Bernoulli **30**(3), 2024, 1945–1970
https://doi.org/10.3150/23-BEJ1659

Asymptotics of discrete Schrödinger bridges via chaos decomposition

ZAID HARCHAOUI[1,a], LANG LIU[1,b] and SOUMIK PAL[2,c]

[1]*Department of Statistics, University of Washington, Seattle, United States,* [a]*zaid@uw.edu,* [b]*liu16@uw.edu*
[2]*Department of Mathematics, University of Washington, Seattle, United States,* [c]*soumik@uw.edu*

Consider the problem of matching two independent i.i.d. samples of size N from two distributions P and Q in $\mathbb{R}^d$. For an arbitrary continuous cost function, the optimal assignment problem looks for the matching that minimizes the total cost. We consider instead in this paper the problem where each matching is endowed with a Gibbs probability weight proportional to the exponential of the negative total cost of that matching. Viewing each matching as a joint distribution with N atoms, we then take a convex combination with respect to the above Gibbs probability measure. We show that this resulting random joint distribution converges, as $N \to \infty$, to the solution of a variational problem, introduced by Föllmer, called the Schrödinger problem. We also prove a limiting Gaussian fluctuation for this convergence in the form of central limit theorems for integrated test functions. This establishes a novel passage for the transition from discrete to continuum in Schrödinger's lazy gas experiment.

Keywords: Chaos decomposition; contiguity; entropy regularization; Hoeffding decomposition; infinite-order U-statistics; optimal matching; optimal transport; Schrödinger bridge

1. Introduction

Consider two probability distributions P and Q on $\mathbb{R}^d$. Let $\{X_i\}_{i\in[N]}$ and $\{Y_i\}_{i\in[N]}$ be two independent i.i.d. samples from P and Q, respectively, where $[N] := \{1,\ldots,N\}$. Consider a continuous *cost function* $c : \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty)$ such that $c(x,y) = 0$ if and only if $x = y$. Let $\mathcal{S}_N$ be the set of permutations of the set $[N] := \{1,2,\ldots,N\}$.

Every permutation can be viewed as a matching between the two sets of random variables. Choose an $\epsilon > 0$ whose significance will be made clear shortly. Suppose we weigh every permutation σ by the (random) weight $w(\sigma) := \exp(-\sum_{i=1}^N c(X_i, Y_{\sigma_i})/\epsilon)$. That is, define a Gibbs measure on $\mathcal{S}_N$,

$$q_\epsilon^*(\sigma) := \frac{w(\sigma)}{\sum_{\tau\in\mathcal{S}_N} w(\tau)} = \frac{\exp\left(-\sum_{i=1}^N c(X_i, Y_{\sigma_i})/\epsilon\right)}{\sum_{\tau\in\mathcal{S}_N} \exp\left(-\sum_{i=1}^N c(X_i, Y_{\tau_i})/\epsilon\right)}, \quad \sigma \in \mathcal{S}_N. \tag{1}$$

Now mix all possible matchings with probabilities given by q_ϵ^* by defining

$$\hat{\mu}_\epsilon^N := \sum_{\sigma\in\mathcal{S}_N} q_\epsilon^*(\sigma) \frac{1}{N} \sum_{i=1}^N \delta_{(X_i, Y_{\sigma_i})}. \tag{2}$$

The random measure $\hat{\mu}_\epsilon^N$ is a joint distribution with marginals given by the two empirical distributions $\hat{P}^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$ and $\hat{Q}^N = \frac{1}{N}\sum_{i=1}^N \delta_{Y_i}$. It is obtained by a convex combination of all possible matchings of atoms. A high cost for a matching results in an exponentially small weight. This paper deals with the limiting behavior of the sequence of random measures $\hat{\mu}_\epsilon^N$ as $N \to \infty$ while $\epsilon > 0$ is

1350-7265 © 2024 ISI/BS

fixed. Concretely, we show that, as $N \to \infty$, $\hat{\mu}_\epsilon^N$ converges weakly to, and has Gaussian fluctuations around, the solution μ_ϵ of the variational problem

$$\mathbf{C}_\epsilon(P,Q) := \min_{\nu \in \Pi(P,Q)} \left[\int c(x,y) d\nu(x,y) + \epsilon \mathrm{KL}(\nu \mid P \otimes Q) \right], \tag{3}$$

where $\Pi(P,Q)$ is the set of *couplings* of (P,Q), i.e., all joint probability distributions over $\mathbb{R}^d \times \mathbb{R}^d$ with marginals given by P and Q, and $\mathrm{KL}(\nu|P \otimes Q) := \int \log \frac{d\nu}{d(P\otimes Q)} d\nu$ if $\nu \ll P \otimes Q$ and infinity otherwise is the Kullback-Leibler divergence. Due to [12,58], the solution μ_ϵ satisfies the following equation: there exist two measurable functions a_ϵ and b_ϵ such that

$$\frac{d\mu_\epsilon}{d(P \otimes Q)}(x,y) = \xi(x,y) := \exp\left[-\frac{1}{\epsilon}\left(c(x,y) - a_\epsilon(x) - b_\epsilon(y)\right) \right]. \tag{4}$$

Schrödinger bridges. The measure μ_ϵ can be viewed as the (static) Schrödinger bridge [11,22,38,59] connecting P to Q at temperature ϵ. Assume that the following Markov transition kernel density is well-defined:

$$p_\epsilon(y \mid x) \propto \exp\left[-\frac{1}{\epsilon} c(x,y) \right].$$

This defines a Markov chain. Suppose (W_0, W_1) is distributed according to this Markov chain, conditioned on "$W_0 \sim P$ and $W_1 \sim Q$". The joint law of (W_0, W_1) is called the Schrödinger bridge connecting P to Q at temperature ϵ. The quoted statement is not an event and is non-trivial to make precise. In continuum, when both P and Q are densities, the Schrödinger bridge can be made precise as the solution of the problem called *the Schrödinger problem* [22,38,59]

$$\min_{\nu \in \Pi(P,Q)} \left[\int c(x,y) d\nu(x,y) + \epsilon H(\nu) \right], \tag{5}$$

where H is the entropy defined as $H(\nu) := \int \nu(x,y) \log \nu(x,y) dx dy$ if ν is a density and infinity otherwise. We mention here two surveys [11,39] on this problem. Since this problem and the problem (3) share the same solution, we call μ_ϵ the *Schrödinger bridge*.

In the same spirit, the random measure $\hat{\mu}_\epsilon^N$ can also be interpreted as the Schrödinger bridge connecting two empirical measures $\hat{P}^N$ and $\hat{Q}^N$ at temperature ϵ. In this interpretation $\hat{\mu}_\epsilon^N$ first appeared in [49, Section 3.2] for a particular cost function. To see this, let $X_i = x_i$ and $Y_i = y_i$ for $i \in [N]$. Then $\hat{P}^N$ and $\hat{Q}^N$ are discrete distributions each supported on exactly N atoms. Imagine N independent Markov chains (or particles) $W(1), \dots, W(N)$, starting from positions $\{W_0(i) = x_i\}_{i=1}^N$, make jumps according to the Markov kernel $\{p_\epsilon(\cdot \mid x_i)\}_{i=1}^N$, respectively. Let $L^N(1) := \frac{1}{N} \sum_{i=1}^N W_1(i)$ denote the empirical distribution of their terminal values and let $L^N(0,1) = \frac{1}{N} \sum_{i=1}^N \delta_{(W_0(i), W_1(i))}$ denote the joint empirical distribution at two time points. The law of $L^N(0,1)$, conditioned on $L^N(1) = \hat{Q}^N$, is given by the mixture formula $\hat{\mu}_\epsilon^N$ in (2) (given $X_i = x_i$ and $Y_i = y_i$ for $i \in [N]$), which solves Schrödinger's problem in the discrete set-up. We refer to $\hat{\mu}_\epsilon^N$ as the *discrete Schrödinger bridge*.

Partition functions in quantum thermodynamics. Although weighted averages of symmetrized empirical distributions (2) and their variations go way back to Feynman's work [20], such quantities also appeared recently in several different contexts. Motivated by the quantum thermodynamics of N non-interacting Boson particles, a variation of (2) where $Y_i = X_i$ for every i has been considered [1–3]. In this setting, the samples are obviously dependent and $P = Q$. One of the goals of these articles is to compute the trace of the exponential of an N particle Hamilton operator for Bose-Einstein statistics. In their language, it can be described as the limit

$$\lim_{N\to\infty}\frac{1}{N}\log\left[\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\exp\left(-\frac{1}{\epsilon}\sum_{i=1}^{N}c(X_i,Y_{\sigma_i})\right)\right]=-\frac{1}{\epsilon}\mathbf{C}_\epsilon(P,Q). \tag{6}$$

The term inside the log is called the partition function and is the denominator appearing in (1) scaled by $N!$. The marginal measure P comes from a Feyman-Kac representation of the trace operator and is taken to be either the uniform density over a compact box or the Lebesgue measure on the entire $\mathbb{R}^d$ in which case it fails to be a probability measure. In a similar vein of work, Trashorras [63] considers the case where $X_i = Y_i = x_i$, $i \in [N]$, are deterministic points such that its empirical measure $\frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ converges weakly to $P = Q$ as $N \to \infty$. If a random permutation σ is chosen uniformly from $\mathcal{S}_N$, one gets a random measure $\frac{1}{N}\sum_{i=1}^N \delta_{(x_i,x_{\sigma_i})}$ which is referred to as the *symmetrized empirical measure*. In [63], a Large Deviation Principle for this sequence of random measures is derived, recovering the limit in (6). One of our key results (Corollary 3) establishes the limit in (6) in the case of independent i.i.d. samples. In fact, this result is obtained from a stronger result (Theorem 2) which gives the exact limit of a scaled version of $(N!)^{-1}\sum_{\sigma\in\mathcal{S}_N}\exp\left(-\epsilon^{-1}\sum_{i=1}^n c(X_i,Y_{\sigma_i})\right)$ via a markedly different proof technique as discussed in Section 1.1. This result can be of independent interest to the literature mentioned above.

Mallows models of random permutations. The Gibbs measure q_ϵ^* itself appears in a more recent work in an entirely different direction studying the limit of Mallows-type models of random permutations [43]. This is done in [47] where the interest is in statistical estimation on Mallows models and in a very recent paper [33] on scaling limits of large random permutations with fixed patterns. In [47, Theorem 1.5] the author obtained the limit (6) for $P = Q = \mathrm{Unif}(0,1)$ in the setting when $X_i = Y_i = i/N$, $i \in [N]$, are deterministic. In this case, the empirical measure $\frac{1}{N}\sum_{i=1}^N \delta_{X_i}$ can be viewed as a deterministic approximation of $\mathrm{Unif}(0,1)$.

Optimal transport and entropic regularization. As shown in [38], when $\epsilon \to 0$, the Schrödinger problem recovers the Monge-Kantorovich optimal transport (OT) problem defined as

$$\mathbf{C}(P,Q)=\inf_{\nu\in\Pi(P,Q)}\int c(x,y)\nu(dxdy). \tag{7}$$

Since the data points are sampled from densities, they are all distinct almost surely. In this case, the empirical measures $\hat{P}^N$ and $\hat{Q}^N$ are discrete measures supported on N atoms. The plug-in estimator $\mathbf{C}(\hat{P}^N,\hat{Q}^N)$ can then be formulated as the linear program

$$\mathbf{C}(\hat{P}^N,\hat{Q}^N)=\min_{M\in\Pi(N^{-1}\mathbf{1},N^{-1}\mathbf{1})}\langle M,C\rangle, \tag{8}$$

where $\Pi(N^{-1}\mathbf{1},N^{-1}\mathbf{1})\subset\mathbb{R}^{N\times N}$ is the set of matrices such that $M\mathbf{1}=M^\top\mathbf{1}=N^{-1}\mathbf{1}$, i.e., NM is doubly stochastic, and $\langle M,C\rangle := \sum_{i=1}^N\sum_{j=1}^N c(X_i,Y_j)M_{ij}$.

The limiting behavior of $\mathbf{C}(\hat{P}^N,\hat{Q}^N)$ towards $\mathbf{C}(P,Q)$ has been studied in combinatorics [4], probability and statistics [23,37,61,68], and applied to economics [24,36]. This problem also arises in nonparametric statistical hypothesis testing [51] where one tests for the null hypothesis $P = Q$ by checking whether $\mathbf{C}(\hat{P}^N,\hat{Q}^N)\approx 0$. This, among other reasons, has spurred a recent interest in the study of asymptotic distributions of $\mathbf{C}(\hat{P}^N,\hat{Q}^N)$, properly scaled with respect to $\mathbf{C}(P,Q)$.

Early works on the large sample behavior of the OT cost were focused on the well-behaved quadratic cost $c(x,y)=|x-y|^2$ ($\sqrt{\mathbf{C}(P,Q)}$ is then called the Wasserstein-2 distance between P and Q) on the real line $\mathbb{R}$; see, e.g. [14,15,48]. These results were built upon the explicit characterization, given by quantile functions, of the Wasserstein distances on measures supported on $\mathbb{R}$. Beyond one dimension, similar results are rather challenging to obtain; see [4,17] for almost sure convergence results. In [56], the authors obtained the limiting law of Wasserstein distances between Gaussian distributions with parameters estimated from data by utilizing the d-form representation in this special case. Recently, normal distributional results have been generalized to $\mathbb{R}^d$ for the quadratic cost [16] and for a general cost on compact domains [32]. Wasserstein distances between discrete probability measures supported on a finite [34,60] and countable [62] metric space have also been investigated.

An entropy-regularized formulation of (8) is particularly attractive both from a computational viewpoint [13] and from a statistical viewpoint [55]. Cuturi [13] defined the following entropy-regularized optimal transport (EOT) problem:

$$\min_{M \in \Pi(N^{-1}\mathbf{1}, N^{-1}\mathbf{1})} [\langle M, C\rangle + \epsilon \mathrm{Ent}(M)], \tag{9}$$

where $\epsilon > 0$ is the regularization parameter and $\mathrm{Ent}(M) = \sum_{i=1}^N \sum_{j=1}^N M_{ij} \log M_{ij}$ is the entropy of M; see also [19]. The solution, although non-explicit, can be efficiently computed using the Sinkhorn algorithm [50, Section 4.2]. Let M_ϵ^N denote the (unique) optimal solution to (9), then the limit behavior of M_ϵ^N and, in particular, the *regularized* cost of transport $\langle C, M_\epsilon^N\rangle$, both as $N \to \infty$ and ϵ either fixed or decaying to zero, becomes important. In fact, M_ϵ^N can be viewed as the plug-in estimator of μ_ϵ since the minimizer of the problem (3) with P and Q replaced by $\hat{P}^N$ and $\hat{Q}^N$ is exactly, in its matrix form, M_ϵ^N. For finite state spaces and $c(x,y) = \|x-y\|^p$ with $p \ge 1$, this has been taken up in [35]. The slightly different but related concept of Sinkhorn divergence has been studied in [9] and later extended in [44] to Euclidean spaces for $p = 2$.

The discrete Schrödinger bridge $\hat{\mu}_\epsilon^N$ is, in fact, the solution of a different discrete EOT problem which explains the surprising appearance of entropy in the limit (5). For a permutation $\sigma \in \mathcal{S}_N$, let A_σ denote the permutation matrix corresponding to σ. By Birkhoff's Theorem [6, Theorem 5.2], every doubly stochastic matrix can be written as a convex combination of permutation matrices. Thus, every coupling M can be expressed as $M = \sum_{\sigma \in \mathcal{S}_N} q_M(\sigma) \frac{1}{N} A_\sigma$, where $q_M(\sigma) \in \mathcal{P}(\mathcal{S}_N)$ is a probability distribution on $\mathcal{S}_N$. Such convex combinations are generally not unique. Nevertheless, for any $q \in \mathcal{P}(\mathcal{S}_N)$, we can get an element in $\Pi(N^{-1}\mathbf{1}, N^{-1}\mathbf{1})$ by defining $M_q := \sum_{\sigma \in \mathcal{S}_N} q(\sigma) \frac{1}{N} A_\sigma$. Moreover, it holds that $\langle M_q, C\rangle = \frac{1}{N} \sum_{\sigma \in \mathcal{S}_N} q(\sigma) \sum_{i=1}^N c(X_i, Y_{\sigma_i})$. For $q \in \mathcal{P}(\mathcal{S}_N)$ we define the entropy of q as $\mathrm{Ent}(q) := \sum_{\sigma \in \mathcal{S}_N} q(\sigma) \log(q(\sigma))$. Consider the problem

$$\min_{q \in \mathcal{P}(\mathcal{S}_N)} \left[\langle M_q, C\rangle + \frac{\epsilon}{N} \mathrm{Ent}(q)\right]. \tag{10}$$

This is a regularization of discrete OT with a different notion of entropy for a doubly stochastic matrix M. We show in Appendix A of the Supplementary Material [30] that the solution to (10) is exactly q_ϵ^* in (1).

The relationship between M_ϵ^N that solves (9) and the matrix $M_{q_\epsilon^*}$ where $q_\epsilon^\star$ solves (10) is not obvious. However, they are connected through the lens of matrix balancing; see [7] and references therein. To see this, we define an $N \times N$ matrix K with (i,j)-th element being $K_{ij} := \exp\left(-\frac{1}{\epsilon} c(X_i, Y_j)\right)$. Let $|K|$ denote the permanent of K, i.e.,

$$|K| = \sum_{\sigma \in \mathcal{S}_N} \prod_{i=1}^N K_{i\sigma_i} = \sum_{\sigma \in \mathcal{S}_N} \exp\left(-\frac{1}{\epsilon} \sum_{i=1}^N c(X_i, Y_{\sigma_i})\right),$$

which is exactly the denominator in (1). Notice that

$$\left(M_{q_\epsilon^*}\right)_{i,j} = \frac{1}{N} \sum_{\sigma:\sigma_i=j} q_\epsilon^*(\sigma) = \frac{1}{N} \frac{\sum_{\sigma:\sigma_i=j} \exp\left(-\sum_{i=1}^N c(X_i, Y_{\sigma_i})/\epsilon\right)}{\sum_{\sigma\in\mathcal{S}_N} \exp\left(-\sum_{i=1}^N c(X_i, Y_{\sigma_i})/\epsilon\right)}.$$

The sum in the numerator is over all permutations $\sigma \in \mathcal{S}_N$ such that $\sigma_i = j$. A little bit of algebra omitted here shows that it is exactly given by $N \exp(-c(X_i, Y_j)/\epsilon)\left|K^{ij}\right|$, where K^{ij} is the minor of K obtained by deleting the ith row and the jth column of the matrix K. Therefore, we get the neat formula $(M_{q_\epsilon^*})_{i,j} = K_{ij}\left|K^{ij}\right|/|K|$. The matrix $M_{q_\epsilon^*}$ is referred to as the *matrix balance* of K [7, Section 3] while the matrix M_ϵ^N is called the *Sinkhorn balance* [7, Section 4]. It is shown in [7, Section 4.1] that the Sinkhorn balance of a 0-1 matrix approximates the matrix balance of it. However, a more in-depth investigation on the relationship of these two objects is needed.

1.1. Main results

We now state our main results regarding the limiting behavior of the discrete Schrödinger bridge where both the dimension d and regularization parameter ϵ are kept fixed. Given a probability measure ν and integer $p \geq 1$, let $\mathbf{L}^p(\nu)$ be the space of functions that have finite p-th norm under ν. We shall keep the same notation for an absolutely continuous measure and its density.

We express our results in their full generality. Let $\mu \in \Pi(P, Q)$ be absolutely continuous w.r.t. $P \otimes Q$ with density $\xi \in \mathbf{L}^1(P \otimes Q)$. Define the random measure

$$\hat{\mu}^N := \frac{\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N} \frac{1}{N}\sum_{i=1}^N \delta_{(X_i, Y_{\sigma_i})} \xi^\otimes(X, Y_\sigma)}{\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N} \xi^\otimes(X, Y_\sigma)}, \tag{11}$$

where $\xi^\otimes(X, Y_\sigma) := \prod_{i=1}^N \xi(X_i, Y_{\sigma_i})$. As a special case, recall from (4) that, if $\xi(x,y)$ is chosen to be $\exp\left(-(c(x,y) - a_\epsilon(x) - b_\epsilon(y))/\epsilon\right)$, then $\mu = \mu_\epsilon$ is the Schrödinger bridge connecting P to Q. Moreover, $\hat{\mu}^N$ recovers the measure defined in (2). Our first result shows that the random measure $\hat{\mu}^N$ converges weakly to its continuous counterpart μ. Let us start by defining two operators on $\mathbf{L}^2(P)$ and $\mathbf{L}^2(Q)$ induced by μ.

Definition 1. Define linear operators $\mathcal{A} : \mathbf{L}^2(P) \to \mathbf{L}^2(Q)$ and its adjoint $\mathcal{A}^* : \mathbf{L}^2(Q) \to \mathbf{L}^2(P)$ by

$$(\mathcal{A}f)(y) = \int f(x)\xi(x,y)dP(x) \quad \text{and} \quad (\mathcal{A}^*g)(x) = \int g(y)\xi(x,y)dQ(y). \tag{12}$$

Call $A : (x,y) \mapsto \xi(x,y)$ the *kernel* of $\mathcal{A}$ and $A^* : (y,x) \mapsto \xi(x,y)$ the kernel of $\mathcal{A}^*$.

We show in Lemma 9 that $\mathcal{A}$ is a well-defined linear operator, and $\mathcal{A}^*\mathcal{A}$ and $\mathcal{A}\mathcal{A}^*$ are two Markov operators defined on $\mathbf{L}^2(P)$ and $\mathbf{L}^2(Q)$, respectively. Moreover, they can be rewritten as two conditional expectations: $(\mathcal{A}f)(y) = \mathbb{E}[f(X) \mid Y](y)$ and $(\mathcal{A}^*g)(x) = \mathbb{E}[g(Y) \mid X](x)$ where $(X,Y) \sim \mu$.

Assumption 1. All the results stated below hold under the following assumptions.

1. $\xi \in \mathbf{L}^2(P \otimes Q)$. As a consequence [8, Appendix A.4], the operator $\mathcal{A}$ is compact. Then the operators $\mathcal{A}^*\mathcal{A}$ and $\mathcal{A}\mathcal{A}^*$ admit eigenvalue decomposition $\mathcal{A}^*\mathcal{A}\alpha_k = s_k^2\alpha_k$ and $\mathcal{A}\mathcal{A}^*\beta_k = s_k^2\beta_k$ for all $k \geq 0$ with $s_0 = 1$, $\alpha_0 = \beta_0 = \mathbf{1}$ and $0 \leq s_k \leq 1$ for all $k \geq 0$. Moreover, it holds that $\mathcal{A}\alpha_k = s_k\beta_k$ and $\mathcal{A}^*\beta_k = s_k\alpha_k$; see [26, Chapter 6.1]. We call $\{s_k\}_{k\geq 0}$ the singular values of $\mathcal{A}$ and $\mathcal{A}^*$, and call $\{\alpha_k\}_{k\geq 0}$ and $\{\beta_k\}_{k\geq 0}$ the singular functions.

2. The operators $\mathcal{A}^*\mathcal{A}$ and $\mathcal{A}\mathcal{A}^*$ have positive eigenvalue gap, *i.e.*, $s_k \le s_1 < 1$ for all $k \ge 1$. By Jentzsch's Theorem [57, Theorem 7.2], a sufficient condition is that ξ is bounded.

Theorem 1. *As $N \to \infty$, $\hat{\mu}^N$ converges weakly to μ, in probability.*

Towards the proof of Theorem 1, a critical result is the limit law of the denominator in (11) which is denoted as D_N. We state it here since it is of independent interest.

Theorem 2. *As $N \to \infty$, the denominator in (11) has the following limiting distribution:*

$$D_N \to_d D := \frac{1}{\sqrt{\prod_{k=1}^{\infty}(1-s_k^2)}} \exp\left\{\frac{1}{2}\sum_{k=1}^{\infty}\left[-\frac{s_k^2}{1-s_k^2}(U_k^2+V_k^2)+\frac{2s_k}{1-s_k^2}U_kV_k\right]\right\}, \tag{13}$$

where $\{U_k\}_{k\ge1}$ and $\{V_k\}_{k\ge1}$ are independent standard normal random variables.

It is noteworthy that D_N is a two-sample U-statistic of infinite order—a generalization of classical U-statistics introduced by Halmos [29] and Hoeffding [31], where the kernel of the U-statistic depends on the sample size. Infinite-order U-statistics were first considered in [28] as a special class of elementary symmetric polynomials of random variables; see also [42,46,64,65] in this line of research. The limiting distribution of general infinite-order U-statistics was obtained in [18, Theorem 1] using randomization of the sample size and multiple Wiener integrals. Theorem 2 extends previous work on one-sample infinite-order U-statistics to *two-sample* infinite-order U-statistics.

Another closely related topic is the asymptotics of random permanents; see the monograph [54] for a review. An elementary symmetric polynomial is the permanent of a random matrix with identical rows [53, Page 2]. The limiting behavior of general random permanents has been studied in the case of i.i.d. entries [52] as well as independent columns [53], where the limit law is the exponential of a Gaussian distribution. The denominator D_N can be viewed as the permanent of the random matrix $(\xi(X_i,Y_j))_{N\times N}$ scaled by $N!$. Hence, Theorem 2 characterizes the asymptotic behavior of the permanent of a random matrix induced by a bivariate function whose rows and columns are dependent—the limit law is given by the exponential of a weighted sum of products of Gaussians.

If we set $\xi(x,y) := \exp(-(c(x,y)-a_\epsilon(x)-b_\epsilon(y))/\epsilon)$, then Theorem 2 yields the limit in (6).

Corollary 3. *As $N \to \infty$, the denominator in (1) has the following limit:*

$$\frac{1}{N}\log\left[\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\exp\left(-\frac{1}{\epsilon}\sum_{i=1}^{N}c(X_i,Y_{\sigma_i})\right)\right] \to_p -\frac{1}{\epsilon}\mathbf{C}_\epsilon(P,Q).$$

To conduct a more refined analysis of the convergence of $\hat{\mu}^N$, we let η be any function on $\mathbb{R}^d\times\mathbb{R}^d$ integrable under μ and consider the convergence of $T_N := T_N(\eta) := \int \eta(x,y)d\hat{\mu}^N(x,y)$ towards $\theta := \int \eta(x,y)d\mu(x,y)$. According to (11),

$$T_N = \frac{\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\frac{1}{N}\sum_{i=1}^{N}\eta(X_i,Y_{\sigma_i})\xi^{\otimes}(X,Y_\sigma)}{\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\xi^{\otimes}(X,Y_\sigma)}. \tag{14}$$

A particularly important example is when $\eta = c$ is the cost function and μ is the Schrödinger bridge. In this case θ is the optimal cost of transport for the regularized problem defined in (3), which is known as the *Sinkhorn distance* [13]. It can be viewed as an approximation to the unregularized optimal transport cost with a convergence rate decaying exponentially in ϵ [41]. On the other hand, most of the previous works consider the optimal value of the problem (3) since their analyses rely heavily on the duality. Moreover, as demonstrated in [40, Chapter 4], the statistic T_N can be used to statistically test for the equality of distributions of two independent samples.

The statistic T_N is a rather complicated function of the two empirical measures $(\hat{P}^N, \hat{Q}^N)$. Our next result shows that it can be well approximated by linear functions of the two measures in a way that is similar to the first order term in a Taylor expansion of smooth functions.

Assumption 2. All the results stated below hold under the following additional assumptions: $\eta^2\xi \in \mathbf{L}^1(P \otimes Q)$ and $\eta\xi \in \mathbf{L}^2(P \otimes Q)$.

We denote by $I_\nu : \mathbf{L}^2(\nu) \to \mathbf{L}^2(\nu)$ the identity operator on $\mathbf{L}^2(\nu)$, and, by convention, its kernel is given by the Dirac delta function. When the context is clear, we will write I for short. Define

$$\eta_{1,0}(x) := \int [\eta(x,y) - \theta]\xi(x,y)dQ(y) \quad \text{and} \quad \eta_{0,1}(y) := \int [\eta(x,y) - \theta]\xi(x,y)dP(x). \tag{15}$$

Theorem 4. *As $N \to \infty$, it holds that $T_N - \theta = \mathcal{L}_1 + o_p\big(1/\sqrt{N}\big)$, where*

$$\mathcal{L}_1 := \frac{1}{N}\sum_{i=1}^{N}[(I - \mathcal{A}^*\mathcal{A})^{-1}(\eta_{1,0} - \mathcal{A}^*\eta_{0,1})(X_i) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(\eta_{0,1} - \mathcal{A}\eta_{1,0})(Y_i)].$$

We call $\mathcal{L}_1$ the first order chaos of T_N.

Corollary 5. *As $N \to \infty$, the sequence $\sqrt{N}(T_N - \theta)$ converges in law to $\mathcal{N}(0, \varsigma^2)$, where $\varsigma^2 = \varsigma^2(\eta)$, as a function of η, is given by*

$$\varsigma^2 := \int \Big((I - \mathcal{A}^*\mathcal{A})^{-1}(\eta_{1,0} - \mathcal{A}^*\eta_{0,1})(x)\Big)^2 dP(x) + \int \Big((I - \mathcal{A}\mathcal{A}^*)^{-1}(\eta_{0,1} - \mathcal{A}\eta_{1,0})(y)\Big)^2 dQ(y).$$

Remark 1. In the arXiv version of this paper (arXiv:2011.08963) we conjectured that the same CLT holds for the solution of the EOT problem (9). This conjecture has been recently verified in [27].

Remark 2. It has been shown in [38,45] that the Schrödinger bridge problem recovers the Monge-Kantorovich OT problem as $\epsilon \to 0$. It is of great interest to verify if the limiting variance ς^2 in Corollary 5 converges to the limiting variance of the OT plan.

Remark 3. When the limiting variance $\varsigma^2 = 0$, we can also establish the second order chaos of T_N and the limiting distribution of $N(T_N - \theta)$. We refer interested readers to [40, Appendix C.5].

The first order chaos $\mathcal{L}_1$ admits a more compact expression using the notion of *tensor products*. Let $\mathcal{A}_1 \in \{\mathcal{A}, \mathcal{A}^*, I_P, I_Q\}$ be an operator mapping from $\mathbf{L}^2(\nu_1)$ to $\mathbf{L}^2(\gamma_1)$ with kernel A_1. And define $\mathcal{A}_2, A_2$ similarly. The tensor product $\mathcal{A}_1 \otimes \mathcal{A}_2 : \mathbf{L}^2(\nu_1 \otimes \nu_2) \to \mathbf{L}^2(\gamma_1 \otimes \gamma_2)$ is defined by

$$(\mathcal{A}_1 \otimes \mathcal{A}_2)f(\nu_1, \nu_2) := \iint f(\nu_1', \nu_2')A_1(\nu_1', \nu_1)A_2(\nu_2', \nu_2)d\nu_1(\nu_1')d\nu_2(\nu_2'), \text{ for all } f \in \mathbf{L}^2(\nu_1 \otimes \nu_2).$$

For instance, $I_P \otimes \mathcal{A} : \mathbf{L}^2(P \otimes P) \to \mathbf{L}^2(P \otimes Q)$ is defined by

$$(I_P \otimes \mathcal{A}) f(v_1, v_2) := \iint f(v_1', v_2') \delta_{v_1}(v_1') \xi(v_2', v_2) dP(v_1') dP(v_2') = \int f(v_1, v_2') \xi(v_2', v_2) dP(v_2'),$$

or as a conditional expectation: $(I_P \otimes \mathcal{A}) f(v_1, v_2) = \mathbb{E}[f(X', X) \mid X', Y](v_1, v_2)$ where $(X, Y) \sim \mu$ is independent of X'. In particular, when $f := f_1 \oplus f_2$, we have $(\mathcal{A}_1 \otimes \mathcal{A}_2)(f_1 \oplus f_2)(v_1, v_2) = \mathcal{A}_1 f_1(v_1) + \mathcal{A}_2 f_2(v_2)$. Finally, define the *swap* operator $\mathcal{T}$ by $\mathcal{T} f(u, v) = f(v, u)$ for any f on $\mathbb{R}^d \times \mathbb{R}^d$. It is clear that $\mathcal{T}(\mathcal{A}_1 \otimes \mathcal{A}_2) = (\mathcal{A}_2 \otimes \mathcal{A}_1)\mathcal{T}$ on $\mathbf{L}^2(\nu_1 \otimes \nu_2)$.

Definition 2. Define the operator $\mathcal{B}$ on the space $\mathbf{L}^2(P \otimes Q)$ as $\mathcal{B} := \mathcal{T}(\mathcal{A} \otimes \mathcal{A}^*) = (\mathcal{A}^* \otimes \mathcal{A})\mathcal{T}$.

With this new operator $\mathcal{B}$, the first order chaos $\mathcal{L}_1$ can be rewritten as (Corollary 12)

$$\mathcal{L}_1 = \frac{1}{N} \sum_{i=1}^{N} (I + \mathcal{B})^{-1} (\eta_{1,0} \oplus \eta_{0,1})(X_i, Y_i).$$

Both expressions of $\mathcal{L}_1$ come from the following system of linear equations. Assume the first order chaos in Theorem 4 is given by $\frac{1}{N} \sum_{i=1}^{N} [f(X_i) + g(Y_i)]$, then f and g are (almost surely) solutions to $\eta_{1,0} = f + \mathcal{A}^* g$ and $\eta_{0,1} = \mathcal{A} f + g$.

1.2. Outline of the paper

Section 2 is devoted to proving Theorem 1. We prove a novel *contiguity* result that allows us to change the model to $\{(X_i, Y_i)\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \mu$ based on the limiting distribution of the denominator in Theorem 2. This change of measure enables a more natural analysis for $\hat{\mu}^N$ and Theorem 1 then follows from the reverse martingale convergence theorem.

In Section 3 we derive the first order approximation of T_N and prove Theorem 4 by a variance bound of the remainder. We show that this approximation is the first order chaos of T_N under the change of measure μ. Each term in the chaos expansion is a polynomial function of the empirical distributions $(\hat{P}^N, \hat{Q}^N)$, which are symmetric under permutations of X_i's or Y_i's, separately. Thus, we obtain symmetric projections on subspaces of $\mathbf{L}^2(\mu^N)$ when X_i and Y_i, under the change of measure, are not independent. Essentially, we extend the classical Hoeffding projection to paired samples, which can be of independent interest.

In Section 4 we derive the asymptotic distribution of the denominator and the variance bound of the remainder used in the previous two sections. The method here is based on a Hoeffding-like decomposition and new variance bounds for a type of U-statistic of increasing order under our original model when X_i and Y_i are independent. The tools developed in this section can also be of independent interest. For readability, we give in Appendix C of the Supplementary Material [30] a table of notation.

2. Weak convergence and contiguity

In this section, we prove the weak convergence of $\hat{\mu}^N$. By definition, it suffices to show the convergence of $T_N := \int \eta d\hat{\mu}^N$ to $\theta := \int \eta d\mu$ for any continuous bounded function η. In fact, the convergence holds for every function η that is integrable under μ.

Recall from (14) that T_N admits a complicated expression, i.e.,

$$T_N = \frac{\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\frac{1}{N}\sum_{i=1}^N \eta(X_i,Y_{\sigma_i})\xi^{\otimes}(X,Y_\sigma)}{\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\xi^{\otimes}(X,Y_\sigma)}.$$

However, it has a rather simple structure under a change of measure—instead of assuming that $\{(X_i,Y_i)\}_{i=1}^N$ is an i.i.d. sample from the product measure $P\otimes Q$, we assume that $\{(X_i,Y_i)\}_{i=1}^N$ is an i.i.d. sample from μ. As Proposition 7 below shows, under this change of measure, T_N is a simple conditional expectation and an unbiased estimator of θ. Hence, it is natural to ask if there is a way to do analysis under the changed measure μ and carry the results over to the original measure $P\otimes Q$. Contiguity [66, Chapter 6] is exactly a tool for such purposes. When $\xi\neq 1$ a.s. under $P\otimes Q$, the laws of the entire i.i.d. sequence $\{(X_i,Y_i)\}_{i\geq 1}$ under the two measures $P\otimes Q$ and μ are singular. But T_N is a function of only $(\hat{P}^N,\hat{Q}^N)$. Restricted to the σ-algebra generated by these marginal empirical distributions, we show that the two measures are contiguous in Theorem 6 below.

We first set-up a measure-theoretic framework. We use the term "under the measure γ" to indicate that the sample $\{(X_i,Y_i)\}_{i=1}^N \overset{i.i.d.}{\sim} \gamma$ and use $\mathbb{E}_\gamma$ to denote the expectation under this model. When $\gamma = P\otimes Q$, we write $\mathbb{E}$ for short. Let $\mathcal{F}_N$ denote the σ-algebra generated by $\{(X_i,Y_i)\}_{i=1}^N$. Let $\mathcal{G}_N$ denote the sub-σ-algebra of $\mathcal{F}_N$ generated by $(\hat{P}^N,\hat{Q}^N)$. Let R^N and S^N be the law of $(\hat{P}^N,\hat{Q}^N)$ under $P\otimes Q$ and μ, respectively. It is clear that $R^N = (P\otimes Q)^N|_{\mathcal{G}_N}$ and $S^N = \mu^N|_{\mathcal{G}_N}$.

According to Le Cam's first lemma [66, page 88], the contiguity holds true if the likelihood ratio dS^N/dR^N converges weakly, under R^N, to an a.s. positive random variable. Before we prove that, we give an explicit expression for the likelihood ratio—it is exactly D_N, i.e., the denominator of T_N.

Fact 1. The likelihood ratio dS^N/dR^N admits the expression:

$$\frac{dS^N}{dR^N} = D_N := \frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\xi^{\otimes}(X,Y_\sigma). \tag{16}$$

Proof. Note that the likelihood ratio of μ^N and $(P\otimes Q)^N$ is given by

$$f_N := \frac{d\mu^N}{d(P\otimes Q)^N} = \prod_{i=1}^N \xi(X_i,Y_i), \quad \text{on } \left(\mathbb{R}^d\times\mathbb{R}^d\right)^N. \tag{17}$$

Hence, by the property of conditional expectation, it holds that $\frac{dS^N}{dR^N} = \frac{d\mu^N|_{\mathcal{G}_N}}{d(P\otimes Q)^N|_{\mathcal{G}_N}} = \mathbb{E}[f_N \mid \mathcal{G}_N]$, where the conditional expectation is under $P\otimes Q$. It follows from exchangeability under $P\otimes Q$ that $\mathbb{E}[f_N\mid\mathcal{G}_N] = \mathbb{E}[\xi^{\otimes}(X,Y_\sigma)\mid\mathcal{G}_N]$ for each $\sigma\in\mathcal{S}_N$. Hence,

$$\mathbb{E}[f_N\mid\mathcal{G}_N] = \mathbb{E}\left[\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\xi^{\otimes}(X,Y_\sigma)\middle|\mathcal{G}_N\right] = \frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\xi^{\otimes}(X,Y_\sigma), \tag{18}$$

where the last equality follows from $\sum_{\sigma\in\mathcal{S}_N}\xi^{\otimes}(X,Y_\sigma)$ is $\mathcal{G}_N$-measurable. □

Recall from Theorem 2 that D_N has a limiting distribution given by the exponential of a weighted sum of products of Gaussians which is almost surely positive. Besides tools such as the Hoeffding decomposition from the U-statistics theory, the proof of Theorem 2 involves a novel approach to control the variance of D_N. We defer it to Section 4. Now we are ready to prove the contiguity result.

Theorem 6. *Under Assumption 1, the sequences* $(R^N)_{N\geq 1}$ *and* $(S^N)_{N\geq 1}$ *are mutually contiguous, i.e.,* $R^N \triangleleft\triangleright S^N$. *Explicitly, for a sequence of events* $(A_N \in \mathcal{G}_N,\ N \geq 1)$, *we have* $\lim_{N\to\infty} S^N(A_N) = 0$ *iff* $\lim_{N\to\infty} R^N(A_N) = 0$.

Proof. According to Le Cam's first lemma [66, page 88], $R^N \triangleleft S^N$, $N \geq 1$, if and only if the following statement holds true: if D_N, under $P \otimes Q$, converges weakly to D, along a sub-sequence, then $P(D > 0) = 1$. This statement follows directly from Theorem 2, so we have $R^N \triangleleft S^N$. By a standard computation, it can be shown that $\mathbb{E}[D] = 1$. Hence, it follows from Le Cam's first lemma again that $S^N \triangleleft R^N$, that is, R^N and S^N are mutually contiguous. □

With Theorem 6 at hand, we can work under the measure μ. The next result rewrites T_N as a simple conditional expectation and verifies its consistency.

Proposition 7. *Assume that* $\{(X_i,Y_i)\}_{i=1}^N \overset{i.i.d.}{\sim} \mu$. *It holds that* $T_N = \mathbb{E}_\mu[\eta(X_1,Y_1) \mid \mathcal{G}_N]$ *for every* $\eta \in \mathbf{L}^1(\mu)$. *Moreover,* T_N *is an unbiased and consistent estimator of* θ. *That is,* $\mathbb{E}_\mu[T_N] = \theta$ *for all* N *and* $\lim_{N\to\infty} T_N = \theta$ *almost surely.*

Proof. For notational simplicity, let $\bar{\eta}(X,Y_\sigma) := \frac{1}{N}\sum_{i=1}^N \eta(X_i,Y_{\sigma_i})$ for each $\sigma \in \mathcal{S}_N$. By exchangeability of $\{(X_i,Y_i)\}_{i=1}^N$, it holds that $\mathbb{E}_\mu[\eta(X_i,Y_i) \mid \mathcal{F}_N] = \mathbb{E}_\mu[\eta(X_j,Y_j) \mid \mathcal{F}_N]$ for all $1 \leq i,j \leq N$ which implies that $\mathbb{E}_\mu[\eta(X_1,Y_1) \mid \mathcal{F}_N] = \mathbb{E}_\mu[\bar{\eta}(X,Y_{\mathrm{id}}) \mid \mathcal{F}_N]$. Since $\bar{\eta}(X,Y_{\mathrm{id}})$ is $\mathcal{F}_N$-measurable, it follows that $\mathbb{E}_\mu[\eta(X_1,Y_1) \mid \mathcal{F}_N] = \bar{\eta}(X,Y_{\mathrm{id}})$. By the tower property of conditional expectations,

$$h_N := \mathbb{E}_\mu[\eta(X_1,Y_1) \mid \mathcal{G}_N] = \mathbb{E}_\mu\left[\mathbb{E}_\mu[\eta(X_1,Y_1) \mid \mathcal{F}_N] \mid \mathcal{G}_N\right] = \mathbb{E}_\mu[\bar{\eta}(X,Y_{\mathrm{id}}) \mid \mathcal{G}_N].$$

By definition, the last expression is the a.s. unique $\mathcal{G}_N$-measurable function such that for any bounded $\mathcal{G}_N$-measurable ϕ, it holds that $\mathbb{E}_\mu[\bar{\eta}(X,Y_{\mathrm{id}})\phi] = \mathbb{E}_\mu[h_N\phi]$. By (17), $\mathbb{E}_\mu[\bar{\eta}(X,Y_{\mathrm{id}})\phi]$ equals

$$\mathbb{E}[f_N\bar{\eta}(X,Y_{\mathrm{id}})\phi] = \mathbb{E}[\mathbb{E}[f_N\bar{\eta}(X,Y_{\mathrm{id}}) \mid \mathcal{G}_N]\phi] = \mathbb{E}_\mu\left[\frac{dR^N}{dS^N}\mathbb{E}[f_N\bar{\eta}(X,Y_{\mathrm{id}}) \mid \mathcal{G}_N]\phi\right],$$

which implies that $h_N = \frac{dR^N}{dS^N}\mathbb{E}[f_N\bar{\eta}(X,Y_{\mathrm{id}}) \mid \mathcal{G}_N]$. Similar to (18), we have

$$\mathbb{E}[f_N\bar{\eta}(X,Y_{\mathrm{id}}) \mid \mathcal{G}_N] = \frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\bar{\eta}(X,Y_\sigma)\xi^{\otimes}(X,Y_\sigma).$$

According to Fact 1, we get $h_N = \frac{1}{D_N}\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\bar{\eta}(X,Y_\sigma)\xi^{\otimes}(X,Y_\sigma) = T_N$.

Hence, the unbiasedness of T_N under μ follows by the tower property of conditional expectations. Now consider the reverse σ-algebra $\overline{\mathcal{G}}_N = \sigma(\mathcal{G}_N, (X_i,Y_i), i \geq N+1)$. Since $\{(X_i,Y_i)\}_{i\geq N+1}$ are independent of $\{(X_i,Y_i)\}_{i=1}^N$, we have $T_N = \mathbb{E}_\mu\left[\eta(X_1,Y_1) \mid \overline{\mathcal{G}}_N\right]$. Consequently, $(T_N,\overline{\mathcal{G}}_N)_{N\geq 1}$ is a reverse martingale and T_N converges almost surely to $\mathbb{E}_\mu[\eta(X_1,Y_1)] = \theta$. □

Proof of Theorem 1. As shown in Proposition 7, for any $\eta \in \mathbf{L}^1(\mu)$, $T_N = T_N(\eta) \to_{a.s.} \theta$ under μ. In particular, Proposition 7 holds for any bounded continuous function η. Thus, except for a null set, the convergence in Proposition 7 holds for a countable collection of bounded continuous functions. By separability of $\mathbb{R}^d$, almost sure weak convergence follows [67, Theorem 3.1] by choosing such a countable collection judiciously. This shows almost sure weak convergence under μ. Weak convergence in probability under $P \otimes Q$ now follows from Theorem 6. □

3. Limit law and chaos decomposition

This section is devoted to the limit law of T_N in (14). Following the standard strategy, our goal is to find the first order approximation $\mathcal{L}_1$ of T_N in the form of a sum of i.i.d. terms. Now, provided that the remainder $T_N - \theta - \mathcal{L}_1 = o_p(N^{-1/2})$, it follows from the CLT that $\sqrt{N}(T_N - \theta)$ converges weakly to a normal distribution. However, there are two main challenges. First, the statistic T_N has a rather complicated expression involving a ratio of two infinite-order U-statistics. This prevents us from utilizing the Hoeffding decomposition to derive the first order approximation. Second, due to its complicated nature, it is extremely challenging to control the remainder—the variance computation for classical U-statistics does not apply here.

To address the first challenge, the key observation is that T_N admits a simple expression under μ as shown in Proposition 7. This allows us to obtain a linear approximation of T_N under μ which we call the first order chaos. Due to the contiguity result in Theorem 6, the first order chaos can be viewed as the first order approximation of T_N under $P \otimes Q$. As for the second challenge, we develop a novel approach to control the remainder using the spectral gap of the operators $\mathcal{A}$ and $\mathcal{A}^*$. Since this approach is also used to establish the limit law of D_N in Theorem 2, we discuss the treatment of D_N and the remainder together in Section 4.

In the remainder of this section, we first give a formal derivation of the first order approximation $\mathcal{L}_1$ and prove the asymptotic normality of T_N in Section 3.1. We then derive, in Section 3.2, $\mathcal{L}_1$ rigorously as the first order chaos of T_N using orthogonal projections in $\mathbf{L}^2(\mu^N)$.

3.1. First order approximation

Recall from Proposition 7 that $T_N = \mathbb{E}_\mu[\eta(X_1,Y_1) \mid \mathcal{G}_N]$. Hence, in order to obtain the first order approximation of T_N, it is natural to approximate $\eta(X,Y) - \theta$ by some linear term $f(X) + g(Y)$ under $(X,Y) \sim \mu$ and then use $\mathbb{E}_\mu[f(X_1) + g(Y_1) \mid \mathcal{G}_N] = \frac{1}{N}\sum_{i=1}^N [f(X_i) + g(Y_i)]$ as the first order approximation of T_N, where the equality can be shown with an argument similar to the proof of Proposition 7. A good linear approximation $f(X) + g(Y)$ should satisfy

$$\begin{aligned} \mathbb{E}_\mu[\eta(X,Y) - \theta \mid X] &= \mathbb{E}_\mu[f(X) + g(Y) \mid X] \\ \mathbb{E}_\mu[\eta(X,Y) - \theta \mid Y] &= \mathbb{E}_\mu[f(X) + g(Y) \mid Y]. \end{aligned} \tag{19}$$

Recall $\frac{d\mu}{d(P\otimes Q)}(x,y) = \xi(x,y)$ and $\eta_{1,0}$ from (15). It holds that

$$\mathbb{E}_\mu[\eta(X,Y) - \theta \mid X](x) = \int [\eta(x,y) - \theta]\xi(x,y)dQ(y) = \eta_{1,0}(x).$$

Similarly, we have $\mathbb{E}_\mu[\eta(X,Y) - \theta \mid Y](y) = \eta_{0,1}(y)$. It then follows from the tower property that

$$\mathbb{E}_P[\eta_{1,0}(X)] = \mathbb{E}_P[\mathbb{E}_\mu[\eta(X,Y) - \theta \mid X]] = 0 \quad \text{and} \quad \mathbb{E}_Q[\eta_{0,1}(Y)] = 0. \tag{20}$$

Moreover, by Definition 1, we obtain

$$\begin{aligned} \mathbb{E}_\mu[g(Y) \mid X](x) &= \int g(y)\xi(x,y)dQ(y) = (\mathcal{A}^* g)(x) \\ \mathbb{E}_\mu[f(X) \mid Y](y) &= \int f(x)\xi(x,y)dP(x) = (\mathcal{A} f)(y). \end{aligned} \tag{21}$$

As a result, the condition (19) becomes

$$\eta_{1,0}(X) = f(X) + \mathcal{A}^* g(X) \quad \text{and} \quad \eta_{0,1}(Y) = \mathcal{A} f(Y) + g(Y). \tag{22}$$

Formally, we can solve the linear system (22) to get

$$f = (I - \mathcal{A}^*\mathcal{A})^{-1}(\eta_{1,0} - \mathcal{A}^*\eta_{0,1}) \quad \text{and} \quad g = (I - \mathcal{A}\mathcal{A}^*)^{-1}(\eta_{0,1} - \mathcal{A}\eta_{1,0}).$$

We will make this rigorous later. This suggests the following first order approximation of T_N:

$$\frac{1}{N}\sum_{i=1}^{N}\left[(I - \mathcal{A}^*\mathcal{A})^{-1}(\eta_{1,0} - \mathcal{A}^*\eta_{0,1})(X_i) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(\eta_{0,1} - \mathcal{A}\eta_{1,0})(Y_i)\right],$$

which is exactly the first order chaos $\mathcal{L}_1$ in Theorem 4. In fact, the next result shows that, after subtracting $\mathcal{L}_1$ from $T_N - \theta$, the variance of the numerator is of order $O(N^{-2})$.

It can be shown that the remainder $T_N - \theta - \mathcal{L}_1 = U_N/D_N$, where D_N is defined in (16) and

$$U_N := \frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\frac{1}{N}\sum_{i=1}^{N}\widetilde{\eta}(X_i, Y_{\sigma_i})\xi^{\otimes}(X, Y_\sigma) \tag{23}$$

with $\widetilde{\eta}$ defined as

$$\widetilde{\eta}(x,y) := \eta(x,y) - \theta - (I - \mathcal{A}^*\mathcal{A})^{-1}(\eta_{1,0} - \mathcal{A}^*\eta_{0,1})(x) - (I - \mathcal{A}\mathcal{A}^*)^{-1}(\eta_{0,1} - \mathcal{A}\eta_{1,0})(y). \tag{24}$$

In fact, for all f and g, it holds that $T_N - \theta - \frac{1}{N}\sum_{i=1}^{N}[f(X_i) + g(Y_i)]$ equals

$$\frac{\sum_{\sigma\in\mathcal{S}_N}\left[\frac{1}{N}\sum_{i=1}^{N}\eta(X_i, Y_{\sigma_i}) - \theta - \frac{1}{N}\sum_{i=1}^{N}[f(X_i) + g(Y_{\sigma_i})]\right]\xi^{\otimes}(X, Y_\sigma)}{\sum_{\sigma\in\mathcal{S}_N}\xi^{\otimes}(X, Y_\sigma)}$$
$$= \frac{\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\frac{1}{N}\sum_{i=1}^{N}[\eta(X_i, Y_{\sigma_i}) - \theta - f(X_i) - g(Y_{\sigma_i})]\xi^{\otimes}(X, Y_\sigma)}{\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\xi^{\otimes}(X, Y_\sigma)}.$$

Proposition 8. *Under Assumptions 1 and 2, we have* $\mathbb{E}[U_N^2] = O(N^{-2})$.

Similar to D_N, the numerator U_N is also a two-sample U-statistic of infinite order. We defer the proof of Proposition 8 to Section 4. Let us prove the main results.

Proof of Theorem 4. According to Theorem 2 and Proposition 8, we have $D_N = O_p(1)$ and $U_N = o_p(N^{-1/2})$. By Slutsky's Lemma, it holds that $T_N - \theta - \mathcal{L}_1 = U_N/D_N = o_p(N^{-1/2})$. Now, Corollary 5 follows from the standard Lindeberg CLT [10, Section 27]. □

3.2. Chaos decomposition for paired samples

We derive the first order chaos $\mathcal{L}_1$ using orthogonal projections in $\mathbf{L}^2(\mu^N)$. We change in this section the measure so that $\{(X_i, Y_i)\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \mu$.

Definition 3. Let $x_{[N]}$ and $y_{[N]}$ be two sets of (random) vectors in $\mathbb{R}^d$. Let $T := T(x_{[N]}, y_{[N]})$. We say T is *permutation symmetric in* x if $T(x_{\sigma_{[N]}}, y_{[N]}) = T(x_{[N]}, y_{[N]})$ for every $\sigma \in \mathcal{S}_N$, where $x_{\sigma_{[N]}} := (x_{\sigma_i})_{i\in[N]}$. We define permutation symmetry in y similarly. We say T is permutation symmetric if it is permutation symmetric in both x and y.

Let $H_0 \subset \mathbf{L}^2(\mu^N)$ be the subspace of constant functions and $H_1 \subset \mathbf{L}^2(\mu^N)$ be the subspace spanned by functions of the type

$$\sum_{i=1}^{N}[f(X_i) + g(Y_i)] \tag{25}$$

that is orthogonal to H_0. By Proposition 7, the (orthogonal) projection of T_N onto H_0 is $\mathrm{Proj}_{H_0}(T_N) = \theta$. Moreover, we show in Appendix B of the Supplementary Material [30] that H_1 is closed so that the projection of T_N onto H_1 uniquely exists. We will compute this projection, which we refer to as the *first order chaos*. Note that the elements in $\mathbf{L}^2$ spaces are only defined up to zero-measure sets (or equivalent classes). For two elements $f, g \in \mathbf{L}^2$, $f = g$ means f equals g up to equivalent classes.

Given a measure ν on $\mathbb{R}^d$, let $\mathbf{L}_0^2(\nu)$ be the subspace of $\mathbf{L}^2(\nu)$ consisting of mean-zero functions. Recall $\mathcal{A}$ and $\mathcal{A}^*$ in Definition 1. We first argue that $(I - \mathcal{A}^*\mathcal{A})^{-1}$ and $(I - \mathcal{A}\mathcal{A}^*)^{-1}$ are well-defined on $\mathbf{L}_0^2(P)$ and $\mathbf{L}_0^2(Q)$, respectively. The proof is deferred to the Supplementary Material [30].

Lemma 9. *Let $(X, Y) \sim \mu$. Under Assumption 1, the following statements hold true:*

(a) *For any $f \in \mathbf{L}^2(P)$ and $g \in \mathbf{L}^2(Q)$, it holds $\mathbb{E}_\mu[f(X) \mid Y](y) = \mathcal{A}f(y)$ and $\mathbb{E}_\mu[g(Y) \mid X](x) = \mathcal{A}^*g(x)$. In particular, $\mathcal{A}f \in \mathbf{L}^2(Q)$ and $\mathcal{A}^*g \in \mathbf{L}^2(P)$.*
(b) *The largest eigenvalue of $\mathcal{A}$ and $\mathcal{A}^*$ is 1, and $\mathcal{A}\mathbf{1} = \mathcal{A}^*\mathbf{1} = \mathbf{1}$.*
(c) *The operator $\mathcal{A}$ maps $\mathbf{L}_0^2(P)$ to $\mathbf{L}_0^2(Q)$, and $\mathcal{A}^*$ maps $\mathbf{L}_0^2(Q)$ to $\mathbf{L}_0^2(P)$.*
(d) *The operators $(I - \mathcal{A}^*\mathcal{A})^{-1} : \mathbf{L}_0^2(P) \to \mathbf{L}_0^2(P)$ and $(I - \mathcal{A}\mathcal{A}^*)^{-1} : \mathbf{L}_0^2(Q) \to \mathbf{L}_0^2(Q)$ are well-defined.*
(e) *It holds that $\mathcal{A}(I - \mathcal{A}^*\mathcal{A})^{-1} = (I - \mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}$ and $\mathcal{A}^*(I - \mathcal{A}\mathcal{A}^*)^{-1} = (I - \mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*$ on their domains defined above. Moreover, for any $f \in \mathbf{L}_0^2(P)$ and $g \in \mathbf{L}_0^2(Q)$, we have*

$$\begin{aligned} &\mathbb{E}_\mu\left[(I - \mathcal{A}^*\mathcal{A})^{-1}(f - \mathcal{A}^*g)(X) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(g - \mathcal{A}f)(Y) \mid X\right] = f(X) \\ &\mathbb{E}_\mu\left[(I - \mathcal{A}^*\mathcal{A})^{-1}(f - \mathcal{A}^*g)(X) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(g - \mathcal{A}f)(Y) \mid Y\right] = g(Y). \end{aligned} \tag{26}$$

Now we are ready to give the first order chaos of T_N, *i.e.*, $\mathrm{Proj}_{H_1}(T_N)$.

Proposition 10. *Under Assumptions 1 and 2, the first order chaos of the statistic T_N is given by*

$$\mathcal{L}_1 := \frac{1}{N}\sum_{i=1}^{N}[(I - \mathcal{A}^*\mathcal{A})^{-1}(\eta_{1,0} - \mathcal{A}^*\eta_{0,1})(X_i) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(\eta_{0,1} - \mathcal{A}\eta_{1,0})(Y_i)]. \tag{27}$$

Proof. By the definition of orthogonal projection, it suffices to show that, for any $i \in [N]$,

$$\mathbb{E}_\mu[T_N - \theta - \mathcal{L}_1 \mid X_i] = 0 \quad \text{and} \quad \mathbb{E}_\mu[T_N - \theta - \mathcal{L}_1 \mid Y_i] = 0$$

almost surely. We will prove it for X_1, and the rest of them can be proved similarly. Recall from (20) that $\eta_{1,0} \in \mathbf{L}_0^2(P)$ and $\eta_{0,1} \in \mathbf{L}_0^2(Q)$. By (c) in Lemma 9, we know $\eta_{1,0} - \mathcal{A}^*\eta_{0,1} \in \mathbf{L}_0^2(P)$ and $\eta_{0,1} - \mathcal{A}\eta_{1,0} \in$

$\mathbf{L}_0^2(Q)$. It then follows from (d) in Lemma 9 that, for every $i \in [N]$,

$$\mathbb{E}_\mu\left[(I - \mathcal{A}^*\mathcal{A})^{-1}(\eta_{1,0} - \mathcal{A}^*\eta_{0,1})(X_i) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(\eta_{0,1} - \mathcal{A}\eta_{1,0})(Y_i)\right] = 0.$$

As a result, $\mathbb{E}_\mu[\mathcal{L}_1 \mid X_1]$ is equal to

$$\frac{1}{N}\mathbb{E}_\mu\left[(I - \mathcal{A}^*\mathcal{A})^{-1}(\eta_{1,0} - \mathcal{A}^*\eta_{0,1})(X_1) + (I - \mathcal{A}\mathcal{A}^*)^{-1}(\eta_{0,1} - \mathcal{A}\eta_{1,0})(Y_1) \mid X_1\right] = \frac{\eta_{1,0}(X_1)}{N},$$

where the last equality follows from (26). We only need to show $\mathbb{E}_\mu[T_N - \theta \mid X_1] = \frac{1}{N}\eta_{1,0}(X_1)$. Let $h(x) := \mathbb{E}_\mu[T_N - \theta \mid X_1](x)$. We will prove that $\mathbb{E}_P[h(X_1)\phi(X_1)] = \mathbb{E}_P[\eta_{1,0}(X_1)\phi(X_1)]/N$ for all $\sigma(X_1)$-measurable ϕ. Fix an arbitrary $\sigma(X_1)$-measurable ϕ. Since $T_N - \theta$ is permutation symmetric in X (see Definition 3), we get $\mathbb{E}_\mu[T_N - \theta \mid X_i](x) \equiv h(x)$ for all $i \in [N]$. As a result, it holds that

$$\mathbb{E}_\mu\left[(T_N - \theta)\sum_{i=1}^N \phi(X_i)\right] = \sum_{i=1}^N \mathbb{E}_\mu[(T_N - \theta)\phi(X_i)] = N\,\mathbb{E}_P[h(X_1)\phi(X_1)].$$

Recall from Proposition 7 that $T_N = \mathbb{E}_\mu[\eta(X_1, Y_1) \mid \mathcal{G}_N]$. Since $\sum_{i=1}^N \phi(X_i)$ is $\mathcal{G}_N$-measurable, by the tower property of conditional expectation, we get

$$\mathbb{E}_\mu\left[(T_N - \theta)\sum_{i=1}^N \phi(X_i)\right] = \mathbb{E}_\mu\left[(\eta(X_1, Y_1) - \theta)\sum_{i=1}^N \phi(X_i)\right] = \mathbb{E}_P[\eta_{1,0}(X_1)\phi(X_1)],$$

where the last equality follows from the independence of $\{(X_i, Y_i)\}_{i=1}^N$ and $\eta_{1,0} \in \mathbf{L}_0^2(P)$. Hence, we have $\mathbb{E}_P[\eta_{1,0}(X_1)\phi(X_1)] = N\mathbb{E}_\mu[h(X_1)\phi(X_1)]$ which completes the proof. □

We then derive a more compact expression of $\mathcal{L}_1$ using $\mathcal{B}$ in Definition 2. We start by providing some properties of $\mathcal{B}$ in the next lemma. The proof is deferred to the Supplementary Material [30].

Lemma 11. *Under Assumption 1, the following statements hold true:*

(a) *Let $(X_1, Y_1), (X_2, Y_2) \overset{i.i.d}{\sim} \mu$. It holds that $\mathbb{E}_\mu[f(X_1, Y_2) \mid X_2, Y_1](x, y) = \mathcal{B}f(x, y)$ for any $f \in \mathbf{L}^2(P \otimes Q)$. In particular, $\mathcal{B}f \in \mathbf{L}^2(P \otimes Q)$.*
(b) *The operator $\mathcal{B}$ maps $\mathbf{L}_0^2(P \otimes Q)$ to $\mathbf{L}_0^2(P \otimes Q)$.*
(c) *For any $f \oplus g \in \mathbf{L}^2(P \otimes Q)$, we have $\mathcal{B}(f \oplus g) = \mathcal{A}^*g \oplus \mathcal{A}f$.*
(d) *The operator $(I + \mathcal{B})^{-1}$ is well-defined on $\mathbf{L}_0^2(P \otimes Q)$.*
(e) *For any $f \in \mathbf{L}_0^2(P)$ and $g \in \mathbf{L}_0^2(Q)$, it holds that*

$$(I + \mathcal{B})^{-1}(f \oplus g) = [(I - \mathcal{A}^*\mathcal{A})^{-1}(f - \mathcal{A}^*g)] \oplus [(I - \mathcal{A}\mathcal{A}^*)^{-1}(g - \mathcal{A}f)]. \tag{28}$$

According to (28), the first order chaos $\mathcal{L}_1$ admits a more compact representation.

Corollary 12. *Under Assumptions 1 and 2, the first order chaos of T_N admits an alternative expression $\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^N (I + \mathcal{B})^{-1}(\eta_{1,0} \oplus \eta_{0,1})(X_i, Y_i)$.*

Remark 4. Note that the above expression of $\mathcal{L}_1$ is permutation symmetric, i.e., $\sum_{i=1}^N (I + \mathcal{B})^{-1}(\eta_{1,0} \oplus \eta_{0,1})(X_i, Y_i) = \sum_{i=1}^N (I + \mathcal{B})^{-1}(\eta_{1,0} \oplus \eta_{0,1})(X_i, Y_{\sigma_i})$ for all $\sigma \in \mathcal{S}_N$.

Remark 5. Another way to see this is: due to (22), $\eta_{1,0} \oplus \eta_{0,1} = f \oplus g + \mathcal{A}^*g \oplus \mathcal{A}f = (I + \mathcal{B})(f \oplus g)$.

4. Analysis of the denominator and the remainder

Recall from (23) that the first order remainder $R_1 := T_N - \theta - \mathcal{L}_1 = U_N / D_N$, where

$$U_N := \frac{1}{N!} \sum_{\sigma \in \mathcal{S}_N} \frac{1}{N} \sum_{i=1}^{N} \widetilde{\eta}(X_i, Y_{\sigma_i}) \xi^{\otimes}(X, Y_\sigma) \quad \text{and} \quad D_N := \frac{1}{N!} \sum_{\sigma \in \mathcal{S}_N} \xi^{\otimes}(X, Y_\sigma), \tag{29}$$

with $\widetilde{\eta}$ defined in (24). We prove in this section the limit law of D_N in Theorem 2 and the variance bound of U_N in Proposition 8. The strategy is to decompose D_N and U_N into orthogonal pieces using the Hoeffding decomposition (Section 4.1), and then bound the higher order terms using the spectral gap of $\mathcal{A}$ and $\mathcal{A}^*$ (Section 4.2). Note that D_N and U_N are infinite-order *two-sample* U-statistics. Techniques for fixed-order U-statistics and infinite-order *one-sample* U-statistics do not apply here. Hence, this section develops new tools to handle two-sample U-statistics of infinite order. We work throughout this section with the original model assuming that $\{(X_i, Y_i)\}_{i=1}^{N} \overset{\text{i.i.d.}}{\sim} P \otimes Q$ and use $\mathbb{E}$ to denote the expectation. The proofs of some technical results are deferred to the Supplementary Material [30].

4.1. Hoeffding decomposition under the product measure

Definition 4. Given $A, B \subset [N]$, we denote by H_{AB} the subspace of $\mathbf{L}^2((P \otimes Q)^N)$ spanned by functions of the form $f(X_A, Y_B)$ such that

$$\mathbb{E}[f(X_A, Y_B) \mid X_C, Y_D] \overset{a.s.}{=} 0, \quad \text{for all } C \subset A, D \subset B \text{ and } |C| + |D| < |A| + |B| \,. \tag{30}$$

We say such an $f(X_A, Y_B)$ is completely degenerate. In particular, when $|A| = |B| = 1$, we write $f \in \mathbf{L}^2_{0,0}(P \otimes Q)$. By definition, for distinct choices of the pair (A, B), the subspaces H_{AB} are orthogonal. Take an arbitrary mean-zero statistic $T \in \mathbf{L}^2_0((P \otimes Q)^N)$. If T can be decomposed as

$$T = \sum_{A, B \subset [N]} T_{AB}, \quad \text{with} \quad T_{AB} \in H_{AB}, \tag{31}$$

then we call it the *Hoeffding decomposition* of T [66, Chapter 11]. Its variance can then be computed as $\mathbb{E}[T^2] = \sum_{A,B \subset [N]} \mathbb{E}[T_{AB}^2]$.

For example, both $\widetilde{\xi}(X_1, Y_1) := \xi(X_1, Y_1) - 1$ and $h(X_1, Y_1) := \widetilde{\eta}(X_1, Y_1)\xi(X_1, Y_1)$ are completely degenerate according to the following lemma.

Lemma 13. *Assume that $\xi, \eta\xi \in \mathbf{L}^2(P \otimes Q)$, then $\widetilde{\xi}, \widetilde{\eta}\xi \in \mathbf{L}^2_{0,0}(P \otimes Q)$.*

We then derive the Hoeffding decompositions of D_N and U_N as defined in (29).

Proposition 14. *Assume that $\xi, \eta\xi \in \mathbf{L}^2(P \otimes Q)$, then the following Hoeffding decompositions hold:*

$$\begin{aligned} D_N &= 1 + \sum_{\substack{A,B \subset [N] \\ |A|=|B|>0}} \frac{1}{N!} \sum_{\sigma \in \mathcal{S}_N : \sigma_A = B} \prod_{i \in A} \widetilde{\xi}(X_i, Y_{\sigma_i}) \\ U_N &= \sum_{\substack{A,B \subset [N] \\ |A|=|B|>0}} \frac{1}{N \cdot N!} \sum_{\sigma \in \mathcal{S}_N : \sigma_A = B} \sum_{i \in A} h(X_i, Y_{\sigma_i}) \prod_{j \in A \setminus \{i\}} \widetilde{\xi}(X_j, Y_{\sigma_j}), \end{aligned} \tag{32}$$

where $\sigma_A := \{\sigma_i : i \in A\}$. *Moreover,*

$$\mathbb{E}[D_N^2] = 1 + \sum_{r=1}^{N} \sum_{\sigma \in \mathcal{S}_r} \mathbb{E}\Big[\prod_{j=1}^{r} \widetilde{\xi}(X_j, Y_j)\widetilde{\xi}(X_j, Y_{\sigma_j})\Big]$$

$$\mathbb{E}[U_N^2] = \frac{1}{N^2}\sum_{r=1}^{N} \frac{r}{r!} \sum_{\sigma \in \mathcal{S}_r} \sum_{i=1}^{r} \mathbb{E}\Big[h(X_1, Y_1)\prod_{j=2}^{r} \widetilde{\xi}(X_j, Y_j) h(X_i, Y_{\sigma_i}) \prod_{j \in [r]\setminus\{i\}} \widetilde{\xi}(X_j, Y_{\sigma_j})\Big].$$

4.2. Variance bounds

We then bound the variances of D_N and U_N using the spectral gap of the operators $\mathcal{A}$ and $\mathcal{A}^*$. Assumption 1 guarantees that such a spectral gap does exist. We first prove a contraction property.

Lemma 15. *Recall* s_1 *from Assumption 1. For any* $f \in \mathbf{L}^2_{0,0}(P \otimes P)$, *we have* $(I_P \otimes \mathcal{A})f \in \mathbf{L}^2_{0,0}(P \otimes Q)$ *and* $\|(I_P \otimes \mathcal{A})f\|_{\mathbf{L}^2(P\otimes Q)} \le s_1 \|f\|_{\mathbf{L}^2(P\otimes P)}$. *Similar results hold for* $I_P \otimes \mathcal{A}^*$, $\mathcal{A} \otimes I_Q$ *and* $\mathcal{A}^* \otimes I_Q$.

According to Proposition 14, the key quantity in the variances of D_N and U_N is

$$\mathbb{E}\Big[f(X_1, Y_1)\prod_{j=2}^{N} \widetilde{\xi}(X_j, Y_j) f(X_i, Y_{\sigma_i}) \prod_{j \in [N]\setminus\{i\}} \widetilde{\xi}(X_j, Y_{\sigma_j})\Big] \tag{33}$$

for some $f \in \mathbf{L}^2_{0,0}(P \otimes Q)$, where $f = \widetilde{\xi} = \xi - 1$ for D_N and $f = h = \widetilde{\eta}\xi$ for U_N. In order to control it, we decompose a permutation into disjoint cycles. By independence, the expectation then equals the product of expectations with respect to each cycle. We first give a simple example to illustrate the idea.

Example 1. Consider the case when $r = 3$, $i = 3$, and σ is given by $\sigma_1 = 2$, $\sigma_2 = 1$ and $\sigma_3 = 3$. We are interested in bounding the expectation

$$\mathbb{E}[f(X_1, Y_1)\widetilde{\xi}(X_2, Y_2)\widetilde{\xi}(X_3, Y_3) f(X_3, Y_3)\widetilde{\xi}(X_1, Y_2)\widetilde{\xi}(X_2, Y_1)]. \tag{34}$$

By construction, σ contains two cycles, $1 \to 2 \to 1$ and $3 \to 3$, and the above expectation reads

$$\mathbb{E}[f(X_1, Y_1)\widetilde{\xi}(X_2, Y_2)\widetilde{\xi}(X_1, Y_2)\widetilde{\xi}(X_2, Y_1)] \cdot \mathbb{E}[f(X_3, Y_3)\widetilde{\xi}(X_3, Y_3)].$$

The second expectation is upper bounded by $\|f\|_{\mathbf{L}^2(P\otimes Q)} \|\widetilde{\xi}\|_{\mathbf{L}^2(P\otimes Q)}$ by the Cauchy-Schwarz inequality. It then suffices to bound the first expectation. We simplify this expectation by iteratively integrating with respect to a single variable while keeping the rest being fixed. We first integrate with respect to X_1 given X_2, Y_1, Y_2. This gives us

$$\mathbb{E}[f(X_1, Y_1)\widetilde{\xi}(X_1, Y_2) \mid X_2, Y_1, Y_2] \cdot \widetilde{\xi}(X_2, Y_2)\widetilde{\xi}(X_2, Y_1) = (\mathcal{A} \otimes I_Q) f(Y_2, Y_1) \cdot \widetilde{\xi}(X_2, Y_2)\widetilde{\xi}(X_2, Y_1),$$

where we used $\mathbb{E}[f(X_1, Y_1)\widetilde{\xi}(X_1, Y_2) \mid X_2, Y_1, Y_2] = \mathbb{E}[f(X_1, Y_1)\xi(X_1, Y_2) \mid Y_1, Y_2] = (\mathcal{A} \otimes I_Q) f(Y_2, Y_1)$ since $f \in \mathbf{L}^2_{0,0}(P \otimes Q)$ and $\widetilde{\xi} = \xi - 1$. We then integrate with respect to Y_2 given X_2 and Y_1. This yields

$$\mathbb{E}[(\mathcal{A} \otimes I_Q) f(Y_2, Y_1)\widetilde{\xi}(X_2, Y_2) \mid X_2, Y_1] \cdot \widetilde{\xi}(X_2, Y_1) = (\mathcal{A}^* \otimes I_Q)(\mathcal{A} \otimes I_Q) f(X_2, Y_1) \cdot \widetilde{\xi}(X_2, Y_1).$$

By the Cauchy-Schwarz inequality and Lemma 15, its expectation is upper bounded by

$$\left\|(\mathcal{A}^* \otimes I_Q)(\mathcal{A} \otimes I_Q)f\right\|_{\mathbf{L}^2(P\otimes Q)} \|\widetilde{\xi}\|_{\mathbf{L}^2(P\otimes Q)} \le s_1^2 \|f\|_{\mathbf{L}^2(P\otimes Q)} \|\widetilde{\xi}\|_{\mathbf{L}^2(P\otimes Q)}.$$

Hence, the expectation in (34) is upper bounded by $s_1^2 \|f\|^2_{\mathbf{L}^2(P\otimes Q)} \|\widetilde{\xi}\|^2_{\mathbf{L}^2(P\otimes Q)}$.

The following lemma generalizes this example to an arbitrary cycle $k_1 \to k_2 \to \cdots \to k_l \to k_1$.

Lemma 16. *Suppose Assumption 1 holds and $f, g \in \mathbf{L}^2_{0,0}(P \otimes Q)$. Define $\varsigma_f := \|f\|_{\mathbf{L}^2(P\otimes Q)}$ and $\varsigma_g := \|g\|_{\mathbf{L}^2(P\otimes Q)}$. For any $l > 0$ and l distinct indices $\{k_1, \ldots, k_l\} \subset [N]$, we have, for all $t, t' \in [l]$,*

$$\mathbb{E}\left[f(X_{k_t}, Y_{k_t})g(X_{k_{t'}}, Y_{k_{t'+1}})\prod_{i\neq t}\widetilde{\xi}(X_{k_i}, Y_{k_i})\prod_{j\neq t'}\widetilde{\xi}(X_{k_j}, Y_{k_{j+1}})\right] \le s_1^{2(l-1)}\varsigma_f\varsigma_g. \tag{35}$$

Now we are ready to control the quantity in (33).

Lemma 17. *Suppose the same assumptions in Lemma 16 hold true. Let $\varsigma_0 := \|\widetilde{\xi}\|_{\mathbf{L}^2(P\otimes Q)}$ and $\varsigma_h := \|h\|_{\mathbf{L}^2(P\otimes Q)}$. For any $N \in \mathbb{N}_+$, $\sigma \in \mathcal{S}_N$ and $i \in [N]$, we have*

$$\mathbb{E}\left[h(X_1, Y_1)\prod_{j=2}^{N}\widetilde{\xi}(X_j, Y_j)h(X_i, Y_{\sigma_i})\prod_{j\in[N]\setminus\{i\}}\widetilde{\xi}(X_j, Y_{\sigma_j})\right] \le s_1^{2(N-\#\sigma)}\varsigma_h^2\varsigma_0^{2(\#\sigma-1)},$$

where $\#\sigma$ is the number of cycles of the permutation σ.

Proof. We first consider the case when $i \neq 1$. It is well-known that every permutation can be decomposed as disjoint cycles. Take a cycle $k_1 \to k_2 \to \cdots \to k_l \to k_1$ of σ. If it contains both 1 and i, then we assume, w.l.o.g., $k_1 = 1$ and $k_2 = i$. Consequently, all the terms that involve $X_{k_{[l]}}$ and $Y_{k_{[l]}}$ are

$$h(X_1, Y_1)h(X_i, Y_{\sigma_i})\prod_{j=2}^{l}\widetilde{\xi}(X_{k_j}, Y_{k_j})\prod_{j\in[l]\setminus\{2\}}\widetilde{\xi}(X_{k_j}, Y_{k_{j+1}}).$$

Using Lemma 16 with $f = h$ and $g = h$, it holds that

$$\mathbb{E}\left[h(X_1, Y_1)h(X_i, Y_{\sigma_i})\prod_{j=2}^{l}\widetilde{\xi}(X_{k_j}, Y_{k_j})\prod_{j\in[l]\setminus\{2\}}\widetilde{\xi}(X_{k_j}, Y_{k_{j+1}})\right] \le s_1^{2(l-1)}\varsigma_h^2.$$

If this cycle only contains 1, then a similar argument gives

$$\mathbb{E}\left[h(X_1, Y_1)\prod_{j=2}^{l}\widetilde{\xi}(X_{k_j}, Y_{k_j})\prod_{j=1}^{l}\widetilde{\xi}(X_{k_j}, Y_{k_{j+1}})\right] \le s_1^{2(l-1)}\varsigma_h\varsigma_0.$$

If this cycle only contains i, with $k_1 = i$, then we have

$$\mathbb{E}\left[h(X_i, Y_{\sigma_i})\prod_{j=1}^{l}\widetilde{\xi}(X_{k_j}, Y_{k_j})\prod_{j=2}^{l}\widetilde{\xi}(X_{k_j}, Y_{k_{j+1}})\right] \le s_1^{2(l-1)}\varsigma_h\varsigma_0.$$

Finally, if this cycle does not contain either 1 or i, then it holds $\mathbb{E}\left[\prod_{j=1}^{l}\widetilde{\xi}(X_{k_j},Y_{k_j})\widetilde{\xi}(X_{k_j},Y_{k_{j+1}})\right] \leq s_1^{2(l-1)}\varsigma_0^2$. Here we are invoking Lemma 16 with $f = g = \widetilde{\xi}$. Putting all together, we obtain

$$\mathbb{E}\left[h(X_1,Y_1)\prod_{j=2}^{N}\widetilde{\xi}(X_j,Y_j)h(X_i,Y_{\sigma_i})\prod_{j\in[N]\setminus\{i\}}\widetilde{\xi}(X_j,Y_{\sigma_j})\right] \leq s_1^{2(N-\#\sigma)}\varsigma_h^2\varsigma_0^{2(\#\sigma-1)}.$$

When $i = 1$, we can invoke Lemma 16 to get the same bound, since we allow $t = t'$ in this lemma. □

Now we are ready to give an upper bound for the variance of U_N and prove Proposition 8.

Proof of Proposition 8. Recall from Proposition 14 that $\mathbb{E}[U_N^2]$ is equal to

$$\frac{1}{N^2}\sum_{r=1}^{N}\frac{r}{r!}\sum_{\sigma\in\mathcal{S}_r}\sum_{i=1}^{r}\mathbb{E}\left[h(X_1,Y_1)\prod_{j=2}^{r}\widetilde{\xi}(X_j,Y_j)h(X_i,Y_{\sigma_i})\prod_{j\in[N]\setminus\{i\}}\widetilde{\xi}(X_j,Y_{\sigma_j})\right]. \tag{36}$$

By Lemma 17, we know

$$\mathbb{E}[U_N^2] \leq \frac{1}{N^2}\sum_{r=1}^{N}\frac{r}{r!}\sum_{\sigma\in\mathcal{S}_r} r s_1^{2(r-\#\sigma)}\varsigma_0^{2(\#\sigma-1)}\varsigma_h^2. \tag{37}$$

If $s_1 = 0$ or $\varsigma_0 = 0$, then $\xi = 1$ $P\otimes Q$-a.s. It follows from (36) that $\mathbb{E}[U_N^2] = 0$ which completes the proof. Hence, we assume in the following that $s_1 > 0$ and $\varsigma_0 > 0$.

Now, let σ^* be a random permutation uniformly sampled from $\mathcal{S}_r$. It is known [5, Chapter 1] that the moment generating function of $\#\sigma^*$ is given by $\mathbb{E}[u^{\#\sigma^*}] = \prod_{i=1}^{r}(1-\frac{1}{i}+\frac{u}{i})$. Thus,

$$\frac{r}{r!}\sum_{\sigma\in\mathcal{S}_r} r s_1^{2(r-\#\sigma)}\varsigma_0^{2(\#\sigma-1)} = r^2\,\mathbb{E}\left[s_1^{2(r-\#\sigma^*)}\varsigma_0^{2(\#\sigma^*-1)}\right] = r^2 s_1^{2r}\varsigma_0^{-2}\prod_{i=1}^{r}\left(1-\frac{1}{i}+\frac{\varsigma_0^2}{s_1^2 i}\right).$$

Let $m := \lceil \varsigma_0^2/s_1^2 - 1\rceil$. Then, for every $r \geq m$,

$$\prod_{i=1}^{r}\left(1-\frac{1}{i}+\frac{\varsigma_0^2}{s_1^2 i}\right) \leq \prod_{i=1}^{r}\left(1+\frac{m}{i}\right) = \frac{\prod_{i=1}^{r}(i+m)}{r!} = \frac{\prod_{i=r-m+1}^{r}(i+m)}{m!} \leq \frac{(r+m)^m}{m!},$$

and thus $\sum_{r=m}^{N}\frac{r}{r!}\sum_{\sigma\in\mathcal{S}_r} r s_1^{2(r-\#\sigma)}\varsigma_0^{2(\#\sigma-1)} \leq \sum_{r=m}^{N} r^2 s_1^{2r}\varsigma_0^{-2}\frac{(r+m)^m}{m!}$ converges as $N\to\infty$ since $s_1 < 1$. It follows from (37) that $\mathbb{E}[U_N^2] = O(N^{-2})$. □

A similar result holds for D_N. Recall from Proposition 14 that $D_N = 1 + \sum_{r=1}^{N} D_{N,r}$ where

$$D_{N,r} := \frac{1}{N!}\sum_{|A|=|B|=r}\sum_{\sigma\in\mathcal{S}_N:\sigma_A=B}\prod_{i\in A}\widetilde{\xi}(X_i,Y_{\sigma_i}). \tag{38}$$

Proposition 18. *Under Assumption 1, we have, for any integer $R\in[0,N]$,*

$$\mathbb{E}\left[\left(D_N - 1 - \sum_{r=1}^{R} D_{N,r}\right)^2\right] \leq \sum_{r=R+1}^{N}\frac{1}{r!}\sum_{\sigma\in\mathcal{S}_r} s_1^{2(r-\#\sigma)}\varsigma_0^{2\#\sigma},$$

which can be arbitrarily small for sufficiently large R.

4.3. Limit law of the denominator

Finally, we prove Theorem 2 regarding the limiting distribution of D_N. According to the singular value decomposition in Assumption 1, it holds that

$$\xi(x,y) = 1 + \sum_{k=1}^{\infty} s_k \alpha_k(x)\beta_k(y), \quad \text{in } \mathbf{L}^2(P \otimes Q),$$

where $0 \le s_k < 1$ is decreasing in k. Hence, we start by considering a truncated version of ξ, i.e., $\xi^K(x,y) := 1 + \sum_{k=1}^K s_k \alpha_k(x)\beta_k(y)$ for some integer K and derive the limit law of

$$D_N^K := \frac{1}{N!} \sum_{\sigma \in \mathcal{S}_N} \prod_{i=1}^N \xi^K(X_i, Y_{\sigma_i}).$$

Note that all the results for D_N in Sections 4.1 and 4.2 hold for D_N^K with ξ being replaced by ξ^K.

Proposition 19. *Under Assumption 1, it holds that*

$$D_N^K \to_d D^K := \frac{1}{\sqrt{\prod_{k=1}^K (1 - s_k^2)}} \exp\left\{ \frac{1}{2} \sum_{k=1}^K \left[-\frac{s_k^2}{1 - s_k^2}(U_k^2 + V_k^2) + \frac{2s_k}{1 - s_k^2} U_k V_k \right] \right\}, \tag{39}$$

where $\{U_k\}_{k=1}^K$ and $\{V_k\}_{k=1}^K$ are independent standard normal random variables.

Proof. We will prove the convergence using characteristic functions, i.e., $\mathbb{E}[e^{itD_N^K}] \to \mathbb{E}[e^{itD^K}]$.

Step 1. Truncation. Recall from (38) that $D_N = 1 + \sum_{r=1}^N D_{N,r}$. Applying it to D_N^K yields $D_N^K = 1 + \sum_{r=1}^N D_{N,r}^K$ where $D_{N,r}^K$ is $D_{N,r}$ with ξ being replaced by ξ^K. We further truncate D_N^K so that it becomes a two-sample U-statistic of fixed order $R > 0$, that is, we consider $D_N^{K,R} := 1 + \sum_{r=1}^R D_{N,r}^K$. We then truncate the limit D^K. By the multi-linear Mehler formula (see, e.g., [21]), we have

$$D^K = \sum_{p_1,\dots,p_K \ge 0} \prod_{k=1}^K \frac{s_k^{p_k}}{p_k!} H_{p_k}(U_k) H_{p_k}(V_k), \tag{40}$$

where $\{H_p\}_{p \ge 0}$ are the Hermite polynomials satisfying

$$\int H_p(x) H_q(x) e^{-x^2/2} dx = \sqrt{2\pi} p! \mathbb{1}\{p = q\}. \tag{41}$$

Therefore, it is natural to define

$$D^{K,R} := 1 + \sum_{r=1}^R \sum_{p_1 + \dots + p_K = r} \prod_{k=1}^K \frac{s_k^{p_k}}{p_k!} H_{p_k}(U_k) H_{p_k}(V_k).$$

By the triangle inequality, $\left|\mathbb{E}[e^{itD_N^K}] - \mathbb{E}[e^{itD^K}]\right| \le C_1 + C_2 + C_3$ where

$$C_1 := \left|\mathbb{E}[e^{itD_N^K} - e^{itD_N^{K,R}}]\right|,\ C_2 := \left|\mathbb{E}[e^{itD_N^{K,R}} - e^{itD^{K,R}}]\right|,\ C_3 := \left|\mathbb{E}[e^{itD^{K,R}} - e^{itD^K}]\right|.$$

We fix some arbitrary $\delta > 0$ and show that $C_1, C_2, C_3 \le \delta$ for sufficiently large N and R.

Step 2. Control C_1 and C_3. Using the inequality $\left|e^{iz} - 1\right| \le |z|$, we get

$$C_1 \le \mathbb{E}\left|e^{itD_N^K} - e^{itD_N^{K,R}}\right| \le |t|\,\mathbb{E}\left|D_N^K - D_N^{K,R}\right| \le |t|\sqrt{\mathbb{E}(D_N^K - D_N^{K,R})^2}.$$

Invoking Proposition 18 for D_N^K implies that, for sufficiently large R, we have $C_1 \le \delta$. Similarly, it holds that $C_3 \le |t|\sqrt{\mathbb{E}(D^{K,R} - D^K)^2}$ where

$$\begin{aligned}\mathbb{E}(D^{K,R} - D^K)^2 &= \mathbb{E}\left|\sum_{r=R+1}^{\infty}\sum_{p_1+\cdots+p_K=r}\prod_{k=1}^{K}\frac{s_k^{p_k}}{p_k!}H_{p_k}(U_k)H_{p_k}(V_k)\right|^2 \\ &= \sum_{r=R+1}^{\infty}\sum_{p_1+\cdots+p_K=r}\prod_{k=1}^{K}s_k^{2p_k} \le \sum_{r=R+1}^{\infty}s_1^{2r}, \quad \text{since } s_k \le s_1.\end{aligned}$$

Here the two equations follow from (40) and (41), respectively. Since $s_1 < 1$, we have $C_3 \le \delta$ for sufficiently large R.

Step 3. Control C_2. It suffices to show that $D_N^{K,R} \to_d D^{K,R}$ as $N \to \infty$ for any $R > 0$. Note that

$$\begin{aligned}D_{N,r}^K &= \frac{1}{N!}\sum_{|A|=|B|=r}\sum_{\sigma_A=B}\prod_{i\in A}\tilde{\xi}^K(X_i, Y_{\sigma_i}) = \frac{(N-r)!}{N!}\sum_{\substack{1\le i_1<\cdots<i_r\le N\\ 1\le j_1<\cdots<j_r\le N}}\sum_{\sigma\in\mathcal{S}_r}\prod_{t=1}^{r}\tilde{\xi}^K(X_{i_t}, Y_{j_{\sigma_t}}) \\ &= \frac{(N-r)!}{N!}\sum_{\substack{1\le i_1<\cdots<i_r\le N\\ j_1\ne\cdots\ne j_r}}\prod_{t=1}^{r}\tilde{\xi}^K(X_{i_t}, Y_{j_t}) = \frac{(N-r)!}{r!N!}\sum_{\substack{i_1\ne\cdots\ne i_r\\ j_1\ne\cdots\ne j_r}}\prod_{t=1}^{r}\tilde{\xi}^K(X_{i_t}, Y_{j_t}) \\ &= \frac{(N-r)!}{r!N!}\sum_{\substack{i_1\ne\cdots\ne i_r\\ j_1\ne\cdots\ne j_r}}\prod_{t=1}^{r}\left[\sum_{k=1}^{K}s_k\alpha_k(X_{i_t})\beta_k(Y_{j_t})\right] \\ &= \frac{(N-r)!}{r!N!}\sum_{\substack{i_1\ne\cdots\ne i_r\\ j_1\ne\cdots\ne j_r}}\sum_{k_1,\ldots,k_r=1}^{K}\prod_{t=1}^{r}s_{k_t}\alpha_{k_t}(X_{i_t})\beta_{k_t}(Y_{j_t}) \\ &= \frac{1}{r!}\sum_{k_1,\ldots,k_r=1}^{K}\left(\prod_{t=1}^{r}s_{k_t}\right)\frac{(N-r)!}{N!}\left[\sum_{i_1\ne\cdots\ne i_r}\prod_{t=1}^{r}\alpha_{k_t}(X_{i_t})\right]\left[\sum_{j_1\ne\cdots\ne j_r}\prod_{t=1}^{r}\beta_{k_t}(X_{j_t})\right].\end{aligned}$$

The last term above can be rewritten as follows. Take an arbitrary sequence $\mathbf{k} := (k_t)_{t=1}^r \subset [K]^r$. For each $k \in [K]$, let $p_k(\mathbf{k})$ be the number of times k appears among $(k_t)_{t=1}^r$. By [66, Theorem 12.10],

$$\sqrt{\frac{(N-r)!}{N!}}\sum_{i_1\ne\cdots\ne i_r}\prod_{t=1}^{r}\alpha_{k_t}(X_{i_t}) = \prod_{k=1}^{K}H_{p_k(\mathbf{k})}(\mathbb{G}_N^{(X)}\alpha_k) + o_p(1)$$

$$\sqrt{\frac{(N-r)!}{N!}} \sum_{j_1 \neq \cdots \neq j_r} \prod_{t=1}^{r} \beta_{k_t}(Y_{j_t}) = \prod_{k=1}^{K} H_{p_k(\mathbf{k})}(\mathbb{G}_N^{(Y)} \beta_k) + o_P(1),$$

where $\mathbb{G}_N^{(X)} \alpha := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \alpha(X_i)$ and $\mathbb{G}_N^{(Y)} \beta$ is defined similarly. Consequently, we have

$$D_{N,r}^{K} = \frac{1}{r!} \sum_{k_1,\ldots,k_r=1}^{K} \prod_{k=1}^{K} s_k^{p_k(\mathbf{k})} H_{p_k(\mathbf{k})}(\mathbb{G}_N^{(X)} \alpha_k) H_{p_k(\mathbf{k})}(\mathbb{G}_N^{(Y)} \beta_k) + o_P(1).$$

Moreover, for any permutation symmetric $f : [K]^r \to \mathbb{R}$, we have

$$\frac{1}{r!} \sum_{k_1,\ldots,k_r=1}^{K} f(k_1, \ldots, k_r) = \sum_{p_1+\cdots+p_K=r} \frac{1}{p_1! \ldots p_K!} f(l_1, \ldots, l_r),$$

where $l_1, \ldots, l_r$ is any sequence such that k appears exactly p_k times for all $k \in [K]$. As a result,

$$D_{N,r}^{K} = \sum_{p_1+\cdots+p_K=r} \prod_{k=1}^{K} \frac{s_k^{p_k}}{p_k!} H_{p_k}(\mathbb{G}_N^{(X)} \alpha_k) H_{p_k}(\mathbb{G}_N^{(Y)} \beta_k) + o_P(1),$$

and thus $D_N^{K,R} = 1 + \sum_{r=1}^{R} \sum_{p_1+\cdots+p_K=r} \prod_{k=1}^{K} \frac{s_k^{p_k}}{p_k!} H_{p_k}(\mathbb{G}_N^{(X)} \alpha_k) H_{p_k}(\mathbb{G}_N^{(Y)} \beta_k) + o_P(1)$. Due to the multivariate CLT [10, Section 29], the random vector $(\mathbb{G}_N^{(X)} \alpha_k, \mathbb{G}_N^{(Y)} \beta_k)_{k=1}^{K}$ converges weakly to $\mathcal{N}_{2K}(0, I_{2K})$ by the orthonormality of $\{\alpha_k\}_{k=1}^{K}$ and $\{\beta_k\}_{k=1}^{K}$. By the continuous mapping theorem,

$$D_N^{K,R} \to_d 1 + \sum_{r=1}^{R} \sum_{p_1+\cdots+p_K=r} \prod_{k=1}^{K} \frac{s_k^{p_k}}{p_k!} H_{p_k}(U_k) H_{p_k}(V_k) = D^{K,R},$$

which completes the proof. □

Proof of Theorem 2. We again prove the convergence using the characteristic functions. *Step 0. Verify the validity of the limit.* We first show $1/\prod_{k=1}^{\infty}(1 - s_k^2) < \infty$. In fact,

$$\frac{1}{\prod_{k=1}^{\infty}(1 - s_k^2)} = \exp\left\{\sum_{k=1}^{\infty} \log \frac{1}{1 - s_k^2}\right\} \leq \exp\left\{\sum_{k=1}^{\infty} \frac{s_k^2}{1 - s_k^2}\right\} \leq \exp\left\{\frac{\sum_{k=1}^{\infty} s_k^2}{1 - s_1^2}\right\} < \infty, \tag{42}$$

where the first inequality follows from $\log(1 + x) \geq \frac{x}{1+x}$ for all $x > -1$ and the last inequality follows from the square summability of $\{s_k\}_{k\geq 1}$. It suffices to show that $D \in \mathbf{L}^2(P \otimes Q)$. For any $k \geq 1$, let

$$Z_k := \frac{1}{\sqrt{1 - s_k^2}} \exp\left\{-\frac{s_k^2}{2(1 - s_k^2)}(U_k^2 + V_k^2) + \frac{s_k}{1 - s_k^2} U_k V_k\right\}. \tag{43}$$

Then $\{Z_k\}_{k\geq 1}$ are mutually independent and $D = \prod_{k=1}^{\infty} Z_k$. By a standard computation, we get $\mathbb{E}[Z_k^2] = 1/(1 - s_k^2)$. Therefore, by (42), $\mathbb{E}[D^2] = \prod_{k=1}^{\infty} \mathbb{E}[Z_k^2] = 1/\prod_{k=1}^{\infty}(1 - s_k^2) < \infty$.

Step 1. Control the difference between the characteristic functions. Recall D_N^K and D^K from Proposition 19. By the triangle inequality, we have $\left|\mathbb{E}[e^{itD_N}] - \mathbb{E}[e^{itD}]\right| \le C_1 + C_2 + C_3$ where

$$C_1 := \left|\mathbb{E}[e^{itD_N}] - \mathbb{E}[e^{itD_N^K}]\right|,\ C_2 := \left|\mathbb{E}[e^{itD_N^K}] - \mathbb{E}[e^{itD^K}]\right|,\ C_3 := \left|\mathbb{E}[e^{itD^K}] - \mathbb{E}[e^{itD}]\right|.$$

Fix $\delta > 0$. By Proposition 19, $C_2 \le \delta$ for sufficiently large N. It then remains to control C_1 and C_3.

Step 2. Control C_1. By construction, it holds that

$$D_N - D_N^K = \sum_{r=1}^{N} \frac{1}{N!} \sum_{|A|=|B|=r} \sum_{\sigma_A = B} \prod_{i\in A} \xi^{-K}(X_i, Y_{\sigma_i}),$$

where $\xi^{-K} := \xi - \xi^K \in \mathbf{L}^2_{0,0}(P\otimes Q)$ and $\varsigma_K^2 := \mathbb{E}_{P\otimes Q}[(\xi^{-K}(X,Y))^2] = \sum_{k\ge K+1} s_k^2$. Invoking Proposition 18 for ξ^{-K}, we obtain $\mathbb{E}[(D_N - D_N^K)^2] \le \sum_{r=1}^N \frac{1}{r!}\sum_{\sigma\in\mathcal{S}_r} s_1^{2(r-\#\sigma)}\varsigma_K^{2\#\sigma}$. As shown in the proof of Proposition 8, the sum $\sum_{r=1}^N \frac{1}{r!}\sum_{\sigma\in\mathcal{S}_r} s_1^{2(r-\#\sigma)}$ converges. Moreover, for sufficiently large K, since ς_K^2 can be arbitrarily small, we have $C_1 \le |t|\,\mathbb{E}[(D_N - D_N^K)^2] \le \delta$.

Step 3. Control C_3. Again, it suffices to control $\mathbb{E}[(D^K - D)^2]$. Recall Z_k in (43). By independence,

$$\mathbb{E}[(D^K - D)^2] = \mathbb{E}\left[\left(\prod_{k=1}^K Z_k - \prod_{k=1}^\infty Z_k\right)^2\right] = \mathbb{E}\left[\prod_{k=1}^K Z_k^2\right]\mathbb{E}\left[\left(1 - \prod_{k\ge K+1} Z_k\right)^2\right]$$
$$= \frac{1}{\prod_{k=1}^K (1-s_k^2)}\left[\frac{1}{\prod_{k\ge K+1}(1-s_k^2)} - 1\right], \quad \text{since } \mathbb{E}[Z_k] = 1.$$

It follows from (42) that $\prod_{k=1}^K (1-s_k^2)^{-1} < \infty$ and

$$1 \le \frac{1}{\prod_{k\ge K+1}(1-s_k^2)} \le \exp\left\{\frac{1}{1-s_1^2}\sum_{k\ge K+1} s_k^2\right\} \to 1, \quad \text{as } K\to\infty.$$

Hence, we have $\mathbb{E}[(D^K - D)^2] \to 0$ as $K\to\infty$, which completes the proof. □

Proof of Corollary 3. Recall from (4) that the Schrödinger bridge μ_ϵ which solves (3) is given by $\mu_\epsilon(x,y) = \xi(x,y)P(x)Q(y)$ where $\xi(x,y) = \exp(-(c(x,y) - a_\epsilon(x) - b_\epsilon(y))/\epsilon)$. Moreover, it follows from the strong duality that [25, Proposition 2.1] (a_ϵ, b_ϵ) solve the dual problem

$$\max_{a,b\in C(\mathbb{R}^d)} \left[\int a(x)dP(x) + \int b(y)dQ(y) + \epsilon - \epsilon\int \exp\left(-\frac{c(x,y) - a(x) - b(y)}{\epsilon}\right) dP(x)dQ(y)\right],$$

where $C(\mathbb{R}^d)$ is the set of continuous functions on $\mathbb{R}^d$. Consequently, $\mathbf{C}_\epsilon(P,Q) = \int a_\epsilon(x)dP(x) + \int b_\epsilon(y)dQ(y)$. By some algebra, we have

$$\frac{1}{N}\log\left[\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\exp\left(-\frac{\sum_{i=1}^N c(X_i,Y_{\sigma_i})}{\epsilon}\right)\right]$$
$$= \frac{1}{N}\log\left[\frac{1}{N!}\sum_{\sigma\in\mathcal{S}_N}\frac{\prod_{i=1}^N \xi(X_i,Y_{\sigma_i})}{\exp\left(\sum_{i=1}^N (a_\epsilon(X_i) + b_\epsilon(Y_{\sigma_i}))/\epsilon\right)}\right] = -\frac{1}{\epsilon N}\sum_{i=1}^N [a_\epsilon(X_i) + b_\epsilon(Y_i)] + \frac{1}{N}\log D_N.$$

Now the claim follows from the facts that $\frac{1}{N}\sum_{i=1}^{N}[a_\epsilon(X_i)+b_\epsilon(Y_i)] \to_p \mathbf{C}_\epsilon(P,Q)$ (by LLN) and $\frac{1}{N}\log D_N = o_p(1)$ (by Theorem 2) as $N\to\infty$. □

Acknowledgments

The authors thank Jon Wellner for fruitful discussions. The authors also thank the anonymous referees and the Associate Editor for their constructive comments that improved the quality of this paper. Thanks to PIMS Kantorovich Initiative for facilitating this collaboration.

Funding

Z.H. acknowledges support from NSF grant DMS-1810975 and CCF-1740551. L.L. acknowledges support from NSF grant DMS-1612483 and CCF-1740551. S.P. acknowledges support from NSF grant DMS-1612483, DMS-2052239 and DMS-2133244. Part of this work was done while Z.H. was visiting the Simons Institute for the Theory of Computing.

Supplementary Material

Supplement to "Asymptotics of discrete Schrödinger bridges via chaos decomposition" (DOI: 10.3150/23-BEJ1659SUPP; .pdf). This Appendix contains the following contents: (a) Additional proofs for some results used in the main paper; (b) closedness of the subspace H_1 defined in (25); (c) a table of notations.

References

[1] Adams, S., Bru, J.-B. and König, W. (2006). Large deviations for trapped interacting Brownian particles and paths. *Ann. Probab.* **34** 1370–1422. MR2257650 https://doi.org/10.1214/009117906000000214

[2] Adams, S. and Dorlas, T. (2008). Asymptotic Feynman-Kac formulae for large symmetrised systems of random walks. *Ann. Inst. Henri Poincaré Probab. Stat.* **44** 837–875. MR2453847 https://doi.org/10.1214/07-AIHP132

[3] Adams, S. and König, W. (2008). Large deviations for many Brownian bridges with symmetrised initial-terminal condition. *Probab. Theory Related Fields* **142** 79–124. MR2413267 https://doi.org/10.1007/s00440-007-0099-5

[4] Ajtai, M., Komlós, J. and Tusnády, G. (1984). On optimal matchings. *Combinatorica* **4** 259–264. MR0779885 https://doi.org/10.1007/BF02579135

[5] Arratia, R., Barbour, A.D. and Tavaré, S. (2003). *Logarithmic Combinatorial Structures: A Probabilistic Approach*. *EMS Monographs in Mathematics*. Zürich: European Mathematical Society (EMS). MR2032426 https://doi.org/10.4171/000

[6] Barvinok, A. (2002). *A Course in Convexity*. *Graduate Studies in Mathematics* **54**. Providence, RI: Amer. Math. Soc. MR1940576 https://doi.org/10.1090/gsm/054

[7] Beichl, I. and Sullivan, F. (1999). Approximating the permanent via importance sampling with application to the dimer covering problem. *J. Comput. Phys.* **149** 128–147. MR1669817 https://doi.org/10.1006/jcph.1998.6149

[8] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer. Reprint of the 1993 original. MR1623559

[9] Bigot, J., Cazelles, E. and Papadakis, N. (2019). Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electron. J. Stat.* **13** 5120–5150. MR4041704 https://doi.org/10.1214/19-EJS1637

[10] Billingsley, P. (1995). *Probability and Measure*, 3rd ed. *Wiley Series in Probability and Mathematical Statistics*. New York: Wiley. MR1324786
[11] Chen, Y., Georgiou, T.T. and Pavon, M. (2021). Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge. *SIAM Rev.* **63** 249–313. MR4253788 https://doi.org/10.1137/20M1339982
[12] Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158. MR0365798 https://doi.org/10.1214/aop/1176996454
[13] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of Advances in Neural Information Processing Systems*.
[14] del Barrio, E., Giné, E. and Matrán, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.* **27** 1009–1071. MR1698999 https://doi.org/10.1214/aop/1022677394
[15] del Barrio, E., Giné, E. and Utzet, F. (2005). Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli* **11** 131–189. MR2121458 https://doi.org/10.3150/bj/1110228245
[16] del Barrio, E. and Loubes, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *Ann. Probab.* **47** 926–951. MR3916938 https://doi.org/10.1214/18-AOP1275
[17] Dobrić, V. and Yukich, J.E. (1995). Asymptotics for transportation cost in high dimensions. *J. Theoret. Probab.* **8** 97–118. MR1308672 https://doi.org/10.1007/BF02213456
[18] Dynkin, E.B. and Mandelbaum, A. (1983). Symmetric statistics, Poisson point processes, and multiple Wiener integrals. *Ann. Statist.* **11** 739–745. MR0707925 https://doi.org/10.1214/aos/1176346241
[19] Ferradans, S., Papadakis, N., Peyré, G. and Aujol, J.-F. (2014). Regularized discrete optimal transport. *SIAM J. Imaging Sci.* **7** 1853–1882. MR3264566 https://doi.org/10.1137/130929886
[20] Feynman, R.P. (1953). Atomic theory of the λ transition in helium. *Phys. Rev.* **91** 1291–1301.
[21] Foata, D. (1981). Some Hermite polynomial identities and their combinatorics. *Adv. in Appl. Math.* **2** 250–259. MR0626861 https://doi.org/10.1016/0196-8858(81)90006-3
[22] Föllmer, H. (1988). Random fields and diffusion processes. In *École D'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*. *Lecture Notes in Math.* **1362** 101–203. Berlin: Springer. MR0983373 https://doi.org/10.1007/BFb0086180
[23] Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields* **162** 707–738. MR3383341 https://doi.org/10.1007/s00440-014-0583-7
[24] Galichon, A. and Salanié, B. (2010). Matching with trade-offs: Revealed preferences over competing characteristics. CEPR Discussion Paper No. DP7858.
[25] Genevay, A., Cuturi, M., Peyré, G. and Bach, F. (2016). Stochastic optimization for large-scale optimal transport. In *Proceedings of Advances in Neural Information Processing Systems*.
[26] Gohberg, I., Goldberg, S. and Kaashoek, M.A. (1990). *Classes of Linear Operators Vol. I*. Basel: Birkhäuser.
[27] González-Sanz, A., Loubes, J.-M. and Niles-Weed, J. (2022). Weak limits of entropy regularized optimal transport; potentials, plans and divergences. arXiv preprint.
[28] Halász, G. and Székely, G.J. (1976). On the elementary symmetric polynomials of independent random variables. *Acta Math. Acad. Sci. Hung.* **28** 397–400. MR0423491 https://doi.org/10.1007/BF01896806
[29] Halmos, P.R. (1946). The theory of unbiased estimation. *Ann. Math. Stat.* **17** 34–43. MR0015746 https://doi.org/10.1214/aoms/1177731020
[30] Harchaoui, Z., Liu, L. and Pal, S. (2024). Supplement to "Asymptotics of discrete Schrödinger bridges via chaos decomposition." https://doi.org/10.3150/23-BEJ1659SUPP
[31] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* **19** 293–325. MR0026294 https://doi.org/10.1214/aoms/1177730196
[32] Hundrieser, S., Klatt, M., Staudt, T. and Munk, A. (2022). A unifying approach to distributional limits for empirical optimal transport. arXiv preprint.
[33] Kenyon, R., Kráľ, D., Radin, C. and Winkler, P. (2020). Permutations with fixed pattern densities. *Random Structures Algorithms* **56** 220–250. MR4052852 https://doi.org/10.1002/rsa.20882
[34] Klatt, M., Munk, A. and Zemel, Y. (2022). Limit laws for empirical optimal solutions in random linear programs. *Ann. Oper. Res.* **315** 251–278. MR4458612 https://doi.org/10.1007/s10479-022-04698-0
[35] Klatt, M., Tameling, C. and Munk, A. (2020). Empirical regularized optimal transport: Statistical theory and applications. *SIAM J. Math. Data Sci.* **2** 419–443. MR4105566 https://doi.org/10.1137/19M1278788

[36] Kosowsky, J.J. and Yuille, A.L. (1994). The invisible hand algorithm: Solving the assignment problem with statistical physics. *Neural Netw.* **7** 477–490.

[37] Lei, J. (2020). Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli* **26** 767–798. MR4036051 https://doi.org/10.3150/19-BEJ1151

[38] Léonard, C. (2012). From the Schrödinger problem to the Monge-Kantorovich problem. *J. Funct. Anal.* **262** 1879–1920. MR2873864 https://doi.org/10.1016/j.jfa.2011.11.026

[39] Léonard, C. (2014). A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst.* **34** 1533–1574. MR3121631 https://doi.org/10.3934/dcds.2014.34.1533

[40] Liu, L. (2022). Statistical Divergences for Learning and Inference: Limit Laws and Non-Asymptotic Bounds. Ph.D. thesis, Univ. Washington.

[41] Luise, G., Rudi, A., Pontil, M. and Ciliberto, C. (2018). Differential properties of Sinkhorn approximation for learning with Wasserstein distance. In *Proceedings of Advances in Neural Information Processing Systems*.

[42] Major, P. (1999). The limit behavior of elementary symmetric polynomials of i.i.d. random variables when their order tends to infinity. *Ann. Probab.* **27** 1980–2010. MR1742897 https://doi.org/10.1214/aop/1022677557

[43] Mallows, C.L. (1957). Non-null ranking models. I. *Biometrika* **44** 114–130. MR0087267 https://doi.org/10.1093/biomet/44.1-2.114

[44] Mena, G. and Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. In *Proceedings of Advances in Neural Information Processing Systems*.

[45] Mikami, T. (2004). Monge's problem with a quadratic cost by the zero-noise limit of h-path processes. *Probab. Theory Related Fields* **129** 245–260. MR2063377 https://doi.org/10.1007/s00440-004-0340-4

[46] Móri, T.F. and Székely, G.J. (1982). Asymptotic behaviour of symmetric polynomial statistics. *Ann. Probab.* **10** 124–131. MR0637380

[47] Mukherjee, S. (2016). Estimation in exponential families on permutations. *Ann. Statist.* **44** 853–875. MR3476619 https://doi.org/10.1214/15-AOS1389

[48] Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 223–241. MR1625620 https://doi.org/10.1111/1467-9868.00121

[49] Pal, S. and Wong, T.-K.L. (2020). Multiplicative Schrödinger problem and the Dirichlet transport. *Probab. Theory Related Fields* **178** 613–654. MR4146546 https://doi.org/10.1007/s00440-020-00987-6

[50] Peyré, G. and Cuturi, M. (2019). *Computational Optimal Transport: With Applications to Data Science*. Found. Trends Mach. Learn.

[51] Ramdas, A., García Trillos, N. and Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy* **19** Paper No. 47. MR3608466 https://doi.org/10.3390/e19020047

[52] Rempała, G.A. and Wesołowski, J. (1999). Limiting behavior of random permanents. *Statist. Probab. Lett.* **45** 149–158. MR1718443 https://doi.org/10.1016/S0167-7152(99)00054-1

[53] Rempała, G.A. and Wesołowski, J. (2005). Approximation theorems for random permanents and associated stochastic processes. *Probab. Theory Related Fields* **131** 442–458. MR2123252 https://doi.org/10.1007/s00440-004-0380-9

[54] Rempała, G.A. and Wesołowski, J. (2008). *Symmetric Functionals on Random Matrices and Random Matchings Problems*. *The IMA Volumes in Mathematics and Its Applications* **147**. New York: Springer. MR2368286

[55] Rigollet, P. and Weed, J. (2018). Entropic optimal transport is maximum-likelihood deconvolution. *C. R. Math. Acad. Sci. Paris* **356** 1228–1235. MR3907589 https://doi.org/10.1016/j.crma.2018.10.010

[56] Rippl, T., Munk, A. and Sturm, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *J. Multivariate Anal.* **151** 90–109. MR3545279 https://doi.org/10.1016/j.jmva.2016.06.005

[57] Rugh, H.H. (2010). Cones and gauges in complex spaces: Spectral gaps and complex Perron-Frobenius theory. *Ann. of Math.* **171** 1707–1752. MR2680397 https://doi.org/10.4007/annals.2010.171.1707

[58] Rüschendorf, L. and Thomsen, W. (1993). Note on the Schrödinger equation and I-projections. *Statist. Probab. Lett.* **17** 369–375. MR1237783 https://doi.org/10.1016/0167-7152(93)90257-J

[59] Schrödinger, E. (1932). Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. *Ann. Inst. Henri Poincaré* **2** 269–310. MR1508000

[60] Sommerfeld, M. and Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 219–238. MR3744719 https://doi.org/10.1111/rssb.12236

[61] Talagrand, M. (1992). Matching random samples in many dimensions. *Ann. Appl. Probab.* **2** 846–856. MR1189420

[62] Tameling, C., Sommerfeld, M. and Munk, A. (2019). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *Ann. Appl. Probab.* **29** 2744–2781. MR4019874 https://doi.org/10.1214/19-AAP1463

[63] Trashorras, J. (2008). Large deviations for symmetrised empirical measures. *J. Theoret. Probab.* **21** 397–412. MR2391251 https://doi.org/10.1007/s10959-007-0121-y

[64] van Es, A.J. and Helmers, R. (1988). Elementary symmetric polynomials of increasing order. *Probab. Theory Related Fields* **80** 21–35. MR0970469 https://doi.org/10.1007/BF00348750

[65] van Es, B. (1986). On the weak limits of elementary symmetric polynomials. *Ann. Probab.* **14** 677–695. MR0832030

[66] van der Vaart, A.W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge Univ. Press.

[67] Varadarajan, V.S. (1958). Weak convergence of measures on separable metric spaces. *Sankhyā* **19** 15–22. MR0094838

[68] Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli* **25** 2620–2648. MR4003560 https://doi.org/10.3150/18-BEJ1065

Received July 2022 and revised May 2023