Uniform Inference for Kernel Density Estimators with Dyadic Data

Matias D. Cattaneo\textsuperscript{1} \hspace{1em} Yingjie Feng\textsuperscript{2} \hspace{1em} William G. Underwood\textsuperscript{1}\textsuperscript{*}

February 16, 2022

Abstract

Dyadic data is often encountered when quantities of interest are associated with the edges of a network. As such it plays an important role in statistics, econometrics and many other data science disciplines. We consider the problem of uniformly estimating a dyadic Lebesgue density function, focusing on nonparametric kernel-based estimators taking the form of dyadic empirical processes. Our main contributions include the minimax-optimal uniform convergence rate of the dyadic kernel density estimator, along with strong approximation results for the associated standardized and Studentized \( t \)-processes. A consistent variance estimator enables the construction of valid and feasible uniform confidence bands for the unknown density function. A crucial feature of dyadic distributions is that they may be “degenerate” at certain points in the support of the data, a property making our analysis somewhat delicate. Nonetheless our methods for uniform inference remain robust to the potential presence of such points. For implementation purposes, we discuss procedures based on positive semi-definite covariance estimators, mean squared error optimal bandwidth selectors and robust bias-correction techniques. We illustrate the empirical finite-sample performance of our methods both in simulations and with real-world data. Our technical results concerning strong approximations and maximal inequalities are of potential independent interest.

Keywords: dyadic data, networks, kernel density estimation, minimaxity, strong approximation.

\textsuperscript{1}Department of Operations Research and Financial Engineering, Princeton University
\textsuperscript{2}School of Economics and Management, Tsinghua University
\textsuperscript{*}Corresponding author: wgu2@princeton.edu
1 Introduction

Dyadic (or graphon) data plays an important role in the statistical, social, behavioral and biomedical sciences. In network settings, this type of dependent data captures interactions between the units of study, and its analysis is of interest in statistics (Kolaczyk and Csárdi, 2014), economics (Graham, 2020), psychology (Kenny et al., 2020), public health (Luke and Harris, 2007), and many other data science disciplines. For $n \geq 2$, a typical dyadic data set contains $\frac{1}{2}n(n-1)$ observed real-valued dyadic random variables

$$W_n = (W_{ij} : 1 \leq i < j \leq n), \quad W_{ij} = W(A_i, A_j, V_{ij}),$$

where $W$ is an unknown function, $A_n = (A_i : 1 \leq i \leq n)$ are independent and identically distributed (i.i.d.) latent random variables, and $V_n = (V_{ij} : 1 \leq i < j \leq n)$ are i.i.d. latent random variables independent of $A_n$. A natural interpretation of this data is as a complete undirected network on $n$ vertices, with the latent variable $A_i$ associated with node $i$ and the observed variable $W_{ij}$ associated with the edge between nodes $i$ and $j$. The data generating process above is justified without loss of generality by the celebrated Aldous–Hoover representation theorem for exchangeable arrays (Aldous, 1981; Hoover, 1979). Consequently, the analysis of dyadic data poses specific challenges due to its inherent lack of statistical independence (Davezies et al., 2021; Chiang et al., 2022).

Various distributional features of dyadic data are of interest in applications. Most of the statistical literature focuses on parametric analysis, almost exclusively considering moments of (transformations of) the identically distributed $W_{ij}$. See Davezies et al. (2021), Gao and Ma (2021), Matsushita and Otsu (2021), and references therein, for contemporary contributions and overviews. More recently, however, a few nonparametric procedures for dyadic data have been proposed in the literature, with Lebesgue density estimation being a leading example (Graham et al., 2019). This paper contributes to the literature by offering an array of uniform estimation and inference results for kernel density
estimators with dyadic data which can be used in empirical analysis of network data.

More precisely, with the aim of estimating nonparametric density-like functions associated with $W_{ij}$ using kernel-based methods, we investigate the statistical properties of a class of local stochastic processes given by

$$w \mapsto \hat{f}_W(w) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_h(W_{ij}, w),$$

where $k_h(\cdot, w)$ is a kernel-like function that can change with the $n$-varying bandwidth parameter $h = h(n)$ and the evaluation point $w \in \mathcal{W} \subseteq \mathbb{R}$. For each $w \in \mathcal{W}$ and with an appropriate choice of kernel function (e.g. $k_h(\cdot, w) = K((\cdot - w)/h)/h$ for $w$ an interior point of $\mathcal{W}$ and $K$ a fixed symmetric integrable kernel function), the statistic $\hat{f}_W(w)$ becomes a kernel density estimator for the Lebesgue density function

$$f_W(w) = \mathbb{E}[f_{W|AA}(w \mid A_i, A_j)],$$

where $f_{W|AA}(w \mid A_i, A_j)$ denotes the conditional Lebesgue density of $W_{ij}$ given $A_i$ and $A_j$. The estimand $f_W(w)$ is useful in applications because it characterizes the distribution of the dyadic data and because it forms the basis of other parameters of interest. Setting $k_h(\cdot, w) = K((\cdot - w)/h)/h$, Graham et al. (2019) recently introduced the dyadic point estimator $\hat{f}_W(w)$ and studied its large sample properties pointwise in $w \in \mathcal{W} = \mathbb{R}$, while Chiang and Tan (2020) established its rate of convergence uniformly in $w \in \mathcal{W}$ for a compact interval $\mathcal{W}$ strictly contained in the support of the dyadic data $W_{ij}$. Chiang et al. (2022) obtained a distributional approximation for the supremum statistic $\sup_{w \in \mathcal{W}} |\hat{f}_W(w)|$ over a finite collection $\mathcal{W}$ of design points.

Allowing for a compact domain $\mathcal{W}$, which may or may not coincide with the support of $W_{ij}$, and employing boundary-adaptive kernel-like functions $k_h(\cdot, w)$ if needed, we contribute to the emerging literature on nonparametric smoothing methods for dyadic data with two main results. Firstly, we derive the minimax rate of uniform convergence for density estimation with dyadic data and show that the kernel density estimator $\hat{f}_W$ in (1) is minimax-optimal under appropriate conditions. Secondly,
we present a complete set of uniform distributional approximation results for the entire stochastic process \( \hat{f}_W(w) : w \in \mathcal{W} \). We illustrate the usefulness of our main results by constructing feasible and valid confidence bands for \( f_W \). Furthermore, our results lay the foundation for studying the uniform distributional properties of other non-/semiparametric estimators based on dyadic data. Importantly, our inference results cannot be deduced from the existing U-statistic, empirical process and U-process theory available in the literature (van der Vaart and Wellner, 1996; van der Vaart, 1998; Giné and Nickl, 2021) because, as explained in detail below, \( \hat{f}_W(w) \) is not a standard U-statistic, nor is the stochastic process \( \hat{f}_W \) Donsker in general, and the underlying dyadic data \( W_n \) exhibits statistical dependence due to its network structure.

Section 2 outlines the setup and presents the main assumptions imposed throughout the paper. We first discuss a Hoeffding-type decomposition of the U-statistic-like \( \hat{f}_W \) which is more general than the standard Hoeffding decomposition for second-order U-statistics (van der Vaart, 1998, Chapter 12) due to its dyadic data structure. In particular, (2) shows that \( \hat{f}_W(w) \) decomposes into a sum of four terms \( B_n(w), L_n(w), E_n(w) \) and \( Q_n(w) \), and we note that the \( E_n(w) \) term is not present in classical second-order U-statistic theory. The first term \( B_n(w) \) captures the usual smoothing bias, the second term \( L_n(w) \) is akin to the Hájek projection for second-order U-statistics, the third term \( E_n(w) \) is a mean-zero double average of conditionally independent terms, and the fourth term \( Q_n(w) \) is a negligible totally degenerate second-order U-process. Both \( L_n \) and \( E_n \) capture the leading stochastic fluctuations of the process \( \hat{f}_W \), and both are known to be asymptotically distributed as Gaussian random variables pointwise in \( w \in \mathcal{W} \) (Graham et al., 2019). However, the Hájek projection term \( L_n \) will often be “degenerate” at some or possibly all evaluation points \( w \in \mathcal{W} \). Section 2 formalizes and illustrates these phenomena, highlighting the importance of accounting for the potential degeneracy of \( L_n \) in our uniform analysis of \( \hat{f}_W \).

Section 3 studies minimax convergence rates for point estimation of \( f_W \) uniformly over \( \mathcal{W} \) and gives precise conditions under which the kernel-based density estimator \( \hat{f}_W \) is minimax-optimal. Firstly, in Theorem 3.1 we establish the uniform rate of convergence of \( \hat{f}_W \) for \( f_W \). This result improves upon the
recent paper of Chiang and Tan (2020) by allowing for compactly supported dyadic data and generic
kernel-like functions $k_n$ (including boundary-adaptive kernels), while also explicitly accounting for
possible degeneracy of the Hájek projection term $L_n$ at some or possibly all points $w \in \mathcal{W}$. Secondly,
in Theorem 3.2 we derive the minimax uniform convergence rate for estimating $f_W$, again allowing
for possible degeneracy, and verify that it is achieved by the kernel-based estimator $\hat{f}_W$. This result
appears to be new to the literature, complementing recent work on parametric moment estimation
using graphon data (Gao and Ma, 2021) and on nonparametric kernel-based regression using dyadic
data (Graham et al., 2021).

Section 4 presents a comprehensive distributional analysis of the stochastic process $\hat{f}_W$ uniformly
in $w \in \mathcal{W}$. Because $\hat{f}_W$ is not asymptotically tight (i.e., non-Donsker) in general, it does not converge
weakly in the space of uniformly bounded real functions supported on $\mathcal{W}$ and equipped with the uniform
norm (van der Vaart and Wellner, 1996). To circumvent this problem, we employ strong approximation
methods to characterize the distributional properties of $\hat{f}_W$. Up to the smoothing bias term $B_n$ and
the negligible term $Q_n$, it is enough to consider the stochastic process $w \mapsto L_n(w) + E_n(w)$. Since
$L_n$ can be degenerate at some or possibly all points $w \in \mathcal{W}$, and also because under some bandwidth
choices both $L_n$ and $E_n$ can be of comparable order, it is crucial to analyze the joint distributional
properties of $L_n$ and $E_n$. To do so, we employ a carefully crafted conditioning approach where we first
establish an unconditional strong approximation for $L_n$ and a conditional-on-$A_n$ strong approximation
for $E_n$. We then combine these to obtain a strong approximation for $L_n + E_n$.

The stochastic process $L_n$ is an empirical process indexed by an $n$-varying class of functions
depending only on the i.i.d. random variables $A_n$. Thus we use the celebrated Hungarian construction
(Komlos et al., 1975), building on ideas in Giné et al. (2004) and Giné and Nickl (2010). The resulting
rate of strong approximation is optimal, and follows from a generic strong approximation result of
potential independent interest in Lemma A.1. Our main result for $L_n$ is given as Lemma 4.1, and
makes explicit the potential presence of degenerate points.

The stochastic process $E_n$ is an empirical process depending on the dyadic variables $W_{ij}$ and indexed
by an \( n \)-varying class of functions. When conditioning on \( A_n \), the variables \( W_{ij} \) are independent but not necessarily identically distributed (i.n.i.d.), and thus we establish a conditional-on-\( A_n \) strong approximation for \( E_n \) based on the Yurinskii coupling (Yurinskii, 1978), leveraging a recent refinement obtained by Belloni et al. (2019, Lemma 38). This result follows from Lemma A.2, a generic strong approximation result which may also be of independent interest as it gives a novel rate of strong approximation for (local) empirical processes based on i.n.i.d. data. Lemma 4.2 gives our conditional strong approximation for \( E_n \).

Once the unconditional strong approximation for \( L_n \) and the conditional-on-\( A_n \) strong approximation for \( E_n \) are established, we show how to properly “glue” them together to deduce a final unconditional strong approximation for \( L_n + E_n \) and hence also for \( \hat{f}_W \) and its associated \( t \)-process. This final step requires some additional technical work. Firstly, building on our conditional strong approximation for \( E_n \), we establish an unconditional strong approximation for \( E_n \) in Lemma 4.3. We then employ a generalization of the celebrated Vorob’ev–Berkes–Philipp theorem (Dudley, 1999) given as Lemma A.3, which might also be of independent interest, to deduce a joint strong approximation for \( (L_n, E_n) \) and, in particular, for \( L_n + E_n \). Thus we obtain our main result in Theorem 4.1, which establishes a valid strong approximation for \( \hat{f}_W \) and its associated \( t \)-process. This uniform inference result complements the recent contribution of Davezies et al. (2021), which is not applicable in our context because the U-process-like statistic \( \hat{f}_W \) is not Donsker in general.

We illustrate the applicability of our strong approximation result for \( \hat{f}_W \) and its associated \( t \)-process by constructing valid standardized confidence bands for the unknown density function \( f_W \). Instead of relying on extreme value theory (e.g. Giné et al., 2004), we employ anti-concentration methods to deduce a pre-asymptotic coverage error rate for the confidence bands, following Chernozhukov et al. (2014a). This illustration improves on the recent work of Chiang et al. (2022), which obtained simultaneous confidence intervals for the dyadic density \( f_W \) based on a high-dimensional central limit theorem over rectangles, following prior work in Chernozhukov et al. (2017). The distributional approximation therein is applied to the Hájek projection term \( L_n \) only, whereas our main construction leading to Theorem 4.1
gives a strong approximation for the entire U-process-like $\hat{f}_W$ and its associated $t$-process, uniformly on $\mathcal{W}$. As a consequence, our uniform inference theory is robust to potential unknown degeneracies in $L_n$ by virtue of our strong approximation of $L_n + E_n$ and the use of proper standardization, delivering a “rate-adaptive” inference procedure. In the setting of dyadic density estimation, our result appears to be the first to provide confidence bands that are valid uniformly over $w \in \mathcal{W}$ rather than over some finite collection of design points. Moreover, our results provide distributional approximations for the whole $t$-statistic process of $\hat{f}_W$, which can be useful in applications where functionals other than the supremum are of interest.

Section 5 addresses outstanding issues of implementation. Firstly, we discuss estimation of the covariance function of the Gaussian process underlying our strong approximation results. We present two estimators, one based on the plug-in method, and the other based on a positive semi-definite regularization thereof (Laurent and Rendl, 2005). We derive the uniform convergence rates for both estimators in Lemma 5.1, which we then use to justify Studentization of $\hat{f}_W$ and a feasible simulation-based approximation of the infeasible Gaussian process underlying our strong approximation results. Secondly, we discuss integrated mean squared error (IMSE) bandwidth selection and provide a simple rule-of-thumb implementation for applications (Wand and Jones, 1994; Simonoff, 2012). Thirdly, we develop feasible, valid uniform inference for $f_W$ employing robust bias-correction methods (Calonico et al., 2018, 2022).

Section 6 reports empirical evidence for our proposed feasible robust bias-corrected confidence bands for $f_W$. In Section 6.1 we use simulations to show that these confidence bands are robust to potential unknown degenerate points in the underlying data generating process, while in Section 6.2 we illustrate our methods with a real dyadic data set recording bilateral trade between countries.

Finally, Appendix A collects several technical results that may be of independent interest, including two generic strong approximation theorems for empirical processes, a generalized Vorob’ev–Berkes–Philipp theorem and a maximal inequality for i.n.i.d. random variables. The online supplemental appendix includes other technical and methodological results, complete proofs and additional details.
omitted here to conserve space.

### 1.1 Notation

We use the following standard notation and conventions throughout the paper. See the supplemental appendix for more details and further references.

**Norms.** The total variation norm of a real-valued function $g$ of a single real variable is defined as

$$
\|g\|_{TV} = \sup_{n \geq 1} \sup_{x_1 \leq \ldots \leq x_n} \sum_{i=1}^{n-1} |g(x_{i+1}) - g(x_i)|. 
$$

**Sets.** For an integer $m \geq 0$, denote by $C^m(X)$ the space of all $m$-times continuously differentiable functions on $X$. For $\beta > 0$ and $C > 0$, define the Hölder class on $X$

$$
H^\beta_C(X) = \left\{ g \in C^\beta(X) : \max_{1 \leq r \leq \beta} |g^{(r)}(x) - g^{(r)}(x')| \leq C|x - x'|^{\beta - \tilde{\beta}}, \ \forall x, x' \in X \right\}
$$

where $\tilde{\beta}$ denotes the largest integer which is strictly less than $\beta$. For $a \in \mathbb{R}$ and $b \geq 0$, we write $[a \pm b]$ for the interval $[a - b, a + b]$.

**Inequalities.** For non-negative sequences $a_n$ and $b_n$, write $a_n \lesssim b_n$ or $a_n = O(b_n)$ to indicate that $a_n / b_n$ is bounded for $n \geq 1$. Write $a_n \ll b_n$ or $a_n = o(b_n)$ if $a_n / b_n \to 0$. If $a_n \lesssim b_n \lesssim a_n$, write $a_n \asymp b_n$. For random non-negative sequences $A_n$ and $B_n$, write $A_n \lesssim \mathbb{P} B_n$ or $A_n = O_{\mathbb{P}}(B_n)$ if $A_n / B_n$ is eventually bounded in probability. Write $A_n = o_{\mathbb{P}}(A_n)$ if $A_n / B_n \to 0$ in probability.

### 2 Setup

We impose the following two assumptions throughout this paper.

**Assumption 2.1 (Data generation)**

Let $A_n = (A_i : 1 \leq i \leq n)$ be i.i.d. random variables supported on $A \subseteq \mathbb{R}$ and let $V_n = (V_{ij} : 1 \leq i < j \leq n)$ be i.i.d. random variables with a Lebesgue density $f_V$ on $\mathbb{R}$, with $A_n$ independent of $V_n$. Let $W_{ij} = W(A_i, A_j, V_{ij})$ and $W_n = (W_{ij} : 1 \leq i < j \leq n)$, where $W$ is an unknown real-valued function which is symmetric in its first two arguments. Let $W \subseteq \mathbb{R}$ be a compact interval with positive Lebesgue...
measure \text{Leb}(W). The conditional distribution of \( W_{ij} \) given \( A_i \) and \( A_j \) admits a Lebesgue density \( f_{W|AA}(w \mid A_i, A_j) \). For \( C_H > 0 \) and \( \beta \geq 1 \), \( f_W \in \mathcal{H}^\beta_{C_H}(W) \) where \( f_W(w) = \mathbb{E}[f_{W|AA}(w \mid A_i, A_j)] \) and \( f_{W|AA}(\cdot \mid a, a') \in \mathcal{H}^1_{C_H}(W) \) for all \( a, a' \in A \). Suppose \( \sup_{w \in W} \|f_{W|A}(w \mid \cdot)\|_{TV} < \infty \) where \( f_{W|A}(w \mid a) = \mathbb{E}[f_{W|AA}(w \mid A_i, a)] \).

In Assumption 2.1 we require the density \( f_W \) be in a \( \beta \)-smooth Hölder class of functions on the compact interval \( W \). As such we cover not only distributions with everywhere-smooth densities such as the Gaussian distribution, but also those with smooth densities up to a boundary such as uniform and exponential distributions. Under Assumption 2.1, the (conditional) densities \( f_W, f_{W|A} \) and \( f_{W|AA} \) are all uniformly bounded by \( C_d := 2\sqrt{C_H} + 1/\text{Leb}(W) \).

If \( W(a_1, a_2, v) \) is strictly monotonic and continuously differentiable in its third argument, we can give the conditional density of \( W_{ij} \) explicitly using the usual change-of-variables formula: with \( w = W(a_1, a_2, v) \), \( f_{W|AA}(w \mid a_1, a_2) = f_V(v)|\partial W(a_1, a_2, v)/\partial v|^{-1} \). However this is not necessary for our results.

**Assumption 2.2 (Kernels and bandwidth)**

Let \( h = h(n) > 0 \) be a sequence of bandwidths satisfying \( h \log n \to 0 \) and \( \frac{\log n}{n^d} \to 0 \). For each \( w \in W \), let \( k_h(\cdot, w) \) be a real-valued function supported on \([w \pm h] \cap W\). For an integer \( p \geq 1 \), let \( k_h \) belong to a family of boundary bias-corrected kernels of order \( p \), i.e.,

\[
\int_W (s - w)^r k_h(s, w) \, ds \begin{cases} 
= 1 & \text{for all } w \in W \text{ if } r = 0, \\
= 0 & \text{for all } w \in W \text{ if } 1 \leq r \leq p - 1, \\
\neq 0 & \text{for some } w \in W \text{ if } r = p.
\end{cases}
\]

Also, for \( C_L > 0 \), suppose \( k_h(s, \cdot) \in \mathcal{H}^1_{C_L h^{-2}}(W) \) for all \( s \in W \).

This assumption allows for all standard compact-supported, possibly boundary-corrected, kernel functions (Wand and Jones, 1994; Simonoff, 2012). Assumption 2.2 implies that if \( h \leq 1 \) then \( k_h \) is uniformly bounded by \( C_k h^{-1} \) where \( C_k := 2C_L + 1 + 1/\text{Leb}(W) \).
2.1 Hoeffding-type decomposition and degeneracy

The estimator $\hat{f}_W(w)$ is akin to a U-statistic and thus admits a Hoeffding-type decomposition which is the starting point for our analysis. We have

$$\hat{f}_W(w) - f_W(w) = B_n(w) + L_n(w) + E_n(w) + Q_n(w)$$  \hspace{1cm} (2)

with

$$B_n(w) = \mathbb{E}[\hat{f}_W(w)] - f_W(w),$$

$$L_n(w) = \frac{2}{n} \sum_{i=1}^{n} l_i(w), \quad E_n(w) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} e_{ij}(w), \quad Q_n(w) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} q_{ij}(w),$$

where

$$l_i(w) = \mathbb{E}[k_h(W_{ij}, w) | A_i] - \mathbb{E}[k_h(W_{ij}, w)],$$

$$e_{ij}(w) = k_h(W_{ij}, w) - \mathbb{E}[k_h(W_{ij}, w) | A_i, A_j],$$

$$q_{ij}(w) = \mathbb{E}[k_h(W_{ij}, w) | A_i, A_j] - \mathbb{E}[k_h(W_{ij}, w) | A_i] - \mathbb{E}[k_h(W_{ij}, w) | A_j] + \mathbb{E}[k_h(W_{ij}, w)].$$

The non-random term $B_n$ captures the smoothing (or misspecification) bias, while the three stochastic processes $L_n$, $E_n$ and $Q_n$ capture the variance of the estimator. These processes are mean-zero: $\mathbb{E}[L_n(w)] = \mathbb{E}[Q_n(w)] = \mathbb{E}[E_n(w)] = 0$ for all $w \in \mathcal{W}$, and mutually orthogonal in $L^2(\mathbb{P})$: $\mathbb{E}[L_n(w)Q_n(w')] = \mathbb{E}[L_n(w)E_n(w')] = \mathbb{E}[Q_n(w)E_n(w')] = 0$ for all $w, w' \in \mathcal{W}$.

The stochastic process $L_n$ is akin to the Hájek projection of a U-process, which can (and often will) exhibit degeneracy at some or possibly all points $w \in \mathcal{W}$. To characterize different types of degeneracy, we introduce the following non-negative lower and upper degeneracy constants:

$$D_{lo}^2 := \inf_{w \in \mathcal{W}} \text{Var} \left[ f_{W|A}(w | A_i) \right] \quad \text{and} \quad D_{up}^2 := \sup_{w \in \mathcal{W}} \text{Var} \left[ f_{W|A}(w | A_i) \right].$$
The following lemma describes the order of different terms in the Hoeffding-type decomposition, explicitly accounting for potential degeneracy. For $a, b \in \mathbb{R}$, define $a \wedge b = \min\{a, b\}$.

**Lemma 2.1** (Bias and variance)

Suppose that Assumptions 2.1 and 2.2 hold. Then the bias term satisfies

$$\sup_{w \in \mathcal{W}} |B_n(w)| \lesssim h^{p_3}$$

and the variance terms satisfy

$$\mathbb{E} \left[ \sup_{w \in \mathcal{W}} |L_n(w)| \right] \lesssim \frac{D_{up}}{\sqrt{n}}, \quad \mathbb{E} \left[ \sup_{w \in \mathcal{W}} |E_n(w)| \right] \lesssim \sqrt{\frac{\log n}{n^2 h}}, \quad \mathbb{E} \left[ \sup_{w \in \mathcal{W}} |Q_n(w)| \right] \lesssim \frac{1}{n}.$$

Lemma 2.1 captures the potential total degeneracy of $L_n$ by showing that if $D_{up} = 0$ then $L_n = 0$ everywhere on $\mathcal{W}$ almost surely. The following lemma captures the potential partial degeneracy of $L_n$, where $D_{up} > D_{lo} = 0$. Define the covariance function of the dyadic kernel density estimator: for $w, w' \in \mathcal{W}$,

$$\Sigma_n(w, w') = \mathbb{E} \left[ (\hat{f}_W(w) - \mathbb{E}[\hat{f}_W(w)]) (\hat{f}_W(w') - \mathbb{E}[\hat{f}_W(w')]) \right].$$

**Lemma 2.2** (Variance bounds)

Suppose that Assumptions 2.1 and 2.2 hold. Then for sufficiently large $n$,

$$\frac{D_{lo}^2}{n} + \frac{1}{n^2 h} \inf_{w \in \mathcal{W}} f_W(w) \lesssim \inf_{w \in \mathcal{W}} \Sigma_n(w, w) \lesssim \sup_{w \in \mathcal{W}} \Sigma_n(w, w) \lesssim \frac{D_{up}^2}{n} + \frac{1}{n^2 h}.$$

Combining Lemmas 2.1 and 2.2, we have the following trichotomy for degeneracy of dyadic distributions based on $D_{lo}$ and $D_{up}$:

(i) Total degeneracy: $D_{up} = D_{lo} = 0$,

(ii) Partial degeneracy: $D_{up} > D_{lo} = 0$,

(iii) No degeneracy: $D_{lo} > 0$. 

10
In the case of no degeneracy, it can be shown that \( \inf_{w \in \mathcal{W}} \text{Var}[L_n(w)] \gtrsim n^{-1} \), while in the case of total degeneracy, \( L_n(w) = 0 \) for all \( w \in \mathcal{W} \) almost surely. When the dyadic distribution is partially degenerate, there exists at least one point \( w \in \mathcal{W} \) such that \( \text{Var}[f_{W|A}(w | A_i)] = 0 \) and \( \text{Var}[L_n(w)] \lesssim h n^{-1} \), and there also exists at least one point \( w' \in \mathcal{W} \) such that \( \text{Var}[f_{W|A}(w' | A_i)] > 0 \) and \( \text{Var}[L_n(w')] \geq \frac{2}{n} \text{Var}[f_{W|A}(w' | A_i)] \) for sufficiently large \( n \). We say \( w \) is a degenerate point if \( \text{Var}[f_{W|A}(w | A_i)] = 0 \), and otherwise say it is a non-degenerate point.

As a simple example, consider the family of dyadic distributions \( \mathbb{P}_\pi \) indexed by \( \pi = (\pi_1, \pi_2, \pi_3) \) with \( \sum_{i=1}^{3} \pi_i = 1 \) and \( \pi_i \geq 0 \), generated by

\[
W_{ij} = A_i A_j + V_{ij},
\]

where \( A_i \) equals \(-1\) with probability \( \pi_1 \), equals \( 0 \) with probability \( \pi_2 \) and equals \(+1\) with probability \( \pi_3 \), and \( V_{ij} \) is standard Gaussian. In line with Assumption 2.1, \( A_n \) and \( V_n \) are i.i.d. sequences independent of each other. Then with \( \phi \) denoting the probability density function of the standard normal distribution,

\[
f_{W|AA}(w | A_i, A_j) = \phi(w - A_i A_j),
\]

\[
f_{W|A}(w | A_i) = \pi_1 \phi(w + A_i) + \pi_2 \phi(w) + \pi_3 \phi(w - A_i),
\]

\[
f_W(w) = (\pi_1^2 + \pi_3^2) \phi(w - 1) + \pi_2 (2 - \pi_2) \phi(w) + 2 \pi_1 \pi_3 \phi(w + 1).
\]

Note that \( f_W(w) \) is strictly positive for all \( w \in \mathbb{R} \). Consider the parameter choices

(i) \( \pi = (\frac{1}{2}, 0, \frac{1}{2}) \): \( \mathbb{P}_\pi \) is degenerate at all \( w \in \mathbb{R} \),

(ii) \( \pi = (\frac{1}{4}, 0, \frac{3}{4}) \): \( \mathbb{P}_\pi \) is degenerate only at \( w = 0 \),

(iii) \( \pi = (\frac{1}{5}, \frac{1}{5}, \frac{3}{5}) \): \( \mathbb{P}_\pi \) is non-degenerate for all \( w \in \mathbb{R} \).

Figure 1 demonstrates these phenomena, plotting the unconditional density \( f_W \) and the standard
deviation of the conditional density $f_{W|A}$ over $W = [-2, 2]$ for each choice of the parameter $\pi$.

![Graphs showing density and standard deviation for different degeneracy cases](image)

Figure 1: Density $f_W$ and standard deviation of $f_{W|A}$ for the family of distributions $P_{\pi}$.

Notes. Panel (a): $\pi = \left(\frac{1}{2}, 0, \frac{1}{2}\right)$, and $P_{\pi}$ is degenerate for all $w \in \mathbb{R}$. Panel (b): $\pi = \left(\frac{1}{4}, 0, \frac{3}{4}\right)$, and $P_{\pi}$ is degenerate only at $w = 0$. Panel (c): $\pi = \left(\frac{1}{5}, \frac{1}{5}, \frac{3}{5}\right)$, and $P_{\pi}$ is non-degenerate for all $w \in \mathbb{R}$.

The trichotomy of total/partial/no degeneracy is useful for understanding the distributional properties of the dyadic kernel density estimator $\hat{f}_W(w)$. Crucially, our need for uniformity in $w$ complicates the simpler degeneracy/no degeneracy dichotomy observed previously in the literature (Graham et al., 2019). More specifically, from a pointwise-in-$w$ perspective, partial degeneracy causes no issues, while it is a fundamental problem when conducting inference uniformly over $w \in W$. In this paper, we develop inference methods that are valid uniformly over $w \in W$, regardless of the presence of partial or total degeneracy.

3 Point estimation results

We now study the uniform point estimation properties of the dyadic kernel density estimator $\hat{f}_W$. As an immediate implication of Lemma 2.1, the next theorem establishes the rate of uniform convergence of $\hat{f}_W$.

**Theorem 3.1** (Uniform convergence rate)
Suppose that Assumptions 2.1 and 2.2 hold. Then

\[
\mathbb{E} \left[ \sup_{w \in W} |\hat{f}_W(w) - f_W(w)| \right] \lesssim h^{p \wedge \beta} + \frac{D_{\text{up}}}{\sqrt{n}} + \sqrt{\frac{\log n}{n^2 h}}.
\]

The constant in Theorem 3.1 depends only on $W$, $\beta$, $C_H$ and the choice of kernel. We interpret this result in light of the degeneracy trichotomy.

(i) Partial or no degeneracy: $D_{\text{up}} > 0$. Any bandwidth sequence satisfying $n^{-1} \log n \lesssim h \lesssim n^{-\frac{1}{2(p \wedge \beta)}}$ yields

\[
\mathbb{E} \left[ \sup_{w \in W} |\hat{f}_W(w) - \mathbb{E}[\hat{f}_W(w)]| \right] \lesssim \frac{1}{\sqrt{n}},
\]

the “parametric” bandwidth-independent rate noted by Graham et al. (2019).

(ii) Total degeneracy: $D_{\text{up}} = 0$. Minimizing the upper bound in Theorem 3.1 by setting $h \asymp \left( \frac{\log n}{n^2} \right)^{\frac{1}{2(p \wedge \beta)+1}}$ yields

\[
\mathbb{E} \left[ \sup_{w \in W} |\hat{f}_W(w) - f_W(w)| \right] \lesssim \left( \frac{\log n}{n^2} \right)^{\frac{p \wedge \beta}{2(p \wedge \beta)+1}}.
\]

These results generalize Chiang and Tan (2020, Theorem 1) by allowing for compactly supported data and more general kernel-like functions $k_h(\cdot, w)$, enabling boundary-adaptive density estimation.

### 3.1 Minimax optimality

We establish the minimax rate under the supremum norm for density estimation with dyadic data. This implies minimax optimality of the kernel density estimator $\hat{f}_W$, regardless of the degeneracy type of the dyadic distribution.

**Theorem 3.2 (Uniform minimax rate)**

Fix $\beta \geq 1$ and $C_H > 0$, and take $W$ a compact interval with positive Lebesgue measure. Define
$\mathcal{P} = \mathcal{P}(W, \beta, C_H)$ as the class of dyadic distributions satisfying Assumption 2.1. Define $\mathcal{P}_d$ as the subclass of $\mathcal{P}$ containing only those dyadic distributions which are totally degenerate on $W$ in the sense that $\sup_{w \in W} \text{Var}_{f_{W|A}(w | A_i)} = 0$. Then

\[
\inf_{\hat{f}_W \in \mathcal{P}} \sup_{f_W \in \mathcal{P}} \mathbb{E}_\mathbb{P} \left[ \sup_{w \in W} \left| \hat{f}_W(w) - f_W(w) \right| \right] \propto \frac{1}{\sqrt{n}},
\]

\[
\inf_{\hat{f}_W \in \mathcal{P}_d} \sup_{f_W \in \mathcal{P}_d} \mathbb{E}_\mathbb{P} \left[ \sup_{w \in W} \left| \hat{f}_W(w) - f_W(w) \right| \right] \propto \left( \frac{\log n}{n^2} \right)^{\frac{\beta}{2\beta+1}},
\]

where $\hat{f}_W$ is any estimator depending only on the data $W_n = (W_{ij} : 1 \leq i < j \leq n)$ distributed according to the dyadic law $\mathbb{P}$. The constants underlying $\propto$ depend only on $W$, $\beta$ and $C_H$.

Theorem 3.2 shows that the uniform convergence rate of $n^{-1/2}$ obtained in Theorem 3.1 (coming from the $L_n$ term) is minimax-optimal in general. When attention is restricted to totally degenerate dyadic distributions, $\hat{f}_W$ also achieves the minimax rate of uniform convergence (assuming a kernel of sufficiently high order $p \geq \beta$), which is on the order of $\left( \frac{\log n}{n^2} \right)^{\frac{\beta}{2\beta+1}}$ and is determined by the bias $B_n$ and the leading variance term $E_n$ in (2).

Combining Theorems 3.1 and 3.2, we conclude that the estimator $\hat{f}_W(w)$ achieves the minimax-optimal rate of uniform convergence for estimating $f_W(w)$ if

\[
h \propto \left( \frac{\log n}{n^2} \right)^{\frac{1}{2\beta+1}}
\]

and $p \geq \beta$, whether or not there are any degenerate points in the underlying data generating process. This result appears to be new to the literature on nonparametric estimation with dyadic data. See Gao and Ma (2021) for a contemporaneous review.

### 4 Distributional results

Next we investigate the distributional properties of the standardized $t$-statistic process

\[
T_n(w) = \frac{\hat{f}_W(w) - f_W(w)}{\sqrt{\Sigma_n(w, w)}}, \quad w \in W.
\]
The stochastic process \( (T_n(w) : w \in \mathcal{W}) \) is not necessarily asymptotically tight, and hence it may not converge weakly on the space of uniformly bounded real functions supported on \( \mathcal{W} \) and equipped with the uniform norm (van der Vaart and Wellner, 1996). Therefore, to approximate the distribution of the entire \( t \)-statistic process, as well as specific functionals thereof, we rely on a novel strong approximation approach outlined in this section. Our results can be used to perform valid uniform inference irrespective of the degeneracy type of the underlying dyadic distribution.

This section is largely concerned with distributional properties and thus frequently requires copies of stochastic processes. We say that \( X' \) is a copy of \( X \) if they have the same distribution, though they may be defined on different probability spaces. For succinctness of notation, we will not differentiate between a process and its copy, but further details are available in the supplemental appendix. Many of the technical details regarding the copying and embedding of stochastic processes are covered by a generalized Vorob’ev–Berkes–Philipp Theorem, which is stated and discussed in Lemma A.3. In particular, this theorem can be applied to random vectors or to stochastic processes indexed on a compact rectangle in \( \mathbb{R}^d \) with a.s. continuous sample paths.

4.1 Strong approximation

By the Hoeffding-type decomposition (2) and Lemma 2.1, it suffices to consider the distributional properties of the stochastic process \( (L_n(w) + E_n(w) : w \in \mathcal{W}) \). Our approach combines the Kőmlos–Major–Tusnády (KMT) approximation (Komlós et al., 1975) to obtain a strong approximation of \( L_n \) with a Yurinskii approximation (Yurinskii, 1978) to obtain a conditional (on \( A_n \)) strong approximation of \( E_n \). The latter is necessary because \( E_n \) is akin to a local empirical process of i.n.i.d. random variables, conditional on \( A_n \), and therefore the KMT approximation is not applicable. These approximations are then carefully combined to give a final (unconditional) strong approximation for \( L_n + E_n \), and thus for the \( t \)-statistic process \( (T_n(w) : w \in \mathcal{W}) \).

The following lemma is an application of our generic KMT approximation result for empirical processes, Lemma A.1, which builds on earlier work by Giné et al. (2004) and Giné and Nickl (2010).
and may be of independent interest.

**Lemma 4.1 (Strong approximation of $L_n$)**

Suppose that Assumptions 2.1 and 2.2 hold. For each $n$ there exists a mean-zero Gaussian process $Z_n^L$ indexed on $W$ with

$$
\mathbb{E}
\left[
\sup_{w \in W} \left| \sqrt{n}L_n(w) - Z_n^L(w) \right|
\right] \lesssim \frac{D_{\text{up}} \log n}{\sqrt{n}},
$$

where $Z_n^L$ has the same covariance structure as $\sqrt{n}L_n$, i.e. $\mathbb{E}[Z_n^L(w)Z_n^L(w')] = n\mathbb{E}[L_n(w)L_n(w')]$ for all $w, w' \in W$.

We also show that $Z_n^L$ has continuous trajectories and that for any $\delta_n \in (0, 1/2]$,

$$
\mathbb{E}
\left[
\sup_{|w-w'| \leq \delta_n} |Z_n^L(w) - Z_n^L(w')|
\right] \lesssim D_{\text{up}}\delta_n \sqrt{\log 1/\delta_n}.
$$

The process $Z_n^L$ is a function only of $A_n$ and some random noise independent of $(A_n, V_n)$. See Lemma A.1 for details.

The strong approximation result in Lemma 4.1 would be sufficient to develop valid and even optimal uniform inference procedures whenever (i) $D_{\text{lo}} > 0$ (no degeneracy in $L_n$) and (ii) $nh \gg \log n$ ($L_n$ is leading). In this special case, the recent Donsker-type results of Davezies et al. (2021) can be applied to analyze the limiting distribution of the stochastic process $\hat{f}_W$. Alternatively, again only when $L_n$ is non-degenerate and leading, standard empirical process methods could also be used (van der Vaart and Wellner, 1996; Giné and Nickl, 2021). However, even in the special case when $\hat{f}_W(w)$ is asymptotically Donsker, our result in Lemma 4.1 improves upon the literature by providing a rate-optimal strong approximation for $\hat{f}_W$ as opposed to only a weak convergence result. See Theorem 4.2 and the subsequent discussion below for more details.

More importantly, as illustrated above, it is common in the literature to find dyadic distributions which exhibit partial or total degeneracy, making the process $\hat{f}_W$ non-Donsker. Thus approximating only
Lemma 4.2 (Conditional strong approximation of $E_n$)

Suppose that Assumptions 2.1 and 2.2 hold. For each $n$ there exists $\tilde{Z}_n^E$ which is a mean-zero Gaussian process conditional on $A_n$ satisfying

$$\mathbb{E} \left[ \sup_{w \in \mathcal{W}} | \sqrt{n^2 h E_n(w)} - \tilde{Z}_n^E(w) | \right] \lesssim \frac{(\log n)^{3/8}}{n^{1/4} h^{3/8}},$$

where $\tilde{Z}_n^E$ has the same conditional covariance structure as $\sqrt{n^2 h E_n}$, i.e., $\mathbb{E}[\tilde{Z}_n^E(w) \tilde{Z}_n^E(w') \mid A_n] = n^2 h \mathbb{E}[E_n(w)E_n(w') \mid A_n]$ for all $w, w' \in \mathcal{W}$.

We also show that $\tilde{Z}_n^E$ has continuous trajectories and that for any $\delta_n \in (0, 1/(2h))$:

$$\mathbb{E} \left[ \sup_{|w-w'| \leq \delta_n} | \tilde{Z}_n^E(w) - \tilde{Z}_n^E(w') | \right] \lesssim \frac{\delta_n}{h} \left( \frac{1}{h \delta_n} \log \frac{1}{h \delta_n} \right).$$

The process $\tilde{Z}_n^E$ is a Gaussian process conditional on $A_n$ but is not in general a Gaussian process unconditionally. The following lemma further constructs an unconditional Gaussian process $Z_n^E$ that approximates $\tilde{Z}_n^E$.

Lemma 4.3 (Unconditional strong approximation of $E_n$)

Suppose that Assumptions 2.1 and 2.2 hold. For each $n$ there exists a mean-zero Gaussian process $Z_n^E$ satisfying

$$\mathbb{E} \left[ \sup_{w \in \mathcal{W}} | \tilde{Z}_n^E(w) - Z_n^E(w) | \right] \lesssim \frac{(\log n)^{2/3}}{n^{1/6}},$$

where $Z_n^E$ is independent of $A_n$ and has the same (unconditional) covariance structure as $\tilde{Z}_n^E$ and
\[ \sqrt{n^2 h} E_n, \text{ i.e., } \mathbb{E}[Z_n^E(w)Z_n^E(w')] = \mathbb{E}[\bar{Z}_n^E(w)\bar{Z}_n^E(w')] = n^2 h \mathbb{E}[E_n(w)E_n(w')] \text{ for all } w, w' \in W. \]

We also show that \( Z_n^E \) has continuous trajectories and that for any \( \delta_n \in (0, 1/(2h)) \),

\[
\mathbb{E}\left[ \sup_{|w-w'| \leq \delta_n} |Z_n^E(w) - Z_n^E(w'|)\right] \lesssim \frac{\delta_n}{h} \sqrt{\log \frac{1}{h \delta_n}}.
\]

Combining Lemmas 4.2 and 4.3, we obtain an unconditional strong approximation for \( E_n \). The resulting rate of approximation may not be optimal, due to the Yurinskii coupling, but to the best of our knowledge it is the first in the literature for the process \( E_n \), and hence for \( \hat{f}_W \) and its associated \( t \)-process in the context of dyadic data. We note that the approximation rate is sufficiently fast to allow for optimal bandwidth choices; see Section 5 for more details. Classical strong approximation results for local empirical processes (e.g. Giné and Nickl, 2010) are not applicable here because the summands in the non-negligible \( E_n \) are not (conditionally) i.i.d. Likewise, neither standard empirical process and U-process theory (van der Vaart and Wellner, 1996; Giné and Nickl, 2021) nor the recent results in Davezies et al. (2021) are applicable to the non-Donsker process \( E_n \).

Now we are ready to construct the strong approximation for the \( t \)-statistic process \( T_n \). The previous lemmas showed that \( L_n \) is \( \sqrt{n} \)-consistent while \( E_n \) is \( \sqrt{n^2 h} \)-consistent (pointwise in \( w \)), showcasing the importance of careful standardization (cf. Studentization in Section 5) for the purpose of rate adaptivity to the unknown degeneracy type. In other words, a fundamental challenge in conducting uniform inference is that the finite-dimensional distributions of the stochastic process \( L_n + E_n \), and hence those of \( \hat{f}_W \) and its associated \( t \)-process, may converge at different rates at different points \( w \in W \). Nevertheless, the following theorem provides an inference procedure which is fully adaptive to such potential unknown degeneracy.

**Theorem 4.1** (Strong approximation of \( T_n \))

*Suppose that Assumptions 2.1 and 2.2 hold and \( f_W(w) > 0 \) on \( W \). Then for each \( n \) there exists a*
centered Gaussian process $Z_n^T$ such that

$$\mathbb{E} \left[ \sup_{w \in W} |T_n(w) - Z_n^T(w)| \right] \lesssim \frac{n^{-1} \log n + n^{-5/4} h^{-7/8} (\log n)^{3/8} + n^{-7/6} h^{-1/2} (\log n)^{2/3} + h^{p \wedge \beta}}{D_{lo}/\sqrt{n} + 1/\sqrt{n^2 h}},$$

where $Z_n^T$ has the same covariance structure as $T_n$, i.e. $\mathbb{E}[Z_n^T(w) Z_n^T(w')] = \mathbb{E}[T_n(w) T_n(w')]$ for all $w, w' \in W$.

The first term in the numerator corresponds to the strong approximation error for $L_n$ characterized in Lemma 4.1 and the error introduced by $Q_n$. The second and third terms correspond to the conditional and unconditional strong approximation errors for $E_n$ given in Lemmas 4.2 and 4.3 respectively. The fourth term is from the smoothing bias result in Lemma 2.1. The denominator is the lower bound on the standard deviation $\Sigma_n(w, w)^{1/2}$ formulated in Lemma 2.2.

In the absence of degenerate points ($D_{lo} > 0$) and if $n h^{7/2} \gtrsim 1$ up to $\log n$ terms, Theorem 4.1 offers a strong approximation for the $t$-process at the rate $\log(n)/\sqrt{n} + \sqrt{n} h^{p \wedge \beta}$, which matches the celebrated KMT approximation rate for i.i.d. data plus the smoothing bias error. Therefore our novel $t$-process strong approximation can achieve the optimal KMT rate for non-degenerate dyadic distributions provided that $p \wedge \beta \geq 3.5$ (up to $\log n$ terms if equality holds). This is clearly achievable whenever a fourth-order (boundary-adaptive) kernel is used and $f_W$ is sufficiently smooth.

In the presence of partial or total degeneracy ($D_{lo} = 0$), Theorem 4.1 provides a strong approximation for the $t$-process at the rate $\sqrt{h} \log n + n^{-1/4} h^{-3/8} (\log n)^{3/8} + n^{-1/6} (\log n)^{2/3} + n h^{1/2 + p \wedge \beta}$. If for example $n h^{p \wedge \beta} \lesssim \log n$, then our result can achieve a strong approximation rate of $n^{-1/7}$ up to $\log n$ terms. We conjecture that this rate is improvable due to our construction approach (see Lemmas 4.2 and 4.3). Nonetheless, Theorem 4.1 is to the best of our knowledge the first in the literature for nonparametric kernel-based statistics based on dyadic data which is also robust to the presence of (unknown) degenerate points in the underlying dyadic distribution.
4.2 Application: confidence bands

To illustrate the usefulness of our main strong approximation result, Theorem 4.1, we construct standardized (infeasible) confidence bands for \( f_W \). In the next section (Section 5), we will make this inference procedure feasible by proposing a valid estimator of the covariance function \( \Sigma_n \) for Studentization, as well as developing bandwidth selection and robust bias-correction methods.

For \( \alpha \in (0, 1) \), let \( q_{1-\alpha} \) be the quantile of \( \sup_{w \in W} |Z_n^T(w)| \) satisfying

\[
P \left( \sup_{w \in W} |Z_n^T(w)| \leq q_{1-\alpha} \right) = 1 - \alpha.
\]

The following result employs the anti-concentration idea due to Chernozhukov et al. (2014a) to deduce valid standardized confidence bands, where we approximate the quantile of the unknown finite sample distribution of \( \sup_{w \in W} |T_n(w)| \) by the quantile \( q_{1-\alpha} \) of \( \sup_{w \in W} |Z_n^T(w)| \). This approach offers a better rate of convergence than relying on extreme value theory for the distributional approximation, hence improving the finite sample performance of the proposed confidence bands.

**Theorem 4.2** (Infeasible uniform confidence bands)

*Suppose that Assumptions 2.1 and 2.2 hold and \( f_W(w) > 0 \) on \( W \). Then

\[
\left| P \left( f_W(w) \in \left[ \hat{f}_W(w) \pm q_{1-\alpha} \sqrt{\Sigma_n(w, w)} \right] \text{ for all } w \in W \right) - (1 - \alpha) \right| \leq n^{-1/2} (\log n)^{3/4} + n^{-5/8} h^{-7/16} (\log n)^{7/16} + n^{-7/12} h^{-1/4} (\log n)^{7/12} + h^{\nu/2} (\log n)^{1/4} \frac{D'}{n^{1/4} + 1/(n^2 h)^{1/4}}.
\]

For the coverage error rate in Theorem 4.2 to converge to zero in large samples, we need further restrictions on the bandwidth sequence, which depend on the degeneracy type of the dyadic distribution. These are summarized in the following assumption.

**Assumption 4.1** (Rate restriction for uniform confidence bands)

*Assume that one of the following holds:
(i) No degeneracy \((D_0 > 0)\): \(n^{-6/7} \log n \ll h \ll (n \log n)^{-\frac{1}{2(p \land \beta)}}\),

(ii) Partial or total degeneracy \((D_0 = 0)\): \(n^{-2/3}(\log n)^{7/3} \ll h \ll (n^2 \log n)^{-\frac{1}{2(p \land \beta)+1}}\).

By Theorem 3.1, the asymptotically optimal choice of bandwidth for uniform convergence is
\(h \asymp \left(\frac{\log n}{n^2}\right)^{\frac{1}{2(p \land \beta)+1}}\). Similarly, as we show in the next section, the approximate integrated mean squared error optimal bandwidth is
\(h \asymp \left(\frac{1}{n^2}\right)^{\frac{1}{2(p \land \beta)+1}}\). Both bandwidth choices satisfy Assumption 4.1 only in the case of no degeneracy. The degenerate cases in Assumption 4.1(ii), which require \(p \land \beta > 1\), exhibit behavior more similar to that of standard nonparametric kernel-based estimation and so the aforementioned optimal bandwidth choices will lead to a non-negligible smoothing bias in the distributional approximation for \(T_n\). Different approaches are available in the literature to address this issue, including undersmoothing or ignoring the bias (Hall and Kang, 2001), bias correction (Hall, 1992), robust bias correction (Calonico et al., 2018, 2022) and Lepski’s method (Lepskii, 1992; Birgé, 2001), among others. In the next section we develop a feasible uniform inference procedure, based on robust bias-correction methods, which amounts to first selecting an optimal bandwidth for the point estimator \(\hat{f}_W\) using a \(p\)th-order kernel, and then correcting the bias of the point estimator while also adjusting the standardization (Studentization) when forming the \(t\)-statistic \(T_n\). One way to implement this approach is simply by using a kernel of higher order \(p' > p\) to construct the confidence bands.

Importantly, regardless of the specific implementation details, Theorem 4.2 shows that any bandwidth sequence \(h\) satisfying both (i) and (ii) in Assumption 4.1 leads to valid uniform inference which is robust and adaptive to the (unknown) degeneracy type of the underlying dyadic distribution, a crucial feature in network data settings.

5 Implementation

This section is concerned with the outstanding implementation details which make our main uniform inference results feasible in applications. In Section 5.1 we propose a covariance estimator along with a modified version which is positive semi-definite. This allows the construction of fully feasible confidence
bands in Section 5.2. In Section 5.3 we suggest some options for practical bandwidth selection and formalize our procedure for robust bias-correction inference.

5.1 Covariance function estimation

Define the following plug-in covariance function estimator of $\Sigma_n$: for $w, w' \in \mathcal{W}$,

$$\hat{\Sigma}_n(w, w') = \frac{4}{n^2} \sum_{i=1}^{n} S_i(w)S_i(w') - \frac{4}{n^2(n-1)^2} \sum_{i<j} k_h(W_{ij}, w)k_h(W_{ij}, w') - \frac{4n - 6}{n(n-1)} \tilde{f}_W(w)\tilde{f}_W(w'),$$

where

$$S_i(w) = \frac{1}{n-1} \left( \sum_{j=1}^{i-1} k_h(W_{ji}, w) + \sum_{j=i+1}^{n} k_h(W_{ij}, w) \right)$$

is an estimator of $\mathbb{E}[k_h(W_{ij}, w) \mid A_i]$. Though $\hat{\Sigma}_n(w, w')$ is consistent in an appropriate sense as shown in Lemma 5.1 below, it is not necessarily almost surely positive semi-definite, even in the limit. Therefore we propose a modified covariance estimator which is guaranteed to be positive semi-definite.

Specifically, consider the following optimization problem:

$$\min_{M \in \mathcal{W} \times \mathcal{W} \to \mathbb{R}} \sup_{w, w' \in \mathcal{W}} \left| \frac{M(w, w') - \hat{\Sigma}_n(w, w')}{{\sqrt{\hat{\Sigma}_n(w, w) + \hat{\Sigma}_n(w', w')}}} \right|$$

subject to $M$ is symmetric and positive semi-definite,

$$|M(w, w') - M(w, w'')| \leq \frac{4}{nh^3} C_k C_L |w' - w''| \text{ for all } w, w', w'' \in \mathcal{W}. \quad (4)$$

Denote by $\hat{\Sigma}_n^+$ any (approximately) optimal solution to (4). The following lemma establishes uniform convergence rates for both $\hat{\Sigma}_n$ and $\hat{\Sigma}_n^+$. It allows us to use these estimators to construct feasible versions of $T_n$ and its associated Gaussian approximation $Z_n^T$ defined in Theorem 4.1.

Lemma 5.1 (Consistency of $\hat{\Sigma}_n$ and $\hat{\Sigma}_n^+$)
Suppose that Assumptions 2.1 and 2.2 hold and that \( nh \gtrsim \log n \) and \( f_W(w) > 0 \) on \( \mathcal{W} \). Then

\[
\sup_{w, w' \in \mathcal{W}} \left| \frac{\hat{\Sigma}_n(w, w') - \Sigma_n(w, w')}{\sqrt{\Sigma_n(w, w) + \Sigma_n(w', w')}} \right| \lesssim_p \frac{\sqrt{\log n}}{n}.
\]

Also, the optimization problem (4) is a semi-definite program (SDP, Laurent and Rendl, 2005) and has an approximately optimal solution \( \hat{\Sigma}_n^+ \) satisfying

\[
\sup_{w, w' \in \mathcal{W}} \left| \frac{\hat{\Sigma}_n^+(w, w') - \Sigma_n(w, w')}{\sqrt{\Sigma_n(w, w) + \Sigma_n(w', w')}} \right| \lesssim_p \frac{\sqrt{\log n}}{n}.
\]

For finite-size covariance matrices, the semi-definite program defining \( \hat{\Sigma}_n^+(w, w') \) can be solved using a general-purpose SDP solver (e.g. using interior point methods, Laurent and Rendl, 2005).

### 5.2 Feasible confidence bands

Given a choice of the kernel order \( p \) and a bandwidth \( h \), we construct a valid confidence band that is implementable in practice. Define the Studentized \( t \)-statistic process

\[
\hat{T}_n(w) = \frac{\hat{f}_W(w) - f_W(w)}{\sqrt{\hat{\Sigma}_n^+(w, w)}}, \quad w \in \mathcal{W}.
\]

Let \( \hat{Z}_n^T(w) \) be a process which, conditional on the data \( \mathbf{W}_n \), is mean-zero and Gaussian, whose conditional covariance structure is

\[
\mathbb{E}\left[ \hat{Z}_n^T(w) \hat{Z}_n^T(w') \mid \mathbf{W}_n \right] = \frac{\hat{\Sigma}_n^+(w, w')}{\sqrt{\hat{\Sigma}_n^+(w, w) \hat{\Sigma}_n^+(w', w')}}.
\]

For \( \alpha \in (0, 1) \), let \( \hat{q}_{1-\alpha} \) be the conditional quantile satisfying

\[
\mathbb{P}\left( \sup_{w \in \mathcal{W}} \left| \hat{Z}_n^T(w) \right| \leq \hat{q}_{1-\alpha} \mid \mathbf{W}_n \right) = 1 - \alpha,
\]
which is shown to be well-defined in the supplemental appendix. The following theorem establishes the validity of our proposed feasible confidence band for $f_W$, which is adaptive to the unknown degeneracy type.

**Theorem 5.1** (Feasible uniform confidence bands)

*Suppose that Assumptions 2.1, 2.2 and 4.1 hold and $f_W(w) > 0$ on $\mathcal{W}$. Then*

$$
\mathbb{P} \left( f_W(w) \in \left[ \hat{f}_W(w) \pm \hat{q}_{1-\alpha} \sqrt{\hat{\Sigma}_n^+(w,w)} \right] \text{ for all } w \in \mathcal{W} \right) - (1 - \alpha) \ll 1.
$$

Recently, Chiang et al. (2022) derived high-dimensional central limit theorems over rectangles for exchangeable arrays and applied them to construct simultaneous confidence intervals for a sequence of design points. Their inference procedure relies on the multiplier bootstrap, and their conditions for valid inference depend on the number of design points considered. In contrast, Theorem 5.1 constructs a feasible uniform confidence band over the entire domain of inference $\mathcal{W}$ based on our strong approximation results for the whole $t$-statistic process and the covariance estimator $\hat{\Sigma}_n^+$. The required rate condition specified in Assumption 4.1 does not depend on the number of design points. Furthermore, our proposed inference methods are robust to potential unknown degenerate points in the underlying dyadic data generating process.

In practice, suprema over $\mathcal{W}$ can be replaced by maxima over sufficiently many design points in $\mathcal{W}$. The conditional quantile $\hat{q}_{1-\alpha}$ can be estimated by Monte Carlo simulation, resampling from the Gaussian process defined by the law of $\hat{Z}_n^T \mid \mathcal{W}_n$.

The bandwidth restrictions in Theorem 5.1 are the same as those required for the infeasible version given in Theorem 4.2, namely those imposed in Assumption 4.1. This follows from the rates of convergence obtained in Lemma 5.1, coupled with some careful technical work given in the supplemental appendix to handle the potential presence of degenerate points in $\Sigma_n$. 
5.3 Bandwidth selection and robust bias-corrected inference

With an eye towards applications, we propose some simple methods for bandwidth selection. Our procedure begins by selecting (approximately) the bandwidth which minimizes the integrated mean squared error (IMSE) of the point estimator $\hat{f}_W$. We then combine this with robust bias-correction ideas (Calonico et al., 2018, 2022).

Let $\psi(w)$ be a non-negative real-valued function on $W$, and suppose that we use a kernel of order $p < \beta$ of the form $k_h(s, w) = K((s - w)/h)/h$. Note that boundary bias is not an issue from an IMSE perspective. Then, the $\psi$-weighted asymptotic IMSE (AIMSE) is minimized by

$$h_{\text{AIMSE}}^* = \left( p!/(p-1)! \left( \int_W f_W(w)\psi(w)\,dw \right) \left( \int_w K(w)^2\,dw \right)^2 \right)^{1/(2p+1)} \left( \frac{n(n-1)}{2} \right)^{-1/(2p+1)}.$$

This is akin to the AIMSE-optimal bandwidth choice for traditional monadic kernel density estimation with a sample size of $\frac{n(n-1)}{2}$. See, for example, Wand and Jones (1994) and Simonoff (2012) for reviews. The choice $h_{\text{AIMSE}}^*$ is slightly undersmoothed (up to a polynomial $\log(n)$ factor) relative to the uniform minimax-optimal bandwidth choice discussed in Section 3, but it is easier to implement in practice.

The choice $h_{\text{AIMSE}}^*$ is an oracle estimator requiring prior knowledge of $f_W$. In practice, many standard feasible bandwidth selection techniques can be used, such as the following.

(i) Rule-of-thumb (ROT). A popular choice for kernels of order $p = 2$ is Silverman’s rule-of-thumb.

Let $\hat{\sigma}^2$, the sample variance, and $\hat{\text{IQR}}$, the sample interquartile range, be the sample variance and sample interquartile range respectively of the data $W_n$. Then define

$$h_{\text{ROT}} = C(K) \left( \hat{\sigma} \wedge \frac{\hat{\text{IQR}}}{1.349} \right) \left( \frac{n(n-1)}{2} \right)^{-1/5},$$
where
\[
C(K) = \left( \frac{8\sqrt{\pi} \int_{-\infty}^{\infty} K(w)^2 \, dw}{3 \left( \int_{-\infty}^{\infty} w^2 K(w) \, dw \right)^2} \right)^{1/5} = \begin{cases} 
2.576, & \text{triangular kernel } K(w) = (1 - |w|) \vee 0, \\
2.435, & \text{Epanechnikov kernel } K(w) = \frac{3}{4}(1 - w^2) \vee 0.
\end{cases}
\]

(ii) **Second generation direct plug-in methods (DPI).** The unknown function \( f_W \) and its derivatives can be estimated using preliminary consistent nonparametric estimators, which rely on some (approximately optimal) pilot bandwidth. These estimators are then plugged into the formula for \( h^*_{\text{AIMSE}} \), yielding a bandwidth estimator \( \widehat{h}_{\text{DPI}} \). This approach is meant to develop consistent nonparametric bandwidth estimators in the sense that \( \widehat{h}_{\text{DPI}} / h^*_{\text{AIMSE}} \to_p 1 \). This procedure can be iterated to give a multi-stage DPI procedure.

(iii) **Likelihood cross-validation.** The bandwidth could also be selected by maximum likelihood cross-validation, though care must be taken to ensure that the estimator is fitted and evaluated on independent samples. For example, a “leave-one-out” regime might fit the estimator on \( W^{-ij}_n = \{ W_{ij'} : \{i,j\} \cap \{i',j'\} = \emptyset \} \) and evaluate it at \( W_{ij} \). A “batch” version of this can be formulated by choosing some \( \mathcal{I} \subseteq \{1, \ldots, n\} \), fitting the estimator on \( W^{-\mathcal{I}}_n = \{ W_{ij} : i \notin \mathcal{I}, j \notin \mathcal{I} \} \) and evaluating it on \( W^\mathcal{I}_n = \{ W_{ij} : i \in \mathcal{I}, j \in \mathcal{I} \} \).

The AIMSE-optimal bandwidth selector \( h^*_{\text{AIMSE}} \approx n^{-\frac{2}{2p+1}} \) and any of its consistent feasible estimators only satisfy Assumption 4.1 in the case of no degeneracy (\( D_{\text{lo}} > 0 \)). Under partial or total degeneracy, such bandwidths are not valid due to the usual leading smoothing (or misspecification) bias that appears in the distributional approximation. To circumvent this problem and construct simple feasible uniform confidence bands for \( f_W \), we propose the following robust bias-correction approach.

Firstly, estimate the bandwidth \( h^*_{\text{AIMSE}} \approx n^{-\frac{2}{2p+1}} \) using a kernel of order \( p \), which leads to an AIMSE-optimal point estimator \( \widehat{f}_W \) in an \( L^2(\psi) \) sense. Then use this bandwidth and a kernel of order \( p' > p \) to construct the statistic \( \widehat{T}_n \) and the confidence band as detailed in Section 5.2. Importantly, both \( \widehat{f}_W \) and \( \widehat{\Sigma}_n^\dagger \) are recomputed with the new higher-order kernel. The change in centering is
equivalent to a bias correction of the original AIMSE-optimal point estimator, while the change in scale captures the additional variability introduced by the bias correction itself. As shown formally in Calonico et al. (2018, 2022) for the case of kernel-based density estimation with i.i.d. data, this approach leads to higher-order refinements in the distributional approximation whenever additional smoothness is available ($p' \leq \beta$). In the present dyadic setting, this procedure is valid so long as

$$n^{-2/3}(\log n)^{7/3} \ll n^{-2/3+1} \ll (n^2 \log n)^{-2/3+1},$$

which is equivalent to $2 \leq p < p'$. For concreteness, we recommend taking $p = 2$ and $p' = 4$, and using the rule-of-thumb bandwidth choice $\tilde{h}_{\text{ROT}}$ defined above.

In particular, this approach automatically delivers a KMT-optimal strong approximation whenever there are no degeneracies in the underlying dyadic data generating process.

Our feasible robust bias-correction method based on AIMSE-optimal dyadic kernel density estimation for constructing uniform confidence bands for $f_W$ is summarized in Algorithm 1.

**Algorithm 1**: Feasible uniform confidence bands for dyadic kernel density estimation

1. Choose a kernel $k_h$ of order $p \geq 2$ satisfying Assumption 2.2.
2. Select a bandwidth $h \approx h_{\text{AIMSE}}^*$ for $k_h$ as in Section 5.3, perhaps using $h = \tilde{h}_{\text{ROT}}$.
3. Choose another kernel $k_h'$ of order $p' > p$ satisfying Assumption 2.2.
4. For $d \geq 1$, choose a set of $d$ distinct evaluation points $W_d$.
5. For each $w \in W_d$, construct the density estimate $\hat{f}_W(w)$ using $k_h'$ as in Section 1.
6. For $w, w' \in W_d$, construct the covariance estimate $\hat{\Sigma}_n(w, w')$ using $k_h'$ as in Section 5.1.
7. Construct the $d \times d$ positive semi-definite covariance estimate $\hat{\Sigma}_n^+$ as in Section 5.1.
8. For $B \geq 1$, let $(\hat{Z}_{n,r}^T : 1 \leq r \leq B)$ be i.i.d. Gaussian vectors from $\hat{Z}_n^T$ defined in Section 5.2.
9. For $\alpha \in (0, 1)$, set $\hat{q}_{1-\alpha} = \inf\{q \in \mathbb{R} : \#\{r : \max_{w \in W_d} |\hat{Z}_{n,r}^T(w)| \leq q\} \geq B(1-\alpha)\}$.
10. Construct $[\hat{f}_W(w) \pm \hat{q}_{1-\alpha}\hat{\Sigma}_n^+(w, w)^{1/2}]$ for each $w \in W_d$. 

27
6 Empirical studies

In this section we investigate the empirical finite-sample performance of the kernel density estimator with dyadic data. We present both simulations and a real-world study with bilateral trade data.

6.1 Simulations

The family of dyadic distributions defined in Section 2.1, along with its three different parametrizations, is used to generate simulated data sets with different degeneracy types.

We use two different boundary bias-corrected Epanechnikov kernels of orders $p = 2$ and $p = 4$ respectively, on the inference domain $W = [-2, 2]$. We select an optimal bandwidth for $p = 2$ as recommended in Section 5.3, using the rule-of-thumb with $C(K) = 2.435$. The semi-definite program in Section 5.1 is solved with the MOSEK interior point optimizer (ApS, 2021), ensuring covariance estimates are positive semi-definite, and Gaussian vectors are resampled $B = 10000$ times.

In Figure 2 we plot a typical outcome for each of the three degeneracy types (total, partial, none), using the Epanechnikov kernel of order $p = 2$, with sample size $n = 100$ (so $N = 4950$ pairs of nodes) and with $d = 100$ equally-spaced evaluation points. Each plot contains the true density function $f_W$, the dyadic kernel density estimate $\hat{f}_W$ and two different approximate 95% confidence bands for $f_W$. The first is the uniform confidence band (UCB) constructed using one of our main results, Theorem 5.1. The second is a sequence of pointwise confidence intervals (PCI) constructed by finding a confidence interval for each evaluation point separately. We show only 10 pointwise confidence intervals for clarity. In general, the PCIs are too narrow as they fail to provide simultaneous (uniform) coverage over the evaluation points. Note that under partial degeneracy the confidence band narrows near the degenerate point $w = 0$. 
Notes. $f_W(w)$: true density. $\hat{f}_W(w)$: estimated density. UCB: uniform confidence band. PCI: pointwise confidence intervals. The nominal coverage rate is 95%.

Next, Table 1 presents numerical results. For each degeneracy type (total, partial, none) and each kernel order ($p = 2, p = 4$), we run 2000 repeats with sample size $n = 500$ (so $N = 124750$ pairs of nodes) and with $d = 50$ equally-spaced evaluation points. We record the average rule-of-thumb bandwidth $\hat{h}_{\text{ROT}}$ and the average root integrated mean squared error (RIMSE). For both the uniform confidence bands (UCB) and the pointwise confidence intervals (PCI), we report the coverage rate (CR) and the average width (AW). The lower-order kernel ($p = 2$) ignores the bias (IB), leading to good RIMSE performance and acceptable UCB coverage under partial or no degeneracy, but gives invalid inference under total degeneracy. In contrast, the higher-order kernel ($p = 4$) provides robust bias correction (RBC) and hence improves the coverage of the UCB in every regime, particularly under total degeneracy, at the cost of increasing both the RIMSE and the average widths of the confidence bands. As expected, the pointwise (in $w \in W$) confidence intervals (PCIs) severely undercover in every regime. Thus our simulation results show that the proposed feasible inference methods based on robust bias correction and proper Studentization deliver valid uniform inference which is robust to unknown degenerate points in the underlying dyadic distribution.
Table 1: Numerical results for three values of the parameter $\pi$.

<table>
<thead>
<tr>
<th>$\pi$</th>
<th>Degeneracy type</th>
<th>$\hat{h}_{\text{ROT}}$</th>
<th>Method</th>
<th>RIMSE</th>
<th>UCB CR</th>
<th>UCB AW</th>
<th>PCI CR</th>
<th>PCI AW</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(\frac{1}{2}, 0, \frac{1}{2})$</td>
<td>Total</td>
<td>0.329</td>
<td>IB</td>
<td>0.0020</td>
<td>85.4%</td>
<td>0.011</td>
<td>19.9%</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RBC</td>
<td>0.0029</td>
<td>93.5%</td>
<td>0.017</td>
<td>19.1%</td>
<td>0.011</td>
</tr>
<tr>
<td>$(\frac{1}{4}, 0, \frac{3}{4})$</td>
<td>Partial</td>
<td>0.324</td>
<td>IB</td>
<td>0.0058</td>
<td>93.7%</td>
<td>0.029</td>
<td>70.6%</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RBC</td>
<td>0.0062</td>
<td>94.1%</td>
<td>0.034</td>
<td>60.4%</td>
<td>0.023</td>
</tr>
<tr>
<td>$(\frac{1}{5}, \frac{1}{5}, \frac{3}{5})$</td>
<td>None</td>
<td>0.296</td>
<td>IB</td>
<td>0.0051</td>
<td>93.0%</td>
<td>0.027</td>
<td>66.5%</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RBC</td>
<td>0.0055</td>
<td>94.3%</td>
<td>0.033</td>
<td>52.3%</td>
<td>0.022</td>
</tr>
</tbody>
</table>

Notes. IB: ignoring bias using a second-order kernel ($p = 2$). RBC: robust bias correction using a fourth-order kernel ($p = 4$). $\hat{h}_{\text{ROT}}$ corresponds to the rule-of-thumb bandwidth for a second-order kernel ($p = 2$).

6.2 Trade data

We also illustrate the performance of our estimation and inference methods with a real-world data set. We use international bilateral trade data from the International Monetary Fund’s Direction of Trade Statistics (DOTS), which was previously analyzed by Head and Mayer (2014) and Chiang et al. (2022). This data set contains information about the yearly trade flows among $n = 208$ economies ($N = 21528$ pairs), and we focus on the years 1995, 2000 and 2005. We define the trade volume between countries $i$ and $j$ to be the logarithm of the sum of the trade flow (in billions of US dollars) from $i$ to $j$ and the trade flow from $j$ to $i$. In each year several pairs of countries did not trade directly, yielding trade flows of zero and hence a trade volume of $-\infty$. We therefore assume that the distribution of trade volumes is a mixture of a point mass at $-\infty$ and a Lebesgue density on $\mathbb{R}$. We report the empirical percentage of non-zero trade flows $\kappa$ for each year. The local nature of our estimator means that samples taking a value of $-\infty$ can simply be removed from the data set.

To estimate the trade volume density function we use Algorithm 1 with $d = 100$ equally-spaced evaluation points in $[-10, 10]$, using the rule-of-thumb bandwidth selector $\hat{h}_{\text{ROT}}$ from Section 5.3 with $p = 2$ and $C(K) = 2.435$. For estimation and inference we use an Epanechnikov kernel of order $p = 4$, resampling the Gaussian process $B = 10000$ times. For each year, Figure 3 plots the density estimate $\hat{f}_W(w)$ along with a uniform confidence band (UCB) and 10 pointwise confidence intervals (PCI) at the nominal coverage rate of 95%. The empirical results are in line with prior studies.
7 Conclusion

We studied the uniform estimation and inference properties of the dyadic kernel density estimator \( w \mapsto \hat{f}_W(w) \) given in (1), which forms a class of U-process-like estimators indexed by the \( n \)-varying kernel functions \( k_h \) on \( W \). We established uniform minimax-optimal point estimation results and uniform distributional approximations for this estimator based on novel strong approximation strategies. We then applied these results to develop valid and feasible uniform confidence bands for the dyadic density estimand \( f_W \), selecting an IMSE-optimal bandwidth and employing methods for robust bias correction. Empirical studies confirmed our theoretical results. From a technical perspective, the appendices contain several generic results concerning strong approximation methods and maximal inequalities for empirical processes that may be of independent interest.

While our focus in this paper was on kernel density estimation with dyadic data, our results are readily applicable to other statistics that can be approximated by the U-process-like \( \hat{f}_W \) and ratios thereof. Examples include nonparametric regression estimation and two-step semiparametric estimation based on dyadic data. In research underway, we are using the novel strong approximation methods developed in this paper to construct valid uniform inference procedures in those settings.
Acknowledgments

We thank Harold Chiang, Laurent Davezies, Xavier D’Haultfoeuille, Yannick Guyonvarch, Jianqing Fan, Kengo Kato, Jason Klusowski and Ricardo Masini for useful comments. The first author was supported through National Science Foundation grant SES-1947805.

Supplemental material

A supplemental appendix containing technical and methodological details as well as full proofs is available at https://arxiv.org/abs/2201.05967. Replication files for the empirical studies are provided at https://github.com/wgunderwood/DyadicKDE.jl.

A Technical results

We present some technical results which may be of broader interest beyond their specific uses in this paper; consequently this appendix is purposely self-contained. Omitted proofs are given in the online supplemental appendix.

A.1 KMT approximation

The following lemma presents a KMT approximation (Komlós et al., 1975) for a class of empirical processes, building on earlier work by Giné et al. (2004) and Giné and Nickl (2010).

Lemma A.1 (KMT approximation)

Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables and $g_n : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be functions satisfying the total variation bound $\sup_{x, y \in \mathbb{R}} \|g_n(\cdot, x)\|_{TV} < \infty$. Then on some probability space there exist independent copies of $X_1, \ldots, X_n$ denoted $X'_1, \ldots, X'_n$ and a mean-zero Gaussian process $Z_n(x)$ such that for universal positive constants $C_1, C_2$ and $C_3$ and for all $t > 0$,

$$\mathbb{P} \left( \sup_{x \in \mathbb{R}} |G_n(x) - Z_n(x)| > \sup_{x \in \mathbb{R}} \|g_n(\cdot, x)\|_{TV} \frac{t + C_1 \log n}{\sqrt{n}} \right) \leq C_2 e^{-C_3 t},$$

where

$$G_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( g_n(X'_i, x) - \mathbb{E}[g_n(X'_i, x)] \right).$$

Further, $Z_n$ has the same covariance structure as $G_n$ in the sense that for all $x, x' \in \mathbb{R}$,

$$\mathbb{E}[Z_n(x)Z_n(x')] = \mathbb{E}[G_n(x)G_n(x')].$$

By independently sampling from the law of $Z_n$ conditional on $X_1, \ldots, X_n$, we can take $Z_n$ to be a function only of $X_1, \ldots, X_n$ and some independent random noise.

We use this lemma to obtain an unconditional strong approximation for $L_n(w)$ defined in (2).
A.2 Yurinskii approximation

The next lemma presents a Yurinskii approximation (Yurinskii, 1978) for a class of empirical processes, building on earlier work by Pollard (2002) and Belloni et al. (2019).

**Lemma A.2** (Yurinskii approximation)

Let \( X_1, \ldots, X_n \) be independent but not necessarily identically distributed (i.i.d.) random variables taking values in a measurable space \((S, S)\) and let \( \mathcal{X}_n \subseteq \mathbb{R} \) be a compact interval. Let \( g_n \) be a measurable function on \( S \times \mathcal{X}_n \) satisfying \( \sup_{\xi \in S} \sup_{x \in \mathcal{X}_n} |g_n(\xi, x)| \leq M_n \) and the \( L^2 \) bound \( \sup_{x \in \mathcal{X}_n} \max_{1 \leq i \leq n} \text{Var}[g_n(X_i, x)] \leq \sigma_n^2 \). Suppose that \( g_n \) satisfies the uniform Lipschitz condition

\[
\sup_{\xi, x, x' \in \mathcal{X}_n} |g_n(\xi, x) - g_n(\xi, x')| \leq \ell_n,
\]

and also the \( L^2 \) Lipschitz condition

\[
\sup_{x, x' \in \mathcal{X}_n} \mathbb{E} \left[ \left| \frac{1}{n} \sum_{i=1}^{n} \left( g_n(X_i, x) - g_n(X_i, x') \right) \right|^{2} \right]^{1/2} \leq l_{n,2}.
\]

Then there exists a probability space carrying independent copies of \( X_1, \ldots, X_n \) denoted \( X_1', \ldots, X_n' \) and a mean-zero Gaussian process \( Z_n(x) \) such that for all \( t > 0 \),

\[
\mathbb{P} \left( \sup_{x \in \mathcal{X}_n} |G_n(x) - Z_n(x)| > t \right) 
\leq C_1 \sigma_n \sqrt{\text{Leb}(\mathcal{X}_n)} \sqrt{\log n} \sqrt{M_n + \sigma_n \sqrt{\log n}} \left[ \frac{l_{n,2}}{\sqrt{n}} \sqrt{\log \frac{n l_{n,\infty}}{l_{n,2}}} + \frac{l_{n,\infty}}{\sqrt{n}} \log \frac{n l_{n,\infty}}{l_{n,2}} \right],
\]

where \( C_1 > 0 \) is a universal constant and

\[
G_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( g_n(X_i', x) - \mathbb{E}[g_n(X_i', x)] \right).
\]

Further, \( Z_n \) has the same covariance structure as \( G_n \) in the sense that for all \( x, x' \in \mathcal{X}_n \),

\[
\mathbb{E}[Z_n(x)Z_n(x')] = \mathbb{E}[G_n(x)G_n(x')].
\]

We use this lemma to construct a conditional (on \( A_n \)) strong approximation for \( E_n(w) \) in (2).

A.3 Vorob’ev–Berkes–Philipp theorem

Next, we present a generalization of the Vorob’ev–Berkes–Philipp theorem (Dudley, 1999), which allows one to “glue” multiple random variables or stochastic processes onto the same probability space, while preserving some pairwise distributions. This result allows us to obtain a joint strong approximation for \( L_n(w) \) and \( E_n(w) \). We begin by giving some definitions.

**Definition A.1** (Tree)

A tree is an undirected graph with finitely many vertices which is connected and contains no cycles or self-loops.

**Definition A.2** (Polish Borel probability space)

A Polish Borel probability space is a triple \((\mathcal{X}, \mathcal{F}, \mathbb{P})\), where \( \mathcal{X} \) is a Polish space (a topological space...
metrizable by a complete separable metric), $\mathcal{F}$ is the Borel $\sigma$-algebra induced on $\mathcal{X}$ by its topology, and $\mathbb{P}$ is a probability measure on $(\mathcal{X}, \mathcal{F})$.

Important examples of Polish spaces include $\mathbb{R}^d$ and the Skorokhod space $\mathcal{D}[0,1]^d$ for $d \geq 1$. In particular, one can consider vectors of real-valued random variables or stochastic processes indexed by compact subsets of $\mathbb{R}^d$ which have almost surely continuous trajectories.

**Definition A.3** (Projection of a law)

Let $(\mathcal{X}_1, \mathcal{F}_1)$ and $(\mathcal{X}_2, \mathcal{F}_2)$ be measurable spaces, and let $\mathbb{P}_{12}$ be a law on $(\mathcal{X}_1 \times \mathcal{X}_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$. The projection of $\mathbb{P}_{12}$ onto $\mathcal{X}_1$ is the law $\mathbb{P}_1$ defined on $(\mathcal{X}_1, \mathcal{F}_1)$ by $\mathbb{P}_1 = \mathbb{P}_{12} \circ \pi_1^{-1}$ where $\pi_1(x_1, x_2) = x_1$ is the first coordinate projection.

**Lemma A.3** (Vorob’ev–Berkes–Philipp theorem, tree form)

Let $\mathcal{T}$ be a tree with vertex set $\mathcal{V} = \{1, \ldots, n\}$ and edge set $\mathcal{E}$. Suppose that attached to each vertex $i$ is a Polish Borel probability space $(\mathcal{X}_i, \mathcal{F}_i, \mathbb{P}_i)$. Suppose that attached to each edge $(i, j) \in \mathcal{E}$ (where $i < j$ without loss of generality) is a law $\mathbb{P}_{ij}$ on $(\mathcal{X}_i \times \mathcal{X}_j, \mathcal{F}_i \otimes \mathcal{F}_j)$. Assume that these laws are pairwise-consistent in the sense that the projection of $\mathbb{P}_{ij}$ onto $\mathcal{X}_i$ (resp. $\mathcal{X}_j$) is $\mathbb{P}_i$ (resp. $\mathbb{P}_j$) for each $(i, j) \in \mathcal{E}$. Then there exists a law $\mathbb{P}$ on

$$\left( \prod_{i=1}^{n} \mathcal{X}_i, \bigotimes_{i=1}^{n} \mathcal{F}_i \right)$$

such that the projection of $\mathbb{P}$ onto $\mathcal{X}_i \times \mathcal{X}_j$ is $\mathbb{P}_{ij}$ for each $(i, j) \in \mathcal{E}$, and therefore also the projection of $\mathbb{P}$ onto $\mathcal{X}_i$ is $\mathbb{P}_i$ for each $i \in \mathcal{V}$.

The requirement that $\mathcal{T}$ must contain no cycles is necessary in general. To see this, consider the Polish Borel probability spaces given by $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3 = \{0, 1\}$, their respective Borel $\sigma$-algebras, and the pairwise-consistent probability measures:

$$1/2 = \mathbb{P}_1(0) = \mathbb{P}_2(0) = \mathbb{P}_3(0),$$

$$1/2 = \mathbb{P}_{12}(0, 1) = \mathbb{P}_{12}(1, 0) = \mathbb{P}_{13}(0, 1) = \mathbb{P}_{13}(1, 0) = \mathbb{P}_{23}(0, 1) = \mathbb{P}_{23}(1, 0).$$

That is, each measure $\mathbb{P}_i$ places equal mass on 0 and 1, while $\mathbb{P}_{ij}$ asserts that each pair of realizations is a.s. not equal. The graph of these laws forms a triangle, which is not a tree. Suppose that $(X_1, X_2, X_3)$ has distribution given by $\mathbb{P}$, where $X_i \sim \mathbb{P}_i$ and $(X_i, X_j) \sim \mathbb{P}_{ij}$ for each $i, j$. But then by definition of $\mathbb{P}_{ij}$ we have $X_1 = 1 - X_2 = X_3 = 1 - X_1$ a.s., which is a contradiction.

### A.4 Maximal inequalities for i.n.i.d. empirical processes

Finally we provide a maximal inequality for empirical processes of independent but not necessarily identically distributed (i.n.i.d.) random variables indexed by a class of functions (possibly of VC-type). This result is an extension of Theorem 5.2 from Chernozhukov et al. (2014b), which only covers i.i.d. random variables, and is proven in the same manner. Such a result is useful in the study of dyadic data because when conditioning on latent variables, we may encounter random variables that are conditionally independent but do not necessarily follow the same conditional distribution.

**Lemma A.4** (Maximal inequality for i.n.i.d. empirical processes)

Let $X_1, \ldots, X_n$ be independent but not necessarily identically distributed (i.n.i.d.) random variables taking values in a measurable space $(S, \mathcal{S})$. Denote the joint distribution of $X_1, \ldots, X_n$ by $\mathbb{P}$ and the marginal distribution of $X_i$ by $\mathbb{P}_i$, and let $\bar{\mathbb{P}} = n^{-1} \sum_i \mathbb{P}_i$. Let $\mathcal{F}$ be a class of Borel measurable functions from $S$ to $\mathbb{R}$ which is pointwise measurable (i.e. it contains a countable subclass which is

34
dense under pointwise convergence). Let $F$ be a strictly positive measurable envelope function for $F$ (i.e. $|g(s)| \leq |g(s)|$ for all $g \in F$ and $s \in S$). For a distribution $Q$ and $q \geq 1$, define the $(Q,q)$-norm of $g \in F$ as $\|g\|_{Q,q}^q = \mathbb{E}_{X \sim Q}[|g(X)|^q]$ and suppose that $\|F\|_{\mathbb{P},2} < \infty$. For $g \in F$ define the empirical process

$$G_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( g(X_i) - \mathbb{E}[g(X_i)] \right).$$

Let $\sigma > 0$ satisfy $\sup_{g \in F} \|g\|_{\mathbb{P},2} \leq \sigma \leq \|F\|_{\mathbb{P},2}$, and define $M = \max_{1 \leq i \leq n} F(X_i)$. Then with $\delta = \sigma/\|F\|_{\mathbb{P},2} \in (0,1]$,

$$\mathbb{E} \left[ \sup_{g \in F} |G_n(g)| \right] \lesssim \|F\|_{\mathbb{P},2} J(\delta, F, F) + \frac{\|M\|_{\mathbb{P},2} J(\delta, F, F)^2}{\delta^2 \sqrt{n}},$$

where $\lesssim$ is up to a universal constant, and $J(\delta, F, F)$ is the covering entropy integral

$$J(\delta, F, F) = \int_{0}^{\delta} \sqrt{1 + \sup_{Q} \log N(F, \rho_Q, \varepsilon \|F\|_{Q,2})} \, d\varepsilon,$$

with the supremum taken over finite discrete probability measures $Q$ on $(S,S)$.

**Lemma A.5** (VC-type maximal inequality for i.n.i.d. empirical processes)
Assume the same setup as in Lemma A.4, and suppose further that $F$ forms a VC-type class, i.e.,

$$\sup_{Q} \log N(F, \rho_Q, \varepsilon \|F\|_{Q,2}) \leq (C_1/\varepsilon)^{C_2}$$

for all $\varepsilon \in (0,1]$, for constants $C_1 \geq e$ (where $e$ is the standard exponential constant) and $C_2 \geq 1$. Then for $\delta \in (0,1]$, we have the covering entropy integral bound

$$J(\delta, F, F) \leq 3\delta \sqrt{C_2 \log(C_1/\delta)},$$

and thus by Lemma A.4,

$$\mathbb{E} \left[ \sup_{g \in F} |G_n(g)| \right] \lesssim \sigma \sqrt{C_2 \log(C_1/\sigma \|F\|_{\mathbb{P},2})} + \frac{\|M\|_{\mathbb{P},2} C_2 \log(C_1/\sigma \|F\|_{\mathbb{P},2})}{\sqrt{n}}.$$

**References**


