frac{1}{8}\sum_{i=1}^{8} \exp(-\|x - x'\|/(2\sigma_i^2))_2^2$ with $\sigma = (1, 2, 4, 8, 16, 32, 64, 128)$, where the soft ranks are obtained via Sinkhorn's algorithm (with a maximum of 5000 iterations). We update the network via minimizing (18) using stochastic gradient descent with momentum (Bottou, 2012). The minibatch size, learning rate and the number of epochs is set to 500, 0.01, and 100, respectively. The training procedure is detailed in Algorithm E.3.

Below, we briefly describe each distributional setting and other optimal hyperparameters e.g., ε, γ for training.

(a) **Multivariate Gaussian AR1:** An autoregressive model of order one in which $X \sim \mathcal{N}(0, \Sigma)$, $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.5$. We set the decorrelation penalty $\gamma = 1$, entropic regularizer $\varepsilon = 100$.

(b) **Gaussian Mixture Model (GMM)**: $X \sim \sum_{k=1}^{4} \tau_k \mathcal{N}(0, \Sigma_k)$, where the covariance matrix is $(\Sigma_k)_{ij} = \rho_k^{|i-j|}$ for $k = 1, \ldots, 4$. $(\rho_1, \rho_2, \rho_3, \rho_4) = (0.6, 0.4, 0.2, 0.1)$, $(\tau_1, \tau_2, \tau_3, \tau_4) = (0.27, 0.23, 0.23, 0.27)$. We set $\gamma = 1$, and $\varepsilon = 100$.

(c) **Multivariate Student's t-Distribution**: A heavy-tailed distribution with zero mean and $\nu = 3$ degrees of freedom, such that $X = \sqrt{\frac{(\nu-2)}{\nu}}\frac{Z}{\sqrt{\Gamma}}$, where $Z \sim \mathcal{N}(0, \Sigma)$ with Σ as in (a) and Γ is independently drawn from a Gamma distribution with shape and rate parameters both equal to $\nu/2$. We set $\gamma = 1$ and $\varepsilon = 100$.

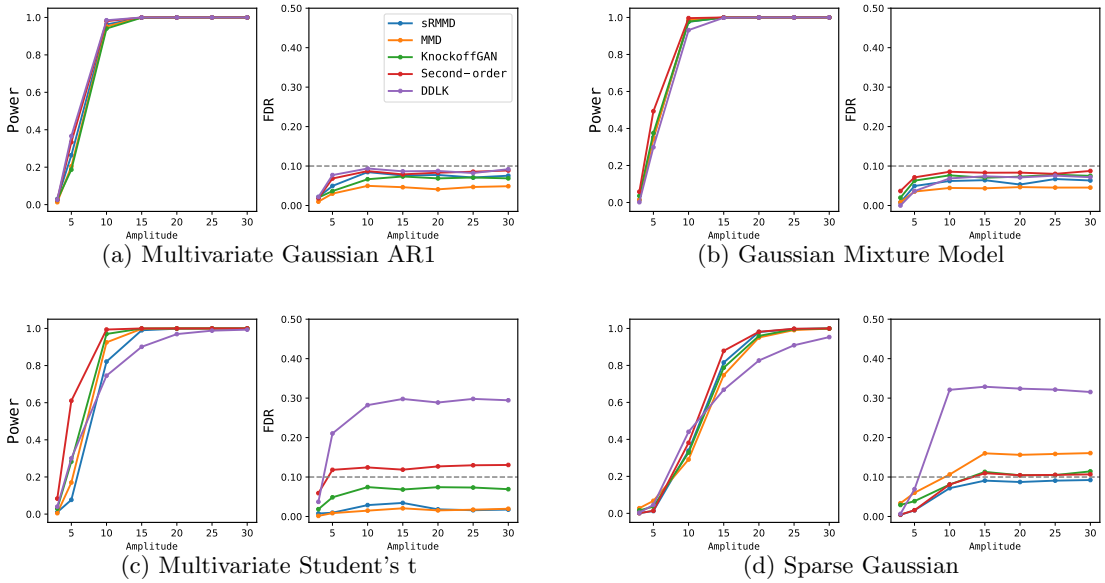
(d) **Sparse Gaussian**: Given $W \sim \mathcal{N}(0, 1)$ and a random subset $A \in \{1, \ldots, p\}$ of size $|A| = L$, we set $X_j = \sqrt{\frac{\binom{L}{p}}{\binom{L-1}{p-1}}} \cdot \begin{cases} W, & \text{if } j \in A, \\ 0, & \text{otherwise.} \end{cases}$ We set $L = 30$, $\gamma = 0.1$ and $\varepsilon = 100$.

After training, we draw $m_t = 200$ new i.i.d. samples as the test set and simulate the outcome as $\boldsymbol{y} = \mathbf{X}_t\beta + \boldsymbol{z}$, where $\mathbf{X}_t \in \mathbb{R}^{m_t \times d}$, $d = 100$, $\boldsymbol{y} \in \mathbb{R}^{m_t}$, $\boldsymbol{z} \sim \mathcal{N}(0, I)$, and $\beta \in \mathbb{R}^d$ is the coefficient vector. The vector β is all zeros except randomly chosen 20 entries, each having an amplitude equal to $\upsilon/\sqrt{m_t}$, where υ is the amplitude parameter. Then we generate the knockoff matrix $\tilde{\mathbf{X}}_t$ and perform LASSO regression (Friedman et al., 2010) on $[\mathbf{X}_t, \tilde{\mathbf{X}}_t] \in \mathbb{R}^{m_t \times 2d}$ via solving

$$\begin{bmatrix} \hat{\beta} \\ \hat{\tilde{\beta}} \end{bmatrix} = \arg\min_{(\beta, \tilde{\beta})} \frac{1}{m_t} \left\| \boldsymbol{y} - \left[\mathbf{X}_t, \tilde{\mathbf{X}}_t\right] \begin{bmatrix} \beta \\ \tilde{\beta} \end{bmatrix} \right\|_2^2 + \alpha_L \left\| \begin{matrix} \beta \\ \tilde{\beta} \end{matrix} \right\|_1, \tag{19}$$

where the $\hat{\beta} \in \mathbb{R}^d$ and $\hat{\tilde{\beta}} \in \mathbb{R}^d$ are the coefficient vectors corresponding to the original variables, and knockoff variables, respectively and α_L is the LASSO penalty. We consider LASSO regression since it works best when the true model is close to linear. We estimate the coefficient vectors using $\alpha_L = 0.01$ and take the absolute difference $W_j = |\hat{\beta}_j| - |\hat{\tilde{\beta}}_j|$ as the knockoff statistic for $1 \le j \le d$. We repeat the experiments 500 times for different values of υ and compare the power versus FDR tradeoff with the benchmarks.

Result: Figure 4 shows the FDR versus power tradeoff with respect to the amplitude parameter υ. In the case of Gaussian AR1 setting, all methods showcase similar detection power and FDR control at level $q = 0.1$ over the entire amplitude region. In the GMM setting, we observe that each method achieves similar power and controls the FDR at



(a) Multivariate Gaussian AR1 (b) Gaussian Mixture Model

(c) Multivariate Student's t (d) Sparse Gaussian

Figure 4: Average FDR and power computed over 500 independent experiments are shown on the y-axis for each synthetic benchmark. The FDR level is set to 0.1. The x-axis represents the amplitude parameter v.

$q = 0.1$ like the multivariate Gaussian setting. This can be explained by the fact that here the GMM and Gaussian settings belong to the same exact family of distributions since the sum of independent Gaussians is a Gaussian. For the heavy-tailed multivariate Student's t-distribution, sRMMD, MMD, and knockoffGAN can control the FDR at $q = 0.1$, however, the second-order method fails to control the FDR, likely because it assumes that the underlying distribution is multivariate Gaussian. DDLK also fails to control the FDR in this case. For the sparse Gaussian setting, sRMMD knockoffs showcase the best FDR versus power tradeoff among all methods over the entire amplitude region. Second-order, and knockoffGAN methods also achieve similar power and control the FDR at $q = 0.1$. In contrast, MMD and DDLK knockoffs fail to control the FDR at $q = 0.1$.

We do acknowledge the fact that in all settings knockoffGAN also achieves high power and controls the FDR at $q = 0.1$. KnockoffGAN is a complex generative adversarial architecture that requires training of four interconnected neural networks. Instead of minimizing a GoF statistic (e.g., MMD or sRMMD), knockoffGAN minimizes the binary cross-entropy loss in order to satisfy the pairwise exchangeability condition. In addition, knockoffGAN uses mutual information loss to make the variables and their knockoffs as independent as possible which is a much stronger notion than decorrelation. In contrast to knockoffGAN, our method is simple and easy to implement, yet achieves comparable FDR versus power tradeoff.

5.2.3 Application to a Real Metabolomics Dataset

We apply the proposed knockoff filter to a publicly available metabolomics dataset in order to discover important biomarkers with FDR guarantees. We use a study titled *Longitudinal Metabolomics of the Human Microbiome in Inflammatory Bowel Disease* (Lloyd-Price et al., 2019) which is available at the Metabolomics Workbench through the National Metabolomics Data Repository (NMDR) website `https://www.metabolomicsworkbench.org/` under the project DOI: 10.21228/M82T15 and sponsored by the Common Fund of the NIH. The study is related inflammatory bowel disease (IBD) and conditions including ulcerative colitis (UC) and Crohn's disease (CD) and seeks to identify important metabolites (biological products produced as intermediates during metabolism) associated with these diseases. We use the *C18 Reverse-Phase Negative Mode* dataset which was collected under this study. The dataset contains 546 samples, each having an average of 91 metabolites. Each sample belongs to one of the three classes (UC, CD, and non-IBD) and assigns the response y to one of $\{0, 1, 2\}$ to reflect this. We preprocess the dataset in three steps: (i) removing the metabolites that have more than 20% missing values which retains only 80 metabolites out of 91, (ii) applying k-nearest neighbor (KNN) missing value imputation technique to fill out the existing missing values, and (ii) standardizing the features by removing the mean and scaling to unit variance. For this dataset, we use the same generative architecture described in Section 5.2.1. We choose entropic regularizer $\varepsilon = 50$ and kernel bandwidth $\sigma = (1, 2, 4, 8, 16, 32, 64, 128)$ to compute sRMMD. We pick $\gamma = 1$. We train the generator on a minibatch of 250 samples according to Algorithm E.3.

After training, we generate the knockoffs, and apply the random forest (RF) classifier (Trainor et al., 2017) to produce knockoff statistics. The two RF parameters—the number of features that are randomly selected at each node and the number of trees—are set to 9 (the closest integer to the recommended $\sqrt{80}$) and 500, respectively. We take the difference between the feature importance scores (Trainor et al., 2017) corresponding to the original variables and the knockoffs as the knockoff statistics. Since the generated knockoffs are random, we repeat the whole procedure 100 times and select those metabolites that appear at least 70 times out of 100 instances, setting the FDR level at $q = 0.05$. In absence of the ground truth, to qualitatively analyze the performance we cross-reference the selected metabolites with published literature. We list the selected metabolites along the references in Table 3.

Out of 80, we have found 18 metabolites to have an impact on IBD in the published literature. Though MMD knockoffs detect most of the significant metabolites, the detection percentage is very low. The second-order method performs better compared to MMD in terms of power though it misses several significant metabolites. On the other hand, sRMMD and KnockoffGAN have almost the same detection power, and outperform second-order and MMD methods.

6. Conclusion and Future Directions

In this paper, we identified some major limitations in use of the recently proposed multivariate rank-based GoF statistics based on the theory of optimal transportation, namely high sample and computational complexity in high dimensions and lack of differentiability, which limits their use in gradient-based machine learning methods. We show that using

Metabolites \ Method (N)	sRMMD (21)	Second-order (16)	MMD (27)	KnockoffGAN (20)	Reference
1.2.3.4-tetrahydro-beta-carboline-1.3-dicarboxylate	✔	✔	✔	✔	(Volkova and Ruggles, 2021)
urobilin	✔	✔	✔	✔	(Qin, 2012)
adrenate	✔	✔	✔	✔	(Lloyd-Price et al., 2019)
12.13-diHOME	✔	✔	✔	✔	(Levan et al., 2019)
salicylate	✔	✔	✔	✔	(Caprilli et al., 2009)
saccharin	✔	✔	✔	✔	(Qin, 2012)
caproate	✔	✔	✔	✔	(Lee et al., 2017)
olmesartan	✔	✔	✔	✔	(Saber et al., 2019)
phenyllactate	✔	✘	✔	✔	(Lavelle and Sokol, 2020)
taurolithocholate	✔	✘	✔	✘	(Bauset et al., 2021)
docosapentaenoate	✔	✔	✔	✔	(Solakivi et al., 2011)
docosahexaenoate	✔	✘	✔	✔	(Solakivi et al., 2011)
dodecanedioate	✔	✘	✔	✘	(Lee et al., 2017)
hydrocinnamate	✔	✔	✔	✔	(Lee et al., 2017)
eicosatrienoate	✔	✘	✔	✔	(Kuroki et al., 1997)
9.10-diHOME	✔	✔	✔	✔	(Lloyd-Price et al., 2019)
arachidonate	✘	✘	✔	✔	(Lloyd-Price et al., 2019)
myristate	✘	✘	✘	✘	(Fretland et al., 1990)
Total = 18	16	11	17	15	
Detection power (%)	76	69	63	75	

Table 3: N is the total number of selected metabolites. DDLK finds almost every metabolites as significant, therefore loses its purpose as a FDR control technique. That is why we refrain from adding it here.

entropic maps derived from entropic regularization of the optimal transportation problem alleviates these issues and leads to efficient statistics for GoF testing. Furthermore, we show that this relaxation allows the use of these GoF statistics for generative modeling in high dimensions.

One future research direction is to evaluate the effect of different distributions as the target distribution in place of $\mathrm{Unif}([0,1]^d)$ such as spherical uniform distribution (Hallin et al., 2021). It may be important to characterize the effect of this choice for the soft rank and soft rank-based statistics proposed in this paper.

A related problem concerns the dependence on ε in the entropic regularization. In particular, the *convergence rate* of the sRE and sRMMD to the RE and RMMD as $\varepsilon \to 0^+$ is the subject of ongoing work. While Proposition 16 provides a coarse convergence result, we conjecture that a more precise bound may hold and that the extra smoothness assumptions on the measures may not be required for this result to be true, and that compactness and together with weaker smoothness assumptions may suffice. Beyond analyzing the convergence rate in ε, it is important to understand the advantages of taking $\varepsilon \gg 0$, beyond the

improvements in statistical and computational complexity. For example, one may wish to understand the impact ε has on the robustness, or stability, of sRE in response to small perturbations in the distributions.

Acknowledgments

This research was sponsored by the U.S. Army DEVCOM Soldier Center, and was accomplished under Cooperative Agreement Number W911QY-19-2-0003. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army DEVCOM Soldier Center, or the U.S. Government. The U. S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

SM is supported by W911QY-19-2-0003. MW is supported by NSF CCF-1553075. JMM acknowledges support from the NSF through grants DMS-1912737, DMS-1924513. SA acknowledges support by NSF CCF-1553075, NSF DRL 1931978, NSF EEC 1937057, and AFOSR FA9550-18-1-0465. All authors acknowledge support through the Tufts TRIPODS Institute, supported by the NSF under grant CCF-1934553.

We thank the two anonymous reviewers for their helpful comments which significantly improved the manuscript.

Appendix A. Lack of Gradient Issue of RE from Section 2.2.2

The issue is to estimate the gradient of $\ell(\theta) = \mathtt{RE}_{m,n}(P_X, (T_\theta)_\# P_Y)^2$. Recalling the definition of $\mathtt{RE}_{m,n}(P_X, (T_\theta)_\# P_Y)^2$, we have:

$$\begin{aligned}
&\mathtt{RE}_{m,n}(P_X, (T_\theta)_\# P_Y)^2 \\
=&\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\|\mathtt{R}^{m+n}(X_i) - \mathtt{R}^{m+n}(T_\theta(Y_j))\| - \frac{1}{m^2}\sum_{i,j=1}^{m}\|\mathtt{R}^{m+n}(X_i) - \mathtt{R}^{m+n}(X_j)\| \\
-&\frac{1}{n^2}\sum_{i,j}^{n}\|\mathtt{R}^{m+n}(T_\theta(Y_i)) - \mathtt{R}^{m+n}(T_\theta(Y_j))\| \\
=&\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\|U_{\sigma_\theta(X_i)} - U_{\sigma_\theta(T_\theta(Y_j))}\| - \frac{1}{m^2}\sum_{i,j=1}^{m}\|U_{\sigma_\theta(X_i)} - U_{\sigma_\theta(X_j)}\| - \\
&\qquad\qquad \frac{1}{n^2}\sum_{i,j}^{n}\|U_{\sigma_\theta(T_\theta(Y_i))} - U_{\sigma_\theta(T_\theta(Y_j))}\|,
\end{aligned}$$

where σ_θ is the optimal permutation for transporting $\{X_1, ..., X_m\} \cup \{T_\theta(Y_1), ..., T_\theta(Y_n)\}$ to $\{U_1, ..., U_{m+n}\}$ where $U_i \sim \mathrm{Unif}([0,1]^d)$. From the last expression it is clear that the expression only changes value when the permutation σ_θ changes. This poses problems for the derivative of the loss functions with respect to θ. If $T_\theta(y)$ varies smoothly with y then any small enough change in θ can either leave the value of $\mathtt{RE}^{m,n}(X_m, T_\theta(Y_n))^2$ unchanged since

the permutation is unchanged, or it causes a jump in the objective when the permutation changes. In the first case the derivative is zero, and in the second it is not well-defined.

Appendix B. Proof of Theorem 13

The proof of this result is quite involved. We first review some background and notation and then build up to the result using a series of lemmas and propositions. To help with the exposition, the proofs of these steps are given at the end of the section. We also will adopt the convention that all constants $C, C_i, C_{j,r,d}$ *do not change values from line to line.* At the end of the section, we include a table of the constants, including their relations to each other or the source from which they are taken.

We first introduce the notation for the entropy-regularized optimal transport distance between P, Q:

$$S_\varepsilon(P,Q) = \inf_{\pi \in \Pi(P,Q)} \int\int \frac{1}{2}\|x-y\|^2 d\pi(x,y) + \varepsilon D_{KL}(\pi||P\otimes Q) \tag{20}$$

$$= \sup_{f\in L^1(P), g\in L^1(Q)} \int f dP + \int g dQ$$
$$-\varepsilon \int\int \exp\left(\frac{1}{\varepsilon}\left[f(x)+g(y)-\frac{1}{2}\|x-y\|^2\right]\right) dP(x)dQ(y) + \varepsilon. \tag{21}$$

The optimal coupling of (P,Q) in (20) will be denoted π_ε, and for (P,Q^n) it will be denoted π_ε^n. These are both guaranteed to exist if P, Q have finite second moment which is always the case for bounded support or subgaussian distributions. The optimal dual potentials of (P,Q) in (21) will be denoted $(f_\varepsilon, g_\varepsilon)$. For (P,Q^n) the optimal dual potentials will be denoted $(f_\varepsilon^n, g_\varepsilon^n)$. As a consequence of (5), we can always choose $f_\varepsilon, g_\varepsilon$ to satisfy:

$$\int \exp\left(\frac{1}{\varepsilon}\left[f_\varepsilon(x)+g_\varepsilon(y)-\frac{1}{2}\|x-y\|^2\right]\right) dP(x) = 1 \text{ for all } y\in\mathbb{R}^d, \tag{22}$$

$$\int \exp\left(\frac{1}{\varepsilon}\left[f_\varepsilon(x)+g_\varepsilon(y)-\frac{1}{2}\|x-y\|^2\right]\right) dQ(y) = 1 \text{ for all } x\in\mathbb{R}^d, \tag{23}$$

and we will make use of this property many times. This is because

$$\exp\left(\frac{1}{\varepsilon}\left[f_\varepsilon(x)+g_\varepsilon(y)-\frac{1}{2}\|x-y\|^2\right]\right) dP(x)$$

is the conditional density of π_ε given y while $\exp\left(\frac{1}{\varepsilon}\left[f_\varepsilon(x)+g_\varepsilon(y)-\frac{1}{2}\|x-y\|^2\right]\right) dQ(y)$ is the conditional density of π_ε given x. For further discussion see (Pooladian and Niles-Weed, 2021).

We also remark that we can add or subtract a constant c from f_ε and g_ε, that is, choose different but still optimal potentials $f = f_\varepsilon + c, g = g_\varepsilon - c$. As a result of this we can enforce that

$$\mathbb{E}_P[f_\varepsilon(X)] = \mathbb{E}_Q[g_\varepsilon(Y)] = \frac{1}{2}S_\varepsilon(P,Q), \tag{24}$$

which will be important for the proof of Proposition 31 below.

For brevity we introduce the notations:

$$c(x,y) \triangleq \frac{1}{2}\|x-y\|^2 \quad \text{and} \quad \gamma(x,y) \triangleq \exp\left(\frac{1}{\varepsilon}\left[f_\varepsilon(x)+g_\varepsilon(y)-c(x,y)\right]\right).$$

A final technical tool that we will use is an alternative definition of a subgaussian random variable. For a *scalar* random variable X we can define its subgaussian norm (Vershynin, 2018) by:

$$\|X\|_{\psi_2} \triangleq \inf\left\{t>0 \ : \ \mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq 2\right\}. \tag{25}$$

Note that for scalar random variables, that is when $d=1$, being σ^2-subgaussian implies $\|X\|_{\psi_2} \leq \sqrt{2}\sigma$, and $\|X\|_{\psi_2} \leq \sigma$ implies X is $\frac{\sigma^2}{2}$-subgaussian. This shows that the two notions are equivalent for scalar random variables up to a constant. The norm notation is much more convenient for describing the concentration of certain random variables we will use in the proofs below.

To start we borrow a result which shows that one can take the supremum in (21) over an even larger space of functions.

Proposition 19 ((Pooladian and Niles-Weed, 2021) Proposition 1)**.** *Letting π_ε denote the optimal coupling in (20) between P and Q. Then:*

$$S_\varepsilon(P,Q) = \sup_{\eta \in L^1(\pi_\varepsilon)} \int \eta(x,y)d\pi_\varepsilon(x,y) - \varepsilon\left(\iint \exp\left(\frac{1}{\varepsilon}\left[\eta(x,y)-\frac{1}{2}\|x-y\|^2\right]\right)dQ(y)dP(x)-1\right).$$

This is a very useful formula when combined with a clever restriction of the class of test functions. The following result does this by considering functions η of the form

$$\eta(x,y) = \varepsilon\chi(x,y) + f_\varepsilon(x) + g_\varepsilon(y),$$

where $\chi \in L^1(\pi_\varepsilon^n)$.

Proposition 20. *Let π_ε^n be the optimal coupling in (20) between P and Q^n and let $(f_\varepsilon, g_\varepsilon)$ be the optimal dual potentials for (P,Q). Then:*

$$\begin{aligned}&\sup_{\chi \in L^1(\pi_\varepsilon^n)} \int \chi(x,y)d\pi_\varepsilon^n(x,y) - \left(\int\int \exp(\chi(x,y))\gamma(x,y)dQ^n(y)dP(x) - 1\right)\\ \leq &\frac{1}{\varepsilon}\left(S_\varepsilon(P,Q^n) - S_\varepsilon(P,Q) + \int g_\varepsilon(y)d(Q^n-Q)(y)\right).\end{aligned}$$

For convenience we define $G \triangleq S_\varepsilon(P,Q^n) - S_\varepsilon(P,Q) + \int g_\varepsilon(y)d(Q^n-Q)(y)$. Importantly, G is a *random* scalar variable, determined by the random batch of samples observed.

Proposition 20 is useful for two reasons. The first reason is that both terms on the right hand side have been studied before and can be shown to have good convergence properties as n grows. Indeed, the first term has been analyzed in (Pooladian and Niles-Weed, 2021) and the second term is easily controlled using standard ideas in Monte Carlo integration. Therefore, the right hand side is generally easy to control. The second reason is that there

is still sufficient flexibility for choosing a test function on the left hand side. In particular we will further restrict the supremum to χ of the form

$$\chi(x,y) = h(x)^T(y - T_\varepsilon^n(x)) - a||h(x)||^2$$

for $a > 0$ and $h : \mathbb{R}^d \to \mathbb{R}^d$. The motivation for this choice is that we don't know the exact form of $T_\varepsilon^n(x) - T_\varepsilon(x)$, since T_ε^n is itself random, and instead we aim for uniform control over a function class which surely contains it, regardless of the random samples drawn. The extra variable a allows us to establish a few useful inequalities later.

Proposition 21. *Let π_ε^n be the optimal coupling in (20) between P and Q^n and let $(f_\varepsilon, g_\varepsilon)$ be the optimal dual potentials for (P, Q). Then for any $a > 0$,*

$$\begin{aligned}
&\sup_h \int (h(x)^T(y - T_\varepsilon^n(x)) - a||h(x)||^2) d\pi_\varepsilon^n(x,y) \\
&- \left(\int \int \exp\left(h(x)^T(y - T_\varepsilon^n(x)) - a||h(x)||^2\right) \gamma(x,y) dQ^n(y) dP(x) - 1 \right) \\
&\leq \frac{1}{\varepsilon} G,
\end{aligned}$$

where the supremum is over all $h : \mathbb{R}^d \to \mathbb{R}^d$ such that $h(x)^T(y - T_\varepsilon^n(x)) - a||h(x)||^2 \in L^1(\pi_\varepsilon^n)$.

Essentially the proof is just restricting the supremum in Proposition 20 to the class of test functions of the form $h(x)^T(y - T_\varepsilon^n(x)) - a||h(x)||^2 \in L^1(\pi_\varepsilon^n)$ for some h. In principle one could go directly from Proposition 19 to Proposition 21 by considering functions of the form $\eta(x,y) = \varepsilon(h(x)^T(y - T_\varepsilon^n(x)) - a||h(x)||^2) + f_\varepsilon(x) + g_\varepsilon(y)$, however the proof using this approach is incredibly cumbersome.

Importantly, we have managed to insert T_ε^n into the left hand side above. The next result marks a large amount of progress and boils down to a choice of h and a large number of simplifications.

Lemma 22. *Define $h_0 : \mathbb{R}^d \to \mathbb{R}^d$ by $h_0(x) = \frac{1}{2a}\left(T_\varepsilon^n(x) - T_\varepsilon(x)\right)$. Then in the setting of Proposition 21 we have:*

$$\frac{1}{4a}||T_\varepsilon^n - T_\varepsilon||_{L^2(P)}^2 \leq \frac{1}{\varepsilon} G + \left(\iint \exp\left(h_0(x)^T(y - T_\varepsilon(x)) - a||h_0(x)||^2\right) \gamma(x,y) dQ^n(y) dP(x) - 1 \right).$$

The next step is to swap the trailing "-1" in the bound above for a term which will make the convergence more explicit. This is done in the following lemma.

Lemma 23. *Suppose $Q \in \mathcal{P}(B_2^d(0,r))$. For $a \geq C_0 r^2$ and h_0 defined as above it holds that:*

$$\int \int \exp\left(h_0(x)^T(y - T_\varepsilon(x)) - a||h_0(x)||^2\right) \gamma(x,y) dQ(y) dP(x) \leq 1,$$

where C_0 is an absolute constant.

Combining Lemmas 22 and 23 we obtain the following.

Proposition 24. *Suppose $Q \in \mathcal{P}(B_2^d(0,r))$. For $a \geq C_0 r^2$ and h_0 defined as above we have:*

$$\frac{1}{4a}||T_\varepsilon^n - T_\varepsilon||^2_{L^2(P)} \leq \frac{1}{\varepsilon}G + \int\int \exp\left(h_0(x)^T(y - T_\varepsilon(x)) - a||h_0(x)||^2\right)\gamma(x,y)[dQ^n - dQ](y)dP(x).$$

If for a moment one ignores the randomness in h_0, the right hand side would be expected to converge to 0 at rate $n^{-1/2}$, since it would be the error in Monte Carlo integration of a single variable function. However, since the function h_0 is random one needs to work a bit harder and introduce tools from empirical process theory.

Lemma 25. *Assume P is σ^2-subgaussian and $Q \in \mathcal{P}(B_2^d(0,r))$. Let $C_1 = \max(C_0, 2)$. Then with h_0 defined as in Lemma 22 and for $a \geq C_1 r^2$ we have:*

$$\begin{aligned}&\mathbb{E}\int\int \exp\left(h_0(x)^T(y - T_\varepsilon(x)) - a||h_0(x)||^2\right)\gamma(x,y)[dQ^n - dQ](y)dP(x)\\ &\leq \frac{C_2\sqrt{d}}{\sqrt{n}}\exp\left(\frac{1}{\varepsilon}\left(d\sigma^2 + \sqrt{2d}\sigma r + r^2\right)\right)\left(\exp\left(\frac{8rd\sigma^2}{\varepsilon^2}\right) + 2\right),\end{aligned}$$

where C_2 is an absolute constant.

Having controlled the right most term in Proposition 24, we now turn our attention to controlling G. This has already been done in the literature and we state these results for completeness. The first term of G is controlled by the following result.

Proposition 26 ((Mena and Niles-Weed, 2019))**.** *Let P, Q be σ^2-subgaussian. Let $\varepsilon > 0$. Then:*

$$\mathbb{E}S_\varepsilon(P, Q^n) - S_\varepsilon(P, Q) \leq K_{d,0} \cdot \varepsilon\left(1 + \frac{\sigma^{\lceil 5d/2\rceil + 6}}{\varepsilon^{\lceil 5d/4\rceil + 3}}\right)\frac{1}{\sqrt{n}}.$$

Note that if $\sigma \geq \frac{r}{\sqrt{2d\log 2}}$ and $Q \in \mathcal{P}(B_2^d(0,r))$ then Q always satisfies the condition of Proposition 26, as can be shown by a direct calculation.

For the second term in G we have simply that:

$$\mathbb{E}\int g_\varepsilon(y)d(Q^n - Q)(y) = 0.$$

We can now state the "zero-to-one" sample theorem for the convergence of the entropic map, which follows by combining Proposition 24, Lemma 25, and Proposition 26.

Theorem 27. *Assume P is σ^2-subgaussian and $Q \in \mathcal{P}(B_2^d(0,r))$. Then:*

$$\mathbb{E}||T_\varepsilon^n - T_\varepsilon||^2_{L^2(P)} \leq \exp\left(\frac{1}{\varepsilon}\left(d\sigma^2 + \sqrt{2d}\sigma r + r^2\right)\right)\left(\exp\left(\frac{8rd\sigma^2}{\varepsilon^2}\right) + 2\right)\frac{C_3 r^2\sqrt{d}}{\sqrt{n}} + \left(1 + \frac{\sigma_0^{\lceil 5d/2\rceil + 6}}{\varepsilon^{\lceil 5d/4\rceil + 3}}\right)\frac{K_{d,1}r^2}{\sqrt{n}},$$

where $\sigma_0 = \max\left(\sigma, \frac{r}{\sqrt{2d\log 2}}\right)$.

A much easier proof is also possible for the "one-to-two" sample theorem for the convergence of the entropic map.

Theorem 28. *Let P be σ^2-subgaussian and $Q \in \mathcal{P}(B_2^d(0,r))$, let $T_\varepsilon^{n,n}$ be the entropic map from P^n to Q^n, and let T_ε^n be the entropic map from P to Q^n. Then $T_\varepsilon^{n,n}$ satisfies:*

$$\mathbb{E}\left\|T_\varepsilon^n - T_\varepsilon^{n,n}\right\|_{L^2(P)}^2 \leq r^2 K_{d,2} \cdot \left(1 + \frac{\sigma_0^{\lceil 5d/2 \rceil + 6}}{\varepsilon^{\lceil 5d/4 \rceil + 3}}\right) \frac{1}{\sqrt{n}},$$

where $\sigma_0 = \max\left(\sigma, \frac{r}{\sqrt{2d \log 2}}\right)$.

As noted above, this result automatically covers bounded support measures since all bounded support measures are σ^2-subgaussian with σ^2 controlled by the radius of the ball containing the support of Q.

Combining Theorems 27 and 28 we have our result.

Proof (Theorem 13) Applying Theorems 27 and 28, we have:

$$\begin{aligned}
\mathbb{E}||T_\varepsilon^{n,n} - T_\varepsilon||_{L^2(P)}^2 &= \mathbb{E}||(T_\varepsilon^{n,n} - T_\varepsilon^n) + (T_\varepsilon^n - T_\varepsilon)||_{L^2(P)}^2 \\
&\leq \mathbb{E}\left(||(T_\varepsilon^{n,n} - T_\varepsilon^n)||_{L^2(P)} + ||(T_\varepsilon^n - T_\varepsilon)||_{L^2(P)}\right)^2 \\
&\leq 2\mathbb{E}||(T_\varepsilon^{n,n} - T_\varepsilon^n)||_{L^2(P)}^2 + 2\mathbb{E}||(T_\varepsilon^n - T_\varepsilon)||_{L^2(P)}^2 \\
&\leq \frac{2C_3 r^2 \sqrt{d}}{\sqrt{n}} \exp\left(\frac{1}{\varepsilon}\left(d\sigma^2 + \sqrt{2d}\sigma r + r^2\right)\right)\left(\exp\left(\frac{8rd\sigma^2}{\varepsilon^2}\right) + 2\right) \\
&\qquad + 2(K_{d,1} + K_{d,2}) r^2 \left(1 + \frac{\sigma_0^{\lceil 5d/2 \rceil + 6}}{\varepsilon^{\lceil 5d/4 \rceil + 3}}\right) \frac{1}{\sqrt{n}}. \qquad (26)
\end{aligned}$$

■

B.1 Proof of Proposition 20

Proof Let $\chi \in L^1(\pi_\varepsilon^n)$ and define η by:

$$\eta(x,y) = \varepsilon\chi(x,y) + f_\varepsilon(x) + g_\varepsilon(y).$$

Note that $\eta \in L^1(\pi_\varepsilon^n)$ since $\chi, (f_\varepsilon + g_\varepsilon) \in L^1(\pi_\varepsilon^n)$. By Proposition 19 and the fact that P, Q^n are the marginals of π_ε^n it follows:

$$\begin{aligned}
S_\varepsilon(P, Q^n) &\geq \int \eta(x,y) d\pi_\varepsilon^n(x,y) - \varepsilon\left(\iint \exp\left(\frac{1}{\varepsilon}\left[\eta(x,y) - \frac{1}{2}\|x-y\|^2\right]\right) dQ^n(y) dP(x) - 1\right) \\
&= \int \left[\varepsilon\chi(x,y) + f_\varepsilon(x) + g_\varepsilon(y)\right] d\pi_\varepsilon^n(x,y) \\
&\quad - \varepsilon\left(\iint \exp\left(\frac{1}{\varepsilon}\left[\left[\varepsilon\chi(x,y) + f_\varepsilon(x) + g_\varepsilon(y)\right] - \frac{1}{2}\|x-y\|^2\right]\right) dQ^n(y) dP(x) - 1\right) \\
&= \varepsilon \int \chi(x,y) d\pi_\varepsilon^n(x,y) + \int f_\varepsilon(x) dP(x) + \int g_\varepsilon(y) dQ^n(y) \\
&\quad - \varepsilon\left(\iint \exp\left(\chi(x,y) + \frac{1}{\varepsilon}\left[f_\varepsilon(x) + g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2\right]\right) dQ^n(y) dP(x) - 1\right)
\end{aligned}$$

$$= \varepsilon \int \chi(x,y) d\pi_\varepsilon^n(x,y) + \int f_\varepsilon(x) dP(x) + \int g_\varepsilon(y) dQ^n(y)$$
$$- \varepsilon \left(\iint \exp(\chi(x,y)) \gamma(x,y) dQ^n(y) dP(x) - 1 \right).$$

Rearranging, we have:

$$\int \chi(x,y) d\pi_\varepsilon^n(x,y) - \left(\int \int \exp(\chi(x,y)) \gamma(x,y) dQ^n(y) dP(x) - 1 \right)$$
$$\leq \frac{1}{\varepsilon} S_\varepsilon(P, Q^n) - \frac{1}{\varepsilon} \left(\int f_\varepsilon(x) dP(x) + \int g_\varepsilon(y) dQ^n(y) \right). \tag{27}$$

Dropping for a moment the $\frac{1}{\varepsilon}$, we have:

$$S_\varepsilon(P, Q^n) - \left(\int f_\varepsilon(x) dP(x) + \int g_\varepsilon(y) dQ^n(y) \right)$$
$$= (S_\varepsilon(P, Q^n) - S_\varepsilon(P, Q)) + \left(S_\varepsilon(P, Q) - \int f_\varepsilon(x) dP(x) + \int g_\varepsilon(y) dQ(y) \right) + \int g_\varepsilon(y) d(Q^n - Q)(y)$$
$$= S_\varepsilon(P, Q^n) - S_\varepsilon(P, Q) + \int g_\varepsilon(y) d(Q^n - Q)(y). \tag{28}$$

In the last line we have used that by (22),

$$-\varepsilon \int \int \exp\left(\frac{1}{\varepsilon} \left[f_\varepsilon(x) + g_\varepsilon(y) - \frac{1}{2} \|x - y\|^2 \right] \right) dQ(y) dP(x) + \varepsilon = -\varepsilon \int 1 dQ(y) + \varepsilon = 0,$$

and therefore:

$$S_\varepsilon(P, Q) = \int f_\varepsilon dP + \int g_\varepsilon dQ - \varepsilon \iint \exp\left(\frac{1}{\varepsilon} \left[f_\varepsilon(x) + g_\varepsilon(y) - \frac{1}{2} \|x - y\|^2 \right] \right) dQ(y) dP(x) + \varepsilon$$
$$= \int f_\varepsilon dP + \int g_\varepsilon dQ.$$

This implies

$$S_\varepsilon(P, Q) - \int f_\varepsilon dP + \int g_\varepsilon dQ = 0.$$

Plugging (28) into (27) we obtain:

$$\int \chi(x,y) d\pi_\varepsilon^n(x,y) - \left(\int \int \exp(\chi(x,y)) \gamma(x,y) dQ^n(y) dP(x) - 1 \right)$$
$$\leq \frac{1}{\varepsilon} \left[S_\varepsilon(P, Q^n) - S_\varepsilon(P, Q) + \int g_\varepsilon(y) d(Q^n - Q)(y) \right].$$

Since χ was chosen arbitrarily in $L^1(\pi_\varepsilon^n)$, this bound holds uniformly over the class. Therefore, we have:

$$\sup_{\chi \in L^1(\pi_\varepsilon^n)} \int \chi(x,y) d\pi_\varepsilon^n(x,y) - \left(\int \int \exp(\chi(x,y)) \gamma(x,y) dQ^n(y) dP(x) - 1 \right)$$
$$\leq \frac{1}{\varepsilon} (S_\varepsilon(P, Q^n) - S_\varepsilon(P, Q)) + \frac{1}{\varepsilon} \int g_\varepsilon(y) d(Q^n - Q)(y),$$

as desired. ■

B.2 Proof of Proposition 21

Proof Define the set

$$\mathcal{H} = \left\{ H(x,y) = h(x)^T(y - T_\varepsilon^n(x)) - a||h(x)||^2 \;\middle|\; h : \mathbb{R}^d \to \mathbb{R}^d; \int \left| h(x)^T(y - T_\varepsilon^n(x)) - a||h(x)||^2 \right| d\pi_\varepsilon^n < \infty \right\}.$$

Then $\mathcal{H} \subset L^1(\pi_\varepsilon^n)$ and by Proposition 20 we have

$$\begin{aligned}
&\sup_h \int h(x)^T(y - T_\varepsilon^n(x)) - a||h(x)||^2 d\pi_\varepsilon^n(x,y) \\
&- \left(\int \int \exp\left(h(x)^T(y - T_\varepsilon^n(x)) - a||h(x)||^2 \right) \gamma(x,y) dQ^n(y) dP(x) - 1 \right) \\
&= \sup_{H \in \mathcal{H}} \int H(x,y) d\pi_\varepsilon^n(x,y) - \left(\int \int \exp(H(x,y)) \gamma(x,y) dQ^n(y) dP(x) - 1 \right) \\
&\leq \sup_{\chi \in L^1(\pi_\varepsilon^n)} \int \chi(x,y) d\pi_\varepsilon^n(x,y) - \left(\int \int \exp(\chi(x,y)) \gamma(x,y) dQ^n(y) dP(x) - 1 \right) \\
&\leq \frac{1}{\varepsilon}(S_\varepsilon(P,Q^n) - S_\varepsilon(P,Q)) + \frac{1}{\varepsilon} \int g_\varepsilon(y) d(Q^n - Q)(y) = \frac{1}{\varepsilon} G.
\end{aligned}$$

■

B.3 Proof of Lemma 22

Proof Using that the first marginal of π_ε^n is P we have

$$\int (h_0(x)^T(y - T_\varepsilon(x)) - a||h_0(x)||^2) d\pi_\varepsilon^n(x,y) = \int h_0(x)^T(y - T_\varepsilon(x)) d\pi_\varepsilon^n(x,y) - \int a||h_0(x)||^2 dP(x). \quad (29)$$

Now let us consider the two integrals separately. For the first we have the following chain:

$$\begin{aligned}
&\int h_0(x)^T(y - T_\varepsilon(x)) d\pi_\varepsilon^n(x,y) \\
&= \int \left(\frac{1}{2a}(T_\varepsilon^n(x) - T_\varepsilon(x)) \right)^T (y - T_\varepsilon(x)) d\pi_\varepsilon^n(x,y) \\
&= \frac{1}{2a} \int (T_\varepsilon^n(x) - T_\varepsilon(x))^T y d\pi_\varepsilon^n(x,y) - \frac{1}{2a} \int (T_\varepsilon^n(x) - T_\varepsilon(x))^T T_\varepsilon(x) d\pi_\varepsilon^n(x,y) \\
&= \frac{1}{2a} \int (T_\varepsilon^n(x) - T_\varepsilon(x))^T y d\pi_\varepsilon^n(x,y) - \frac{1}{2a} \int (T_\varepsilon^n(x) - T_\varepsilon(x))^T T_\varepsilon(x) dP(x) \\
&= \frac{1}{2a} \int (T_\varepsilon^n(x) - T_\varepsilon(x))^T T_\varepsilon^n(x) dP(x) - \frac{1}{2a} \int (T_\varepsilon^n(x) - T_\varepsilon(x))^T T_\varepsilon(x) dP(x) \quad (30) \\
&= \frac{1}{2a} \int (T_\varepsilon^n(x) - T_\varepsilon(x))^T (T_\varepsilon^n(x) - T_\varepsilon(x)) dP(x) \\
&= \frac{1}{2a} \int ||T_\varepsilon^n(x) - T_\varepsilon(x)||^2 dP(x) = \frac{1}{2a} ||T_\varepsilon^n - T_\varepsilon||^2_{L^2(P)}. \quad (31)
\end{aligned}$$

The third equality uses that P is the marginal of π_ε^n. To see (30), note:

$$\int (T_\varepsilon^n(x) - T_\varepsilon(x))^T y d\pi_\varepsilon^n(x,y) = \int \left[\int (T_\varepsilon^n(x) - T_\varepsilon(x))^T y d\pi_\varepsilon^n(y|x) \right] dP(x)$$

$$= \int (T_\varepsilon^n(x) - T_\varepsilon(x))^T \left[\int y d\pi_\varepsilon^n(y|x) \right] dP(x) \quad \text{(Linearity)}$$
$$= \int (T_\varepsilon^n(x) - T_\varepsilon(x))^T T_\varepsilon^n(x) dP(x). \quad \text{(Eq12)}$$

For the second integral in (29), we have:

$$\int a||h_0(x)||^2 dP(x) = \int a||\frac{1}{2a}[T_\varepsilon^n(x) - T_\varepsilon(x)]||^2 dP(x)$$
$$= \frac{1}{4a}||T_\varepsilon^n - T_\varepsilon||^2_{L^2(P)}. \quad (32)$$

Plugging (31),(32) into (29) we have

$$\int h_0(x)^T(y - T_\varepsilon(x)) - a||h(x)||^2 d\pi_\varepsilon^n(x,y) = \frac{1}{2a}||T_\varepsilon^n - T_\varepsilon||^2_{L^2(P)} - \frac{1}{4a}||T_\varepsilon^n - T_\varepsilon||^2_{L^2(P)}$$
$$= \frac{1}{4a}||T_\varepsilon^n - T_\varepsilon||^2_{L^2(P)}.$$

Now by Proposition 21 we have:

$$\begin{aligned}
\frac{1}{\varepsilon}G &\geq \sup_h \int h(x)^T(y - T_\varepsilon(x)) - a||h(x)||^2 d\pi_\varepsilon^n(x,y) \\
&\quad - \left(\int \int \exp\left(h(x)^T(y - T_\varepsilon(x)) - a||h(x)||^2\right) \gamma(x,y) dQ^n(y) dP(x) - 1 \right) \\
&\geq \int h_0(x)^T(y - T_\varepsilon(x)) - a||h_0(x)||^2 d\pi_\varepsilon^n(x,y) \\
&\quad - \left(\int \int \exp\left(h_0(x)^T(y - T_\varepsilon(x)) - a||h_0(x)||^2\right) \gamma(x,y) dQ^n(y) dP(x) - 1 \right) \\
&= \frac{1}{4a}||T_\varepsilon^n - T_\varepsilon||^2_{L^2(P)} - \left(\iint \exp\left(h_0(x)^T(y - T_\varepsilon(x)) - a||h_0(x)||^2\right) \gamma(x,y) dQ^n(y) dP(x) - 1 \right).
\end{aligned}$$

If we re-arrange the first and last inequality we have:

$$\frac{1}{4a}||T_\varepsilon^n - T_\varepsilon||^2_{L^2(P)} \leq \frac{1}{\varepsilon}G + \left(\iint \exp\left(h_0(x)^T(y - T_\varepsilon(x)) - a||h_0(x)||^2\right) \gamma(x,y) dQ^n(y) dP(x) - 1 \right),$$

which proves the result. ■

B.4 Proof of Lemma 23

In order to prove this result we first collect two further facts. The first is proved by a direct calculation.

Lemma 29. *In the setting of Proposition 21, for any h it holds for every x*

$$\int h(x)^T(y - T_\varepsilon(x)) \gamma(x,y) dQ(y) = 0.$$

Proof By linearity, we have:

$$\int h(x)^T(y - T_\varepsilon(x))\gamma(x,y)dQ(y) = h(x)^T\left[\int (y - T_\varepsilon(x))\gamma(x,y)dQ(y)\right]$$

and the integral on the inside can be expressed as:

$$\begin{aligned}
&\int (y - T_\varepsilon(x))\gamma(x,y)dQ(y) \\
&= \int y\gamma(x,y) - \gamma(x,y)\left[\frac{\int y_0 \exp\left(\frac{1}{\varepsilon}\left[g_\varepsilon(y_0) - c(x,y_0)\right]\right)dQ(y_0)}{\int \exp\left(\frac{1}{\varepsilon}\left[g_\varepsilon(y_1) - c(x,y_1)\right]\right)dQ(y_1)}\right]dQ(y) && \text{(Def. of } T_\varepsilon\text{)} \\
&= \int y\gamma(x,y) - \gamma(x,y)\left[\frac{\int y_0 \exp\left(\frac{1}{\varepsilon}\left[g_\varepsilon(y_0) + f_\varepsilon(x) - c(x,y_0)\right]\right)dQ(y_0)}{\int \exp\left(\frac{1}{\varepsilon}\left[g_\varepsilon(y_1) + f_\varepsilon(x) - c(x,y_1)\right]\right)dQ(y_1)}\right]dQ(y) \\
&= \int y\gamma(x,y) - \gamma(x,y)\left[\int y_0 \exp\left(\frac{1}{\varepsilon}\left[g_\varepsilon(y_0) + f_\varepsilon(x) - c(x,y_0)\right]\right)dQ(y_0)\right]dQ(y) && \text{(Eq. 23)} \\
&= \int y\gamma(x,y)dQ(y) - \int \gamma(x,y)\left[\int y_0\gamma(x,y_0)dQ(y_0)\right]dQ(y) \\
&= \int y\gamma(x,y)dQ(y) - \left[\int y_0\gamma(x,y_0)dQ(y_0)\right]\int \gamma(x,y)dQ(y) \\
&= \int y\gamma(x,y)dQ(y) - \int y_0\gamma(x,y_0)dQ(y_0) = 0. && \text{(Eq 22)}
\end{aligned}$$

■

In particular this result holds when $h = h_0$, and going a step further, it holds *independent* of both the choice of x and the form of the random function h_0. This is critical for establishing uniform control. The mean-zero condition also enables us to use a key inequality which only holds for mean-zero random variables.

Lemma 30. *There exists an absolute constant C_0 such that for all $a \geq C_0 r^2$ and h_0 defined as in Lemma 22, it holds for all x that:*

$$\int \exp(h_0(x))^T(y - T_\varepsilon(x))\gamma(x,y)dQ(y) \leq \exp(a||h_0(x)||^2).$$

Proof We need a few facts from (Vershynin, 2018). First, if X is a bounded random variable such that $||X||_\infty < B$ then:

$$||X||_{\psi_2} \leq C_4 B,$$

where $||X||_\psi$ denotes the subgaussian norm of X and C_4 is an absolute constant (see Example 2.5.8.iii in (Vershynin, 2018)). Second, if X is mean-zero, then for all $\lambda \in \mathbb{R}$:

$$\mathbb{E}\exp(\lambda X) \leq \exp(C_5\lambda^2||X||_{\psi_2}^2), \tag{33}$$

where C_5 is another absolute constant (See Proposition 2.5.2 or (2.16) in (Vershynin, 2018)).

Now for a fixed x, let Y^x be the random variable whose law is the conditional distribution of π_ε with $X = x$. Further define:

$$Z^x \triangleq (T_\varepsilon^n(x) - T_\varepsilon(x))^T (Y^x - T_\varepsilon(x)).$$

First note that since $T_\varepsilon(x), T_\varepsilon^n(x), Y^x \in B_2^d(0, r)$ we have:

$$\begin{aligned} |Z^x| &= |(T_\varepsilon^n(x) - T_\varepsilon(x))^T (Y^x - T_\varepsilon(x))| \\ &\leq ||T_\varepsilon^n(x) - T_\varepsilon(x)||||Y^x - T_\varepsilon(x)|| \\ &\leq (2r)||T_\varepsilon^n(x) - T_\varepsilon(x)|| \\ \implies ||Z^x||_{\psi_2} &\leq 2rC_4||T_\varepsilon^n(x) - T_\varepsilon(x)||. \end{aligned} \tag{34}$$

Next, we have by Lemma 29

$$\mathbb{E}[Z^x] = \int (T_\varepsilon^n(x) - T_\varepsilon(x))^T (Y^x - T_\varepsilon(x)) d\pi_\varepsilon^x(y) = \int (T_\varepsilon^n(x) - T_\varepsilon(x))^T (Y^x - T_\varepsilon(x)) \gamma(x, y) dQ(y) = 0.$$

This shows that Z^x satisfies the conditions to use (33). Doing so we obtain:

$$\begin{aligned} &\int \exp(h_0(x))^T (y - T_\varepsilon(x)) \gamma(x, y) dQ(y) \\ &= \int \exp\left(\frac{1}{2a} (T_\varepsilon^n(x) - T_\varepsilon(x))^T (Y^x - T_\varepsilon(x))\right) \gamma(x, y) dQ(y) \\ &= \mathbb{E}\left[\exp\left(\frac{1}{2a} Z^x\right)\right] \\ &\leq \exp\left(C_5 \frac{1}{4a^2} 4C_4^2 r^2 ||T_\varepsilon^n(x) - T_\varepsilon(x)||^2\right) \qquad \text{(Eq.(33) and (34))} \\ &= \exp\left(C_6 \frac{r^2}{a^2} ||T_\varepsilon^n(x) - T_\varepsilon(x)||^2\right), \end{aligned}$$

where $C_6 = C_4^2 C_5$.

From here it is enough to show that

$$\exp\left(C_6 \frac{r^2}{a^2} ||T_\varepsilon^n(x) - T_\varepsilon(x)||^2\right) \leq \exp\left(\frac{1}{4a} ||T_\varepsilon^n(x) - T_\varepsilon(x)||^2\right) = \exp(a||h_0(x)||^2).$$

By monotonicity of exp it is enough to show that

$$C_6 \frac{r^2}{a^2} ||T_\varepsilon^n(x) - T_\varepsilon(x)||^2 \leq \frac{1}{4a} ||T_\varepsilon^n(x) - T_\varepsilon(x)||^2,$$

and this holds true if $a \geq 4C_6 r^2$. Letting $C_0 = 4C_6$ proves the result. ■

With the preceding lemma in hand, the proof of Lemma 23 becomes straightforward.

Proof (Lemma 23) For $a \geq C_0 r^2$ we have:

$$\int \int \exp\left(h_0(x)^T (y - T_\varepsilon(x)) - a||h_0(x)||^2\right) \gamma(x, y) dQ(y) dP(x)$$

$$
\begin{aligned}
&= \int \exp(-a||h_0(x)||^2) \left[\int \exp\left(h_0(x)^T (y - T_\varepsilon(x)) \right) \gamma(x,y) dQ(y) \right] dP(x) \\
&\leq \int \exp(-a||h_0(x)||^2) \left[\exp(a||h_0(x)||^2) \right] dP(x) \qquad \text{(Lemma 30)} \\
&= 1.
\end{aligned}
$$

■

B.5 Proof of Lemma 25

To start we have the following result, which is essentially contained in Proposition A.1 of (Mena and Niles-Weed, 2019). We include its proof here for completeness and to specify the bounds to our setting.

Proposition 31. *Let P be σ^2-subgaussian. Let $Q \in \mathcal{P}(B_2^d(0,r))$. Then, there exist smooth optimal potentials $(f_\varepsilon, g_\varepsilon)$ for $S_\varepsilon(P,Q)$ such that,*

$$
g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2 \leq d\sigma^2 + (\|x\| + \sqrt{2}d\sigma)\|y\| - \|x\|^2,
$$

$$
f_\varepsilon(x) \leq \frac{1}{2}(r + \|x\|)^2.
$$

Proof Let $X \sim P$ and $Y \sim Q$. We note that by Equations (22), (23), and (24) one can chose $f_\varepsilon, g_\varepsilon$ such that:

$$
\begin{aligned}
f_\varepsilon(x) &= -\varepsilon \log\left(\mathbb{E} \exp\left(\frac{1}{\varepsilon}(g_\varepsilon(Y) - 1/2\|x-Y\|^2) \right) \right), \\
g_\varepsilon(y) &= -\varepsilon \log\left(\mathbb{E} \exp\left(\frac{1}{\varepsilon}(f_\varepsilon(X) - 1/2\|X-y\|^2) \right) \right),
\end{aligned}
$$

and $\mathbb{E}[f_\varepsilon(X)] = \mathbb{E}[g_\varepsilon(Y)] = \frac{1}{2}S_\varepsilon(P,Q) \geq 0$. Given these choices, we note that by the convexity of $-\varepsilon \log$ and Jensen's inequality that:

$$
\begin{aligned}
g_\varepsilon(y) &= -\varepsilon \log\left(\mathbb{E} \exp\left(\frac{1}{\varepsilon}(f_\varepsilon(X) - 1/2\|X-y\|^2) \right) \right) \\
&\leq \mathbb{E}\left[-\varepsilon \log \exp\left(\frac{1}{\varepsilon}\left(f_\varepsilon(X) - 1/2\|X-y\|^2 \right) \right) \right] \\
&= \mathbb{E}\left[\frac{1}{2}\|X-y\|^2 - f_\varepsilon(X) \right] \\
&\leq \frac{1}{2}\mathbb{E}\|X-y\|^2.
\end{aligned}
$$

This implies that:

$$
\begin{aligned}
g_\varepsilon(y) - \frac{1}{2}\|x-y\|^2 &\leq \frac{1}{2}\mathbb{E}\|X-y\|^2 - \frac{1}{2}\|x-y\|^2 \\
&= \frac{1}{2}(\mathbb{E}\|X\|^2 + \|y\|^2) - \mathbb{E}[X]^\top y - \frac{1}{2}(\|x\|^2 + \|y\|^2) + x^\top y
\end{aligned}
$$

$$\leq d\sigma^2 + \sqrt{2d}\sigma\|y\| - \frac{\|x\|^2}{2} + \|x\|\|y\|.$$

Similarly, applying Jensen's inequality as above but to f_ε we obtain:

$$f_\varepsilon(x) \leq \frac{1}{2}\mathbb{E}\|Y - x\|^2 \leq \frac{1}{2}(r + \|x\|)^2.$$

■

For our analysis this proposition implies the following bound on $\gamma(x, y)$ for P a.e. x and for Q a.e. y.

Lemma 32. *Under the assumptions of Proposition 31,*

$$\gamma(x, y) \leq \exp\left(\frac{1}{\varepsilon}\left[d\sigma^2 + \sqrt{2d}\sigma r + r^2 + 2r\|x\|\right]\right).$$

Proof Using Proposition 31 we have

$$\begin{aligned}
\varepsilon \log \gamma(x, y) &= f_\varepsilon(x) + g_\varepsilon(y) - \frac{1}{2}\|x - y\|^2 \\
&\leq d\sigma^2 + \sqrt{2d}\sigma\|y\| - \frac{\|x\|^2}{2} + \|x\|\|y\| + \frac{1}{2}(r + \|x\|)^2 \\
&= d\sigma^2 + \sqrt{2d}\sigma r + \|x\|r + \frac{1}{2}(r^2 + 2\|x\|r) \\
&= d\sigma^2 + \sqrt{2d}\sigma r + r^2 + 2r\|x\|.
\end{aligned}$$

Dividing by ε and exponentiating proves the result. ■

Before continuing to the main proof we require one more basic tool. The following result is a standard bound in empirical process theory which will help us to control the randomness of h_0. To state it we must introduce the notation $\mathcal{N}(T, d_T, \delta)$ which is the covering number of the metric space (T, d_T) by balls of radius at most δ whose centers lie in T.

Theorem 33 (Dudley's Inequality, (Vershynin, 2018) Theorem 8.1.3)**.** *Let $(X_t)_{t\in T}$ be a mean-zero random process on a metric space (T, d_T) with subgaussian increments satisfying $||X_t - X_s||_{\psi_2} \leq K d_T(t, s)$ for all $t, s \in T$. Then:*

$$\mathbb{E}\sup_{t\in T} X_t \leq C_7 K \int_0^\infty \sqrt{\log \mathcal{N}(T, d_T, \delta)}d\delta.$$

The following result is a standard bound for the covering numbers of balls in $\mathbb{R}^d$, a proof of which can be found in (Vershynin, 2018).

Lemma 34. *For $\delta \geq r$ we have $\mathcal{N}(B_2^d(0, r), ||\cdot||, r) = 1$ and for $0 < \delta < r$ we have:*

$$\mathcal{N}(B_2^d(0, r), ||\cdot||, \delta) \leq \left(\frac{3r}{\delta}\right)^d.$$

Having collected the necessary results we now proceed to the proof of Lemma 25.

Proof (Lemma 25) First we re-write our integral in a form that is more convenient for us.

$$\begin{aligned}
&\iint \exp\left(h_0(x)^T(y - T_\varepsilon(x)) - a||h_0(x)||^2\right)\gamma(x,y)[dQ^n - dQ](y)dP(x) \\
&= \iint \exp\left(\frac{1}{2a}(T_\varepsilon^n(x) - T_\varepsilon(x))^T(y - T_\varepsilon(x)) - \frac{1}{4a}||T_\varepsilon^n(x) - T_\varepsilon(x)||^2\right)\gamma(x,y)[dQ^n - dQ](y)dP(x).
\end{aligned}$$

Now note that $T_\varepsilon^n(x)$ is a vector contained in $B_2^d(0,r)$ so we always have the following bound (since we can choose $v = T_\varepsilon^n(x)$)

$$\begin{aligned}
&\int\int \exp\left(\tfrac{1}{2a}(T_\varepsilon^n(x) - T_\varepsilon(x))^T(y - T_\varepsilon(x)) - \tfrac{1}{4a}||T_\varepsilon^n(x) - T_\varepsilon(x)||^2\right)\gamma(x,y)[dQ^n - dQ](y)dP(x) \\
&\leq \int\left[\sup_{v\in B_2^d(0,r)}\int \exp\left(\tfrac{1}{2a}(v - T_\varepsilon(x))^T(y - T_\varepsilon(x)) - \tfrac{1}{4a}||v - T_\varepsilon(x)||^2\right)\gamma(x,y)[dQ^n - dQ](y)\right]dP(x).
\end{aligned}$$

Taking the expectation of both sides of this inequality, and subsequently changing the order of integration we have

$$\begin{aligned}
&\mathbb{E}_{Y^n}\int\int \exp\left(h_0(x)^T(y - T_\varepsilon(x)) - a||h_0(x)||^2\right)\gamma(x,y)[dQ^n - dQ](y)dP(x) \\
&\leq \mathbb{E}_{Y^n}\int\left[\sup_{v\in B_2^d(0,r)}\int \exp\left(\tfrac{1}{2a}(v - T_\varepsilon(x))^T(y - T_\varepsilon(x)) - \tfrac{1}{4a}||v - T_\varepsilon(x)||^2\right)\gamma(x,y)[dQ^n - dQ](y)\right]dP(x) \\
&= \int \mathbb{E}_{Y^n}\left[\sup_{v\in B_2^d(0,r)}\int \exp\left(\tfrac{1}{2a}(v - T_\varepsilon(x))^T(y - T_\varepsilon(x)) - \tfrac{1}{4a}||v - T_\varepsilon(x)||^2\right)\gamma(x,y)[dQ^n - dQ](y)\right]dP(x).
\end{aligned}$$

Therefore it is enough to uniformly bound over x the inner expectation, which is what we do now. From here, x is treated as fixed.

Recall that $Q^n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ where $Y_1, ..., Y_n$ are i.i.d. according to Q. We now define the empirical process $(Z_v)_{v\in B_2^d(0,r)}$ where Z_v is defined by:

$$\begin{aligned}
Z_v &\triangleq \int \exp\left(\frac{1}{2a}(v - T_\varepsilon(x))^T(y - T_\varepsilon(x)) - \frac{1}{4a}||v - T_\varepsilon(x)||^2\right)\gamma(x,y)[dQ^n - dQ](y) \\
&= \frac{1}{n}\sum_{i=1}^n \exp\left(\frac{1}{2a}(v - T_\varepsilon(x))^T(Y_i - T_\varepsilon(x)) - \frac{1}{4a}||v - T_\varepsilon(x)||^2\right)\gamma(x,Y_i) \\
&\qquad - \mathbb{E}_Y \exp\left(\frac{1}{2a}(v - T_\varepsilon(x))^T(Y - T_\varepsilon(x)) - \frac{1}{4a}||v - T_\varepsilon(x)||^2\right)\gamma(x,Y) \\
&= \frac{1}{n}\sum_{i=1}^n A_v^i - \mathbb{E}A_v,
\end{aligned}$$

where $A_v^i \triangleq \exp\left(\tfrac{1}{2a}(v - T_\varepsilon(x))^T(Y_i - T_\varepsilon(x)) - \tfrac{1}{4a}||v - T_\varepsilon(x)||^2\right)\gamma(x,Y_i)$.

Our first task is to control the increments of the process Z_v. Namely, we seek a bound of the form:

$$||Z_u - Z_v||_{\psi_2} \leq K||u - v||.$$

By applying Lemma 2.6.8 and then Proposition 2.6.1 from (Vershynin, 2018) we have

$$||Z_u - Z_v||_{\psi_2} = ||(\frac{1}{n}\sum_{i=1}^n A_u^i - \mathbb{E}A_u) - (\frac{1}{n}\sum_{i=1}^n A_v^i - \mathbb{E}A_v)||_{\psi_2}$$

$$
\begin{aligned}
&= ||\frac{1}{n}\sum_{i=1}^{n}(A_u^i - A_v^i) - \mathbb{E}(A_u - A_v)||_{\psi_2} \\
&\leq C_8||\frac{1}{n}\sum_{i=1}^{n}(A_u^i - A_v^i)||_{\psi_2} \\
&\leq C_8C_9\frac{1}{n}\left(\sum_{i=1}^{n}||A_u^i - A_v^i||_{\psi_2}^2\right)^{1/2}, \qquad (35)
\end{aligned}
$$

where C_8 and C_9 are absolute constants. Let $C_{10} = C_8C_9$

Now we need to control $||A_u^i - A_v^i||_{\psi_2}$. Define:

$$\Gamma(x, r, \varepsilon) \triangleq \frac{1}{\varepsilon}\left[d\sigma^2 + \sqrt{2d}\sigma r + r^2 + 2r\|x\|\right].$$

so that by Lemma 32 we have with probability 1

$$\gamma(x, Y) \leq \exp\left(\Gamma(x, r, \varepsilon)\right).$$

For the moment suppressing the dependence on i, this can be done as follows:

$$
\begin{aligned}
|A_u - A_v| = \Bigg| &\exp\left(\frac{1}{2a}\left(u - T_\varepsilon(x)\right)^T\left(Y - T_\varepsilon(x)\right) - \frac{1}{4a}||u - T_\varepsilon(x)||^2\right)\gamma(x, Y) \\
&- \exp\left(\frac{1}{2a}\left(v - T_\varepsilon(x)\right)^T\left(Y - T_\varepsilon(x)\right) - \frac{1}{4a}||v - T_\varepsilon(x)||^2\right)\gamma(x, Y)\Bigg| \\
= \gamma(x, Y)\Bigg| &\exp\left(\frac{1}{2a}\left(u - T_\varepsilon(x)\right)^T\left(Y - T_\varepsilon(x)\right) - \frac{1}{4a}||u - T_\varepsilon(x)||^2\right) \\
&- \exp\left(\frac{1}{2a}\left(v - T_\varepsilon(x)\right)^T\left(Y - T_\varepsilon(x)\right) - \frac{1}{4a}||v - T_\varepsilon(x)||^2\right)\Bigg| \\
\leq \exp(\Gamma(x, r, \varepsilon))\Bigg| &\exp\left(\frac{1}{2a}\left(u - T_\varepsilon(x)\right)^T\left(Y - T_\varepsilon(x)\right) - \frac{1}{4a}||u - T_\varepsilon(x)||^2\right) \\
&- \exp\left(\frac{1}{2a}\left(v - T_\varepsilon(x)\right)^T\left(Y - T_\varepsilon(x)\right) - \frac{1}{4a}||v - T_\varepsilon(x)||^2\right)\Bigg|.
\end{aligned}
$$

Now by the assumption that $a \geq 2r^2$ and the fact that $u, v, Y, T_\varepsilon(x) \in B_2^d(0, r)$ we have both

$$
\begin{aligned}
\frac{1}{2a}\left(u - T_\varepsilon(x)\right)^T\left(Y - T_\varepsilon(x)\right) - \frac{1}{4a}||u - T_\varepsilon(x)||^2 &\leq \frac{1}{2a}\left(u - T_\varepsilon(x)\right)^T\left(Y - T_\varepsilon(x)\right) \\
&\leq \frac{1}{2a}||u - T_\varepsilon(x)||||Y - T_\varepsilon(x)|| \\
&\leq \frac{4r^2}{2a} \leq \frac{4r^2}{4r^2} = 1,
\end{aligned}
$$

and by an analogous computation replacing u be v,

$$\frac{1}{2a}\left(v - T_\varepsilon(x)\right)^T\left(Y - T_\varepsilon(x)\right) - \frac{1}{4a}||v - T_\varepsilon(x)||^2 \leq 1.$$

Using this, and the inequality valid for all $a, b \leq 1$ that $|e^a - e^b| \leq e|a-b|$, we have:

$$
\begin{aligned}
&\left| \exp\left(\frac{1}{2a} \left(u - T_\varepsilon(x)\right)^T \left(Y - T_\varepsilon(x)\right) - \frac{1}{4a}||u - T_\varepsilon(x)||^2 \right) \right. \\
&\left. - \exp\left(\frac{1}{2a} \left(v - T_\varepsilon(x)\right)^T \left(Y - T_\varepsilon(x)\right) - \frac{1}{4a}||v - T_\varepsilon(x)||^2 \right) \right| \\
\leq & e \left| \tfrac{1}{2a} \left(u - T_\varepsilon(x)\right)^T \left(Y - T_\varepsilon(x)\right) - \tfrac{1}{4a}||u - T_\varepsilon(x)||^2 - \tfrac{1}{2a} \left(v - T_\varepsilon(x)\right)^T \left(Y - T_\varepsilon(x)\right) + \tfrac{1}{4a}||v - T_\varepsilon(x)||^2 \right| \\
= & \frac{e}{2a} \left| (u-v)^T (Y - T_\varepsilon(x)) + \frac{1}{2}||v - T_\varepsilon(x)||^2 - \frac{1}{2}||u - T_\varepsilon(x)||^2 \right| \\
= & \frac{e}{2a} \left| (u-v)^T Y + \frac{1}{2}||v||^2 - \frac{1}{2}||u||^2 \right| \\
\leq & \frac{e}{2a}||u-v||||Y|| + \frac{e}{4a} \left| (||u|| - ||v||)(||u|| + ||v||) \right| \\
\leq & \frac{re}{2a}||u-v|| + \frac{re}{2a}||u-v|| \\
= & \frac{re}{a}||u-v||.
\end{aligned}
$$

Using this and the chain above we can conclude

$$|A_u - A_v| \leq \exp(\Gamma(x,r,\varepsilon)) \frac{re}{a}||u-v||,$$

which implies (as in the proof of Lemma 30)

$$||A_u - A_v||_{\psi_2} \leq C_4 \exp(\Gamma(x,r,\varepsilon)) \frac{re}{a}||u-v||. \tag{36}$$

Now plugging (36) into (35) we obtain:

$$
\begin{aligned}
||Z_u - Z_v||_{\psi_2} &\leq C_{10} \frac{1}{n} \left(\sum_{i=1}^n \left(C_4 \exp(\Gamma(x,r,\varepsilon)) \frac{re}{a}||u-v|| \right)^2 \right)^{1/2} \\
&= \frac{1}{\sqrt{n}} C_{11} \exp(\Gamma(x,r,\varepsilon)) \frac{re}{a}||u-v||,
\end{aligned}
$$

where $C_{11} = C_4 C_{10}$.

Using Theorem 33 above with $K(x,r,\varepsilon) = \frac{1}{\sqrt{n}} C_{11} \exp\left(\Gamma(x,r,\varepsilon)\right) \frac{re}{a}$ we have:

$$
\begin{aligned}
&\mathbb{E}_{Y^n} \left[\sup_{v \in B_2^d(0,r)} \int \exp\left(\frac{1}{2a} \left(v - T_\varepsilon(x)\right)^T \left(y - T_\varepsilon(x)\right) - \frac{1}{4a}||v - T_\varepsilon(x)||^2 \right) \gamma(x,y) [dQ^n - dQ](y) \right] \\
= & \mathbb{E} \sup_{v \in B_2^d(0,r)} Z_v \\
\leq & C_6 K(x,r,\varepsilon) \int_0^\infty \sqrt{\log \mathcal{N}(B_2^d(0,r), ||\cdot||, \delta)} d\delta \\
\leq & C_6 K(x,r,\varepsilon) \int_0^\infty \sqrt{d \log(3r/\delta)} d\delta
\end{aligned}
$$

$$\leq C_6 K(x,r,\varepsilon)\int_0^r \sqrt{d\log(3r/\delta)}d\delta$$
$$\leq C_{12}K(x,r,\varepsilon)r\sqrt{d}, \tag{37}$$

where we have used Lemma 34 and that

$$\int_0^r \sqrt{\log(3r/\delta)}d\delta = r\int_0^1 \sqrt{\log(3/\delta)}d\delta$$

and this integral is finite.

The next step is to unfix x and integrate the bound in (37) against the distribution of x. This is done as follows:

$$\begin{aligned}
&\mathbb{E}\int\int \exp\left(h_0(x)^T(y-T_\varepsilon(x)) - a||h_0(x)||^2\right)\gamma(x,y)[dQ^n - dQ](y)dP(x)\\
&\leq \int C_{12}K(x,r,\varepsilon)r\sqrt{d}\; dP(x).\\
&= \int C_{12}\frac{1}{\sqrt{n}}C_{11}\exp\left(\Gamma(x,r,\varepsilon)\right)\frac{re}{a}r\sqrt{d}dP(x)\\
&= C_{13}\frac{r^2e\sqrt{d}}{a\sqrt{n}}\int \exp\left(\frac{1}{\varepsilon}\left(d\sigma^2+\sqrt{2d}\sigma r + r^2 + 2r\|x\|\right)\right)dP(x)\\
&= C_{13}\frac{r^2e\sqrt{d}}{a\sqrt{n}}\exp\left(\frac{1}{\varepsilon}\left(d\sigma^2+\sqrt{2d}\sigma r + r^2\right)\right)\int \exp\left(\frac{2r}{\varepsilon}\|x\|\right)dP(x).
\end{aligned}$$

To conclude the proof we handle the integral by using the subgaussian condition on $\|X\|$:

$$\begin{aligned}
\mathbb{E}\exp\left(\frac{2\|X\|r}{\varepsilon}\right) &= \mathbb{E}\exp\left(\frac{2\|X\|r}{\varepsilon}\right)\mathbf{1}[\|X\| < (4rd\sigma^2/\varepsilon)] + \mathbb{E}\exp\left(\frac{2\|X\|r}{\varepsilon}\right)\mathbf{1}[\|X\| \geq (4rd\sigma^2/\varepsilon)]\\
&\leq \exp\left(\frac{8rd\sigma^2}{\varepsilon^2}\right) + \mathbb{E}\exp\left(\frac{\|X\|^2}{2d\sigma^2}\right)\\
&\leq \exp\left(\frac{8rd\sigma^2}{\varepsilon^2}\right) + 2.
\end{aligned}$$

Plugging this in above and using that $a \geq C_1 r^2$ completes the proof. ■

B.6 Proof of Theorem 27

Proof Note that if $Q \in \mathcal{B}_2^d(0,r)$ then Q is $\frac{r^2}{2d\log 2}$-subgaussian. Letting $\sigma_0 = \max\left(\sigma, \frac{r}{\sqrt{2d\log 2}}\right)$ we have that P and Q are both σ_0^2-subgaussian and therefore we can apply Proposition 26 with constant σ_0. Taking expectations and applying the bounds in Proposition 24 followed by Lemma 25 and Proposition 26 we obtain:

$$\begin{aligned}
&\mathbb{E}\frac{1}{4a}||T_\varepsilon^n - T_\varepsilon||^2_{L^2(P)}\\
&\leq \mathbb{E}\left[\frac{1}{\varepsilon}G + \int\int \exp\left(h_0(x)^T(y-T_\varepsilon(x)) - a||h_0(x)||^2\right)\gamma(x,y)[dQ^n - dQ](y)dP(x)\right]
\end{aligned}$$

$$\leq \frac{1}{\varepsilon}\left(K_{d,0}\cdot\varepsilon\left(1+\frac{\sigma_0^{\lceil 5d/2\rceil+6}}{\varepsilon^{\lceil 5d/4\rceil+3}}\right)\frac{1}{\sqrt{n}}\right)+\frac{C_2\sqrt{d}}{\sqrt{n}}\exp\left(\frac{1}{\varepsilon}\left(d\sigma^2+\sqrt{2d}\sigma r+r^2\right)\right)\left(\exp\left(\frac{8rd\sigma^2}{\varepsilon^2}\right)+2\right)$$
$$= K_{d,0}\left(1+\frac{\sigma_0^{\lceil 5d/2\rceil+6}}{\varepsilon^{\lceil 5d/4\rceil+3}}\right)\frac{1}{\sqrt{n}}+\frac{C_2\sqrt{d}}{\sqrt{n}}\exp\left(\frac{1}{\varepsilon}\left(d\sigma^2+\sqrt{2d}\sigma r+r^2\right)\right)\left(\exp\left(\frac{8rd\sigma^2}{\varepsilon^2}\right)+2\right).$$

Choosing $a = C_1 r^2$ and multiplying by both sides proves the result with $K_{d,1} = 4C_1K_{d,0}$ and $C_3 = 4C_1C_2$. ■

B.7 Proof of Theorem 28

The main body of the proof is spent in establishing a result similar to Proposition 21. It starts with the following result:

Proposition 35. *Let P, Q be σ^2-subgaussian and let P^n, Q^n be the random empirical measures. Then:*

$$\mathbb{E}\left[\sup_{\chi\in L^1(\pi_\varepsilon^n)}\int\chi d\pi_\varepsilon^n-\int\int\left[\exp(\chi(x,y))-1\right]\gamma_\varepsilon^{n,n}dP(x)dQ^n(y)\right]\leq\frac{1}{\varepsilon}\mathbb{E}\left[\int f_\varepsilon^{n,n}[dP^n-dP]\right],$$

where π_ε^n is the random optimal coupling of (P, Q^n), the function $\gamma_\varepsilon^{n,n}(x,y) = \exp(\frac{1}{\varepsilon}[f_\varepsilon^{n,n}(x)+ g_\varepsilon^{n,n}(y) - 1/2\,\|x-y\|^2])$, and $(f_\varepsilon^{n,n}, g_\varepsilon^{n,n})$ are the optimal potentials for P^n, Q^n.

Proof Let $(f_\varepsilon^{n,n}, g_\varepsilon^{n,n})$ be the optimal potentials for (P^n, Q^n) and define $\eta(x,y) = \varepsilon\chi(x,y)+ f_\varepsilon^{n,n}(x) + g_\varepsilon^{n,n}(y)$ for any $\chi \in L^1(\pi_\varepsilon^n)$. Applying Proposition 19 and calculating one has:

$$\begin{aligned}S_\varepsilon(P,Q^n)&\geq\int\eta d\pi_\varepsilon^n-\varepsilon\int\int\exp\left(\frac{1}{\varepsilon}\left[\eta(x,y)-\frac{1}{2}\|x-y\|^2\right]\right)dP(x)dQ^n(y)+\varepsilon\\&=\varepsilon\int\chi(x,y)d\pi_\varepsilon^n+\int f_\varepsilon^{n,n}dP(x)+\int g_\varepsilon^{n,n}(y)dQ^n(y)\\&\quad-\varepsilon\int\int\left[\exp(\chi(x,y))-1\right]\gamma_\varepsilon^{n,n}(x,y)dP(x)dQ^n(y).\end{aligned}$$

Rearranging, one obtains:

$$\int\chi(x,y)d\pi_\varepsilon^n-\int\int\left[\exp(\chi(x,y))-1\right]\gamma_\varepsilon^{n,n}(x,y)dP(x)dQ^n(y)\leq\frac{1}{\varepsilon}\left[S_\varepsilon(P,Q^n)-\int f_\varepsilon^{n,n}dP-\int g_\varepsilon^{n,n}dQ^n\right].$$

Since this holds uniformly over $\chi \in L^1(\pi_\varepsilon^n)$ we can take the supremum on the left as well:

$$\begin{aligned}&\sup_{\chi\in L^1(\pi_\varepsilon^n)}\int\chi(x,y)d\pi_\varepsilon^n-\int\int\left[\exp(\chi(x,y))-1\right]\gamma_\varepsilon^{n,n}(x,y)dP(x)dQ^n(y)\\&\qquad\leq\frac{1}{\varepsilon}\left[S_\varepsilon(P,Q^n)-\int f_\varepsilon^{n,n}dP-\int g_\varepsilon^{n,n}dQ^n\right].\end{aligned}$$

We now turn our attention to the right side. Let $(f_\varepsilon^n, g_\varepsilon^n)$ be optimal for (P, Q^n). By optimality we have:

$$\int f_\varepsilon^{n,n}dP^n+\int g_\varepsilon^{n,n}dQ^n=S_\varepsilon(P^n,Q^n)$$

$$
\begin{aligned}
&\geq \int f_\varepsilon^n dP^n + \int g_\varepsilon^n dQ^n - \varepsilon \int \int \exp\left(\frac{1}{\varepsilon}\left[f_\varepsilon^n(x) + g_\varepsilon^n(y) - \frac{1}{2}\|x-y\|^2\right]\right) dP^n(x) dQ^n(y) + \varepsilon \\
&= \int f_\varepsilon^n dP^n + \int g_\varepsilon^n dQ^n - \varepsilon \int \left[\int \exp\left(\frac{1}{\varepsilon}\left[f_\varepsilon^n(x) + g_\varepsilon^n(y) - \frac{1}{2}\|x-y\|^2\right]\right) dQ^n(y)\right] dP^n(x) + \varepsilon \\
&= \int f_\varepsilon^n dP^n + \int g_\varepsilon^n dQ^n - \varepsilon \int 1 dP^n(x) + \varepsilon \\
&= \int f_\varepsilon^n dP^n + \int g_\varepsilon^n dQ^n.
\end{aligned}
$$

Comparing the first and last we have shown:

$$
\int f_\varepsilon^{n,n} dP^n + \int g_\varepsilon^{n,n} dQ^n \geq \int f_\varepsilon^n dP^n + \int g_\varepsilon^n dQ^n.
$$

From here we can develop a bound as follows:

$$
\begin{aligned}
&S_\varepsilon(P, Q^n) - \int f_\varepsilon^{n,n} dP - \int g_\varepsilon^{n,n} dQ^n \\
&= \int f_\varepsilon^n dP + \int g_\varepsilon^n dQ^n - \int f_\varepsilon^{n,n} dP - \int g_\varepsilon^{n,n} dQ^n \\
&= \int (f_\varepsilon^n - f_\varepsilon^{n,n}) dP + \int g_\varepsilon^n dQ^n + \int f_\varepsilon^n dP^n - \int f_\varepsilon^n dP^n - \int g_\varepsilon^{n,n} dQ^n \\
&\leq \int (f_\varepsilon^n - f_\varepsilon^{n,n}) dP + \int g_\varepsilon^{n,n} dQ^n + \int f_\varepsilon^{n,n} dP^n - \int f_\varepsilon^n dP^n - \int g_\varepsilon^{n,n} dQ^n \\
&= \int f_\varepsilon^{n,n} [dP^n - dP] + \int f_\varepsilon^n [dP - dP^n].
\end{aligned}
$$

Note that f_ε^n is independent of P^n, and so conditioned on Q^n we have:

$$
\mathbb{E}\left[\int f_\varepsilon^n [dP - dP^n]\right] = 0.
$$

Backtracking and taking expectations we have shown:

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\chi \in L(\pi_\varepsilon^n)} \int \chi d\pi_\varepsilon^n - \int \int \left[\exp(\chi(x,y)) - 1\right] \gamma_\varepsilon^{n,n}(x,y) dP(x) d^n Q(y)\right] \\
&\leq \frac{1}{\varepsilon} \mathbb{E}\left[S_\varepsilon(P, Q^n) - \int f_\varepsilon^{n,n} dP - \int g_\varepsilon^{n,n} dQ^n\right] \\
&\leq \frac{1}{\varepsilon} \mathbb{E} \int f_\varepsilon^{n,n} [dP^n - dP] + \int f_\varepsilon^n [dP - dP^n] \\
&= \frac{1}{\varepsilon} \mathbb{E} \int f_\varepsilon^{n,n} [dP^n - dP],
\end{aligned}
$$

which proves the result. ■

The next step is to control $\mathbb{E} \int f_\varepsilon^{n,n} [dP^n - dP]$. Thankfully, this has already been controlled in the literature. While not explicitly stated in this form, the calculations in the proof of Theorem 2 in (Mena and Niles-Weed, 2019) show the following.

Theorem 36. *If P, Q are σ^2-subgaussian, then:*

$$\mathbb{E}\int f_\varepsilon^{n,n}[dP^n - dP] \leq K_{d,0}\cdot\varepsilon\left(1+\frac{\sigma^{\lceil 5d/2\rceil+6}}{\varepsilon^{\lceil 5d/4\rceil+3}}\right)\frac{1}{\sqrt{n}}.$$

We are now ready to proceed to the proof of Theorem 28.

Proof (Theorem 28) Choose $\chi(x,y) = h(x)^T(y - T_\varepsilon^{n,n}(x)) - a\left\|h(x)\right\|^2$ for a and h to be specified. First note that for fixed x we have:

$$\begin{aligned}\int h(x)^T(y - T_\varepsilon^{n,n}(x))\gamma_\varepsilon^{n,n}(x,y)dQ^n(y) &= h(x)^T\left[\int y\gamma_\varepsilon^{n,n}(x,y)dQ^n(y) - T_\varepsilon^{n,n}(x)\right]\\ &= h(x)^T\left[T_\varepsilon^{n,n}(x) - T_\varepsilon^{n,n}(x)\right] = 0.\end{aligned}$$

which means that by Hoeffding's inequality we have for all fixed x:

$$\begin{aligned}\int \exp\left(\chi(x,y)\right)\gamma_\varepsilon^{n,n}(x,y)dQ(y) &= \exp(-a\left\|h(x)\right\|^2)\int \exp(h(x)^T(y - T_\varepsilon^{n,n}(x)))\gamma_\varepsilon^{n,n}(x,y)dQ^n(y)\\ &\leq \exp(-a\left\|h(x)\right\|^2)\int \exp(C_4(h(x)^T(y - T_\varepsilon^{n,n}(x)))^2)\gamma_\varepsilon^{n,n}(x,y)dQ^n(y)\\ &\leq \exp(-a\left\|h(x)\right\|^2)\int \exp(C_4 4r^2\left\|h(x)\right\|_2^2)\gamma_\varepsilon^{n,n}(x,y)dQ^n(y)\\ &= \exp((C_4 4r^2 - a)\left\|h(x)\right\|^2).\end{aligned}$$

In particular for $a \geq C_4 4r^2$ this gives:

$$\int \exp\left(\chi(x,y)\right)\gamma_\varepsilon^{n,n}(x,y)dQ^n(y) \leq 1.$$

This implies for all x:

$$\int \left[\exp\left(\chi(x,y)\right) - 1\right]\gamma_\varepsilon^{n,n}(x,y)dQ^n(y) \leq 0.$$

Now set $h_0(x) = \frac{1}{2a}(T_\varepsilon^n(x) - T_\varepsilon^{n,n}(x))$. A direct calculation gives:

$$\begin{aligned}\int h_0(x)^T(y - T_\varepsilon^{n,n}) - a\left\|h_0(x)\right\|^2 d\pi_\varepsilon^n &= \int \frac{1}{4a}\left\|T_\varepsilon^n(x) - T_\varepsilon^{n,n}(x)\right\|^2 dP(x)\\ &= \frac{1}{4a}\left\|T_\varepsilon^n - T_\varepsilon^{n,n}\right\|_{L^2(P)}^2.\end{aligned}$$

Combining this with the fact that $\int \left[\exp\left(\chi(x,y)\right) - 1\right]\gamma_\varepsilon^{n,n}(x,y)dQ^n(y) \leq 0$, we have after taking expectations and applying the results above that:

$$\begin{aligned}\mathbb{E}\frac{1}{4a}\left\|T_\varepsilon^n - T_\varepsilon^{n,n}\right\|_{L^2(P)}^2 &\leq \mathbb{E}\int \chi d\pi_\varepsilon^n - \int \left[\exp\left(\chi(x,y)\right) - 1\right]\gamma_\varepsilon^{n,n}(x,y)dQ^n(y)dP(x)\\ &\leq \mathbb{E}\sup_\chi \int \chi d\pi_\varepsilon^n - \int \left[\exp\left(\chi(x,y)\right) - 1\right]\gamma_\varepsilon^{n,n}(x,y)dQ^n(y)dP(x)\end{aligned}$$

$$\leq \frac{1}{\varepsilon}\mathbb{E}\int f_\varepsilon^{n,n}[dP^n - dP]$$
$$\leq K_{d,0}\cdot\left(1+\frac{\sigma_0^{\lceil 5d/2\rceil+6}}{\varepsilon^{\lceil 5d/4\rceil+3}}\right)\frac{1}{\sqrt{n}},$$

where $\sigma_0 = \max\left(\sigma, \frac{r}{\sqrt{2d\log 2}}\right)$, valid for all $a \geq 4C_4r^2$. Multiplying both sides by this quantity gives:

$$\mathbb{E}\,\|T_\varepsilon^n - T_\varepsilon^{n,n}\|_{L^2(P)}^2 \leq K_{d,2}r^2\cdot\left(1+\frac{\sigma_0^{\lceil 5d/2\rceil+6}}{\varepsilon^{\lceil 5d/4\rceil+3}}\right)\frac{1}{\sqrt{n}},$$

where $K_{d,2} = 16C_4K_{d,0}$. ■

Constant	First Introduction	Value or Source
C_d	Statement of Definition 3	$(2\Gamma(d/2))^{-1}\sqrt{\pi}(d-1)\Gamma((d-1)/2)$
C_0	Statement of Lemma 23	$4C_5^2C_4$
C_1	Statement of Lemma 25	$\max(C_0, 2)$
C_2	Statement of Lemma 25	eC_{13}/C_1
C_3	Statement of Theorem 27	$4C_1C_2$
C_4	Proof of Lemma 23	(2.5.2) in(Vershynin, 2018)
C_5	Proof of Lemma 23	(2.3.8.iii) in (Vershynin, 2018), can be taken as $(\log 2)^{-1/2}$
C_6	Proof of Lemma 23	$C_5^2C_4$
C_7	Proof of Lemma 25	(8.1.3) in (Vershynin, 2018)
C_8	Proof of Lemma 25	(2.6.8) in (Vershynin, 2018)
C_9	Proof of Lemma 25	(2.6.1) in (Vershynin, 2018)
C_{10}	Proof of Lemma 25	C_8C_9
C_{11}	Proof of Lemma 25	C_5C_{10}
C_{12}	Proof of Lemma 25	$C_7(\sqrt{\log 3}+\sqrt{\pi}/2)$
C_{13}	Proof of Lemma 25	$C_{11}C_{12}$
$K_{d,0}$	Statement of Proposition 26	See (Mena and Niles-Weed, 2019)
$K_{d,1}$	Statement of Theorem 27	$4C_1K_{d,0}$
$K_{d,2}$	Statement of Theorem 28	$16C_4K_{d,0}$

Table 4: Table of constants used in the proofs of the theoretical results.

Appendix C.

C.1 Proof of Proposition 15

Proof

(a) Given $X, X' \overset{i.i.d.}{\sim} P_X$ and $Y, Y' \overset{i.i.d.}{\sim} P_Y$. Using Lemmas 2.2, 2.3 from (Baringhaus and Franz, 2004), we have:

$$2\mathbb{E}\|X-Y\| - \mathbb{E}\|X-X'\| - \mathbb{E}\|Y-Y'\|$$

$$= C_d \int_{\mathcal{S}^{d-1}} \int_{\mathbb{R}} \Big(\mathbb{P}(a^\top X \leq t) - P(a^\top Y \leq t) \Big)^2 \, d\kappa(a) dt.$$

Following the above expression, for the rank transformed random variables $\mathtt{R}_{\lambda,\varepsilon}(X)$, $\mathtt{R}_{\lambda,\varepsilon}(X')$, $\mathtt{R}_{\lambda,\varepsilon}(Y)$, and $\mathtt{R}_{\lambda,\varepsilon}(Y')$ corresponding to X, X', Y, and Y', respectively, we can express the sRE as,

$$\begin{aligned} &2\mathbb{E}\|\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(Y)\| - \mathbb{E}\|\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(X')\| - \mathbb{E}\|\mathtt{R}_{\lambda,\varepsilon}(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y')\| \\ =& C_d \int_{\mathcal{S}^{d-1}} \int_{\mathbb{R}} \Big(\mathbb{P}(a^\top \mathtt{R}_{\lambda,\varepsilon}(X) \leq t) - \mathbb{P}(a^\top \mathtt{R}_{\lambda,\varepsilon}(Y) \leq t) \Big)^2 \, \mathrm{d}\kappa(a) dt. \end{aligned}$$

(b) Following the definition of $\mathtt{sRE}_{\lambda,\varepsilon}$,

$$\begin{aligned} \mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2 &= C_d \int_{\mathbb{R}} \int_{\mathcal{S}^{d-1}} \Big(\mathbb{P}(a^\top \mathtt{R}_{\lambda,\varepsilon}(X) \leq t) - \mathbb{P}(a^\top \mathtt{R}_{\lambda,\varepsilon}(Y) \leq t) \Big)^2 d\kappa(a) dt \\ &= C_d \int_{\mathbb{R}} \int_{\mathcal{S}^{d-1}} \Big(\mathbb{P}(a^\top \mathtt{R}_{\lambda,\varepsilon}(Y) \leq t) - \mathbb{P}(a^\top \mathtt{R}_{\lambda,\varepsilon}(X) \leq t) \Big)^2 d\kappa(a) dt \\ &= \mathtt{sRE}_{\lambda,\varepsilon}(P_Y, P_X)^2. \end{aligned}$$

(c) Assuming $X \overset{d}{=} Y$, we have $\mathbb{P}(a^\top X \leq t) = \mathbb{P}(a^\top Y \leq t)$ for all $a \in \mathcal{S}^{d-1}$ and $t \in \mathbb{R}$ (Baringhaus and Franz, 2004). Since $\mathtt{R}_{\lambda,\varepsilon}(X)$, $\mathtt{R}_{\lambda,\varepsilon}(Y)$ are also mapped random vectors corresponding to X, Y, respectively, we see that $\mathbb{P}\big(a^\top \mathtt{R}_{\lambda,\varepsilon}(X) \leq t\big) = \mathbb{P}\big(a^\top \mathtt{R}_{\lambda,\varepsilon}(Y) \leq t\big)$ and the result follows.

■

C.2 Proof of Proposition 16

Proof It is equivalent to show that:

$$\lim_{\varepsilon \to 0^+} |\mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2 - \mathtt{RE}_\lambda(P_X, P_Y)^2| = 0.$$

Using the original definitions, we have:

$$\begin{aligned} &|\mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2 - \mathtt{RE}_\lambda(P_X, P_Y)^2| \\ =& \Big| 2\mathbb{E}\left[||\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(Y)|| - ||\mathtt{R}_\lambda(X) - \mathtt{R}_\lambda(Y)|| \right] - \mathbb{E}\left[||\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(X')|| - ||\mathtt{R}_\lambda(X) - \mathtt{R}_\lambda(X')|| \right] \\ &- \mathbb{E}\left[||\mathtt{R}_{\lambda,\varepsilon}(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y')|| - ||\mathtt{R}_\lambda(Y) - \mathtt{R}_\lambda(Y')|| \right] \Big| \\ \leq& 2\mathbb{E}\Big| ||\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(Y)|| - ||\mathtt{R}_\lambda(X) - \mathtt{R}_\lambda(Y)|| \Big| + \mathbb{E}\Big| ||\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(X')|| - ||\mathtt{R}_\lambda(X) - \mathtt{R}_\lambda(X')|| \Big| \\ &+ \mathbb{E}\Big| ||\mathtt{R}_{\lambda,\varepsilon}(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y')|| - ||\mathtt{R}_\lambda(Y) - \mathtt{R}_\lambda(Y')|| \Big| \\ \leq& 2\mathbb{E}||[\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(Y)] - [\mathtt{R}_\lambda(X) - \mathtt{R}_\lambda(Y)]|| + \mathbb{E}||[\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(X')] - [\mathtt{R}_\lambda(X) - \mathtt{R}_\lambda(X')]|| \\ &+ \mathbb{E}||[\mathtt{R}_{\lambda,\varepsilon}(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y')] - [\mathtt{R}_\lambda(Y) - \mathtt{R}_\lambda(Y')]|| \end{aligned}$$

$$
\begin{aligned}
&\leq 4\mathbb{E}||\mathtt{R}_\lambda(X) - \mathtt{R}_{\lambda,\varepsilon}(X)|| + 4\mathbb{E}||\mathtt{R}_\lambda(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y)|| \\
&\leq 4\sqrt{\mathbb{E}||\mathtt{R}_\lambda(X) - \mathtt{R}_{\lambda,\varepsilon}(X)||^2} + 4\sqrt{\mathbb{E}||\mathtt{R}_\lambda(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y)||^2} \\
&\leq 8\sqrt{\mathbb{E}||\mathtt{R}_\lambda(X) - \mathtt{R}_{\lambda,\varepsilon}(X)||^2 + \mathbb{E}||\mathtt{R}_\lambda(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y)||^2} \\
&\leq \frac{8}{\min(\sqrt{\lambda}, \sqrt{1-\lambda})}\sqrt{\mathbb{E}[\lambda||\mathtt{R}_\lambda(X) - \mathtt{R}_{\lambda,\varepsilon}(X)||^2 + (1-\lambda)||\mathtt{R}_\lambda(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y)||^2]} \\
&= \frac{8}{\min(\sqrt{\lambda}, \sqrt{1-\lambda})}||\mathtt{R}_\lambda - \mathtt{R}_{\lambda,\varepsilon}||_{L^2(P_\lambda)} \\
&\lesssim \frac{8}{\min(\sqrt{\lambda}, \sqrt{1-\lambda})}\sqrt{\varepsilon^2 I_0(P_\lambda, \mathrm{Unif}([0,1]^d)) + \varepsilon^{\min(\alpha,3)/2}},
\end{aligned}
$$

where for the final inequality we have used Proposition 1 in (Pooladian et al., 2022) and where the implicit constant is independent of ε. Here I_0 is the integrated Fisher information along the geodesic from P_λ and $\mathrm{Unif}([0,1]^d)$ which under the assumptions made is guaranteed to be finite, (Chizat et al., 2020; Pooladian and Niles-Weed, 2021). Taking the limit in the last expression we have:

$$
\begin{aligned}
0 &\leq \lim_{\varepsilon\to 0^+} |\mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2 - \mathtt{RE}_\lambda(P_X, P_Y)^2| \\
&\lesssim \lim_{\varepsilon\to 0^+} \frac{8}{\min(\sqrt{\lambda}, \sqrt{1-\lambda})}\sqrt{\varepsilon^2 I_0(P_\lambda, \mathrm{Unif}([0,1]^d)) + \varepsilon^{\min(\alpha,3)/2}} = 0,
\end{aligned}
$$

which shows indeed $\lim_{\varepsilon\to 0^+} |\mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2 - \mathtt{RE}_\lambda(P_X, P_Y)^2| = 0$. ■

Appendix D.

D.1 Proof of Theorem 17

Proof Using the notation that $\mathtt{R}_{\lambda,\varepsilon}^{m+n}$ is the random independently estimated map and $X^m = (X_1, ..., X_m), Y^n = (Y_1, ..., Y_n)$ are the samples used to evaluate the statistic (which are independent of $\mathtt{R}_{\lambda,\varepsilon}^{m+n}$) we immediately have the following:

$$
\begin{aligned}
&||\mathtt{sRE}_{\lambda,\varepsilon}^{m,n}(P_X, P_Y)^2 - \mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2||_{L^2}^2 \\
&= \mathop{\mathbb{E}}_{\mathtt{R}_{\lambda,\varepsilon}^{m+n}}\left[\mathop{\mathbb{E}}_{X^n,Y^m}\left[\left(\mathtt{sRE}_{\lambda,\varepsilon}^{m,n}(P_X, P_Y)^2 - \mathtt{sRE}_{\lambda,\varepsilon}(P_X, P_Y)^2\right)^2 \middle| \mathtt{R}_{\lambda,\varepsilon}^{m+n}\right]\right].
\end{aligned}
$$

We first consider the inner expectation. For brevity we suppress the conditioning on $\mathtt{R}_{\lambda,\varepsilon}^{m+n}$ and also introduce the following six collections of random variables:

$$
R^X \triangleq \frac{1}{m^2}\sum_{i,j=1}^m ||\mathtt{R}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(X_j)||, \qquad \hat{R}^X \triangleq \frac{1}{m^2}\sum_{i,j=1}^m ||\mathtt{R}_{\lambda,\varepsilon}^{m+n}(X_i) - \mathtt{R}_{\lambda,\varepsilon}^{m+n}(X_j)||,
$$

$$
R^Y \triangleq \frac{1}{n^2}\sum_{i,j=1}^n ||\mathtt{R}_{\lambda,\varepsilon}(Y_i) - \mathtt{R}_{\lambda,\varepsilon}(Y_j)||, \qquad \hat{R}^Y \triangleq \frac{1}{n^2}\sum_{i,j=1}^n ||\mathtt{R}_{\lambda,\varepsilon}^{m+n}(Y_i) - \mathtt{R}_{\lambda,\varepsilon}^{m+n}(Y_j)||,
$$

$$R^{XY} \triangleq \frac{2}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}||\mathtt{R}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(Y_j)||, \qquad \hat{R}^{XY} \triangleq \frac{2}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_j)||.$$

These are further compressed into:

$$R \triangleq R^{XY} - R^X - R^Y,$$
$$\hat{R} \triangleq \hat{R}^{XY} - \hat{R}^X - \hat{R}^Y.$$

We also introduce the notation:

$$E^X \triangleq \mathop{\mathbb{E}}_{X,X'}||\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(X')||,$$
$$E^Y \triangleq \mathop{\mathbb{E}}_{Y,Y'}||\mathtt{R}_{\lambda,\varepsilon}(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y')||,$$
$$E^{XY} \triangleq 2\mathop{\mathbb{E}}_{X,Y}||\mathtt{R}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(Y)||,$$
$$E \triangleq E^{XY} - E^X - E^Y.$$

With this notation we have:

$$\begin{aligned}\mathbb{E}\left[\left(\mathtt{sRE}^{m,n}_{\lambda,\varepsilon}(P_X,P_Y)^2 - \mathtt{sRE}_{\lambda,\varepsilon}(P_X,P_Y)^2\right)^2\right] &= \mathbb{E}(\hat{R}-E)^2 \\ &= \mathbb{E}([\hat{R}-R]-[R-E])^2 \\ &\leq 2\mathbb{E}(\hat{R}-R)^2 + 2\mathbb{E}(R-E)^2.\end{aligned}$$

We now control the two expectations separately:

$$\begin{aligned}\mathbb{E}(\hat{R}-R)^2 &= \mathbb{E}\left([\hat{R}^{XY}-R^{XY}]+[R^X-\hat{R}^X]+[R^Y-\hat{R}^Y]\right)^2 \\ &\leq 3\mathbb{E}(\hat{R}^{XY}-R^{XY})^2 + 3\mathbb{E}(\hat{R}^X-R^X)^2 + 3\mathbb{E}(\hat{R}^Y-R^Y)^2.\end{aligned}$$

Next we control these three expectations separately:

$$\begin{aligned}\mathbb{E}(\hat{R}^{XY}-R^{XY})^2 &= \mathbb{E}\left(\frac{2}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_j)|| - ||\mathtt{R}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(Y_j)||\right)^2 \\ &\leq 4\mathbb{E}\frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_j)|| - ||\mathtt{R}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(Y_j)||\right)^2 \\ &\leq 4\mathbb{E}\frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}||[\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_j)] - [\mathtt{R}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(Y_j)]||^2 \\ &\leq 4\mathbb{E}\frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(||[\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(X_i)|| + ||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_j) - \mathtt{R}_{\lambda,\varepsilon}(Y_j)]||\right)^2 \\ &\leq 8\mathbb{E}\frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}||[\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(X_i)||^2 + ||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_j) - \mathtt{R}_{\lambda,\varepsilon}(Y_j)]||^2 \\ &= 8\mathbb{E}\frac{1}{m}\sum_{i=1}^{m}||[\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(X_i)||^2 + 8\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}||[\mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_i) - \mathtt{R}_{\lambda,\varepsilon}(Y_i)||^2\end{aligned}$$

$$
\begin{aligned}
&= 8\mathbb{E}||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(X)||^2 + ||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y)||^2 \\
&\leq \frac{8}{\min(\lambda, 1-\lambda)}\mathbb{E}\lambda||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(X)||^2 + (1-\lambda)||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y)||^2 \\
&= \frac{8}{\min(\lambda, 1-\lambda)}||\mathtt{R}^{m+n}_{\lambda,\varepsilon} - \mathtt{R}_{\lambda,\varepsilon}||^2_{L^2(P_\lambda)}.
\end{aligned}
$$

In the third and fourth lines we have used the reverse triangle inequality followed by the triangle inequality to re-group the terms. In the second to last line we have made use of the inequality valid for all $a, b \geq 0, \lambda \in (0,1)$:

$$
a + b \leq \frac{\lambda}{\min(\lambda, 1-\lambda)}a + \frac{1-\lambda}{\min(\lambda, 1-\lambda)}b = \frac{1}{\min(\lambda, 1-\lambda)}\left[\lambda a + (1-\lambda)b\right].
$$

The last line follows from the fact that $P_\lambda = \lambda P_X + (1-\lambda)P_Y$. Next we have:

$$
\begin{aligned}
\mathbb{E}(\hat{R}^X - R^X)^2 &= \mathbb{E}\left(\frac{1}{n^2}\sum_{i,j=1}^{n} ||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_j)|| - ||\mathtt{R}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(X_j)||\right)^2 \\
&\leq \mathbb{E}\frac{1}{n^2}\sum_{i,j=1}^{n}\left(||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_j)|| - ||\mathtt{R}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(X_j)||\right)^2 \\
&\leq \mathbb{E}\frac{1}{n^2}\sum_{i,j=1}^{n} ||[\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_j)] - [\mathtt{R}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(X_j)]||^2 \\
&\leq \mathbb{E}\frac{1}{n^2}\sum_{i,j=1}^{n}\left(||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(X_i)|| + ||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_j) - \mathtt{R}_{\lambda,\varepsilon}(X_j)||\right)^2 \\
&\leq 2\mathbb{E}\frac{1}{n^2}\sum_{i,j=1}^{n} ||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(X_i)||^2 + ||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_j) - \mathtt{R}_{\lambda,\varepsilon}(X_j)||^2 \\
&= 4\mathbb{E}\frac{1}{n}\sum_{i=1}^{n} ||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(X_i)||^2 \\
&= 4\mathbb{E}||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(X)||^2.
\end{aligned}
$$

Again in the third and fourth lines we have used the reverse-triangle inequality followed by the triangle inequality. An exactly analogous calculation for Y shows:

$$
\mathbb{E}(\hat{R}^Y - R^Y)^2 \leq 4\mathbb{E}||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y)||^2.
$$

Combining these two we have similarly to above that:

$$
\begin{aligned}
\mathbb{E}(\hat{R}^X - R^X)^2 + \mathbb{E}(\hat{R}^Y - R^Y)^2 &\leq 4\mathbb{E}\left[||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X) - \mathtt{R}_{\lambda,\varepsilon}(X)||^2 + ||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y) - \mathtt{R}_{\lambda,\varepsilon}(Y)||^2\right] \\
&\leq \frac{4}{\min(\lambda, 1-\lambda)}||\mathtt{R}^{m+n}_{\lambda,\varepsilon} - \mathtt{R}_{\lambda,\varepsilon}||^2_{L^2(P_\lambda)}.
\end{aligned}
$$

Combining bounds we have:

$$
\begin{aligned}
\mathbb{E}(\hat{R} - R)^2 &\leq \frac{24}{\min(\lambda, 1-\lambda)}||\mathtt{R}^{m+n}_{\lambda,\varepsilon} - \mathtt{R}_{\lambda,\varepsilon}||^2_{L^2(P_\lambda)} + \frac{12}{\min(\lambda, 1-\lambda)}||\mathtt{R}^{m+n}_{\lambda,\varepsilon} - \mathtt{R}_{\lambda,\varepsilon}||^2_{L^2(P_\lambda)} \\
&= \frac{36}{\min(\lambda, 1-\lambda)}||\mathtt{R}^{m+n}_{\lambda,\varepsilon} - \mathtt{R}_{\lambda,\varepsilon}||^2_{L^2(P_\lambda)}.
\end{aligned}
$$

Now we turn our attention to $\mathbb{E}(R-E)^2$. First we show that $R-E$ is mean-zero:

$$\begin{aligned}\mathbb{E}R &= \mathbb{E}\frac{2}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}||\mathtt{R}_{\lambda,\varepsilon}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(Y_j)|| - \frac{1}{m^2}\sum_{i,j=1}^{m}||\mathtt{R}_{\lambda,\varepsilon}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(X_j)|| \\ &\qquad - \frac{1}{n^2}\sum_{i,j=1}^{n}||\mathtt{R}_{\lambda,\varepsilon}(Y_i)-\mathtt{R}_{\lambda,\varepsilon}(Y_j)|| \\ &= 2\mathbb{E}||\mathtt{R}_{\lambda,\varepsilon}(X)-\mathtt{R}_{\lambda,\varepsilon}(Y)|| - \mathbb{E}||\mathtt{R}_{\lambda,\varepsilon}(X)-\mathtt{R}_{\lambda,\varepsilon}(X')|| - \mathbb{E}||\mathtt{R}_{\lambda,\varepsilon}(Y)-\mathtt{R}_{\lambda,\varepsilon}(Y')|| \\ &= E^{XY}-E^{X}-E^{Y}=E.\end{aligned}$$

Subtracting E from the first and last shows $\mathbb{E}[R-E]=0$. Using this we have:

$$\mathbb{E}[(R-E)^2] = \mathbb{E}[(R-E)^2] - \mathbb{E}[R-E]^2 = \mathrm{Var}(R-E).$$

To control the variance we apply the Efron-Stein inequality ((Boucheron et al., 2013) Theorem 3.1) to the function:

$$\begin{aligned}f(X_1,...,X_m,Y_1,...,Y_n) = \frac{2}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}||\mathtt{R}_{\lambda,\varepsilon}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(Y_j)|| - \frac{1}{m^2}\sum_{i,j=1}^{m}||\mathtt{R}_{\lambda,\varepsilon}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(X_j)|| \\ - \frac{1}{n^2}\sum_{i,j=1}^{n}||\mathtt{R}_{\lambda,\varepsilon}(Y_i)-\mathtt{R}_{\lambda,\varepsilon}(Y_j)||.\end{aligned}$$

First note that we have the bounds:

$$\begin{aligned}&|f(X_1,...,X_{i-1},X_i,X_{i+1},...,X_m,Y_1,...,Y_n) - f(X_1,...,X_{i-1},X_i',X_{i+1},...,X_m,Y_1,...,Y_n)| \\ &= \left|\frac{2}{nm}\sum_{j=1}^{n}\left(||\mathtt{R}_{\lambda,\varepsilon}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(Y_j)|| - ||\mathtt{R}_{\lambda,\varepsilon}(X_i')-\mathtt{R}_{\lambda,\varepsilon}(Y_j)||\right)\right. \\ &\qquad \left. - \frac{1}{m^2}\sum_{j\neq i}^{m}\left(||\mathtt{R}_{\lambda,\varepsilon}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(X_j)|| - ||\mathtt{R}_{\lambda,\varepsilon}(X_i')-\mathtt{R}_{\lambda,\varepsilon}(X_j)||\right)\right| \\ &\leq \frac{2}{nm}\sum_{j=1}^{n}\left|||\mathtt{R}_{\lambda,\varepsilon}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(Y_j)|| - ||\mathtt{R}_{\lambda,\varepsilon}(X_i')-\mathtt{R}_{\lambda,\varepsilon}(Y_j)||\right| \\ &\qquad + \frac{1}{m^2}\sum_{j\neq i}^{m}\left|||\mathtt{R}_{\lambda,\varepsilon}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(X_j)|| - ||\mathtt{R}_{\lambda,\varepsilon}(X_i')-\mathtt{R}_{\lambda,\varepsilon}(X_j)||\right| \\ &\leq \frac{2}{nm}\sum_{j=1}^{n}||\mathtt{R}_{\lambda,\varepsilon}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(X_i')|| + \frac{1}{m^2}\sum_{j\neq i}^{m}||\mathtt{R}_{\lambda,\varepsilon}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(X_i')|| \\ &\leq \frac{3}{m}||\mathtt{R}_{\lambda,\varepsilon}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(X_i')|| \\ &\leq \frac{3\sqrt{d}}{m},\end{aligned}$$

where we have in the last line used the $\mathtt{R}_{\lambda,\varepsilon}$ maps into $[0,1]^d$ and the diameter is $\sqrt{d}$. A completely analogous computation shows:

$$|f(X_1,...,X_m,Y_1,...,Y_{i-1},Y_i,Y_{i+1},...,Y_n)-f(X_1,...,X_m,Y_1,...,Y_{i-1},Y_i',Y_{i+1},...,Y_n)|\leq\frac{3\sqrt{d}}{m}.$$

Using this bound in the Efron-Stein inequality we have:

$$\begin{aligned}
&\mathbb{E}[(R-E)^2]\\
=&\mathrm{Var}(R-E)\\
\leq&\frac{1}{2}\sum_{i=1}^{m}\mathbb{E}\big(f(X_1,...,X_{i-1},X_i,X_{i+1},...,X_m,Y_1,...,Y_n)-f(X_1,...,X_{i-1},X_i',X_{i+1},...,X_m,Y_1,...,Y_n)\big)^2\\
+&\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\big(f(X_1,...,X_m,Y_1,...,Y_{i-1},Y_i,Y_{i+1},...,Y_n)-f(X_1,...,X_m,Y_1,...,Y_{i-1},Y_i',Y_{i+1},...,Y_n)\big)^2\\
\leq&\frac{1}{2}\sum_{i=1}^{m}\mathbb{E}\left(\frac{3\sqrt{d}}{m}\right)^2+\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left(\frac{3\sqrt{d}}{n}\right)^2\\
=&\frac{9d}{2m}+\frac{9d}{2n}\\
=&\frac{9d(m+n)}{2mn}.
\end{aligned}$$

Collecting terms we have, conditionally on the estimate of $\mathtt{R}_{\lambda,\varepsilon}^{m+n}$, that:

$$\begin{aligned}
\mathbb{E}\left[\left(\mathtt{sRE}_{\lambda,\varepsilon}^{m,n}(P_X,P_Y)^2-\mathtt{sRE}_{\lambda,\varepsilon}(P_X,P_Y)^2\right)^2\right]&\leq2\mathbb{E}(\hat{R}-R)^2+2\mathbb{E}(R-E)^2\\
&\leq\frac{72}{\min(\lambda,1-\lambda)}||\mathtt{R}_{\lambda,\varepsilon}^{m+n}-\mathtt{R}_{\lambda,\varepsilon}||_{L^2(P_\lambda)}^2+\frac{9d(m+n)}{mn}.
\end{aligned}$$

Unconditioning and applying either Theorem 13 or Theorem 14, with $Q=\mathrm{Unif}([0,1]^d)\in\mathcal{P}(B(0,\sqrt{d}))$ completes the proof. ∎

D.2 Proof of Theorem 18

The proof is essentially the same as the proof of Theorem 17. The key difference is that we require a kernel analog of the reverse-triangle, followed by triangle inequality trick:

$$\begin{aligned}
&|||\mathtt{R}_{\lambda,\varepsilon}^{m+n}(X_i)-\mathtt{R}_{\lambda,\varepsilon}^{m+n}(Y_j)||-||\mathtt{R}_{\lambda,\varepsilon}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(Y_j)|||\\
&\leq||[\mathtt{R}_{\lambda,\varepsilon}^{m+n}(X_i)-\mathtt{R}_{\lambda,\varepsilon}^{m+n}(Y_j)]-[\mathtt{R}_{\lambda,\varepsilon}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(Y_j)]||\\
&\leq||\mathtt{R}_{\lambda,\varepsilon}^{m+n}(X_i)-\mathtt{R}_{\lambda,\varepsilon}(X_i)||+||\mathtt{R}_{\lambda,\varepsilon}^{m+n}(Y_j)-\mathtt{R}_{\lambda,\varepsilon}(Y_j)||.
\end{aligned}$$

This is achieved through:

$$\begin{aligned}
&|k(\mathtt{R}_{\lambda,\varepsilon}^{m+n}(X_i),\mathtt{R}_{\lambda,\varepsilon}^{m+n}(Y_j))-k(\mathtt{R}_{\lambda,\varepsilon}(X_i),\mathtt{R}_{\lambda,\varepsilon}(Y_j))|\\
=&|[k(\mathtt{R}_{\lambda,\varepsilon}^{m+n}(X_i),\mathtt{R}_{\lambda,\varepsilon}^{m+n}(Y_j))-k(\mathtt{R}_{\lambda,\varepsilon}(X_i),\mathtt{R}_{\lambda,\varepsilon}^{m+n}(Y_j))]
\end{aligned}$$

$$
\begin{aligned}
&\quad + [k(\mathtt{R}_{\lambda,\varepsilon}(X_i), \mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_j)) - k(\mathtt{R}_{\lambda,\varepsilon}(X_i), \mathtt{R}_{\lambda,\varepsilon}(Y_j))]| \\
\leq &|k(\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i), \mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_j)) - k(\mathtt{R}_{\lambda,\varepsilon}(X_i), \mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_j))| \\
&\quad + |k(\mathtt{R}_{\lambda,\varepsilon}(X_i), \mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_j)) - k(\mathtt{R}_{\lambda,\varepsilon}(X_i), \mathtt{R}_{\lambda,\varepsilon}(Y_j))| \\
\leq &l||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(X_i) - \mathtt{R}_{\lambda,\varepsilon}(X_i)|| + l||\mathtt{R}^{m+n}_{\lambda,\varepsilon}(Y_j) - \mathtt{R}_{\lambda,\varepsilon}(Y_j)||.
\end{aligned}
$$

With this key inequality established, one can show the result by following the proof of Theorem 17, substituting this inequality in the two places where the reverse-triangle followed triangle inequality is used. In both places these inequalities are done inside of a squaring so ultimately, we gain a factor of l^2.

Appendix E.

E.1 Related Works on Knockoff Generation

The seminal work (Candès et al., 2016) assumes that the joint feature distribution follows a multivariate Gaussian distribution and satisfies the pairwise exchangeability condition (18) via approximating only the first two moments (mean and covariance). In cases where the distributions are not multivariate Gaussian, the second order method in (Candès et al., 2016) cannot guarantee any control of the FDR. In contrast, methods like knockoffGAN (Salimans et al., 2016), deep knockoff (Romano et al., 2020), auto-encoding knockoff (Liu and Zheng, 2018) focus on learning generative models to sample knockoffs. KnockoffGAN is a complex architecture which consists of four different neural networks and optimizes a difficult minimax problem. A comparatively simpler approach adopted by deep knockoff employs MMD (Gretton et al., 2012) as the discriminating statistic for testing pairwise exchangeability in (18). As we see in Figure 3, the generator learned using the MMD performs poorly in high dimensions, and fails to approximate the input distribution properly. The auto-encoding knockoff method employs a variational auto-encoder (Kingma and Welling, 2019) to learn a low-dimensional latent space for high-dimensional data, assuming that the data lies close to a low-dimensional manifold. However, if the covariates violate this low-dimensional assumption, a more appropriate model may require learning a higher dimensional latent space. This comes with a risk of retaining more information about the original data, which could potentially result in diminished power. Another method, called deep direct likelihood knockoff (DDLK) (Sudarshan et al., 2020) minimizes the KL (Kullback-Leibler) divergence to test for pairwise exchangeability.

E.2 Schematic of the Knockoff Generator

Figure 5 shows the schematic of the deep generative model used for the knockoff generator (Section 5.2.1). The model has a fully connected neural network f_θ, where θ represents the parameters of the network (weights w and biases b). f_θ has 6 has hidden layers, each of them having $6 \cdot d$ units. The first layer of the neural network takes a vector of original variables $X \in \mathbb{R}^d$ and a d-dimensional noise vector $V \sim \mathcal{N}(0, I)$. Each unit in the hidden layers is produced by first taking linear combinations of the input followed by applying to each a nonlinear activation function. A parametric rectified linear unit is used as the activation function (Xu et al., 2015). The output layer returns a d-dimensional knockoff vector as depicted in Figure 5.

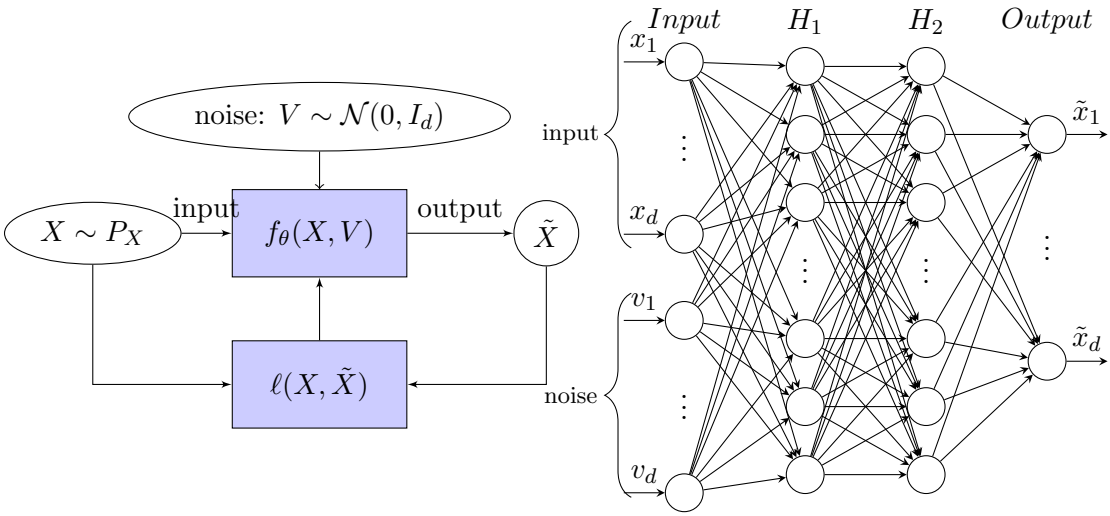


Figure 5: *Left:* schematic of the deep generative model for knockoff generation. *Right:* schematic of f_θ is shown for 2 hidden layers (in applications, we use 6 layers).

E.3 Algorithm to Train Knockoff Generator

Algorithm 1: Training of the Knockoff Generator

Input : training data: $\mathbf{X} \in \mathbb{R}^{n \times d}$, learning rate: η, entropic regularizer: ε, decorrelation parameter: γ, initialization of network parameters: θ_0, no of epochs: T, batch size: m, no of batches: n_b

Output: knockoff generator: f_{θ_T}

for $t \leftarrow 0$ **to** T **do**
  for $j \leftarrow 0$ **to** n_b **do**
    X_i, for all $1 \le i \le m$ `// samples for a minibatch`
    $V_i \sim \mathcal{N}(0, I)$ for all $1 \le i \le m$ `// noise sampling.`
    $\tilde{X}_i \leftarrow f_{\theta_t}(X_i, V_i)$, for all $1 \le i \le m$ `// knockoff generation.`
    $B \subset \{1, \ldots, d\}$ `// picking a random subset.`
    $J_{\theta_t}(\mathbf{X}_m, \tilde{\mathbf{X}}_m) \leftarrow \ell(\mathbf{X}_m, \tilde{\mathbf{X}}_m)$ `// loss calculation using` (18)
    $\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t}(J_{\theta_t})$ `// parameter update`
  end for
end for

E.4 Compute Resources

We use NVIDIA TESLA-K80 24 GB GPU for each simulation. Table 5 shows the comparison of wall-clock time between sRMMD and MMD. Compared to just MMD ($\mathcal{O}(n^2 d)$), sRMMD requires an additional step that is the computation of entropy-regularized optimal transport via the Sinkhorn algorithm ($\mathcal{O}(n^2/\varepsilon^2)$). Therefore, sRMMD can be computed in total $\mathcal{O}(n^2 d + n^2/\varepsilon^2)$ steps. The smaller the ε, the longer time it takes to compute

sRMMD. However, a few recent works (Li et al., 2023; Luo et al., 2023) improve the computational time to $\mathcal{O}(n^2/\varepsilon)$, which we believe will significantly reduce the training time for both MNIST and knockoff generator with sRMMD and will be explored in future work.

Method	Experiment Type		
	MNIST	Knockoff	
		Synthetic	Real
MMD	1155.89	2446.23	750.34
sRMMD	2634.78	7092.96	1979.45

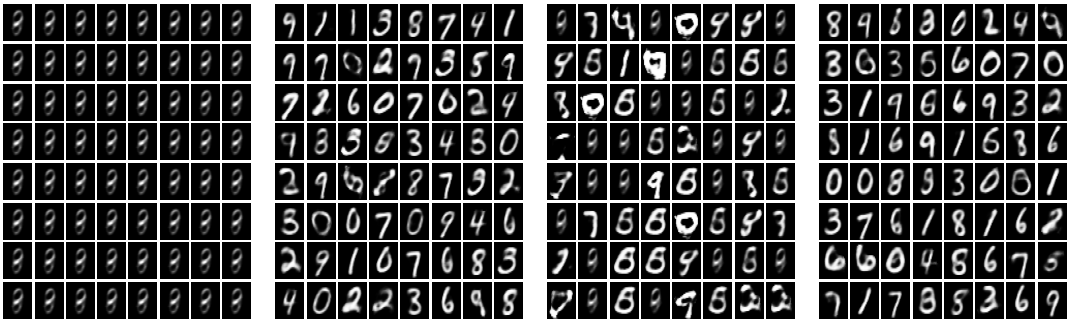
Table 5: Average training time (in seconds) comparison between MMD and sRMMD-based generators to reproduce the main results of the paper (keeping the same batch-size and same generative structure).

E.5 Baseline Models for Knockoff Generation

For KnockoffGAN, code is used from `https://bitbucket.org/mvdschaar/mlforhealthlabpub/src/master/`. For DDLK, code from `https://github.com/rajesh-lab/ddlk` is used. For second-order and MMD, we use code from `https://github.com/msesia/deepknockoffs`.

Appendix F. Additional Experiments

F.1 Effect of ε, σ on sRMMD-Based MNIST Image Generator (Section 5.1)



(a) (b) (c) (d)

Figure 6: sRMMD-based MNIST image generator (latent space dimension 8) using (a) $\varepsilon = 10, \sigma = (1, 2, 4, 8, 16, 32)$. (b) $\varepsilon = 10, \sigma = (0.01, 0.02, 0.04, 0.06, 0.08)$ (c) $\varepsilon = 20, \sigma = (0.01, 0.02, 0.04, 0.06, 0.08)$, and (d) $\varepsilon = 20, \sigma = (0.001, 0.002, 0.004, 0.006, 0.008)$.

MNIST-image generator minimizing the sRMMD loss produces a lot of ambiguous digits of similar shape when $\varepsilon = 10$ and $\sigma = (1, 2, 4, 8, 16, 32)$ are used (Figure 6). To understand why this may be the case, we plot the generator's loss over the training epochs (Figure 7). We observe that the loss is nearly zero at the beginning of the training, does not decrease smoothly and becomes very unstable after few epochs. This instability consequently leads to a poorly trained generator. In contrast, for $\varepsilon = 10$, using $\sigma = (0.01, 0.02, 0.04, 0.06, 0.08)$ maintains a smooth and decreasing loss over the epochs and the generator converges rapidly which brings about an improved image generator (Figure 6(b)). We also observe that if we increase ε further e.g., $\varepsilon = 20$, using $\sigma =$

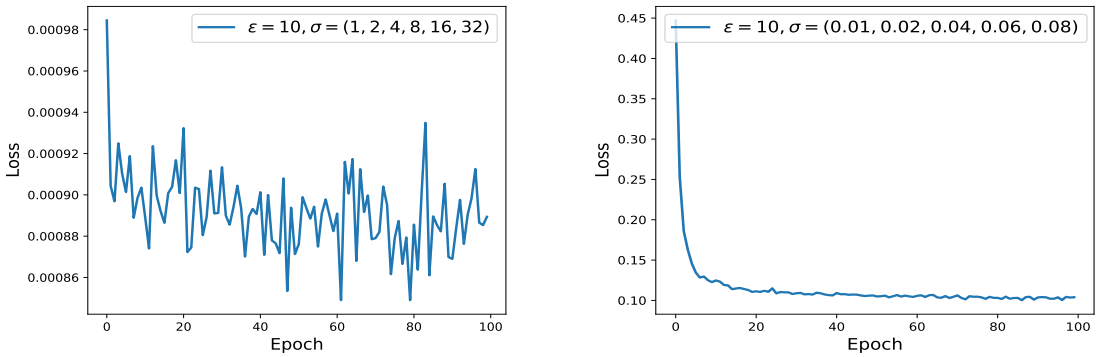


Figure 7: The plots show the generator's loss (sRMMD) on the y-axis and the training epoch on the x-axis, for $\varepsilon = 10$ with two different values of σ. On the left, where σ is large, the loss fluctuates and the model fails to converge. On the right, where σ is small, the loss gradually decreases with each epoch and the model converges faster.

$(0.01, 0.02, 0.04, 0.06, 0.08)$ still leads to a poor generator (Figure 6c). This situation can be avoided by reducing σ to $(0.001, 0.002, 0.004, 0.006, 0.008)$. Based on this empirical evidence, it can be said that using smaller bandwidth in case of larger ε leads to a better generator. Note that the choices of ε and σ we mentioned are specific to our case where we use latent space with dimension 8. It may require finding the optimal combination of ε and σ to get a good generator in case a different latent space dimension is used.

F.2 Why sRMMD not sRE?

Figure 8 shows that unlike sRMMD, the generator minimizing sRE in (18) cannot capture all four modes perfectly with $\varepsilon = 100$ when $d = 100$. Though similar to sRMMD, the reconstruction ability gets better with ε, sRE still fails to capture every mode even when ε is doubled. Moreover, a direct comparison between sRE and sRMMD when $\varepsilon = 20, 50, 200$ indicates that sRMMD outperforms sRE in reconstructing the original distribution.

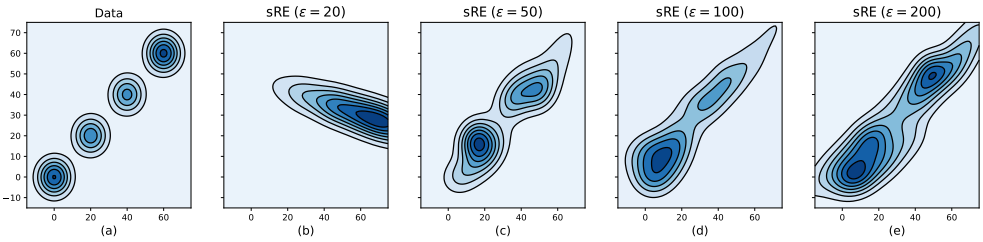
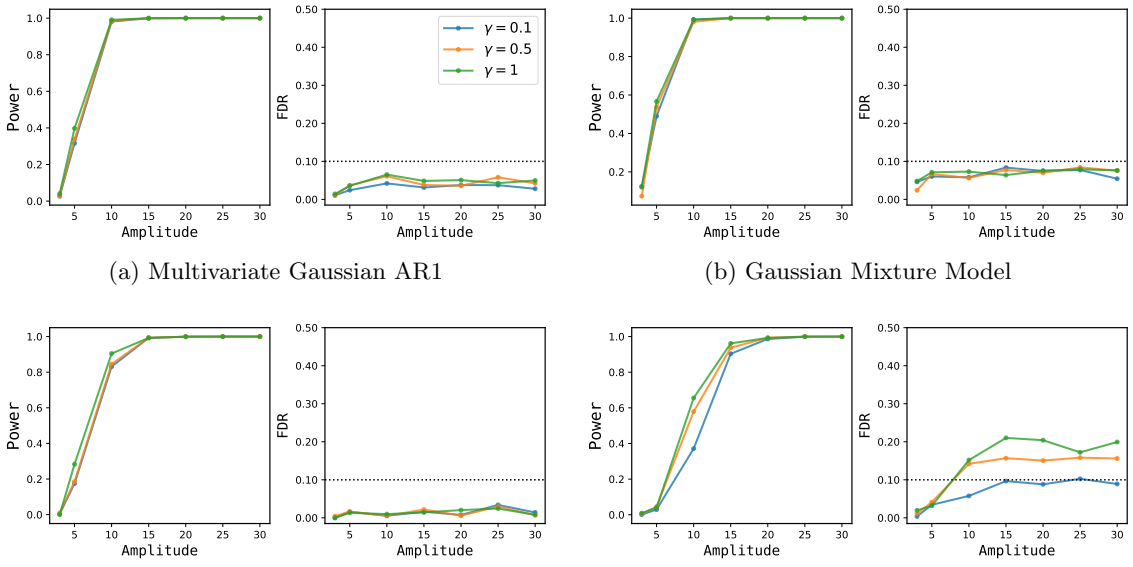


Figure 8: Visualizing two randomly selected dimensions of the original data used in Figure 3 and the generated knockoffs. The sRE-based knockoff generator did not converge when using $\varepsilon = 1$ and 10 (we encountered 'nan' after few epochs). That is why we refrain from adding it here.

F.3 Impact of Choice of Decorrelation Parameter γ

The selection of the optimal decorrelation parameter γ is difficult as we cannot perform cross-validation due to lack of access to the ground truth. Therefore, we pick the optimal γ for each distributional setting by investigating the sensitivity of the results to different values of γ. Figure 9 shows the FDR versus power tradeoff w.r.t. the amplitude parameter for several values of γ for each distributional setting considered in Section 5.2.2.



(a) Multivariate Gaussian AR1

(b) Gaussian Mixture Model

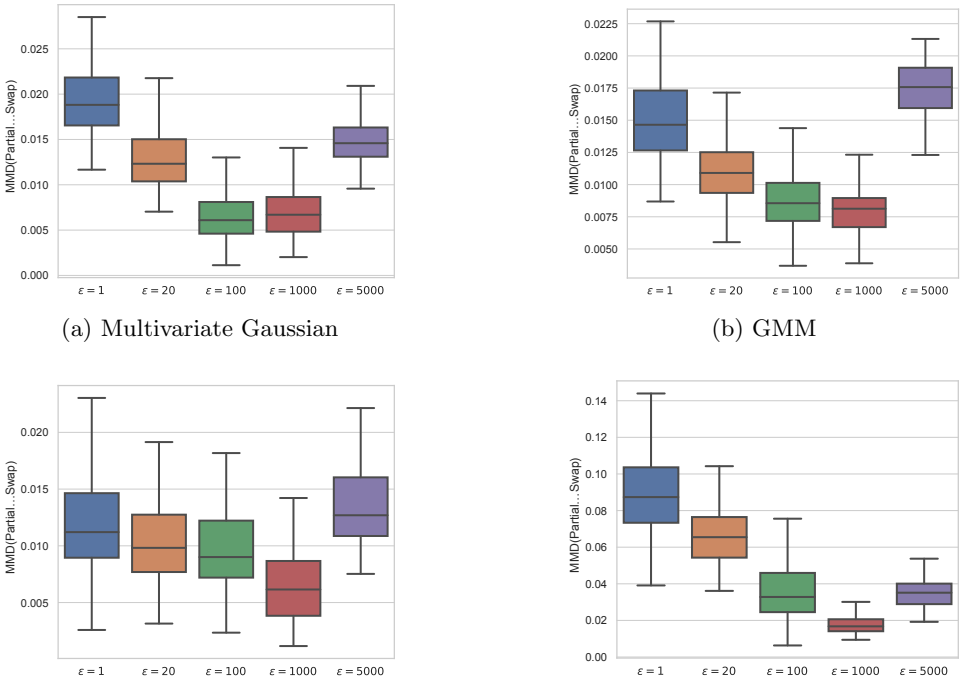
(c) Multivariate Student's t

(d) Sparse Gaussian

Figure 9: Average FDR and power computed over 200 independent experiments are shown on the y-axis for each synthetic benchmark. The FDR level is set to 0.1. The x-axis represents the amplitude parameter υ.

For multivariate Gaussian, GMM and multivariate Student's t distributions, sRMMD-based knockoff generator is not sensitive to the value of γ. In these cases, each γ, sRMMD controls the FDR at $q = 0.1$ and achieves nearly identical power over the entire amplitude region. In the case of sparse Gaussian setting, $\gamma = 0.1$ controls the FDR at $q = 0.1$. As γ increases (e.g., $0.5, 1$), the power also increases but fails to keep the FDR below 0.1.

F.4 Quality of the sRMMD Knockoffs with Respect to ε



(a) Multivariate Gaussian

(b) GMM

(c) Multivariate Student's T

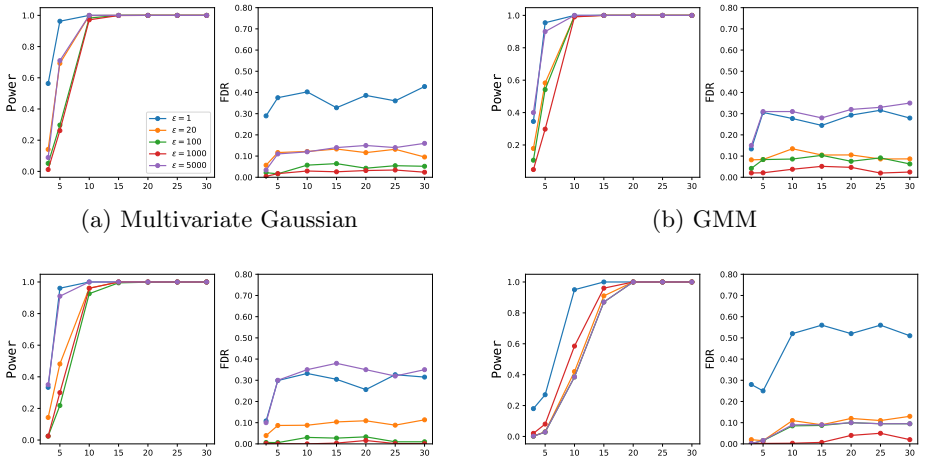
(d) Sparse Gaussian

Figure 10: Boxplot showing the quality of the knockoff generator w.r.t. ε. Lower MMD indicates better quality knockoffs compared to higher value of MMD.

In this section, we measure the GoF of knockoffs for each distributional setting considered in Section 5.2.2 w.r.t. different entropic regularization parameters ε. For a fixed covariate distribution P_X and corresponding conditional distribution $P_{\tilde{X}|X}$ of the knockoffs, we test the following hypothesis:

$$H_0^{\text{partial}} : P_{(X,\tilde{X})} = P_{(X,\tilde{X})_{\text{partial}(B)}},$$

where B is a random subset of $\{1, \ldots, d\}$, such that each $j \in B$ with probability $1/2$ independent of other elements. We draw n independent observations from $P_{(X,\tilde{X})}$ and $P_{(X,\tilde{X})_{\text{partial}(B)}}$ and generate two matrices $\mathbf{Z}, \mathbf{Z}' \in \mathbf{R}^{n \times 2d}$, $n = 200, d = 100$, respectively. Then, we compute



(a) Multivariate Gaussian (b) GMM

(c) Multivariate Student's T (d) Sparse Gaussian

Figure 11: Average FDR and power computed over 200 independent experiments are shown on the y-axis for each synthetic benchmark. The FDR level is set to 0.1. The x-axis represents the amplitude parameter ν.

an estimate of MMD to measure the GoF via:

$$\texttt{MMD}(\mathbf{Z}, \mathbf{Z}') \triangleq \frac{1}{n(n-1)} \sum_{i,j=1}^{n} k(Z_i, Z_j) - \frac{2}{n^2} \sum_{i,j=1}^{n,n} k(Z_i, Z'_j) + \frac{1}{n(n-1)} \sum_{i,j=1}^{n} k(Z'_i, Z'_j),$$

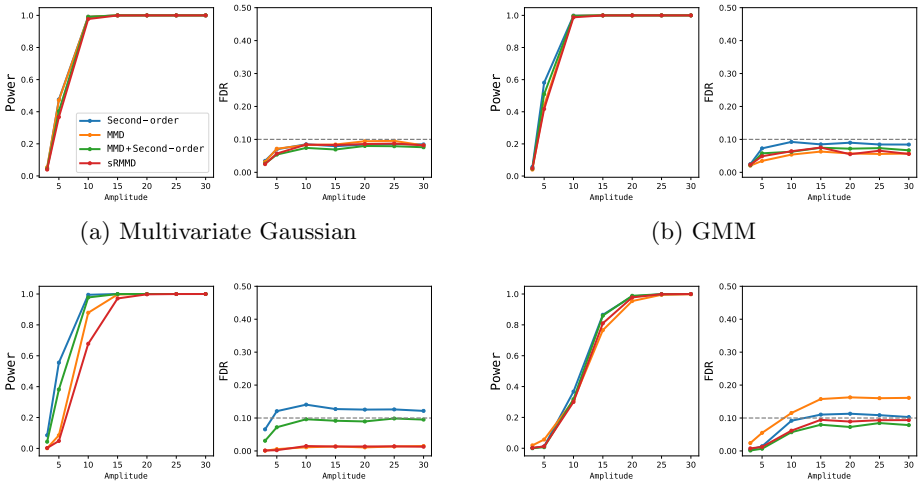
where $k(\cdot, \cdot)$ is a Gaussian mixture kernel with $\sigma = (1, 2, 4, 8, 16, 32, 64, 128)$. A small value of MMD indicates that the knockoffs achieve greater pairwise exchangeability (16) compared to the knockoffs that yield higher values of MMD.

Figure 10 shows the quality of the knockoffs produced by the sRMMD-based knockoff generator w.r.t. different ε. For each distributional setting, we observe that knockoffs generated with smaller entropic regularizer (e.g., $\varepsilon = 1, 20$) have poor quality, hence fail to control the FDR below the predefined level (Figure 11). As we increase the entropic regularizer (e.g., $\varepsilon = 100, 1000$), the quality of the knockoffs improves, and consequently FDR is controlled below the predefined level. However, as ε increases significantly, for example when $\varepsilon = 5000$, the quality of the knockoffs begins to degrade, and sRMMD loses its ability to control the FDR across most distributional settings.

F.5 Comparison Between Deep Knockoff and sRMMD Knockoff

The effect of adding the second-order term to the MMD knockoff, hence known as the Deep knockoff and its comparison with the sRMMD knockoffs are shown in Figure 12. In the case of multivariate Gaussian, GMM, and multivariate Student's t distributional settings, the addition of the second-order term to the MMD knockoff does not improve the FDR vs. power tradeoff to a great extent.

However, in the sparse Gaussian setting, a noticeable improvement is achieved in FDR control compared to the case where we do not add the second-order term to the MMD



(a) Multivariate Gaussian (b) GMM

(c) Multivariate Student's T (d) Sparse Gaussian

Figure 12: Average FDR and power computed over 200 independent experiments are shown on the y-axis for each synthetic benchmark. The FDR level is set to 0.1. The x-axis represents the amplitude parameter v.

knockoff (Figure 4). On the contrary, sRMMD knockoffs without the second-order term control the FDR and achieve either comparable or better FDR vs. power tradeoff than Deep knockoff in all cases.

References

Ryan Prescott Adams and Richard S. Zemel. Ranking via Sinkhorn propagation, 2011.

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J. Zico Kolter. Differentiable convex optimization layers. *Advances in neural information processing systems*, 32, 2019.

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085, 2015.

Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

Cristina Bauset, Laura Gisbert-Ferrándiz, and Jesús Cosín-Roger. Metabolomics as a promising resource identifying potential biomarkers for inflammatory bowel disease. *Journal of Clinical Medicine*, 10(4):622, 2021.

Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramer distance as a solution to biased Wasserstein gradients. *arXiv*, 2017.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959. PMLR, 2020.

Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*, 2016.

Renzo Caprilli, Monica Cesarini, Erika Angelucci, and Giuseppe Frieri. The long journey of salicylates in ulcerative colitis: The past and the future. *Journal of Crohn's and Colitis*, 3(3):149–156, 2009.

Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge–Kantorovich depth, quantiles, ranks and signs. *Annals of Statistics*, 45(1):223–256, 2017.

Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

Nabarun Deb and Bodhisattva Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, 118(541):192–207, 2023.

Richard Mansfield Dudley. The Speed of Mean Glivenko-Cantelli Convergence. *The Annals of Mathematical Statistics*, 40(1):40 – 50, 1969.

Patrick Emami and Sanjay Ranka. Learning permutations with Sinkhorn policy gradient. *arXiv preprint arXiv:1805.07010*, 2018.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

Andre Franke, Dermot PB McGovern, Jeffrey C Barrett, Kai Wang, Graham L Radford-Smith, Tariq Ahmad, Charlie W Lees, Tobias Balschun, James Lee, Rebecca Roberts, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics*, 42(12):1118–1125, 2010.

D.J. Fretland, D.L. Widomski, S. Levin, and T.S. Gaginella. Colonic inflammation in the rabbit induced by phorbol-12-myristate-13-acetate. *Inflammation*, 14(2):143–150, 1990.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

Yulia Gavrilov, Yoav Benjamini, and Sanat K Sarkar. An adaptive step-down procedure with proven FDR control under independence. *The Annals of Statistics*, 37(2):619–629, 2009.

Aude Genevay. *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres (ComUE), 2019.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Arthur Gretton, Karsten M Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Marc Hallin. On distribution and quantile functions, ranks and signs in $\mathbb{R}^d$. *ECARES Working Papers*, 2017.

Marc Hallin. Measure transportation and statistical decision theory. *Annual Review of Statistics and its Application*, 9, 2021.

Marc Hallin, Eustasio del Barrio, Juan Cuesta-Albertos, and Carlos Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139–1165, 2021.

Wassily Hoeffding. A non-parametric test of independence. *The Collected Works of Wassily Hoeffding*, pages 214–226, 1994.

Roswitha Hofer. On the distribution properties of Niederreiter–Halton sequences. *Journal of Number Theory*, 129(2):451–463, 2009.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166 – 1194, 2021.

James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Knockoffgan: Generating knockoffs for feature selection using generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

Fumitoshi Kuroki, Mitsuo Iida, Takayuki Matsumoto, Kunihiko Aoyagi, Kohki Kanamoto, and Masatoshi Fujishima. Serum n3 polyunsaturated fatty acids are depleted in Crohn's disease. *Digestive Diseases and Sciences*, 42(6):1137–1141, 1997.

Aonghus Lavelle and Harry Sokol. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nature Reviews Gastroenterology & Hepatology*, 17(4):223–237, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Thomas Lee, Thomas Clavel, Kirill Smirnov, Annemarie Schmidt, Ilias Lagkouvardos, Alesia Walker, Marianna Lucio, Bernhard Michalke, Philippe Schmitt-Kopplin, Richard Fedorak, et al. Oral versus intravenous iron replacement therapy distinctly alters the gut microbiota and metabolome in patients with IBD. *Gut*, 66(5):863–871, 2017.

Sophia R. Levan, Kelsey A. Stamnes, Din L. Lin, Ariane R. Panzer, Elle Fukui, Kathryn McCauley, Kei E. Fujimura, Michelle McKean, Dennis R. Ownby, Edward M. Zoratti, et al. Elevated faecal 12, 13-dihome concentration in neonates at high risk for asthma is produced by gut bacteria and impedes immune tolerance. *Nature Microbiology*, 4(11): 1851–1861, 2019.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017.

Gen Li, Yanxi Chen, Yuejie Chi, H Vincent Poor, and Yuxin Chen. Fast computation of optimal transport via entropy-regularized extragradient methods. *arXiv preprint arXiv:2301.13006*, 2023.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727. PMLR, 2015.

Ying Liu and Cheng Zheng. Auto-encoding knockoff generator for FDR controlled variable selection. *arXiv preprint arXiv:1809.10765*, 2018.

Jason Lloyd-Price, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, Nadim J Ajami, Kevin S Bonham, Colin J Brislawn, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, 2019.

Yiling Luo, Yiling Xie, and Xiaoming Huo. Improved rate of first order algorithms for entropic optimal transport. *arXiv preprint arXiv:2301.09675*, 2023.

Robert J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309–324, 1995.

Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

Gaspard Monge. Memory on the theory of cuttings and embankments. *Mem. Math. Phys. Acad. Royal Sci.*, pages 666–704, 1781.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch, 2017.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Jeff M Phillips and Suresh Venkatasubramanian. A gentle introduction to the kernel distance. *arXiv preprint arXiv:1103.1625*, 2011.

Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

Aram-Alexandre Pooladian, Marco Cuturi, and Jonathan Niles-Weed. Debiaser beware: Pitfalls of centering regularized transport maps. In *International Conference on Machine Learning*, pages 17830–17847. PMLR, 2022.

Xiaofa Qin. Etiology of inflammatory bowel disease: a unified hypothesis. *World Journal of Gastroenterology*, 18(15):1708, 2012.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

Philippe Rigollet and Austin J Stromme. On the sample complexity of entropic optimal transport. *arXiv preprint arXiv:2206.13472*, 2022.

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, 2020.

Sameh Saber, Rania M Khalil, Walied S Abdo, Doaa Nassif, and Eman El-Ahwany. Olmesartan ameliorates chemically-induced ulcerative colitis in rats via modulating NFκb and Nrf-2/HO-1 signaling crosstalk. *Toxicology and Applied Pharmacology*, 364:120–132, 2019.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. *arXiv preprint arXiv:1606.03498*, 2016.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55 (58-63):94, 2015.

Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my GAN? In *Proceedings of the European conference on computer vision (ECCV)*, pages 213–229, 2018.

Nikolai V. Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2):3–14, 1939.

Tiina Solakivi, Katri Kaukinen, Tarja Kunnas, Terho Lehtimäki, Markku Mäki, and Seppo Tapio Nikkari. Serum fatty acid profile in subjects with irritable bowel syndrome. *Scandinavian Journal of Gastroenterology*, 46(3):299–303, 2011.

Mukund Sudarshan, Wesley Tansey, and Rajesh Ranganath. Deep direct likelihood knockoffs. *arXiv preprint arXiv:2007.15835*, 2020.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Patrick J. Trainor, Andrew P. DeFilippis, and Shesh N. Rai. Evaluation of classifier performance for multiclass phenotype discrimination in untargeted metabolomics. *Metabolites*, 7(2):30, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Angelina Volkova and Kelly V. Ruggles. Predictive metagenomic analysis of autoimmune disease identifies robust autoimmunity and disease specific microbial signatures. *Frontiers in Microbiology*, 12:418, 2021.

Abraham Wald and Jacob Wolfowitz. On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162, 1940.

Frank Wilcoxon. Probability tables for individual comparisons by ranking methods. *Biometrics*, 3(3):119–122, 1947.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.