# Estimation of non-linear functionals of densities with confidence

Kumar Sricharan, *Student Member, IEEE,* Raviv Raich *Member, IEEE,* and Alfred O. Hero III, *Fellow, IEEE*

**Abstract**

This paper introduces a class of k-nearest neighbor ($k$-NN) estimators called bipartite plug-in (BPI) estimators for estimating integrals of non-linear functions of a probability density, such as Shannon entropy and Rényi entropy. The density is assumed to be smooth, have bounded support, and be uniformly bounded from below on this set. Unlike previous $k$-NN estimators of non-linear density functionals, the proposed estimator uses data-splitting and boundary correction to achieve lower mean square error. Specifically, we assume that $T$ i.i.d. samples $\mathbf{X}_i \in \mathbb{R}^d$ from the density are split into two pieces of cardinality $M$ and $N$ respectively, with $M$ samples used for computing a $k$-nearest-neighbor density estimate and the remaining $N$ samples used for empirical estimation of the integral of the density functional. By studying the statistical properties of $k$-NN balls, explicit rates for the bias and variance of the BPI estimator are derived in terms of the sample size, the dimension of the samples and the underlying probability distribution. Based on these results, it is possible to specify optimal choice of tuning parameters $M/T$, $k$ for maximizing the rate of decrease of the mean square error (MSE). The resultant optimized BPI estimator converges faster and achieves lower mean squared error than previous $k$-NN entropy estimators. In addition, a central limit theorem is established for the BPI estimator that allows us to specify tight asymptotic confidence intervals.

**Index Terms**

Entropy estimation, bipartite k-NN graphs, adaptive estimators, data-splitting estimators, convergence rates, bias and variance tradeoff, concentration bounds.

## I. Introduction

Non-linear functionals of a multivariate density $f$ of the form $\int g(f(x), x)f(x)dx$ arise in applications including machine learning, signal processing, mathematical statistics, and statistical communication theory. Important examples of such functionals include Shannon and Rényi entropy. Entropy based applications for image matching, image registration and texture classification are developed in [1, 2]. Entropy functional estimation is fundamental to

K. Sricharan, A. O. Hero III are with the Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI, 48109-2122. R. Raich is with the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331-5501.

Manuscript received February 1, 2011; revised February 25, 2012.

independent component analysis in signal processing [3]. Entropy has also been used in Internet anomaly detection [4] and data and image compression applications [5]. Several entropy based nonparametric statistical tests have been developed for testing statistical models including uniformity and normality [6, 7]. Parameter estimation methods based on entropy have been developed in [8, 9]. For further applications, see, for example, Leonenko *etal* [10].

In these applications, the functional of interest must be estimated empirically from sample realizations of the underlying densities. Several estimators of entropy measures have been proposed for general multivariate densities $f$. These include consistent estimators based on entropic graphs [11, 12], gap estimators [13], nearest neighbor distances [14, 10, 15, 16], kernel density plug-in estimators [17, 18, 19, 20, 21, 22], Edgeworth approximations [23], convex risk minimization [24] and orthogonal projections [25].

The class of density-plug-in estimators considered in this paper are based on $k$-nearest neighbor ($k$-NN) distances and, more specifically, bipartite k-nearest neighbor graphs over the random sample. The basic construction of the proposed bipartite plug-in (BPI) estimator is as follows (see Sec. II.A for a precise definition). Given a total of $T$ data samples we split the data into two parts of size $N$ and size $M$, $N + M = T$. On the part of size $M$ a $k$-NN density estimate is constructed. The density functional is then estimated by plugging the $k$-NN density estimate into the functional and approximating the integral by an empirical average over the remaining $N$ samples. This can be thought of as computing the estimator over a bipartite graph with the $M$ density estimation nodes connected to the $N$ integral approximating nodes. The BPI estimator exploits a close relation between density estimation and the geometry of proximity neighborhoods in the data sample. The BPI estimator is designed to automatically incorporate boundary correction, *without* requiring prior knowledge of the support of the density. Boundary correction compensates for bias due to distorted $k$-NN neighborhoods that occur for points near the boundary of the density support set. Furthermore, this boundary correction is *adaptive* in that we achieve the same MSE rate of convergence that can be attained using an oracle BPI estimator having knowledge of boundary of the support. Since the rate of convergence relates the number of samples $T = N + M$ to the performance of the estimator, convergence rates have great practical utility. A statistical analysis of the bias and variance, including rates of convergence, is presented for this class of boundary compensated BPI estimators. In addition, results on weak convergence (CLT) of BPI estimators are established. These results are applied to optimally select estimator tuning parameters $M/T, k$ and to derive confidence intervals. For arbitrary smooth functions $g$, we show that by choosing $k$ increasing in $T$ with order $O(T^{-2/(2+d)})$, an optimal MSE rate of order $O(T^{-4/(2+d)})$ is attained by the BPI estimator. For certain specific functions $g$ including Shannon entropy ($g(u) = \log(u)$) and Rényi entropy ($g(u) = u^{\alpha-1}$), a faster MSE rate of order $O(((\log T)^6/T)^{4/d})$ is achieved by BPI estimators by correcting for bias.

### A. Previous work on $k$-NN functional estimation

The authors of [26, 14, 10, 15] propose $k$-NN estimators for Shannon entropy ($g(u) = \log(u)$) and Rényi entropy($g(u) = u^{\alpha-1}$). Evans *etal* [27] consider positive moments of the $k$-NN distances ($g(u) = u^k, k \in \mathbb{N}$). Recently, Baryshnikov *etal* [28] proposed $k$-NN estimators for estimating $f$-divergence $\int \phi(f_0(x)/f(x))f(x)dx$

between an unknown density $f$, from which sample realizations are available, and a known density $f_0$. Because $f_0$ is known, the $f$-divergence $\int \phi(f_0(x)/f(x))f(x)dx$ is equivalent to a entropy functional $\int g(f(x),x)dx$ for a suitable choice of $g$. Wang *etal* [16] developed a $k$-NN based estimator of $\int g(f_1(x)/f_2(x),x)f_2(x)dx$ when both $f_1$ and $f_2$ are unknown. The authors of these works [26, 14, 27, 16] sestablish that the estimators they propose are asymptotically unbiased and consistent. The authors of [15] analyze estimator bias for $k$-NN estimation of Shannon and Rényi entropy. For smooth functions $g(.)$, Evans *etal* [29] show that the variance of the sums of these functionals of $k$-NN distances is bounded by the rate $O(k^5/T)$. Baryshnikov *etal* [28] improved on the results of Evans *etal* by determining the exact variance up to the leading term ($c_k/T$ for some constant $c_k$ which is a function of $k$). Furthermore, Baryshnikov *etal* show that the entropy estimator they propose converges weakly to a normal distribution. However, Baryshnikov *etal* do not analyze the bias of the estimators, nor do they show that the estimators they propose are consistent. Using the results obtained in this paper, we provide an expression for this bias in Section III-E and show that the optimal MSE for Baryshnikov's estimators is $O(T^{-2/(1+d)})$.

In contrast, the main contribution of this paper is the analysis of a general class of BPI estimators of smooth density functionals. We provide asymptotic bias and variance expressions and a central limit theorem. The bipartite nature of the BPI estimator enables us to correct for bias due to truncation of $k$-NN neighborhoods near the boundary of the support set; a correction that does not appear straightforward for previous $k$-NN based entropy estimators. We show that the BPI estimator is MSE consistent and that the MSE is guaranteed to converge to zero as $T \to \infty$ and $k \to \infty$ with a rate that is minimized for a specific choice of $k$, $M$ and $N$ as a function of $T$. Therefore, the thus optimized BPI estimator can be implemented without any tuning parameters. In addition a CLT is established that can be used to construct confidence intervals to empirically assess the quality of the BPI estimator. Finally, our method of proof is very general and it is likely that it can be extended to kernel density plug-in estimators, $f$-divergence estimation and mutual information estimation.

Another important distinction between the BPI estimator and the $k$-NN estimators of Shannon and Rényi entropy proposed by the authors of [26, 14, 10] is that these latter estimators are consistent for finite $k$, while the proposed BPI estimator requires the condition that $k \to \infty$ for MSE convergence. By allowing $k \to \infty$, the BPI estimators of Shannon and Rényi entropy achieve MSE rate of order $O(((\log T)^6/T)^{4/d})$. This asymptotic rate is faster than the $O(T^{-2/d})$ MSE convergence rate [15] of the previous $k$-NN estimators [26, 14, 10] that use a fixed value of $k$. It is shown by simulation that BPI's asymptotic performance advantages, predicted by our theory, also hold for small sample regimes.

### *B. Organization*

The remainder of the paper is organized as follows. Section II formulates the entropy estimation problem and introduces the BPI estimator. The main results concerning the bias, variance and asymptotic distribution of these estimators are stated in Section III and the consequences of these results are discussed. The proofs are given in the Appendix. We discuss bias correction of the BPI estimator for the case of Shannon and Rényi entropy estimation

in Section IV. We numerically validate our theory by simulation in Section V. A conclusion is given in Section VI.

*Notation:* Bold face type will indicate random variables and random vectors and regular type face will be used for non-random quantities. Denote the expectation operator by the symbol $\mathbb{E}$ and conditional expectation given $\mathbf{Z}$ by $\mathbb{E}_{\mathbf{Z}}$. Also define the variance operator as $\mathbb{V}[\mathbf{X}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2]$ and the covariance operator as $Cov[\mathbf{X}, \mathbf{Y}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])]$. Denote the bias of an estimator by $\mathbb{B}$.

## II. Preliminaries

We are interested in estimating non-linear functionals $G(f)$ of $d$-dimensional multivariate densities $f$ with support $\mathcal{S}$, where $G(f)$ has the form

$$G(f) = \int g(f(x), x) f(x) d\mu(x) = \mathbb{E}[g(f(x), x)],$$

for some smooth function $g(f(x), x)$. Let $\mathcal{B}$ denote the boundary of $\mathcal{S}$. Here, $\mu$ denotes the Lebesgue measure and $\mathbb{E}$ denotes statistical expectation w.r.t density $f$. We assume that i.i.d realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_N, \mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$ are available from the density $f$. Neither $f$ nor its support set are known.

The plug-in estimator is constructed using a data splitting approach as follows. The data is randomly subdivided into two parts $\mathcal{X}_N = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ and $\mathcal{X}_M = \{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$ of $N$ and $M$ points respectively. In the first stage, a boundary compensated $k$-NN density estimator $\tilde{\mathbf{f}}_k$ is estimated at the $N$ points $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ using the $M$ realizations $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$. Subsequently, the $N$ samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ are used to approximate the functional $G(f)$ to obtain the basic Bipartite Plug-In (BPI) estimator:

$$\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) \quad = \quad \frac{1}{N} \sum_{i=1}^{N} g(\tilde{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i). \tag{II.1}$$

As the above estimator performs an average over the $N$ variables $X_i$ of the function $g(\tilde{f}(X_i), X_i)$, which is estimated from the other $M$ variables, this estimator can be viewed as averaging over the edges of a bipartite graph with $N$ and $M$ nodes on its left and right parts.

### A. Boundary compensated $k$-NN density estimator

Since the probability density $f$ is bounded above, the observations will lie strictly on the interior of the support set $\mathcal{S}$. However, some observations that occur close to the boundary of $\mathcal{S}$ will have $k$-NN balls that intersect the boundary. This leads to significant bias in the $k$-NN density estimator. In this section we describe a method that compensates for this bias. The method can be interpreted as extrapolating the location of the boundary from extreme points in the sample and suitably reducing the volumes of their $k$-NN balls.

Let $d(X, Y)$ denote the Euclidean distance between points $X$ and $Y$ and $\mathbf{d}_k(X)$ denote the Euclidean distance between a point X and its $k$-th nearest neighbor amongst the $M$ realizations $\mathbf{X}_{N+1}, .., \mathbf{X}_{N+M}$. Define a ball with radius $r$ centered at $X$ contained in the support $\mathcal{S}$: $S_r(X) = \{Y \in \mathcal{S} : d(X, Y) \leq r\}$. The $k$-NN region is

$\mathbf{S}_k(X) = \{Y : d(X,Y) \leq \mathbf{d}_k(X)\}$ and the volume of the $k$-NN region is $\mathbf{V}_k(X) = \int_{\mathbf{S}_k(X)} dZ$. The standard $k$-NN density estimator [30] is defined as

$$\hat{\mathbf{f}}_k(X) = \frac{k-1}{M\mathbf{V}_k(X)}.$$

If a probability density function has bounded support, the $k$-NN balls $\mathbf{S}_k(X)$ centered at points $X$ close to the boundary may intersect with the boundary $\mathcal{B}$, or equivalently $\mathbf{S}_k(X) \cap \mathcal{S}^c \neq \phi$, where $\mathcal{S}^c$ is the complement of $\mathcal{S}$. As a consequence, the $k$-NN ball volume $\mathbf{V}_k(X)$ will tend to be higher for points $X$ close to the boundary leading to significant bias of the $k$-NN density estimator.

Let $R_k(X)$ correspond to the coverage value $(1+p_k)k/M$, i. e. , $R_k(X) = \inf\{r : \int_{S_r(X)} f(Z)dZ = (1+p_k)k/M\}$, where $p_k = \sqrt{6}/(k^{\delta/2})$ for some fixed $\delta \in (2/3, 1)$. Define

$$\epsilon_{BC} = N\exp(-3k^{(1-\delta)}).$$

Define $N_k(X)$ as the region corresponding to the coverage value $(1+p_k)k/M$, i.e. $N_k(X) = \{Y : d(X,Y) \leq R_k(X)\}$. Finally, define the interior region $\mathcal{S}_I$

$$\mathcal{S}_I = \{X \in \mathcal{S} : N_k(X) \cap \mathcal{S}^c = \phi\}. \tag{II.2}$$

We show in Appendix B that the bias of the standard $k$-NN density estimate is of order $O((k/M)^{(2/d)})$ for points $X \in \mathcal{S}_I$ and is of order $O(1)$ at points $X \in \mathcal{S} - \mathcal{S}_I$. This motivates the following method for compensating for this bias. This compensation is done in two stages: (i) the set of interior points $\mathcal{I}_N \subset \mathcal{X}_N$ are identified using variation in $k$-nearest neighbor distances in Algorithm 1 (see Appendix B for details) and it is show that $\mathcal{I}_N \notin \mathcal{S} - \mathcal{S}_I$ with probability $1 - O(\epsilon_{BC})$; and (ii) the density estimator at points in $\mathcal{B}_N = \mathcal{X}_N - \mathcal{I}_N$ are corrected by extrapolating to the density estimates at interior points $\mathcal{I}_N$ that are close to the boundary points. We emphasize that this nonparametric correction strategy does not assume knowledge about the support of the density $f$.

For each boundary point $\mathbf{X}_i \in \mathcal{B}_N$, let $\mathbf{X}_{n(i)} \in \mathcal{I}_\mathcal{N}$ be the interior sample point that is closest to $\mathbf{X}_i$. The corrected density estimator $\tilde{\mathbf{f}}_k$ is defined as follows.

$$\tilde{\mathbf{f}}_k(\mathbf{X}_i) = \begin{cases} \hat{\mathbf{f}}_k(\mathbf{X}_i) & \{\mathbf{X}_i \in \mathcal{I}_\mathcal{N}\} \\ \hat{\mathbf{f}}_k(\mathbf{X}_{n(i)}) & \{\mathbf{X}_i \in \mathcal{B}_N\} \end{cases} \tag{II.3}$$

## III. Main results

Let $\mathbf{Z}$ denote an independent realization drawn from $f$. Also, define $\mathbf{Z}_{-1} \in \mathcal{S}_I$ to be $\mathbf{Z}_{-1} = \arg\min_{x \in \mathcal{S}_I} d(x, \mathbf{Z})$. Define $h(X) = \Gamma^{(2/d)}((d+2)/2)f^{-2/d}(X)tr[\nabla^2(f(X))]$. Denote the $n$-th partial derivative of $g(x,y)$ wrt $x$ by $g^{(n)}(x,y)$. Also, let $g'(x,y) := g^{(1)}(x,y)$ and $g''(x,y) := g^{(2)}(x,y)$. For some fixed $0 < \epsilon < 1$, define $p_l = ((k-1)/M)(1-\epsilon)\epsilon_0$ and $p_u = ((k-1)/M)(1+\epsilon)\epsilon_\infty$. Also define $\epsilon_1 = 1/(c_d\mathcal{D}^d)$, where $\mathcal{D}$ is the diameter of the bounded set $\mathcal{S}$ and define $q_l = ((k-1)/M)\epsilon_1$ and $q_u = (1+\epsilon)\epsilon_\infty$. Let $\mathbf{p}$ be a beta random variable with parameters $k, M-k+1$.

### A. Assumptions

$(\mathcal{A}.0)$ : Assume that $M$, $N$ and $T$ are linearly related through the proportionality constant $\alpha_{frac}$ with: $0 < \alpha_{frac} < 1$, $M = \alpha_{frac} T$ and $N = (1-\alpha_{frac})T$. $(\mathcal{A}.1)$ : Let the density $f$ be uniformly bounded away from 0 and finite on the set $\mathcal{S}$, i.e., there exist constants $\epsilon_0$, $\epsilon_\infty$ such that $0 < \epsilon_0 \leq f(x) \leq \epsilon_\infty < \infty \; \forall x \in \mathcal{S}$. $(\mathcal{A}.2)$: Assume that the density $f$ has continuous partial derivatives of order $2\nu$ in the interior of the set $\mathcal{S}$ where $\nu$ satisfies the condition $(k/M)^{2\nu/d} = o(1/M)$, and that these derivatives are upper bounded. $(\mathcal{A}.3)$: Assume that the function $g(x,y)$ has $\lambda$ partial derivatives w.r.t. $x$, where $\lambda$ satisfies the conditions $k^{-\lambda} = o(1/M)$ and $O((\lambda^2((k/M)^{2/d} + 1/M))/M) = o(1/M)$. $(\mathcal{A}.4)$: Assume that $\max\{6, 2\lambda\} < k <= M$. $(\mathcal{A}.5)$: Assume that the absolute value of the functional $g(x,y)$ and its partial derivatives are strictly bounded away from $\infty$ in the range $\epsilon_0 < x < \epsilon_\infty$ for all $y$. $(\mathcal{A}.6)$: Assume that $\sup_{x \in (q_l, q_u)} |(g^{(r)}/r!)^2(x,y)| e^{-3k^{(1-\delta)}} < \infty$, $\mathbb{E}[\sup_{x \in (p_l, p_u)} |(g^{(r)}/r!)^2(x/\mathbf{p}, y)|] < \infty$, for $r = 3, \lambda$.

### B. Bias and Variance

Below the asymptotic bias and variance of the BPI estimator of general functionals of the density $f$ are specified. These asymptotic forms will be used to establish a form for the asymptotic MSE.

**Theorem III.1.** *The bias of the BPI estimator $\hat{\mathbf{G}}_k(f)$ is given by*

$$\mathbb{B}[\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)] \;\; = \;\; c_1 \left(\frac{k}{M}\right)^{2/d} + c_2 \left(\frac{1}{k}\right) + c_3(k,M,N) + O(\epsilon_{BC}) + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{2/d}\right),$$

*where $c_3(k,M,N) = \mathbb{E}[1_{\{\mathbf{Z} \in \mathcal{S} - \mathcal{S}_I\}}(g(f(\mathbf{Z}_{-1}), \mathbf{Z}_{-1}) - g(f(\mathbf{Z}), \mathbf{Z}))] = O(k/M)^{2/d}$, and the constants $c_1 = \mathbb{E}[g'(f(\mathbf{Z}), \mathbf{Z}) h(\mathbf{Z})]$, $c_2 = \mathbb{E}[f^2(\mathbf{Z}) g''(f(\mathbf{Z}), \mathbf{Z})/2]$.*

The leading terms $c_1(k/M)^{2/d} + c_2/k$ arise due to the bias and variance of $k$-NN density estimates respectively (see Appendix A), while the term $c_3(k,M,N)$ arises due to boundary correction (see Appendix B). Henceforth, we will refer to $c_3(k,M,N)$ by $c_3$. It is shown in Appendix B that $c_3 = O((k/M)^{2/d})$ (B.11). The term $O(\epsilon_{BC})$ arises from a concentration inequality that gives the probability of the event $\mathcal{I}_N \notin \mathcal{S} - \mathcal{S}_I$ as $1 - O(\epsilon_{BC})$. Observe that if $k$ increases logarithmically in $M$, specifically $(\log(M))^{2/(1-\delta)}/k \to 0$, then $O(\epsilon_{BC}) = o(N/M^3) = o(1/T)$.

**Theorem III.2.** *The variance of the BPI estimator $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ is given by*

$$\mathbb{V}[\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)] \;\; = \;\; c_4 \left(\frac{1}{N}\right) + c_5 \left(\frac{1}{M}\right) + O(\epsilon_{BC}) + o\left(\frac{1}{M} + \frac{1}{N}\right),$$

*where the constants $c_4 = \mathbb{V}[g(f(\mathbf{Z}), \mathbf{Z})]$ and $c_5 = \mathbb{V}[f(\mathbf{Z}) g'(f(\mathbf{Z}), \mathbf{Z})]$.*

The term $c_4/N$ is due to approximation of the integral $\int g(f(x),x) f(x) dx$ by the sample mean $(1/N)\sum_{i=1}^N g(f(\mathbf{X}_i), \mathbf{X}_i)$. The term $c_5/M$ on the other hand is due to the covariance between density estimates $\tilde{\mathbf{f}}(\mathbf{X}_i)$ and $\tilde{\mathbf{f}}(\mathbf{X}_j)$, $i \neq j$.

### C. Optimized parameter tuning

Theorem III.1 implies that $k \to \infty$ and $k/M \to 0$ in order that the BPI estimator $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ be asymptotically unbiased. Likewise, Theorem III.2 implies that $N \to \infty$ and $M \to \infty$ in order that the variance of the estimator

converge to $0$. It is clear from Theorem III.1 that the MSE is minimized when $k$ grows in polynomially in $M$. Throughout this section, we assume that $k = k_0 M^r$ for some $r \in (0,1)$. This implies that $O(\epsilon_{BC}) = O(N\mathcal{C}(k)) = o(1/M) = o(1/T)$.

*1) Assumptions:* Under the condition $k = k_0 M^r$, the assumptions $(\mathcal{A}.2)$ and $(\mathcal{A}.3)$ reduce to the following equivalent conditions: $(\mathcal{A}.2)$: Let the density $f$ have continuous partial derivatives of order $2r$ in the interior of the set $\mathcal{S}$ where $r$ satisfies the condition $2r(1-t)/d > 1$. $(\mathcal{A}.3)$: Let the functional $g(x,y)$ have $\lambda$ partial derivatives w.r.t. $x$, where $\lambda$ satisfies the conditions $t\lambda > 1$.

*2) Optimal choice of $k$:* Theorems III.1 and III.2 provide an optimal choice of $k$ that minimizes asymptotic MSE. Minimizing the MSE over $k$ is equivalent to minimizing the square of the bias over $k$. Define $c_o = c_1 + c_3/(k/M)^{2/d}$. The optimal choice of $k$ is given by

$$k_{opt} \quad = \quad \arg\min_k \mathbb{B}(\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)) = \lfloor k_0 M^{\frac{2}{2+d}} \rfloor, \tag{III.1}$$

where $\lfloor x \rfloor$ is the closest integer to $x$, and the constant $k_0$ is defined as $k_0 = (|c_2|d/2|c_0|)^{\frac{d}{d+2}}$ when $c_0 c_2 > 0$ and as $k_0 = (|c_2|/|c_0|)^{\frac{d}{d+2}}$ when $c_0 c_2 < 0$.

Observe that the constants $c_0$ and $c_2$ can possibly have opposite signs. When $c_0 c_2 > 0$, the bias evaluated at $k_{opt}$ is $b_0^+ M^{\frac{-2}{2+d}}(1+o(1))$ where $b_0^+ = c_0 k_0^{2/d} + c_2/k_0$. Let $k_{frac} = k_0 M^{\frac{2}{2+d}} - k_{opt}$. When $c_0 c_2 < 0$, observe that $c_0((k_{frac}+k_{opt})/M)^{2/d} + c_2/(k_{frac}+k_{opt})$ is equal to zero. When $c_0 c_2 < 0$, a higher order asymptotic analysis is required to specify the bias at the optimal value of $k$ (see Page 10, [31]). The bias evaluated at $k_{opt}$ in this case is given by $b_0^- M^{\frac{-4}{2+d}}(1+o(1))$ where $b_0^-$ is a constant which depends on the underlying density $f$.

Even though the optimal choice $k_{opt}$ depends on the unknown density $f$ (via the constant $k_0$), we observe from simulations that simply matching the rates, i.e. choosing $k = \bar{k} = M^{2/(2+d)}$, leads to significant MSE improvement. This is illustrated in Section V.

*3) Choice of $\alpha_{frac} = M/T$:* Observe that the MSE of $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ is dominated by the squared bias $(O(M^{-4/(2+d)}))$ as contrasted to the variance $(O(1/N + 1/M))$. This implies that the MSE rate of convergence is invariant to the choice of $\alpha_{frac}$. This is corroborated by the experimental results shown in Fig. 6.

*4) Discussion on optimal choice of $k$:* The optimal choice of $k$ grows at a smaller rate as compared to the total number of samples $M$ used for the density estimation step. Furthermore, the rate at which $k/M$ grows decreases as the dimension $d$ increases. This can be explained by observing that the choice of $k$ primarily controls the bias of the entropy estimator. For a fixed choice of $k$ and $M$ $(k < M)$, one expects the bias in the density estimates (and correspondingly in the estimates of the functional $G(f)$) to increase as the dimension increases. For increasing dimension an increasing number of the $M$ points will be near the boundary of the support set. This in turn requires choosing a smaller $k$ relative to $M$ as the dimension $d$ grows.

*5) Optimal rate of convergence:* Observe that the optimal bias decays as $b_0^+(T^{\frac{-2}{2+d}})(1+o(1))$ when $c_0 c_2 > 0$ and $b_o^-(T^{\frac{-4}{2+d}})(1+o(1))$ when $c_0 c_2 < 0$. The variance decays as $\Theta(1/T)(1+o(1))$.

### D. Central limit theorem

In addition to the results on bias and variance shown in the previous section, it is shown here that the BPI estimator, appropriately normalized, weakly converges to the normal distribution. The asymptotic behavior of the BPI estimator is studied under the following limiting conditions: (a) $k/M \to 0$, (b) $k \to \infty$ and (c) $N \to \infty$. As shorthand, the above limiting assumptions will be collectively denoted by $\Delta \to 0$.

**Theorem III.3.** *The asymptotic distribution of the BPI estimator* $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ *is given by*

$$\lim_{\Delta \to 0} Pr\left( \frac{\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) - \mathbb{E}[\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)]}{\sqrt{\mathbb{V}[\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)]}} \leq \alpha \right) = Pr(\mathbf{S} \leq \alpha),$$

*where* $\mathbf{S}$ *is a standard normal random variable.*

### E. Comparison with results by Baryshnikov etal

Recently, Baryshnikov *etal* [28] have developed asymptotic convergence results for estimators of $f$-divergence $G(f_0, f) = \int f(x)\phi(f_0(x)/f(x))dx$ for the case where $f_0$ is known. Their estimators are based on sums of functionals of $k$-NN distances. They assume that they have $T$ i.i.d realizations from the unknown density $f$, and that $f$ and $f_0$ are bounded away from 0 and $\infty$ on their support. The general form of the estimator of Baryshnikov *etal* is given by

$$\tilde{\mathbf{G}}_N(\hat{\mathbf{f}}_{kS}) = \frac{1}{T}\sum_{i=1}^{T} g(\hat{\mathbf{f}}_{kS}(\mathbf{X}_i)),$$

where $\hat{\mathbf{f}}_{kS}(\mathbf{X}_i)$ is the standard $k$-NN density estimator [32] estimated using the $T-1$ samples $\{\mathbf{X}_1, .., \mathbf{X}_T\} - \{\mathbf{X}_i\}$.

Baryshnikov *etal* do not show that their estimator is consistent and do not analyze the bias of their estimator. They show that the leading term in the variance is given by $c_k/T$ for some constant $c_k$ which is a function of the number of nearest neighbors $k$. Finally they show that their estimator, when suitably normalized, is asymptotically normal. In contrast, we assume higher order conditions on continuity of the density $f$ and the functional $g$ (see Section 3) as compared to Baryshnikov *etal* and provide results on bias, variance and asymptotic distribution of data-split $k$-NN functional estimators of entropies of the form $G(f) = \int g(f(x))f(x)dx$. Note that we also require the assumption that $f$ is bounded away from 0 and $\infty$ on its support. Because we are able to establish expressions on both the bias and variance of the BPI estimator, we are able to specify optimal choice of free parameters $k, N, M$ for minimum MSE.

For estimating the functional $G(f) = \int g(f(x))f(x)dx$, the estimator of Baryshnikov can be used by restricting $f_0$ to be uniform. In Appendix C it is shown that under the additional assumption that $(\mathcal{A}.6)$ is satisfied by $\tilde{g} = g$, the bias of $\tilde{\mathbf{G}}_N(\hat{\mathbf{f}}_{kS})$ is

$$\mathbb{B}(\tilde{\mathbf{G}}_N(\hat{\mathbf{f}}_{kS})) = O((k/T)^{1/d}) + O(1/k). \tag{III.2}$$

In contrast, Theorem III. 1 establishes that the bias of the BPI estimator $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ decays as $\Theta((k/M)^{2/d} + 1/k) + O(\epsilon_{BC})$ and the variance decays as $\Theta(1/T)$. The bias of the BPI estimator has a higher exponent ($2/d$ as opposed to $1/d$) and this is a direct consequence of using the boundary compensated density estimator $\tilde{\mathbf{f}}_k$ in place of $\hat{\mathbf{f}}_k$.

It is clear from III.2 that the estimator of Baryshnikov will be unbiased iff $k \to \infty$ as $T \to \infty$. Furthermore, the optimal rate of growth of $k$ is given by $k = T^{1/(1+d)}$. Furthermore, $c_k = \Theta(1)$ and therefore the overall optimal bias and variance of $\tilde{\mathbf{G}}_N(\hat{\mathbf{f}}_{kS})$ is given by $\Theta(T^{-1/(1+d)})$ and $\Theta(T^{-1})$ respectively. On the other hand, the optimal bias of the BPI estimator decays as $b_0^+(T^{\frac{-2}{2+d}})(1+o(1))$ when $c_1 c_2 > 0$ and $b_o^-(T^{\frac{-4}{2+d}})(1+o(1))$ when $c_1 c_2 < 0$ and the optimal variance decays as $\Theta(1/T)$. The BPI estimator therefore has faster rate of MSE convergence. Experimental MSE comparison of Baryshnikov's estimator against the proposed BPI estimator is shown in Fig. 6.

## IV. Bias correction factors

When the density functional of interest is the Shannon entropy ($g(u) = -\log(u)$) or the Rényi -$\alpha$ entropy($g(u) = u^{\alpha-1}$), a bias correction can be added to the BPI estimator that accelerates rate of convergence. Goria et.al. [10] and Leonenko et.al. [14] developed consistent Shannon and Rényi estimators with bias correction. The authors of [15] analyzed the bias for these estimators. When combined with the results of Baryshnikov *etal*, one can easily deduce the variance of these estimators and establish a CLT.

Let $\hat{\mathbf{H}}_S$ be the Shannon entropy estimate $\tilde{\mathbf{G}}_N(\hat{\mathbf{f}}_{kS})$ with the choice of functional $g(x) = -\log(x)$. Let $\hat{\mathbf{I}}_{\alpha,S}$ be the estimate of the Rényi $\alpha$-integral estimate $\tilde{\mathbf{G}}_N(\hat{\mathbf{f}}_{kS})$ with the choice of functional $g(x) = x^{\alpha-1}$. Define $\tilde{\mathbf{H}}_S = \hat{\mathbf{H}}_S + [\log(k-1) - \Psi(k)]$, where $\psi(.)$ is the digamma function, and $\tilde{\mathbf{I}}_{\alpha,S} = [(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}]^{-1}\hat{\mathbf{I}}_{\alpha,S}$. Also define the Rényi entropy estimator to be $\tilde{\mathbf{H}}_{\alpha,S} = (1-\alpha)^{-1}\log(\tilde{\mathbf{I}}_{\alpha,S})$. The estimators $\tilde{\mathbf{H}}_S$ and $\tilde{\mathbf{H}}_{\alpha,S}$ are the Shannon and Rényi entropy estimators of Goria *etal* [14] and Leonenko *etal* [10] respectively. In [15], it is shown that the bias of $\tilde{\mathbf{H}}_S$ and $\tilde{\mathbf{I}}_{\alpha,S}$ is given by $\Theta((k/T)^{1/d})$, while the variance was shown by Baryshnikov *etal* to be $O(1/T)$. In contrast, by (III.2), the bias of $\hat{\mathbf{H}}_S$ and $\hat{\mathbf{I}}_{\alpha,S}$ is given by $\Theta((k/T)^{1/d} + (1/k))$ (III.2). This can be understood as follows. From the results by [15], we have

$$\mathbb{E}[\hat{\mathbf{H}}_S] = I - [\log(k-1) - \Psi(k)] + c_{0,0}(k/T)^{1/d} + o((k/T)^{1/d}) \tag{IV.1}$$

and

$$\mathbb{E}[\hat{\mathbf{I}}_{\alpha,S}] = [(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}]I_\alpha + c_{0,\alpha}(k/T)^{1/d} + o((k/T)^{1/d}) \tag{IV.2}$$

for some functionals of the density $c_{0,0}$ and $c_{0,\alpha}$. Note that $[(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}] = 1 + O(1/k)$ and $\Psi(k) = \log(k-1) + O(1/k)$ as $k \to \infty$. From the above equations, the scale factor $[(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}]$ and the additive factor $[\log(k-1) - \Psi(k)]$ account for the $O(1/k)$ terms in the expressions for bias of $\hat{\mathbf{H}}_S$ and $\hat{\mathbf{I}}_{\alpha,S}$, thereby removing the requirement that $k \to \infty$ for asymptotic unbiasedness. These bias corrections can be incorporated into the BPI estimator as follows.

### A. Main results

For a general function $g(x,y)$, if there exist functions $g_1(k,M)$ and $g_2(k,M)$, such that

$$\begin{aligned}
&(i) \quad \mathbb{E}[g((k-1)x/M\mathbf{p},y)] = g(x,y)g_1(k,M) + g_2(k,M) + o(1/M),\\
&(ii) \quad ((k-1)/M)\mathbb{E}[g'((k-1)x/M\mathbf{p},y)\mathbf{p}^{2/d-1}] = g'(x,y)(k/M)^{2/d} + o((k/M)^{2/d}),\\
&(iii) \quad \lim_{k\to\infty} g_1(k,M) = 1,\\
&(iv) \quad \lim_{k\to\infty} g_2(k,M) = 0,
\end{aligned} \tag{IV.3}$$

then define the BPI estimator with bias correction as

$$\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k) \quad = \quad \frac{\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) - g_2(k,M)}{g_1(k,M)}. \tag{IV.4}$$

*1) Bias and Variance:* In addition to the assumptions listed in section III-A, assume that $k = O((\log(M))^{2/(1-\delta)})$. Below the asymptotic bias and variance of the BPI estimator with bias correction are specified.

**Theorem IV.1.** *The bias of the BPI estimator $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$ is given by*

$$\mathbb{B}[\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)] \quad = \quad c_1\left(\frac{k}{M}\right)^{2/d} + c_3(k,M,N) + o\left(\left(\frac{k}{M}\right)^{2/d}\right). \tag{IV.5}$$

**Theorem IV.2.** *The variance of the BPI estimator $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$ is given by*

$$\mathbb{V}[\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)] \quad = \quad c_4\left(\frac{1}{N}\right) + c_5\left(\frac{1}{M}\right) + o\left(\frac{1}{M}+\frac{1}{N}\right).$$

*2) CLT:*

**Theorem IV.3.** *The asymptotic distribution of the BPI estimator $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$ is given by*

$$\lim_{\Delta\to 0} Pr\left(\frac{\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k) - \mathbb{E}[\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)]}{\sqrt{\mathbb{V}[\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)]}} \leq \alpha\right) = Pr(\mathbf{S} \leq \alpha),$$

*where $\mathbf{S}$ is a standard normal random variable.*

*3) MSE:* Theorem IV. 1 specifies the bias of the BPI estimator, $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$, as $\Theta((k/M)^{2/d})$. Theorem IV. 2 specifies the variance as $\Theta(1/N+1/M)$. By making $k$ increase logarithmically in $M$, specifically, $k = O((\log(M))^{2/(1-\delta)})$ for any value $\delta \in (2/3, 1)$, the MSE is given by the rate $\Theta(((\log(T))^{2/(1-\delta)}/T)^{4/d})$. The BPI estimator therefore has a faster rate of convergence in comparison to both Baryshnikov *etal*'s estimators $\hat{\mathbf{H}}_S$ and $\hat{\mathbf{I}}_{\alpha,S}$ (MSE $= \Theta(T^{-2/(1+d)})$) and Leonenko *etal*'s and Goria *etal*'s estimators $\tilde{\mathbf{H}}_S$ and $\tilde{\mathbf{I}}_{\alpha,S}$ (MSE $= \Theta(T^{-2/d})$). Experimental MSE comparison of Leonenko's estimator against the BPI estimator in Section V shows the MSE of the BPI estimator to be significantly lower. Finally, note that such bias correction cannot be applied for general entropy functionals, and the bias correction factors cannot in general be incorporated. In the next section, the application of BPI estimators for estimation of Shannon and Rényi entropies is illustrated.

### *B. Shannon and Rényi entropy estimation*

For the case of Shannon entropy ($g(u) = -\log(u)$), it can be verified that $g_1(k, M) = 1$, $g_2(k, M) = \psi(k) - \log(k-1)$ satisfy (IV.3). Similarly, for the case of Rényi entropy ($g(u) = u^{\alpha-1}$), $g_1(k, M) = (\Gamma(k)/\Gamma(k+1-\alpha))(1/(k-1)^{\alpha-1})$, $g_2(k, M) = 0$ satisfy (IV.3).

For Shannon entropy ($g(u) = -\log(u)$) and Rényi entropy ($g(u) = u^{\alpha-1}$), the assumptions in Section III-A reduce to the following under the condition $k = O((\log(M))^{2/(1-\delta)})$. Assumption $(\mathcal{A}.1)$ is unchanged. Assumption $(\mathcal{A}.2)$ holds for any $r$ such that $2r > d$. The assumption $(\mathcal{A}.3)$ is satisfied by the choice of $\lambda = \log(M)$. Assumption $(\mathcal{A}.4)$ holds for ($g(u) = -\log(u)$) and ($g(u) = u^{\alpha-1}$). Next, it will be shown that $(\mathcal{A}.5)$ is also satisfied by ($g(u) = -\log(u)$) and ($g(u) = u^{\alpha-1}$).

We note that $\tilde{g} = (g^{(3)}/6)^2$ for the choice of $g(u) = -\log(u)$ is given by $\tilde{g} = cu^{-6}$ for some constant $c$. Therefore,

$$\begin{aligned}\sup_{x\in(q_l,q_u)} |\tilde{g}(x,y)|e^{-3k^{(1-\delta)}} &= |c\epsilon_1^{-6}|(M/k)^6 O(e^{-3k^{(1-\delta)}}) \\ &= |c\epsilon_1^{-6}|(M/k)^6 O(e^{-3(\log(M))^2}) \\ &= |c\epsilon_1^{-6}|O(e^{-3(\log(M))^2+6\log(M)-6\log(k)}) = o(1),\end{aligned}$$

and by (A.7), $\mathbb{E}[\sup_{x\in(p_l,p_u)} |\tilde{g}(x/\mathbf{p}, y)|] = |c|((1-\epsilon)\epsilon_0)^{-6}\mathbb{E}[(M\mathbf{p}/(k-1))^6] = |c|((1-\epsilon)\epsilon_0)^{-6}O(1) = O(1)$. Similarly, $\tilde{g} = (g^{(\lambda)}/(\lambda!))^2$ for the choice of $g(u) = -\log(u)$ is given by $\tilde{g} = \lambda^{-2}u^{-2\lambda}$. Then,

$$\begin{aligned}\sup_{x\in(q_l,q_u)} |\tilde{g}(x,y)|e^{-3k^{(1-\delta)}} &= O((M/k)^{2\lambda} e^{-3k^{(1-\delta)}}) \\ &= O((M/k)^{2\lambda} e^{-3(\log(M))^2}) \\ &= O(e^{-3(\log(M))^2+2(\log(M))^2-2\log(M)\log(k)}) = o(1),\end{aligned}$$

and by (A.7), $\mathbb{E}[\sup_{x\in(p_l,p_u)} |\tilde{g}(x/\mathbf{p}, y)|] = O(\mathbb{E}[(M\mathbf{p}/(k-1))^{2\lambda})] = O(1)$. In an identical manner, $(\mathcal{A}.5)$ is satisfied when $g(u) = u^{\alpha-1}$.

To summarize, for functions $g(u) = -\log(u)$ and $g(u) = u^{\alpha-1}$, Theorem IV.1, IV.2 and IV.3 hold under the following assumptions: (i) $(\mathcal{A}.0)$, (ii) $(\mathcal{A}.1)$, (iii) the density $f$ has bounded continuous partial derivatives of order greater than $d$ and (iv) $k = O((\log(M))^{2/(1-\delta)})$. Furthermore the proposed BPI estimator $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$ can be used to estimate Shannon entropy ($g(u) = -\log(u)$) and Rényi entropy ($g(u) = u^{\alpha-1}$) at MSE rate of $\Theta(((\log(T))^{2/(1-\delta)}/T)^{4/d})$.

## V. EXPERIMENTS

Here the theory established in Section 3 and Section 4 is validated. A three dimensional vector $\underline{X} = [X_1, X_2, X_3]^T$ was generated on the unit cube according to the i.i.d. Beta plus i.i.d. uniform mixture model:

$$f(x_1, x_2, x_3) = (1-\epsilon)\prod_{i=1}^{3} f_{a,b}(x_i) + \epsilon, \tag{V.1}$$

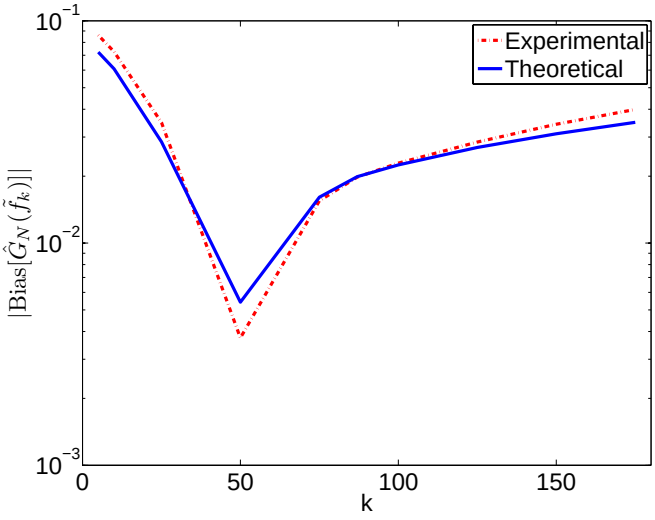


Fig. 1. Comparison of theoretically predicted bias of BPI estimator $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ against experimentally observed bias as a function of $k$. The Shannon entropy ($g(u) = -\log(u)$) is estimated using the BPI estimator $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ on $T = 10^4$ i. i. d. samples drawn from the $d = 3$ dimensional uniform-beta mixture density (V.1). $N, M$ were fixed as $N = 3000$, $M = 7000$ respectively. The theoretically predicted bias agrees well with experimental observations. The predictions of our asymptotic theory therefore extend to the finite sample regime. The theoretically predicted optimal choice of $k_{opt} = 52$ also minimizes the empirical bias.

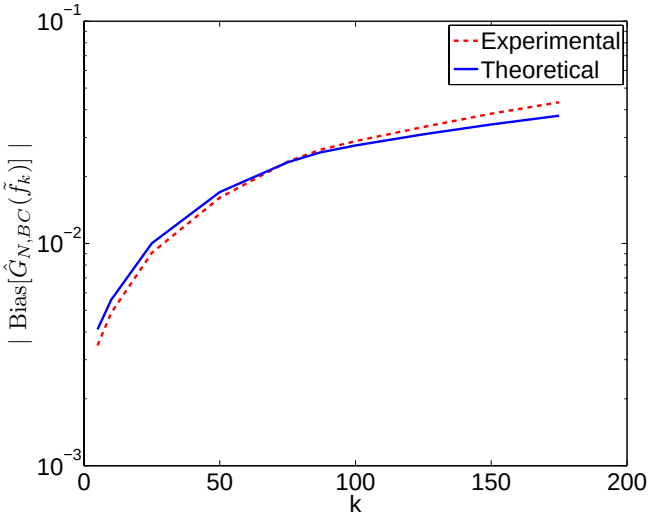


Fig. 2. Comparison of theoretically predicted bias of BPI estimator $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$ against experimentally observed bias as a function of $k$. The Shannon entropy ($g(u) = -\log(u)$) is estimated using the proposed BPI estimator $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$ on $T = 10^4$ i. i. d. samples drawn from the $d = 3$ dimensional uniform-beta mixture density (V.1). $N, M$ were fixed as $N = 3000$, $M = 7000$ respectively. The empirical bias is in agreement with the bias approximations of Theorem IV. 1 and monotonically increases with $k$.

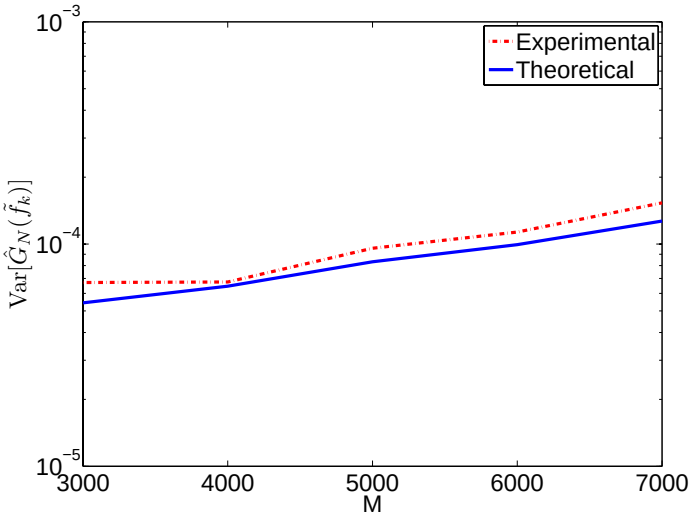


Fig. 3. Comparison of theoretically predicted variance of BPI estimator $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ against experimentally observed variance as a function of $M$. The Shannon entropy ($g(u) = -\log(u)$) is estimated using the proposed BPI estimator $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ on $T = 10^4$ i. i. d. samples drawn from the $d = 3$ dimensional uniform-beta mixture density (V.1). $k$ is chosen to be $k_{opt} = k_0 M^{2/(2+d)}$. The theoretically predicted variance agrees well with experimental observations.

where $f_{a,b}(x)$ is a univariate Beta density with shape parameters $a$ and $b$. For the experiments the parameters were set to $a = 4, b = 4$, and $\epsilon = 0.2$. The Shannon entropy ($g(u) = -\log(u)$) is estimated using the BPI estimators $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ and $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$.

In Fig. 1, the bias approximations of Theorem III. 1 are compared to the empirically determined estimator bias of $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$. $N$ and $M$ are fixed as $N = 3000$, $M = 7000$. Note that the theoretically predicted optimal choice of $k_{opt} = 52$ minimizes the experimentally obtained bias curve. Thus, even though our theory is asymptotic it provides useful predictions for the case of finite sample size, specifying bandwidth parameters that achieve minimum bias. Further note that by matching rates, i.e. choosing $k = \bar{k} = M^{2/(2+d)} = 83$ also results in significantly lower MSE when compared to choosing $k$ arbitrarily ($k < 10$ or $k > 150$). In Fig. 2, the bias approximations of Theorem IV. 1 are compared to the empirically determined estimator bias of $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$. Observe that the empirical bias, in agreement with the bias approximations of Theorem IV. 1, monotonically increases with $k$.

In Fig. 3, the empirically determined variance of $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ is compared with the variance expressed by Theorem III. 2 for varying choices of $N$ and $M$, with fixed $N + M = 10,000$. The theoretically predicted variance agrees well with experimental observations. A Q-Q plot of the normalized BPI estimate $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ and the standard normal distribution is shown in Fig. 4. The linear Q-Q plot validates the Central Limit Theorem III. 3 on the uncompensated BPI estimator. For Shannon entropy ($g(u) = -\log(u)$), the uncompensated and compensated BPI estimators are related by

$$\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k) = \hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) + \log(k-1) - \psi(k).$$

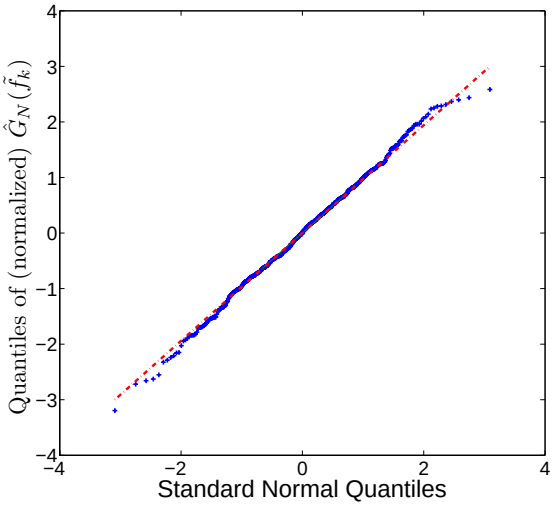


Fig. 4. Q-Q plot comparing the quantiles of the BPI estimator $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ (with $g(u) = -\log(u)$) on the vertical axis to a standard normal population on the horizontal axis. The Shannon entropy ($g(u) = -\log(u)$) is estimated using the proposed BPI estimator$\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ on $T = 10^4$ i. i. d. samples drawn from the $d = 3$ dimensional uniform-beta mixture density (V.1). $k, N, M$ are fixed as $k = k_{opt} = 52$, $N = 3000$ and $M = 7000$ respectively. The approximate linearity of the points validates our central limit theorem III.3.

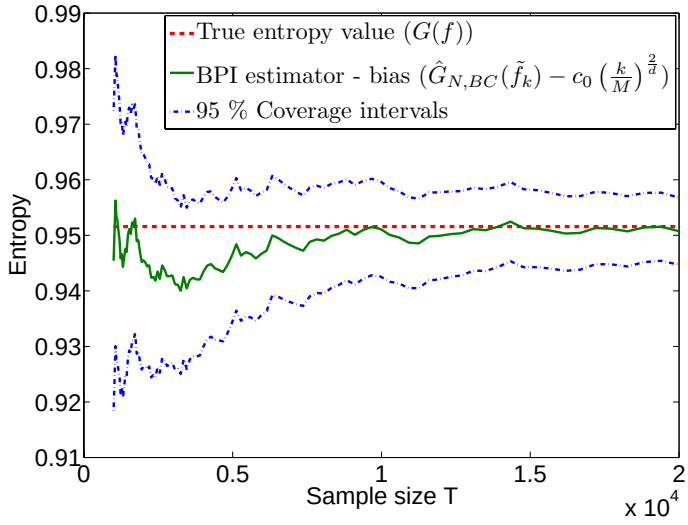


Fig. 5. $95\%$ coverage intervals of BPI estimator $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$, predicted using the Central limit theorem III.3, as a function of sample size $T$. The Shannon entropy ($g(u) = -\log(u)$) is estimated using the proposed BPI estimator $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$ on $T$ i. i. d. samples drawn from the $d = 3$ dimensional uniform-beta mixture density (V.1). The lengths of the coverage intervals are accurate to within $12\%$ of the empirical confidence intervals obtained from the empirical distribution of the BPI estimator.

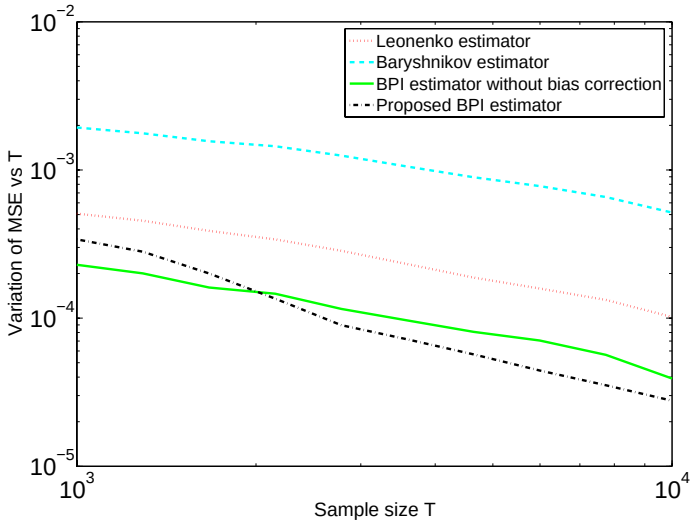


Fig. 6. Variation of MSE of $k$-nearest neighbor estimator of Leonenko *etal* [14] and the $k$-nearest neighbor estimator of Baryshnikov *etal* [28] and BPI estimators with and without boundary correction, as a function of sample size $T$. The Rényi entropy ($g(u) = u^{\alpha-1}$) is estimated for $\alpha = 0.5$ using these estimators on $T$ i. i. d. samples drawn from the $d = 3$ dimensional uniform-beta mixture density (V.1). The figure shows that the proposed BPI estimator has the fastest rate of convergence.

The variance and normalized distribution of these estimators are therefore identical. Consequently, Fig. 3 and Fig. 4 also validate Theorem IV. 2 and Theorem IV. 3 respectively.

Finally, using the CLT, the $95\%$ coverage intervals of the BPI estimator $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$ are shown as a function of sample size $T$ in Fig. 5. The lengths of the predicted confidence intervals are accurate to within $12\%$ of the true confidence intervals (determined by simulation over the range of $80\%$ to $100\%$ coverage - data not shown). These coverage intervals can be interpreted as confidence intervals on the true entropy, provided that the constants $c_1, .., c_5$ can be accurately estimated.

### *A. Experimental comparison of estimators*

The Rényi $\alpha$-entropy ($g(u) = u^{\alpha-1}$) is estimated for $\alpha = 0.5$, with the same underlying 3 dimensional mixture of the beta and uniform densities defined above. Several estimators are compared: Baryshnikov's estimator $\hat{\mathbf{I}}_{\alpha,S}$, the $k$-NN estimator $\tilde{\mathbf{I}}_{\alpha,S}$ of Leonenko *etal* [14], the BPI estimator without bias correction $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ and the proposed BPI estimator with bias correction $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$. The results are shown in Fig. 6. It is clear from the figure that the BPI estimator $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$ has the fastest rate of convergence, consistent with our theory. Note that, in agreement with our analysis in Section III-E, the bias uncompensated BPI estimator $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ outperforms Baryshnikov's estimator $\hat{\mathbf{I}}_{\alpha,S}$.

## VI. CONCLUSIONS

A new class of boundary compensated bipartite k-NN density plug-in estimators was proposed for estimation of smooth non-linear functionals of densities that are strictly bounded strictly away from 0 on their finite support. These estimators, called bipartite plug-in (BPI) estimators, correct for bias due to boundary effects and outperform previous $k$-NN entropy estimators in terms of MSE convergence rate. Expressions for asymptotic bias and variance of the estimator were derived estimator in terms of the sample size, the dimension of the samples and the underlying probability distribution. In addition, a central limit theorem was developed for the proposed BPI estimators. The accuracy of these asymptotic results were validated through simulation and it was established that the theory can be used to specify optimal finite sample estimator tuning parameters such as bandwidth and optimal partitioning of data samples.

Using the theory presented in the paper, one can tune the parameters of the plug-in estimator to achieve minimum asymptotic estimation MSE. Furthermore, the theory can be used to specify the minimum necessary sample size required to obtain requisite accuracy. This in turn can be used to predict and optimize performance in applications like structure discovery in graphical models and dimension estimation for support sets of low intrinsic dimension. The reader can refer to [31] for details on these and other applications.

For the reader's convenience, the notation used in this paper is listed in the table below.

| Notation | Description |
|---|---|
| $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ | BPI estimator (II.1) |
| $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)$ | BPI estimator with bias compensation (IV.4) |
| $g_1(k,M), g_2(k,M)$ | Bias correction factors |
| $\mathcal{S}$ | Support of density $f$ |
| $d$ | dimension of support $\mathcal{S}$ |
| $c_d$ | unit ball volume in $d$ dimensions |
| $\{\mathbf{X}_1,\ldots,\mathbf{X}_T,\mathbf{Y},\mathbf{Z}\}$ | $T+2$ independent realizations drawn from $f$ |
| $\mathcal{X}_N$ | $\{\mathbf{X}_1,\ldots,\mathbf{X}_N\}$ |
| $\mathcal{X}_M$ | $\{\mathbf{X}_{N+1},\ldots,\mathbf{X}_{N+M}\}$ |
| $\mathcal{S}_I$ | Interior of support |
| $\mathcal{I}_N$ | Interior points subset of $\mathcal{X}_N$ |
| $\mathcal{B}_N$ | Boundary points subset of $\mathcal{X}_N$ |
| $\mathbf{Z}_{-1}$ | Closest interior point to $\mathbf{Z}$; $\mathbf{Z}_{-1} = \arg\min_{x\in\mathcal{S}_I} d(x,\mathbf{Z})$ |
| $\mathbf{X}_{n(i)}$ | $\mathbf{X}_{n(i)} \in \mathcal{I}_N$ is the interior sample point that is closest to $\mathbf{X}_i \in \mathcal{B}_N$ |
| $\delta$ | Constant; $\delta \in (2/3, 1)$ |
| $\epsilon_{BC} = N\exp(-3k^{(1-\delta)})$ | Probability of misclassification of $x \in \mathcal{S} - \mathcal{S}_I$ as interior point |
| $\mathbf{d}_k(X)$ | $k$-NN ball radius |
| $\mathbf{S}_k(X)$ | $k$-NN ball |
| $\mathbf{V}_k(X)$ | $k$-NN ball volume |
| $\mathbf{P}(X)$ | Coverage function |
| $\hat{\mathbf{f}}_k(X)$ | $k$-NN density estimate |
| $\tilde{\mathbf{f}}_k(X)$ | Boundary corrected $k$-NN density estimate |
| $g^{(n)}(x,y)$ | $n$-th derivative of $g(x,y)$ wrt $x$ |
| $\mathbf{p}$ | beta random variable with parameters $k, M-k+1$ |
| $\alpha_{frac}$ | Proportionality constant; $M = \alpha_{frac}T$ and $N = (1-\alpha_{frac})T$ |
| $\epsilon_0$, $\epsilon_\infty$ | constants such that $\epsilon_0 \le f(x) \le \epsilon_\infty\ \forall x \in \mathcal{S}$ |
| $2\nu$ | Number of times $f$ is assumed to be differentiable |
| $\lambda$ | Number of times $g(x,y)$ is assumed to be differentiable wrt $x$ |
| $c_1,..,c_5$ | Constants appearing in Theorems III.1, III.2, III.3 and IV.1, IV.2, IV.3 |
| $\mathcal{C}(k)$ | Function which satisfies the rate of decay condition $\mathcal{C}(k) = O(e^{-3k^{(1-\delta)}})$ |
| $k_M$ | $k_M = (k-1)/M$ |
| $\natural(X)$ | The event $\mathbf{P}(X) > (1-p_k)k_M$ |
| $\natural_{-1}(X)$ | The event $\mathbf{P}(X) < (1+p_k)k_M$ |
| $\natural\natural(X)$ | The event $(1-p_k)k_M < \mathbf{P}(X) < (1+p_k)k_M$ |
| $\mathbf{e}_k(X)$ | Error function $\mathbf{e}_k(X) = \hat{\mathbf{f}}_k(X) - \mathbb{E}[\hat{\mathbf{f}}_k(X) \mid X]$ |
| $\mathbf{e}(X)$ | Error function $\mathbf{e}(X) = \tilde{\mathbf{f}}_k(X) - \mathbb{E}[\tilde{\mathbf{f}}_k(X) \mid X]$ |

## APPENDIX A
## $k$-NN DENSITY ESTIMATES

In this appendix, moment properties of the standard $k$-NN density estimate $\hat{\mathbf{f}}_k(X)$ are derived conditioned on $X_1, \ldots, X_N$. As the samples $X_1, \ldots, X_N, X_{N+1}, \ldots, X_T$, $T = M + N$ are i.i.d., these conditional moments are independent of the $N$ samples $\mathbf{X}_1, .., \mathbf{X}_N$.

### *A. Preliminaries*

Let $d(X, Y)$ denote the Euclidean distance between points $X$ and $Y$ and $\mathbf{d}_X^{(k)}$ denote the Euclidean distance between a point X and its $k$-th nearest neighbor amongst $\mathbf{X}_{N+1}, .., \mathbf{X}_{N+M}$. Let $c_d$ denote the unit ball volume in $d$ dimensions. The $k$-NN region is

$$\mathbf{S}_k(X) = \{Y : d(X, Y) \leq \mathbf{d}_X^{(k)}\}$$

and the volume of the $k$-NN region is

$$\mathbf{V}_k(X) = \int_{\mathbf{S}_k(X)} dZ.$$

The standard $k$-NN density estimator [30] is defined as

$$\hat{\mathbf{f}}_k(X) = \frac{k-1}{M\mathbf{V}_k(X)}.$$

Define the coverage function as

$$\mathbf{P}(X) = \int_{\mathbf{S}_k(X)} f(Z)dZ.$$

Define spherical regions

$$S_r(X) = \{Y \in \mathbb{R}^d : d(X, Y) \leq r\}.$$

*1) Concentration inequality for coverage probability:* It has been previously established that $\mathbf{P}(X)$ has a beta distribution with parameters $k$, $M - k + 1$ [33]. Using Chernoff inequalities, we can then establish the following concentration inequality (Section B.1, [31]). For some $0 < p < 1/2$,

$$\begin{aligned} Pr(\mathbf{P}(X) > (1+p)(k-1)/M) &= O(e^{-p^2 k/2(1+p)}) \\ Pr(\mathbf{P}(X) < (1-p)(k-1)/M) &= O(e^{-p^2 k/2(1-p)}). \end{aligned} \tag{A.1}$$

Define

$$k_M = (k-1)/M.$$

Let $\natural(X)$ denote the event

$$\mathbf{P}(X) < (p_k + 1)k_M, \tag{A.2}$$

where $p_k = \sqrt{6}/(k^{\delta/2})$. Then, $1 - Pr(\natural(X)) = O(e^{-p_k^2 k/2}) = O(e^{-3k^{(1-\delta)}})$. Equivalently,

$$1 - Pr(\natural(X)) = O(\mathcal{C}(k)), \tag{A.3}$$

where $\mathcal{C}(k)$ is a function which satisfies the rate of decay condition $\mathcal{C}(k) = O(e^{-3k^{(1-\delta)}})$. Similarly, let $\natural_{-1}(X)$ denote the event

$$\mathbf{P}(X) > (1 - p_k)k_M, \tag{A.4}$$

Then

$$1 - Pr(\natural_{-1}(X)) = O(\mathcal{C}(k)), \tag{A.5}$$

Also let $\natural\natural(X) = \natural(X) \cap \natural_{-1}(X)$. Then

$$1 - Pr(\natural\natural(X)) = O(\mathcal{C}(k)), \tag{A.6}$$

Finally, we note that $\Gamma(x + a)/\Gamma(x) = x^a + o(x^a)$. Then for any $a < k$, $\mathbb{E}[\mathbf{P}^{-a}(X)]$ exists and is given by

$$\mathbb{E}[\mathbf{P}^{-a}(X)] = \frac{\Gamma(k - a)\Gamma(M + 1)}{\Gamma(k)\Gamma(M + 1 - a)} = \Theta((k_M)^{-a}). \tag{A.7}$$

*2) Interior points:* Let $\mathcal{S}'$ to be any arbitrary subset of $\mathcal{S}_I$ (II.2) satisfying the condition $Pr(\mathbf{Y} \notin \mathcal{S}') = o(1)$ where $\mathbf{Y}$ is random variable with density $f$. This implies that given the event $\natural(X)$, the $k$-NN neighborhoods $\mathbf{S}_k(X)$ of points $X \in \mathcal{S}'$ will lie completely inside the domain $\mathcal{S}$. Therefore the density $f$ has continuous partial derivatives of order $2\nu$ in the $k$-NN ball neighborhood $\mathbf{S}_k(X)$ for each $X \in \mathcal{S}'$ (assumption $(\mathcal{A}.2)$). We will now derive moments for the interior set of points $X \in \mathcal{S}'$. This excludes the set of points $X$ close to the boundary of the support whose $k$-NN neighborhoods $\mathbf{S}_k(X)$ intersect with the boundary of the support. We will deal with these points in Appendix B.

*3) Taylor series expansion of coverage probability:* Let $X \in \mathcal{S}'$. Given the event $\natural(X)$, the coverage function $\mathbf{P}(X)$ can be represented in terms of the volume of the $k$-NN ball $\mathbf{V}_k(X)$ by expanding the density $f$ in a Taylor series about $X$ as follows. In particular, for some fixed $x \in \mathcal{S}'$, let

$$p(u) = \int_{S_u(x)} f(z)dz.$$

Using $(\mathcal{A}.2)$, we can write, by a Taylor series expansion of $f$ around $x$ using multi-index notation [34]

$$f(z) = \sum_{0 \leq |\alpha| \leq 2\nu} \frac{(z - x)^\alpha}{\alpha!}(\partial^\alpha f)(x) + o(||z - x||^{2\nu}) \tag{A.8}$$

Assuming $S_u(x) \subset \mathcal{S}$, we can then write

$$\begin{aligned} p(u) &= \int_{S_u(x)} f(z)dz \\ &= \int_{S_u(x)} \left( \sum_{|0 \leq \alpha \leq 2\nu|} \frac{(z - x)^\alpha}{\alpha!}(\partial^\alpha f)(x) \right) dz + o(u^{d+2\nu}) \\ &= f(x)c_d u^d + \sum_{i=1}^{\nu-1} c_i(x) c_d^{1+2i/d} u^{d+2i} + o(u^{d+2\nu}). \end{aligned} \tag{A.9}$$

where $c_i(x)$ are functionals of the derivatives of $f$. Now, denote $v(u) = \int_{S_u(x)} dz$ to be the volume of $S_u(x)$. Let $u^{inv}(v)$ be the inverse function of $v(u)$. Note that this inverse is well-defined since $v(u)$ is monotonic in $u$. Since $S_u(x) \subset \mathcal{S}$, $v(u) = c_d u^d$. This gives $u^{inv}(v) = (v/c_d)^{1/d}$. Define

$$P(v) = \int_{S_{u^{inv}(v)}(x)} f(z)dz.$$

Using (A.9),

$$P(v) \quad = \quad f(X)v + \sum_{i=1}^{\nu-1} c_i(X) v^{1+2i/d} + o(v^{1+2\nu/d}). \tag{A.10}$$

Now denote $V(p) = P^{inv}(p)$ to be the inverse of $P(.)$. Note that this inverse is well-defined since $P(v)$ is monotonic in $v$. Dividing (A.10) by $vP(v)$ on both sides, we get

$$\frac{1}{v} \quad = \quad \frac{f(X)}{P(v)} + \sum_{i=1}^{\nu-1} \frac{c_i(X)}{P(v)} v^{2i/d} + o(v^{2\nu/d} P^{-1}(v)) \tag{A.11}$$

By repeatedly substituting the LHS of (A.11) in the RHS of (A.11), we can obtain (A.12):

$$\frac{1}{V(p)} \quad = \quad \frac{f(X)}{p} + \sum_{i=1}^{\nu-1} \frac{h_i(X)}{p^{1-2i/d}} + o(p^{2\nu/d-1}), \tag{A.12}$$

From our derivation of (A.12) using (A.10), it is clear that $h_i(X)$ are of the form

$$h_i(X) = \sum_{\{a_i\}=A; A \in \mathcal{A}} \frac{\prod_{i=1}^{\nu-1} c_i^{a_i}}{f^{a_0}(X)}$$

where $A$ is a $\nu$-tuple of positive real numbers $a_0, .., a_{\nu-1}$ and the cardinality of $\mathcal{A}$ is finite. By assumptions $(\mathcal{A}.1)$ and $(\mathcal{A}.2)$, this implies that the constants $h_i(X)$ are *bounded*. Also, we note that $h(X) = h_1(X) = c(X) f^{-2/d}(X)$ [33], where $c(X) := c_1(X) = \Gamma^{(2/d)}(\frac{d+2}{2}) tr[\nabla^2(f(X))]$. This then implies that under the event $\natural(X)$

$$\frac{1}{\mathbf{V}_k(X)} = \frac{f(X)}{\mathbf{P}(X)} + \sum_{t \in \mathcal{T}} \frac{h_t(X)}{\mathbf{P}^{1-t}(X)} + \mathbf{h_r}(X), \tag{A.13}$$

where $\mathcal{T} = \{2/d, 4/d, 6/d.., 2\nu/d\}$ and $\mathbf{h_r}(X) = o(\mathbf{P}^{2\nu/d-1}(X))$. Now, by $(\mathcal{A}.2)$, we have $(k/M)^{2\nu/d} = o(1/M)$. This implies that $2\nu/d > 1$. Under the event $\natural(X)$, we have $\mathbf{P}(X) \leq (p_k + 1)k/M$, which, in conjunction with the condition $2\nu/d > 1$ implies that

$$\mathbf{h_r}(X) \quad = \quad o(\mathbf{P}^{2\nu/d-1}(X)) = o((k/M)^{2\nu/d-1}) = o(1/k_M M). \tag{A.14}$$

On the other hand, under the event, $\natural^c(X)$, $(p_k + 1)k/M \leq \mathbf{P}(X) \leq 1$, which gives

$$\mathbf{h_r}(X) \quad = \quad O(1). \tag{A.15}$$

*4) Approximation to the $k$-NN density estimator:* Define the *coverage* density estimate to be,

$$\hat{\mathbf{f}}_c(X) = f(X)\frac{k-1}{M}\frac{1}{\mathbf{P}(X)}.$$

The estimate $\hat{\mathbf{f}}_c(X)$ is clearly not implementable. Note also that the two estimates - $\hat{\mathbf{f}}_c(X)$ and $\hat{\mathbf{f}}_k(X)$ - are identical in the case of the uniform density.

$$\frac{1}{\mathbf{V}_k(X)} = \frac{f(X)}{\mathbf{P}(X)} + \frac{h(X)}{\mathbf{P}^{1-2/d}(X)} + \mathbf{h_s}(X), \tag{A.16}$$

where $\mathbf{h_s}(X) = o(1/\mathbf{P}^{1-2/d}(X))$. This gives,

$$\hat{\mathbf{f}}_k(X) \quad = \quad \hat{\mathbf{f}}_c(X) + \left(\frac{k-1}{M}\right)\frac{h(X)}{\mathbf{P}^{1-2/d}(X)} + \frac{k-1}{M}\mathbf{h_s}(X). \tag{A.17}$$

whenever $\natural(X)$ is true.

*5) Bounds on $k$-NN density estimates:* Let $X$ be a Lebesgue point of $f$, i.e., an $X$ for which

$$\lim_{r\to 0}\frac{\int_{S_r(X)} f(y)dy}{\int_{S_r(x)} dy} = f(X).$$

Because $f$ is an density, we know that almost all $X \in \mathcal{S}$ satisfy the above property. Now, fix $\epsilon \in (0,1)$ and find $\delta > 0$ such that

$$\sup_{0<r\leq\delta}\frac{\int_{S_r(X)} f(y)dy}{\int_{S_r(x)} dy} - f(X) \leq \epsilon f(X).$$

This in turn implies that, for $\mathbf{P}(X) \leq P(\delta)$,

$$\frac{\mathbf{P}(X)}{(1+\epsilon)f(X)} \leq \mathbf{V}_k(X) \leq \frac{\mathbf{P}(X)}{(1-\epsilon)f(X)} \tag{A.18}$$

and in turn implies

$$(1-\epsilon)\hat{\mathbf{f}}_c(X) \leq \quad \hat{\mathbf{f}}_k(X) \quad \leq (1+\epsilon)\hat{\mathbf{f}}_c(X). \tag{A.19}$$

Also, because $\delta > 0$ is fixed, we note that the event $\mathbf{P}(X) \leq P(\delta)$ is a subset of $\natural(X)$ and therefore (A.18) holds under $\natural(X)$.

Under the event $\natural^c(X)$, we can bound $\mathbf{V}_k(X)$ from above by $c_d\mathcal{D}^d$. Also, since $\mathbf{V}_k(X)$ is monotone in $\mathbf{P}(X)$, under the event $\natural^c(X)$, we can bound $\mathbf{V}_k(X)$ from below by $(1+p_k)(k-1)/M(1-\epsilon)f(X)$ and therefore by $(k-1)/M(1-\epsilon)f(X)$. Written explicitly,

$$\frac{(k-1)}{M(1-\epsilon)f(X)} \leq \mathbf{V}_k(X) \leq c_d\mathcal{D}^d \tag{A.20}$$

and in turn implies

$$(k-1)/(Mc_d\mathcal{D}^d) \leq \quad \hat{\mathbf{f}}_k(X) \quad \leq (1-\epsilon)f(X). \tag{A.21}$$

Finally, note that $k_M/\mathbf{P}(X)$ is bounded above by $O(1)$ under the event $\natural(X)$. This implies that for any $a < k$,

$$\mathbb{E}[\natural^c(X)]k_M^a\mathbf{P}^{-a}(X) \leq O(1)Pr(\natural^c(X)) = O(\mathcal{C}(k)). \tag{A.22}$$

### B. Bias of the $k$-NN density estimates

Let $X \in \mathcal{S}'$. We can analyze the bias of $k$-NN density estimates as follows by using (A.17)

$$\begin{aligned}
\mathbb{E}[1_{\natural(X)}\hat{\mathbf{f}}_k(X)] &= \mathbb{E}[1_{\natural(X)}\hat{\mathbf{f}}_c(X)] + \mathbb{E}\left[1_{\natural(X)}\left(\frac{k-1}{M}\right)\frac{h(X)}{\mathbf{P}^{1-2/d}(X)}\right] + \mathbb{E}\left[1_{\natural(X)}\frac{k-1}{M}\mathbf{h}_{\mathbf{s}}(X)\right] \\
&= \mathbb{E}[1_{\natural(X)}\hat{\mathbf{f}}_c(X)] + \mathbb{E}\left[1_{\natural(X)}\left(\frac{k-1}{M}\right)\frac{h(X)}{\mathbf{P}^{1-2/d}(X)}\right] + o\left(\mathbb{E}\left[1_{\natural(X)}\frac{k-1}{M}\mathbf{P}^{2/d-1}(X)\right]\right) \\
&= \mathbb{E}[\hat{\mathbf{f}}_c(X)] + \mathbb{E}\left[\left(\frac{k-1}{M}\right)\frac{h(X)}{\mathbf{P}^{1-2/d}(X)}\right] + o\left(\frac{k}{M}\right)^{2/d} + O(\mathcal{C}(k)) \\
&= f(X) + h(X)\left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{2/d}, \qquad \text{(A.23)}
\end{aligned}$$

where we used the fact that under the event $\natural^c(X)$, $((k-1)/M)\mathbf{P}^{1-t}(X) = O(1)$ for any $t >= 0$, which in turn gives $\mathbb{E}[1_{\natural^c(X)}((k-1)/M)\mathbf{P}^{1-t}(X)] = O(Pr(\natural^c(X))) = O(\mathcal{C}(k))$. This implies that

$$\begin{aligned}
\mathbb{E}[\hat{\mathbf{f}}_k(X)] - f(X) &= \mathbb{E}[1_{\natural(X)}\hat{\mathbf{f}}_k(X)] + \mathbb{E}[1_{\natural^c(X)}\hat{\mathbf{f}}_k(X)] - f(X) \\
&= h(X)\left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{2/d} + O(\mathcal{C}(k)) + \mathbb{E}[1_{\natural^c(X)}\hat{\mathbf{f}}_k(X)] \\
&= h(X)\left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{2/d} + O(\mathcal{C}(k)), \qquad \text{(A.24)}
\end{aligned}$$

where the last step follows because , by (A.21), $1_{\natural^c(X)}\hat{\mathbf{f}}_k(X) = O(1)$. This expression is true for $k >= 3$ by (A.7).

Next, assuming that (IV.3) holds, we evaluate $\mathbb{E}[g(\hat{\mathbf{f}}_k(X), X)]$ in an identical fashion to the derivation of (A.24).

$$\begin{aligned}
&\mathbb{E}[1_{\natural(X)}g(\hat{\mathbf{f}}_k(X), X)] = \mathbb{E}\left[1_{\natural}(X)g\left(\hat{\mathbf{f}}_c(X) + k_M h(X)(\mathbf{P}(X))^{2/d-1} + k_M \mathbf{h}_{\mathbf{s}}(X), X\right)\right] \\
&= \mathbb{E}\left[1_{\natural(X)}g\left(\hat{\mathbf{f}}_c(X) + k_M h(X)(\mathbf{P}(X))^{2/d-1} + k_M o((\mathbf{P}(X))^{2/d-1}), X\right)\right] \\
&= \mathbb{E}\left[g\left(\hat{\mathbf{f}}_c(X) + k_M h(X)(\mathbf{P}(X))^{2/d-1} + k_M o((\mathbf{P}(X))^{2/d-1}), X\right)\right] + O(\mathcal{C}(k)) \\
&= \mathbb{E}\left[g(\hat{\mathbf{f}}_c(X), X) + g'(\hat{\mathbf{f}}_c(X), X)k_M h(X)(\mathbf{P}(X))^{2/d-1} + o(k_M \mathbf{P}(X))^{2/d-1})\right] + O(\mathcal{C}(k)) \\
&= g(f(X), X)g_1(k, M) + g_2(k, M) + g'(f(X), X)h(X)(k/M)^{2/d} + o((k/M)^{2/d}) + O(\mathcal{C}(k)).
\end{aligned}$$

This gives,

$$\begin{aligned}
&\mathbb{E}[g(\hat{\mathbf{f}}_k(X), X)] = \mathbb{E}[1_{\natural(X)}g(\hat{\mathbf{f}}_k(X), X)] + \mathbb{E}[1_{\natural^c(X)}g(\hat{\mathbf{f}}_k(X), X)] \\
&= g(f(X), X)g_1(k, M) + g_2(k, M) + g'(f(X), X)h(X)(k/M)^{2/d} + o((k/M)^{2/d}) + O(\mathcal{C}(k)). \qquad \text{(A.25)}
\end{aligned}$$

### C. Moments of error function

Let $\gamma_1(X)$, $\gamma_2(X)$ be arbitrary continuous functions satisfying the condition: $\sup_X[\gamma_i(X)]$ is finite, $i = 1, 2$. Also let $\gamma(X) = \gamma_1(X)$. Let $\mathbf{X}_1, .., \mathbf{X}_M, \mathbf{X}, \mathbf{Y}$ denote $M + 2$ i.i.d realizations of the density $f$. Let $q$, $r$ be arbitrary positive integers less than $k$. Define the error function

$$\mathbf{e}_k(X) = \hat{\mathbf{f}}_k(X) - \mathbb{E}[\hat{\mathbf{f}}_k(X) \mid X].$$

Then,

**Lemma A.1.**

$$\mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})\big] = O(k^{-q\delta/2}) + o(1/M) + O(\mathcal{C}(k)). \tag{A.26}$$

**Lemma A.2.**

$$\begin{aligned} Cov\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e}_k^q(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e}_k^r(\mathbf{Y})\big] &= O\left(\frac{1}{k^{((q+r)\delta/2-1)}M}\right) + O(k_M^{2/d}/M) \\ &+ O(1/M^2) + O(\mathcal{C}(k)). \end{aligned} \tag{A.27}$$

Define the operator $\mathcal{M}(\mathbf{Z}) = \mathbf{Z} - \mathbb{E}[\mathbf{Z}]$. Let $\beta$ be any positive real number and define

$$\mathbf{E}_\beta(X) = k_M^\beta(\mathcal{M}(\mathbf{P}^{-\beta}(X))). \tag{A.28}$$

Define the terms

$$\mathbf{e}_c(X) = \hat{\mathbf{f}}_c(X) - \mathbb{E}[\hat{\mathbf{f}}_c(X) \mid X], \tag{A.29}$$

$$\mathbf{e}_t(X) = \mathcal{M}\left(\sum_{t\in\mathcal{T}} \frac{k_M h_t(X)}{\mathbf{P}^{1-t}(X))}\right), \tag{A.30}$$

$$\mathbf{e}_r(X) = \mathcal{M}(k_M \mathbf{h_r}(X)). \tag{A.31}$$

Note that

$$\mathbf{e}_c(X) = f(X)\mathbf{E}_1(X) \tag{A.32}$$

and

$$\mathbf{e}_t(X) = (\sum_{t\in\mathcal{T}} k_M^t h_t(X)(\mathbf{E}_{1-t}(X))). \tag{A.33}$$

Define the event $\{X \in \mathcal{S}'\}\cap\{\natural(X)\}$ by $\dagger(X)$. Note that under the event $\dagger(X)$, $\mathbf{e}_k(X) = \mathbf{e}_c(X)+\mathbf{e}_t(X)+\mathbf{e}_r(X) =: \mathbf{e}_o(X)$. Also, under the event $\natural(X)$, $\mathbf{P}(X) \leq (1+p_k)k_M$, which implies that under the event $\natural(X)$, the following hold

$$\mathbf{E}_\beta(X) = O(1), \mathbf{e}_c(X) = O(1), \mathbf{e}_t(X) = O(1), \mathbf{e}_r(X) = O(1), \mathbf{e}_o(X) = O(1). \tag{A.34}$$

Furthermore, by (A.21), under the event $\natural(X)$,

$$\mathbf{e}_k(X) = O(1). \tag{A.35}$$

*Proof:* of Lemma A.1. Since $\mathbf{P}(X)$ is a beta random variable, the probability density function of $\mathbf{P}(X)$ is given by

$$f(p_X) = \frac{M!}{(k-1)!(M-k)!} p_X^{k-1}(1-p_X)^{M-k}.$$

By (A.7), $\mathbb{E}[\mathbf{P}^{-\beta}(X)] = \Theta((k/M)^{-\beta})$ if $\beta < k$. We will first show that $\mathbb{E}[\mathbf{E}_\beta^q(X)] = O(1)$ if $q\beta < k$. This in turn implies that, by (A.32) and (A.33), $\mathbb{E}[\mathbf{e}_c^q(X)] = O(1)$ and $\mathbb{E}[\mathbf{e}_t^q(X)] = O(1)$ for any $q < k$.

$$\begin{aligned}
\mathbb{E}[\mathbf{E}_\beta^q(X)] &= \mathbb{E}\left[k_M^{q\beta}(\mathbf{P}^{-\beta}(X) - \mathbb{E}[\mathbf{P}^{-\beta}(X)])^q\right] \\
&= k_M^{q\beta}\sum_{i=1}^{q}\binom{q}{i}(-1)^{q-i}\mathbb{E}[\mathbf{P}^{-i\beta}(X)]\mathbb{E}[\mathbf{P}^{-(q-i)\beta}(X)] \\
&= k_M^{q\beta}\sum_{i=1}^{q}\binom{q}{i}(-1)^{q-i}\Theta((k/M)^{-i\beta})\Theta((k/M)^{-(q-i)\beta}) \\
&= \sum_{i=1}^{q}\binom{q}{i}(-1)^{q-i}\Theta(1) = O(1). \qquad \text{(A.36)}
\end{aligned}$$

By (A.6) and (A.36),

$$\mathbb{E}[1_{\natural\natural^c(X)}\mathbf{E}_\beta^q(X)] = O(\mathcal{C}(k)).$$

By the definition of $\natural\natural(X)$,

$$1_{\natural\natural(X)}\mathbf{E}_\beta^q(X) = O\left(k^{-(\delta q/2)}\right), \qquad \text{(A.37)}$$

and therefore

$$\mathbb{E}[1_{\natural\natural(X)}\mathbf{E}_\beta^q(X)] = O\left(k^{-(\delta q/2)}\right).$$

This gives,

$$\mathbb{E}[\mathbf{E}_\beta^q(X)] = O(k^{-\delta q/2}) + O(\mathcal{C}(k)). \qquad \text{(A.38)}$$

From this analysis on $\mathbf{E}_\beta(X)$, it trivially follows from (A.32) that

$$\mathbb{E}[\mathbf{e}_c^l(X)] = O(k^{-\delta l/2}) + O(\mathcal{C}(k)). \qquad \text{(A.39)}$$

Also observe that by (A.14) and (A.15),

$$\mathbb{E}[\mathbf{e}_r^l(X)] = \mathbb{E}[1_{\natural(X)}\mathbf{e}_r^l(X)] + \mathbb{E}[1_{\natural^c(X)}\mathbf{e}_r^l(X)] = o(1/M^l) + O(\mathcal{C}(k)). \qquad \text{(A.40)}$$

We will now bound $\mathbf{e}_t^l(X)$. Let $L = \sum_{t\in\mathcal{T}} l_t t$. Now, using (A.33), $\mathbf{e}_t^l(X)$ can be expressed as a *sum* of terms of the form $(k/M)^L\binom{l}{l_1,..,l_t}\prod_{t\in\mathcal{T}}(h_t^l(X)\mathbf{E}_t^{l_t}(X))$ where $\sum_t l_t = l$. Now, we can bound each of these summands using (A.37) as follows:

$$\begin{aligned}
(k/M)^l\mathbb{E}[\prod_{t\in\mathcal{T}}\mathbf{E}_t^{l_t}(X)] &= (k/M)^L\mathbb{E}[1_{\natural\natural(X)}\prod_{t\in\mathcal{T}}\mathbf{E}_t^{l_t}(X)] + (k/M)^L\mathbb{E}[1_{\natural\natural^c(X)}\prod_{t\in\mathcal{T}}\mathbf{E}_t^{l_t}(X)] \\
&= (k/M)^L\prod_{t\in\mathcal{T}}O(k^{-l_t\delta/2}) + O(\mathcal{C}(k)) \\
&= (k/M)^L O(k^{-l\delta/2}) + O(\mathcal{C}(k)) \\
&= o(k^{-l\delta/2}) + O(\mathcal{C}(k)). \qquad \text{(A.41)}
\end{aligned}$$

This implies that

$$\mathbb{E}[\mathbf{e}_t^l(X)] = o(k^{-l\delta/2}) + O(\mathcal{C}(k)). \qquad \text{(A.42)}$$

Note that $\mathbf{e}_o^q(X)$ will contain terms of the form $(\mathbf{e}_c(X)+\mathbf{e}_t(X))^l(\mathbf{e}_r(X))^{q-l}$. If $l<q$, the expectation of this term can be bounded as follows

$$\begin{aligned}
&|\mathbb{E}[(\mathbf{e}_c(X)+\mathbf{e}_t(X))^l(\mathbf{e}_r(X))^{q-l}]| \\
&\leq \sqrt{\mathbb{E}[(\mathbf{e}_c(X)+\mathbf{e}_t(X))^{2l}]\mathbb{E}[(\mathbf{e}_r(X))^{2(q-l)}]} \\
&= \sqrt{O(1)^{2l}(o(1/M))^{2(q-l)}} \\
&= O(1)\times(o(1/M))^{q-l} = o(1/M). \qquad \text{(A.43)}
\end{aligned}$$

Let us concentrate on the case $l=q$. In this case, $\mathbf{e}_k^q(X)$ will contain terms of the form $(\mathbf{e}_c(X))^m(\mathbf{e}_t(X))^{q-m}$. For $m<q$,

$$\begin{aligned}
&|\mathbb{E}[(\mathbf{e}_c(X))^m(\mathbf{e}_t(X))^{q-m}]| \\
&\leq \sqrt{\mathbb{E}[(\mathbf{e}_c(X))^{2l}]\mathbb{E}[(\mathbf{e}_t(X))^{2(q-l)}]} \\
&= \left(O(k^{-m\delta/2})\times o(k^{-(q-m)\delta/2})\right)+\mathcal{C}(k) = o(k^{-q\delta/2})+O(\mathcal{C}(k)). \qquad \text{(A.44)}
\end{aligned}$$

This therefore implies that, by (A.39), (A.40), (A.42), (A.43) and (A.44),

$$\begin{aligned}
\mathbb{E}[\mathbf{e}_o^q(X)] &= \mathbb{E}[\mathbf{e}_c^q(X)]+o(k^{-q\delta/2})+\mathcal{C}(k) \\
&= O(k^{-q\delta/2})+o(k^{-q\delta/2})+o(1/M)+\mathcal{C}(k) \\
&= O(k^{-q\delta/2})+o(1/M)+\mathcal{C}(k). \qquad \text{(A.45)}
\end{aligned}$$

This finally implies that

$$\begin{aligned}
\mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})\big] &= \mathbb{E}\big[1_{\dagger(\mathbf{X})}\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})\big]+O(\mathcal{C}(k)) \quad (by(A.35)) \\
&= \mathbb{E}\big[1_{\dagger(\mathbf{X})}\gamma(\mathbf{X})\mathbf{e}_o^q(\mathbf{X})\big]+O(\mathcal{C}(k)) \\
&= \mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_o^q(\mathbf{X})\big]+O(\mathcal{C}(k)) \quad (by(A.34)) \\
&= O(k^{-q\delta/2})+o(1/M)+O(\mathcal{C}(k)). \qquad \text{(A.46)}
\end{aligned}$$

This concludes the proof.

■

Before proving Lemma A.2, we seek to answer the following question: for which set of pair of points $\{X,Y\}$ are the $k$-NN balls disjoint?

*1) Intersecting and disjoint balls:* Define $\Psi_\epsilon := \{X,Y\}\in\mathcal{S}' : ||X-Y||\geq R_\epsilon(X)+R_\epsilon(Y)$ where $R_\epsilon(X)$ and $R_\epsilon(Y)$ are the ball radii of the spherical regions $S_u(X)$ and $S_u(Y)$, such that $\int_{S_u(X)}f(z)dz=\int_{S_u(Y)}f(z)dz=(1+p_k)k_M$. We will now show that for $\{X,Y\}\in\Psi_\epsilon$, the $k$-NN balls will be disjoint with exponentially high probability. Let $\mathbf{d}_{\mathbf{X}}^{(\mathbf{k})}$ and $\mathbf{d}_{\mathbf{Y}}^{(\mathbf{k})}$ denote the $k$-NN distances from $X$ and $Y$ and let $\boldsymbol{\Upsilon}$ denote the event that the $k$-NN

balls intersect. For $\{X,Y\} \in \Psi_\epsilon$,

$$\begin{aligned} Pr(\boldsymbol{\Upsilon}) &= Pr(\mathbf{d}_{\mathbf{X}}^{(\mathbf{k})} + \mathbf{d}_{\mathbf{Y}}^{(\mathbf{k})} \geq ||X-Y||) \\ &\leq Pr(\mathbf{d}_{\mathbf{X}}^{(\mathbf{k})} + \mathbf{d}_{\mathbf{Y}}^{(\mathbf{k})} \geq R_\epsilon(X) + R_\epsilon(Y)). \\ &\leq Pr(\mathbf{d}_{\mathbf{X}}^{(\mathbf{k})} \geq R_\epsilon(X)) + Pr(\mathbf{d}_{\mathbf{Y}}^{(\mathbf{k})} \geq R_\epsilon(Y)) \\ &= Pr(\mathbf{P}(X) \geq (p_k+1)((k-1)/M)) \\ &\quad + Pr(\mathbf{P}(Y) \geq (p_k+1)((k-1)/M)) \\ &= 2\mathcal{C}(k), \end{aligned}$$

where the last inequality follows from the concentration inequality (A.1). We conclude that for $\{X,Y\} \in \Psi_\epsilon$, the probability of intersection of $k$-NN balls centered at $X$ and $Y$ decays exponentially in $p_k^2 k$. Stated in a different way, we have shown that for a given pair of points $\{X,Y\}$, if the $\epsilon$ balls around these points are disjoint, then the $k$-NN balls will be disjoint with exponentially high probability. Let $\Delta_\epsilon(X,Y)$ denote the event $\{X,Y\} \in \Psi_\epsilon^c$. From the definition of the region $\Psi_\epsilon$, we have $Pr(\{\mathbf{X},\mathbf{Y}\} \in \Psi_\epsilon^c) = O(k/M)$.

Let $\{X,Y\} \in \Psi_\epsilon$ and let $q, r$ be non-negative integers satisfying $q + r > 1$. The event that the $k$-NN balls intersect is given by $\boldsymbol{\Upsilon} := \{\mathbf{d}_{\mathbf{X}}^{(\mathbf{k})} + \mathbf{d}_{\mathbf{Y}}^{(\mathbf{k})} > ||X-Y||\}$. The joint probability distribution of $\mathbf{P}(X)$ and $\mathbf{P}(Y)$ when the $k$-NN balls do not intersect $=: \boldsymbol{\Upsilon}^{\mathbf{c}}$ is given by

$$f_{\boldsymbol{\Upsilon}^{\mathbf{c}}}(p_X, p_Y) = M! \frac{(p_X p_Y)^{k-1}}{(k-1)!^2} \frac{(1-p_X-p_Y)^{M-2k}}{(M-2k)!}.$$

Define

$$i(p_X, p_Y) = \frac{\Gamma(t)\Gamma(u)\Gamma(v)}{\Gamma(t+u+v)} p_X^{t-1} p_Y^{u-1} (1-p_X-p_Y)^{v-1},$$

and note that

$$\int_{p_X=0}^{1} \int_{p_Y=0}^{1} 1_{\{p_X+p_Y \leq 1\}} i(p_X, p_Y) dp_X dp_Y = 1.$$

Now note that $i(p_X, p_Y)$ corresponds to the density function $f_{\boldsymbol{\Upsilon}^{\mathbf{c}}}(p_X, p_Y)$ for the choices $t = k$, $u = k$ and $v = M - 2k + 1$. Furthermore, for $\{X,Y\} \in \Psi_\epsilon$, the set $\mathcal{Q} := \{p_X, p_Y\} : p_X, p_Y \leq (1+p_k)(k-1)/M$ is a subset of the region $\mathcal{T} := \{p_X, p_Y\} : 0 \leq p_X, p_Y \leq 1;\ p_X + p_Y \leq 1$. Note that $\mathbb{E}[1_{\mathcal{Q}}] = 1 - \mathcal{C}(k)$. This implies that expectations over the region $\mathcal{R} := \{p_X, p_Y\} : 0 \leq p_X, p_Y \leq 1$; should be of the same order as the expectations over $\mathcal{T}$ with differences of order $\mathcal{C}(k)$. In particular, for $t, u < k$,

$$\mathbb{E}[\mathbf{P}^{-t}(X)\mathbf{P}^{-u}(Y)] = \mathbb{E}[1_{\mathcal{T}}\mathbf{P}^{-t}(X)\mathbf{P}^{-u}(Y)] + \mathcal{C}(k).$$

From the joint distribution representation, it follows that

$$\frac{\mathbb{E}[1_{\mathcal{T}}\mathbf{P}^{-t}(X)\mathbf{P}^{-u}(Y)]}{\mathbb{E}[\mathbf{P}^{-t}(X)]\mathbb{E}[\mathbf{P}^{-u}(Y)]} = \frac{\Gamma(M-t)\Gamma(M-u)}{\Gamma(M-t-u)\Gamma(M)} = -\frac{tu}{M} + O(1/M^2). \tag{A.47}$$

Now observe that

$$\begin{aligned}
&(k_M)^{t+u} Cov(\mathbf{P}^{-t}(X), \mathbf{P}^{-u}(Y)) \\
&= (k_M)^{t+u} [\mathbb{E}[\mathbf{P}^{-t}(X)\mathbf{P}^{-u}(Y)] - \mathbb{E}[\mathbf{P}^{-t}(X)]\mathbb{E}[\mathbf{P}^{-u}(Y)]] \\
&= (k_M)^{t+u} \mathbb{E}[\mathbf{P}^{-t}(X)]\mathbb{E}[\mathbf{P}^{-u}(Y)] \left[ \frac{\mathbb{E}[\mathbf{P}^{-t}(X)\mathbf{P}^{-u}(Y)]}{\mathbb{E}[\mathbf{P}^{-t}(X)]\mathbb{E}[\mathbf{P}^{-u}(Y)]} - 1 \right] \\
&= (k_M)^{t+u} \Theta(k_M^{-t}) \Theta(k_M^{-u}) \left[ 1 - \frac{tu}{M} + o(1/M^2) - 1 \right] \qquad \text{(by (A.7) and (A.47))} \\
&= -\left( \frac{tu}{M} \right) + O(1/M^2). \qquad \text{(A.48)}
\end{aligned}$$

Then, the covariance between the powers of the error function $\mathbf{E}_\beta$, for $qt, ru < k$ is given by

$$\begin{aligned}
Cov(\mathbf{E}_t^q(X), \mathbf{E}_u^r(Y)) &= k_M^{(tq+ur)} Cov\left( \left[\mathbf{P}^{-t}(X) - \mathbb{E}\left[\mathbf{P}^{-t}(X)\right]\right]^q, \left[\mathbf{P}^{-u}(Y) - \mathbb{E}\left[\mathbf{P}^{-u}(Y)\right]\right]^r \right) \\
&= \sum_{a=1}^{q} \sum_{b=1}^{r} \binom{q}{a} \binom{r}{b} [(-1)^{a+b} + o(1)] k_M^{(ta+ub)} Cov(\mathbf{P}^{-ta}(X), \mathbf{P}^{-ub}(Y)) \\
&= -tu \sum_{a=1}^{q} \sum_{b=1}^{r} \binom{q}{a} \binom{r}{b} \frac{(-1)^a a (-1)^b b}{M} + O\left( \frac{1}{M^2} \right) \\
&= 1_{\{q=1,r=1\}} \left( \frac{-tu}{M} \right) + O(1/M^2). \qquad \text{(A.49)}
\end{aligned}$$

*Proof:* of Lemma A.2. Let $\mathbf{X}_1, .., \mathbf{X}_M, \mathbf{X}, \mathbf{Y}$ denote $M+2$ i.i.d realizations of the density $f$. Then, identical to the derivation of (A.46) in the proof of Lemma A.1,

$$\begin{aligned}
&Cov\big[1_{\{\mathbf{X} \in \mathcal{S}'\}} \gamma_1(\mathbf{X}) \mathbf{e}_k^q(\mathbf{X}), 1_{\{\mathbf{Y} \in \mathcal{S}'\}} \gamma_2(\mathbf{Y}) \mathbf{e}_k^r(\mathbf{Y})\big] \\
&= Cov\big[1_{\{\mathbf{X} \in \mathcal{S}'\}} \gamma_1(\mathbf{X}) \mathbf{e}_o^q(\mathbf{X}), 1_{\{\mathbf{Y} \in \mathcal{S}'\}} \gamma_2(\mathbf{Y}) \mathbf{e}_o^r(\mathbf{Y})\big] + O(\mathcal{C}(k)).
\end{aligned}$$

Using the exact same arguments as in proof of Lemma A.1, it can be shown that the contribution of terms $\mathbf{e}_r(\mathbf{X})$,$\mathbf{e}_r(\mathbf{Y})$ to the R.H.S. of the above equation is $o(1/M)$. Define $\sharp(\mathbf{X}, \mathbf{Y}) := \gamma_1(\mathbf{X})\gamma_2(\mathbf{Y}) Cov_{\{\mathbf{X},\mathbf{Y}\}}[(\mathbf{e}_c(X) + \mathbf{e}_t(X))^q, (\mathbf{e}_c(Y) + \mathbf{e}_t(Y))^r]$. Thus,

$$\begin{aligned}
&Cov\big[1_{\{\mathbf{X} \in \mathcal{S}'\}} \gamma_1(\mathbf{X}) \mathbf{e}_k^q(\mathbf{X}), 1_{\{\mathbf{Y} \in \mathcal{S}'\}} \gamma_2(\mathbf{Y}) \mathbf{e}_k^r(\mathbf{Y})\big] \\
&= \mathbb{E}[1_{\{\mathbf{X},\mathbf{Y} \in \mathcal{S}'\}} \sharp(\mathbf{X}, \mathbf{Y})] + O(\mathcal{C}(k)) \\
&= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon{}^c(\mathbf{X},\mathbf{Y})} \sharp(\mathbf{X}, \mathbf{Y})] + \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})} \sharp(\mathbf{X}, \mathbf{Y})] + O(\mathcal{C}(k)) \\
&= I + II + O(\mathcal{C}(k)).
\end{aligned}$$

*For* $\{X,Y\} \in \Psi_\epsilon^c$*:* The covariance term $Cov_{\{\mathbf{X},\mathbf{Y}\}}[(\mathbf{e}_c(X)+\mathbf{e}_t(X))^q, (\mathbf{e}_c(Y)+\mathbf{e}_t(Y))^r]$ can be shown to be $O(k^{-(q+r)\delta/2})$ for $q, r < k$ by using Cauchy-Schwarz and (A.43), (A.44) as follows.

$$\begin{aligned} |Cov[(\mathbf{e}_c(X)+\mathbf{e}_t(X))^q, (\mathbf{e}_c(Y)+\mathbf{e}_t(Y))^r]| &\leq \sqrt{\mathbb{V}[(\mathbf{e}_c(X)+\mathbf{e}_t(X))^q]\mathbb{V}[(\mathbf{e}_c(Y)+\mathbf{e}_t(Y))^r]} \\ &\leq \sqrt{\mathbb{E}[(\mathbf{e}_c(X)+\mathbf{e}_t(X))^{2q}]\mathbb{E}[(\mathbf{e}_c(Y)+\mathbf{e}_t(Y))^{2r}]} \\ &= \sqrt{O(k^{-(2q)\delta/2})O(k^{-(2r)\delta/2})} \\ &= O(k^{-(q+r)\delta/2}). \end{aligned} \tag{A.50}$$

This implies that

$$II = \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\sharp(\mathbf{X},\mathbf{Y})] = \mathbb{E}\Big[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}O(k^{-(q+r)\delta/2})\Big] = O\left(\frac{1}{k^{((q+r)\delta/2-1)}M}\right),$$

where the last but one step follows since the probability $Pr(\{\mathbf{X},\mathbf{Y}\} \in \Psi_\epsilon^c) = O(k/M)$.

*For* $\{X,Y\} \in \Psi_\epsilon$*:* Now note that $(\mathbf{e}_c(X)+\mathbf{e}_t(X))^q$ will contain terms of the form $(\mathbf{e}_c(X))^m(\mathbf{e}_t(X))^{q-m}$. For $m < q$, the term $(\mathbf{e}_c(X))^m(\mathbf{e}_t(X))^{q-m}$ will be a sum of terms of the form $(k/M)^{(m+u)}\mathbf{P}^{-(m+v)}(X)$ for arbitrary $v < q-m$ with $u - v >= 2/d$. By (A.48), the covariance term $Cov[(\mathbf{e}_c(X))^m(\mathbf{e}_t(X))^{q-m}, (\mathbf{e}_c(Y))^n(\mathbf{e}_t(Y))^{r-m}]$ will be therefore be $O(k_M^{2/d}/M)$ if either $m < q$ or $n < r$.

On the other hand, if $m = q$ and $n = r$, $Cov[(\mathbf{e}_c(X))^q, (\mathbf{e}_c(Y))^r] = 1_{\{q=1,r=1\}}O(1/M) + O(1/M^2)$ by noting that the error $\mathbf{e}_c(X) = f(X)\mathbf{E}_1(X)$ and subsequently invoking (A.49). Therefore

$$\begin{aligned} I &= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon^{\mathbf{c}}(\mathbf{X},\mathbf{Y})}\sharp(\mathbf{X},\mathbf{Y})] \\ &= \mathbb{E}\Big[\mathbf{1}_{\mathbf{\Delta}_\epsilon^{\mathbf{c}}(\mathbf{X},\mathbf{Y})}\left(1_{\{q=1,r=1\}}O(1/M) + O(k_M^{2/d}/M) + O(1/M^2)\right)\Big] \\ &= 1_{\{q=1,r=1\}}O(1/M) + O(k_M^{2/d}/M) + O(1/M^2), \end{aligned}$$

where the last step follows from the fact that probability $Pr(\{\mathbf{X},\mathbf{Y}\} \in \Psi_\epsilon) = 1 - O(k/M) = O(1)$. ■

### *D. Specific cases*

We now focus on evaluating the specific cases

$$\mathbb{E}\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_k^2(\mathbf{X})\big]$$

and

$$Cov\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e}_k(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e}_k(\mathbf{Y})\big],$$

for $k > 2$.

*1) Evaluation of* $\mathbb{E}\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_k^2(\mathbf{X})\right]$*:* $\mathbf{P}(X)$ has a beta distribution with parameters $k, M-k+1$. Therefore for $k>2$

$$\begin{aligned}\mathbb{E}[\mathbf{E}_\beta^2(X)] &= \mathbb{E}\left[k_M^{2\beta}(\mathbf{P}^{-\beta}(X)-\mathbb{E}[\mathbf{P}^{-\beta}(X)])^2\right]\\ &= k_M^{2\beta}\mathbb{E}[\mathbf{P}^{-2\beta}(X)]-\left(\mathbb{E}[\mathbf{P}^{-\beta}(X)]\right)^2\\ &= k_M^{2\beta}\left(\frac{\Gamma(k-2\beta)\Gamma(M+1)}{\Gamma(k)\Gamma(M+1-2\beta)}-\left(\frac{\Gamma(k-\beta)\Gamma(M+1)}{\Gamma(k)\Gamma(M+1-\beta)}\right)^2\right)\\ &= O(1/k) \qquad \text{(A.51)}\end{aligned}$$

where the last step follows by noting that for any $a>0$,

$$\frac{\Gamma(x)}{\Gamma(x+a)} = x^{-a}(1+o(1/x)).$$

From ( A.46),

$$\mathbb{E}\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_k^2(\mathbf{X})\right] = \mathbb{E}\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_o^2(\mathbf{X})\right]+O(\mathcal{C}(k)). \qquad \text{(A.52)}$$

Note that $\mathbf{e}_o^2(X)=(\mathbf{e}_c(X)+\mathbf{e}_t(X)+\mathbf{e}_r(X))^2$ is a sum of terms of the form $(\mathbf{e}_c(X))^{2-l-m}(\mathbf{e}_t(X))^l(\mathbf{e}_r(X))^m$. Also,

$$\begin{aligned}\mathbb{E}[\mathbf{e}_c^2(X)] &= f^2(X)\mathbb{E}\left[k_M^2(\mathbf{P}^{-1}(X)-\mathbb{E}[\mathbf{P}^{-1}(X)])^2\right]\\ &= f^2(X)k_M^2\mathbb{E}[\mathbf{P}^{-2}(X)]-\left(\mathbb{E}[\mathbf{P}^{-1}(X)]\right)^2\\ &= f^2(X)k_M^{2\beta}\left(\frac{\Gamma(k-2)\Gamma(M+1)}{\Gamma(k)\Gamma(M+1-2)}-\left(\frac{\Gamma(k-1)\Gamma(M+1)}{\Gamma(k)\Gamma(M)}\right)^2\right)\\ &= \frac{1}{k}+o\left(\frac{1}{k}\right). \qquad \text{(A.53)}\end{aligned}$$

Using (A.51), identical to the derivation of (A.43) and (A.44), it is clear that if $l+m>0$, $\mathbb{E}[(\mathbf{e}_c(X))^{2-l-m}(\mathbf{e}_t(X))^l(\mathbf{e}_r(X))^m] = o(k^{-1})+o(1/M)+O(\mathcal{C}(k))$. This implies that

$$\begin{aligned}\mathbb{E}\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_k^2(\mathbf{X})\right] &= \mathbb{E}\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_o^2(\mathbf{X})\right]+O(\mathcal{C}(k))\\ &= f^2(X)\left(\frac{1}{k}\right)+o\left(\frac{1}{k}\right). \qquad \text{(A.54)}\end{aligned}$$

*2) Evaluation of* $Cov\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e}_k(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e}_k(\mathbf{Y})\big]$*:* We separately analyze disjoint balls and intersecting balls as follows:

$$
\begin{aligned}
&Cov\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e}_k(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e}_k(\mathbf{Y})\big]\\
&= \mathbb{E}[\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e}_k(\mathbf{X})\mathbf{e}_k(\mathbf{Y})\big]]\\
&= \mathbb{E}[\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e}_o(\mathbf{X})\mathbf{e}_o(\mathbf{Y})\big]] + O(\mathcal{C}(k))\\
&= \mathbb{E}[\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})(\mathbf{e}_c(\mathbf{X})+\mathbf{e}_t(\mathbf{X})+\mathbf{e}_r(\mathbf{X}))(\mathbf{e}_c(\mathbf{Y})+\mathbf{e}_t(\mathbf{Y})+\mathbf{e}_r(\mathbf{Y}))\big]] + O(\mathcal{C}(k))\\
&= \mathbb{E}[\big[1_{\{\mathbf{X}\in\mathcal{S}'\}}1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})(\mathbf{e}_c(\mathbf{X})+\mathbf{e}_t(\mathbf{X}))(\mathbf{e}_c(\mathbf{Y})+\mathbf{e}_t(\mathbf{Y}))\big]] + O(\mathcal{C}(k)) + o(1/M)\\
&= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon{}^c(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}}[(\mathbf{e}_c(X)+\mathbf{e}_t(X))(\mathbf{e}_c(Y)+\mathbf{e}_t(Y))]]\\
&+\mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}}[(\mathbf{e}_c(X)+\mathbf{e}_t(X))(\mathbf{e}_c(Y)+\mathbf{e}_t(Y))]]\\
&+O(\mathcal{C}(k)) + o(1/M)\\
&= I + II + O(\mathcal{C}(k)) + o(1/M).
\end{aligned}
$$

*For* $\{X,Y\}\in\Psi_\epsilon$*:*

$$\mathbb{E}[(\mathbf{e}_c(X))(\mathbf{e}_c(Y))] = Cov[(\mathbf{e}_c(X)),(\mathbf{e}_c(Y))] = \frac{-f(X)f(Y)}{M} + O(1/M^2)$$

by noting that the error $\mathbf{e}_c(X) = \mathbf{E}_1(X)/f(X)$ and subsequently invoking (A.49) in conjunction with the condition $k>2$. Similarly, using (A.32), (A.33) and (A.49),

$$
\begin{aligned}
\mathbb{E}[(\mathbf{e}_c(X))(\mathbf{e}_t(Y))] &= O(k_M^{2/d}/M) + O(1/M^2),\\
\mathbb{E}[(\mathbf{e}_t(X))(\mathbf{e}_c(Y))] &= O(k_M^{2/d}/M) + O(1/M^2),\\
\mathbb{E}[(\mathbf{e}_t(X))(\mathbf{e}_t(Y))] &= O(k_M^{4/d}/M) + O(1/M^2).
\end{aligned}
$$

This implies that

$$
\begin{aligned}
I &= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon^c(\mathbf{X},\mathbf{Y})}\mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}}[(\mathbf{e}_c(X)+\mathbf{e}_t(X))(\mathbf{e}_c(Y)+\mathbf{e}_t(Y))]]\\
&= \mathbb{E}\Big[\mathbf{1}_{\mathbf{\Delta}_\epsilon^c(\mathbf{X},\mathbf{Y})}\left(-f(X)f(Y)(1/M) + O(k_M^{2/d}/M) + O(1/M^2)\right)\Big]\\
&= \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})(f(\mathbf{X})f(\mathbf{Y}))]\left(-1/M + O(k_M^{2/d}/M) + O(1/M^2)\right)\\
&= -\mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})f(\mathbf{X})]\mathbb{E}[1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})f(\mathbf{Y})]\frac{1}{M} + O(k_M^{2/d}/M) + O(1/M^2). \qquad \text{(A.55)}
\end{aligned}
$$

where the last but one step follows from the fact that probability $Pr(\{\mathbf{X},\mathbf{Y}\}\in\Psi_\epsilon) = 1 - O(k/M) = O(1)$.

*For* $\{X,Y\}\in\Psi_\epsilon^c$*:* First observe that by Cauchy Schwarz, and by (A.51) $|\mathbb{E}[\mathbf{E}_t(X)\mathbf{E}_u(X)]| \leq \sqrt{\mathbb{E}[\mathbf{E}_t^2(X)]\mathbb{E}[\mathbf{E}_u^2(X)]} = O(1/k)$. This implies that

$$\mathbb{E}[(\mathbf{e}_c(X)+\mathbf{e}_t(X))(\mathbf{e}_c(Y)+\mathbf{e}_t(Y))] = \mathbb{E}[\mathbf{e}_c(X)\mathbf{e}_c(Y)] + O(k_M^{2/d}/k). \qquad \text{(A.56)}$$

In subsection A-F, we will show Lemma A.5, which states that

$$\begin{aligned}&\mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e_c}(\mathbf{X})\mathbf{e_c}(\mathbf{Y})]\\&= \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\gamma_2(\mathbf{X})f^2(\mathbf{X})]\left(\frac{1}{M}+o\left(\frac{1}{M}\right)\right)\end{aligned}$$

This implies that

$$\begin{aligned}II &= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}}[(\mathbf{e}_c(X)+\mathbf{e}_t(X))(\mathbf{e}_c(Y)+\mathbf{e}_t(Y))]]\\&= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}}[\mathbf{e}_c(X)\mathbf{e}_c(Y)]+O(k_M^{2/d}/k)]\\&= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e_c}(\mathbf{X})\mathbf{e_c}(\mathbf{Y})]+\mathbb{E}\left[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\left(O(k_M^{2/d}/k)\right)\right]\\&= \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\gamma_2(\mathbf{X})/f^2(\mathbf{X})]\left(\frac{1}{M}+O(k_M^{2/d}/M)+o\left(\frac{1}{M}\right)\right)\end{aligned}\tag{A.57}$$

where the last step follows from recognizing that $Pr(\{\mathbf{X},\mathbf{Y}\}\in\Psi_\epsilon^c)=O(k/M)$ and $O(k/M)\times 1/k = O(1/M)$. This implies that

$$\begin{aligned}&Cov\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e}_k(\mathbf{X}),1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e}_k(\mathbf{Y})\right]\\&= I+II+O(\mathcal{C}(k))+o(1/M)\\&= Cov[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})/f(\mathbf{X}),1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})/f(\mathbf{Y})]\left(\frac{1}{M}\right)+o(1/M)+O(\mathcal{C}(k)).\end{aligned}\tag{A.58}$$

*E. Summary*

Noting that $\delta>2/3$, the equations (A.26), (A.2), (A.54), (A.58) imply that for positive integers $q,r<k$,

$$\mathbb{E}\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})\mathbf{e}_k^q(\mathbf{X})\right]=1_{\{q=2\}}\mathbb{E}\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma(\mathbf{X})f^2(\mathbf{X})\right]\left(\frac{1}{k}\right)+o\left(\frac{1}{k}\right)+O(\mathcal{C}(k)),\tag{A.59}$$

$$\begin{aligned}&Cov\left[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\mathbf{e}_k^q(\mathbf{X}),1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})\mathbf{e}_k^r(\mathbf{Y})\right]\\&= 1_{\{q,r=1\}}Cov[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})f(\mathbf{X}),1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})f(\mathbf{Y})]\left(\frac{1}{M}+o(1/M)\right)\\&+1_{\{q+r>2\}}\left(O\left(\frac{1}{k^{((q+r)\delta/2-1)}M}\right)+O(k_M^{2/d}/M)+O(1/M^2)\right)+O(\mathcal{C}(k)).\end{aligned}\tag{A.60}$$

*F. Evaluation of* $\mathbb{E}[\mathbf{e}_c(X)\mathbf{e}_c(Y)]$ *for* $\{X,Y\}\in\Psi_\epsilon^c$

For $\{X,Y\}\in\Psi_\epsilon^c$, it will be shown that the cross-correlations $\mathbb{E}[\mathbf{e}_c(X)\mathbf{e}_c(Y)]$ of the coverage density estimator and an oracle uniform kernel density estimator (defined below) are identical up to leading terms (without explicitly evaluating the cross-correlation between the coverage density estimates) and then derive the correlation of the oracle density estimator to obtain corresponding results for the coverage estimate.

*Oracle $\epsilon$ ball density estimate:* In order to estimate cross moments for the $k$-NN density estimator, the $\epsilon$ *ball* density estimator is introduced. The $\epsilon$-ball density estimator is a kernel density estimator that uses a uniform kernel with bandwidth which depends on the unknown density $f$. Let the volume of the kernel be $V_\epsilon(X)$ and the corresponding kernel region be $S_\epsilon(X) = \{Y \in \mathcal{S} : c_d||X-Y||^d \leq V_\epsilon(X)\}$. The volume is chosen such that the coverage $Q_\epsilon(X) = \int_{S_\epsilon(X)} f(z)dz$ is set to $(1+p_k)k/M$. Let $\mathbf{l}_\epsilon(X)$ denote the number of points among $\{\mathbf{X}_1,..,\mathbf{X}_M\}$ falling in $S_\epsilon(X)$: $\mathbf{l}_\epsilon(\mathbf{X}) = \Sigma_{i=1}^M 1_{\mathbf{X}_i \in S_\epsilon(X)}$. The $\epsilon$ *ball* density estimator is defined as

$$\hat{\mathbf{f}}_\epsilon(X) = \frac{\mathbf{l}_\epsilon(\mathbf{X})}{MV_\epsilon(X)}. \tag{A.61}$$

Also define the error $\mathbf{e}_\epsilon(X)$ as $\mathbf{e}_\epsilon(X) = \hat{\mathbf{f}}_\epsilon(X) - \mathbb{E}[\hat{\mathbf{f}}_\epsilon(X)]$. It is then possible to prove the following lemma using results on the volumes of intersections of hyper spheres (refer Appendix A, [31] for details).

**Lemma A.3.** *Let $\gamma_1(X)$, $\gamma_2(X)$ be arbitrary continuous functions. Let $\mathbf{X}_1,..,\mathbf{X}_M,\mathbf{X},\mathbf{Y}$ denote $M+2$ i.i.d realizations of the density $f$. Then,*

$$\begin{aligned}&\mathbb{E}\big[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\mathbf{e}_\epsilon(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e}_\epsilon(\mathbf{Y})\big]\\ &= \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\gamma_2(\mathbf{X})f^2(\mathbf{X})]\left(\frac{1}{M}+o\left(\frac{1}{M}\right)\right).\end{aligned}$$

Next, the cross-correlations of the coverage density estimator and the $\epsilon$ ball density estimator are shown to be asymptotically equal. In particular,

**Lemma A.4.**

$$\mathbb{E}[\mathbf{e}_c(X)\mathbf{e}_c(Y)] = \mathbb{E}[\mathbf{e}_\epsilon(X)\mathbf{e}_\epsilon(Y)] + o(1/k).$$

*Proof:*

We begin by establishing the conditional density and expectation of $\hat{\mathbf{f}}_\epsilon(X)$ given $\hat{\mathbf{f}}_\mathbf{c}(X)$. We drop the dependence on $X$ and denote $\mathbf{l}_\epsilon = \Sigma_{i=1}^M 1_{\{X_i \in S_\epsilon(X)\}}$, the $k$-NN coverage by $\mathbf{P}$ and the $\epsilon$ ball coverage $Q_\epsilon(X)$ by $Q$. Let $\mathbf{q} = Q/\mathbf{P}$ and $\mathbf{r} = (Q-\mathbf{P})/(1-\mathbf{P})$. The following expressions for conditional densities and expectations are derived in [35]

$$\begin{aligned}&\mathbf{P}r\{\mathbf{l}_\epsilon = l|\mathbf{P};\mathbf{P}>Q\}\\ &= \begin{cases}\binom{k-1}{l}\mathbf{q}^l(1-\mathbf{q})^{k-1-l} & l=0,1,\ldots,k-1\\ 0 & l=k,k+1,\ldots,M\end{cases}\end{aligned}$$

$$\begin{aligned}&\mathbf{P}r\{\mathbf{l}_\epsilon = l|\mathbf{P};\mathbf{P}\leq Q\}\\ &= \begin{cases}0 & l=0,1,\ldots,k-1\\ \binom{M-k}{l-k}\mathbf{r}^{l-k}(1-\mathbf{r})^{M-l} & l=k,k+1,\ldots,M\end{cases}\end{aligned}$$

which implies

$$\mathbb{E}[\mathbf{l}_\epsilon = l|\mathbf{P}; \mathbf{P} > Q] = (k-1)Q/\mathbf{P}$$
$$\mathbb{E}[\mathbf{l}_\epsilon = l|\mathbf{P}; \mathbf{P} \leq Q] = \left(\frac{1-Q}{1-\mathbf{P}}\right)(k-M) + M$$

Using the above expressions for conditional expectations, the following marginal expectation are obtained. Denote the density of the coverage $\mathbf{P}$ by $f_{k,M}(p)$. Also let $\hat{\mathbf{P}}$ be the coverage corresponding to the $k-2$ nearest neighbor in a total field of $M-3$ points. Then

$$\begin{aligned}
\mathbb{E}[\tilde{\mathbf{e}}_c(X)\tilde{\mathbf{e}}_\epsilon(X)] &= \mathbb{E}[\hat{\mathbf{f}}_\epsilon(X)\hat{\mathbf{f}}_\mathbf{c}(X)] - \mathbb{E}[\hat{\mathbf{f}}_\mathbf{c}(X)]\mathbb{E}[\hat{\mathbf{f}}_\epsilon(X)] \\
&= \mathbb{E}\left[\left(\left(\frac{1-Q}{\mathbf{P}(1-\mathbf{P})}\right)(k-M) + M/\mathbf{P}\right)1_{\mathbf{P}\leq Q}\right] \\
&\quad + \frac{f^2(X)(k-1)}{kM}\mathbb{E}\left[\left((k-1)Q/\mathbf{P}^2\right)1_{\mathbf{P}>Q}\right] - \frac{f^2(X)}{k}MQ. \\
&= \frac{f^2(X)}{k}\frac{(M-1)(M-2)}{(k-2)(M-k)} \times \\
&\quad \mathbb{E}[(1-Q\hat{\mathbf{P}})(k-M) + M\hat{\mathbf{P}}(1-\hat{\mathbf{P}})] - \frac{f^2(X)}{k}MQ \\
&\quad + \mathbb{E}[((k-1)Q(1-\hat{\mathbf{P}}) - (1-Q\hat{\mathbf{P}})(k-M) + M\hat{\mathbf{P}}(1-\hat{\mathbf{P}}))(1_{\hat{\mathbf{P}}>Q})] \\
&= C \times (I - II + III).
\end{aligned}$$

It can be shown that $C \times (I - II) = \frac{f^2(X)}{k}(1-Q)$ using the fact that $\hat{\mathbf{P}}$ has a beta distribution. Note that from the definition of $Q = ((1+p_k)(k-1)/M)$, from the concentration inequality we have that $\mathbb{E}[1_{\hat{\mathbf{P}}>Q}] = \mathcal{C}(M)$. The remainder ($C \times III$) can be simplified and bounded using the Cauchy-Schwarz inequality and the concentration inequality to show $C \times III = o(1/M)$.

Therefore,

$$\begin{aligned}
\mathbb{E}[\mathbf{e}_c(X)\mathbf{e}_\epsilon(X)] &= \frac{f^2(X)}{k}(1-Q) + \mathcal{C}(M). \\
&= \frac{f^2(X)}{k} - \frac{f^2(X)}{M} + o\left(\frac{1}{M}\right) \\
&= f^2(X)\left(\frac{1}{k} + o\left(\frac{1}{k}\right)\right). \qquad \text{(A.62)}
\end{aligned}$$

Now denote $\mathbf{E}(X) = (\mathbf{e}_c(X) - \mathbf{e}_\epsilon(X))$. Note that $\mathbb{E}[\mathbf{E}^2(X)] = \mathbb{E}[\mathbf{e}_c(X)^2] - 2E[\mathbf{e}_c(X)\mathbf{e}_\epsilon(X)] + \mathbb{E}[\mathbf{e}_\epsilon(X)^2]$. Since $E[\mathbf{e}_c(X)^2] = f^2(X)\frac{1}{k} + o(1/k)$ and $E[\mathbf{e}_\epsilon(X)^2] = f^2(X)(1/k + o(1/k))$ it follows from (A.62) that $\mathbb{E}[E(X)] = o(1/k)$. This result means $\mathbf{e}_c(X)$ and $\mathbf{e}_\epsilon(X)$ are almost perfectly correlated. Next express the covariance between

the coverage density estimates in terms of the covariance between the $\epsilon$ ball estimates as follows:

$$\begin{aligned}
&\mathbb{E}[\mathbf{e}_c(X)\mathbf{e}_c(Y)] \\
&= \mathbb{E}[(\mathbf{e}_\epsilon(X) + \mathbf{E}(X))(\mathbf{e}_\epsilon(Y) + \mathbf{E}(Y))] \\
&= \mathbb{E}[\mathbf{e}_\epsilon(X)\mathbf{e}_\epsilon(Y)] + \mathbb{E}[\mathbf{e}_\epsilon(X)(\mathbf{E}(Y))] \\
&+\mathbb{E}[\mathbf{e}_\epsilon(Y)(\mathbf{E}(X))] + \mathbb{E}[(\mathbf{E}(X))(\mathbf{E}(Y))] \\
&= I + II + III + IV.
\end{aligned}$$

Using Cauchy-Schwarz, a bound on each of the terms $II$, $III$ and $IV$ is obtained in terms of $\mathbb{E}[\mathbf{E}(X)]$: $|II| \leq \sqrt{\mathbb{E}[\mathbf{E}(Y)]\mathbb{E}[\mathbf{e}_\epsilon{}^2(X)]}$, $|III| \leq \sqrt{\mathbb{E}[\mathbf{E}(X)]\mathbb{E}[\mathbf{e}_\epsilon{}^2(Y)]}$ and $|IV| \leq \sqrt{\mathbb{E}[\mathbf{E}(X)]\mathbb{E}[\mathbf{E}(Y)]}$. Note that the above application of Cauchy-Schwarz *decouples* the problem of joint expectation of density estimates located at two *different* points $X$and $Y$ to a problem of estimating the error $\mathbf{E}$ between two different density estimates at the *same* point(s). Therefore all the three terms $II$, $III$ and $IV$ are $o(1/k)$. This concludes the proof of Lemma A.4. ■

For Lemma A.4 to be useful, $\mathbb{E}[\mathbf{e}_\epsilon(X)\mathbf{e}_\epsilon(Y)]$ must be orders of magnitude larger than the error $o(1/k)$, which is indeed the case for $\{X, Y\} \in {\Psi_\epsilon}^c$ since $\mathbb{E}[\mathbf{e}_\epsilon(X)\mathbf{e}_\epsilon(Y)] = O(1/k)$ (Lemma A.2, Appendix .1) for such $X$ and $Y$. This lemma can be used along with previously established results on co-variance of $\epsilon$-ball density estimates (Lemma A.3) to obtain the following result:

**Lemma A.5.** *Let $\gamma_1(X)$, $\gamma_2(X)$ be arbitrary continuous functions. Let $\mathbf{X}_1, .., \mathbf{X}_M, \mathbf{X}, \mathbf{Y}$ denote $M + 2$ i.i.d realizations of the density $f$. Then,*

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e_c}(\mathbf{X})\mathbf{e_c}(\mathbf{Y})] \\
&= \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\gamma_2(\mathbf{X})f^2(\mathbf{X})]\left(\frac{1}{M} + o\left(\frac{1}{M}\right)\right)
\end{aligned}$$

*Proof:*

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbb{E}_{\mathbf{X},\mathbf{Y}}[\mathbf{e_c}(X)\mathbf{e_c}(Y)]] \\
&= \mathbb{E}[\mathbf{1}_{\mathbf{\Delta}_\epsilon(\mathbf{X},\mathbf{Y})}\gamma_1(\mathbf{X})\gamma_2(\mathbf{Y})\mathbf{e}_\epsilon(\mathbf{X})\mathbf{e}_\epsilon(\mathbf{Y})] + o(1/k) \\
&= \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})\gamma_2(\mathbf{X})f^2(\mathbf{X})]\left(\frac{1}{M} + o\left(\frac{1}{M}\right)\right).
\end{aligned}$$

In the second to last step, $o(1/M)$ is obtained for the second term by recognizing that $Pr(\{\mathbf{X}, \mathbf{Y}\} \in {\Psi_\epsilon}^c) = O(k/M)$ and $O(k/M) \times o(1/k) = o(1/M)$. ■

## APPENDIX B
## BOUNDARY EXTENSION

In the previous section, moment results were established for the standard $k$-NN density estimate $\hat{\mathbf{f}}_k(X)$ for points $X$ in any deterministic set $\mathcal{S}'$ with respect to the samples $\mathcal{X}_M = \{\mathbf{X}_{N+1}, .., \mathbf{X}_{N+M}\}$ satisfying the condition

$Pr(\mathbf{X} \notin \mathcal{S}') = o(1)$ and $\mathcal{S}' \subset \mathcal{S}_I$, where $\mathbf{X}$ is an realization from density $f$. In this section, these moment results are extended to boundary corrected $k$-NN density estimate $\tilde{\mathbf{f}}_k(X)$ for all $X \in \mathcal{S}$ as follows.

Specify the set $\mathcal{S}'$ to be $\mathcal{S}' = \mathcal{S}_I$ as defined in (II.2). Exclusively using the set $\mathcal{X}_N = \{\mathbf{X}_1, .., \mathbf{X}_N\}$, a set of interior points $\mathcal{I}_\mathcal{N} \subset \mathcal{X}_\mathcal{N}$ are determined such that $\mathcal{I}_\mathcal{N} \subset \mathcal{S}'$ with high probability $1 - O(N\mathcal{C}(k))$. Define the set of boundary points $\mathcal{B}_\mathcal{N} = \mathcal{X}_\mathcal{N} - \mathcal{I}_\mathcal{N}$. For points $X \in \mathcal{I}_N$, the boundary corrected $k$-NN density estimate $\tilde{\mathbf{f}}_k(X)$ is defined to be the standard $k$-NN estimate $\hat{\mathbf{f}}_k(X)$, and we invoke the moment properties of the standard $k$-NN density estimate $\hat{\mathbf{f}}_k(X)$ derived in the previous section. For points $X \in \mathcal{B}_\mathcal{N}$, the density estimate $\tilde{\mathbf{f}}_k(X)$ is defined as $\hat{\mathbf{f}}_k(Y_n)$ for points $Y_n \in \mathcal{I}_N$, and we invoke the moment properties of the standard $k$-NN density estimate $\hat{\mathbf{f}}_k(X)$ derived in the previous section.

### *A. Bias in the $k$-NN density estimator near boundary*

If a probability density function has bounded support, the $k$-NN balls centered at points close to the boundary are often truncated at the boundary. Let

$$\alpha_k(X) = \frac{\int_{\mathbf{S}_k(X) \cap \mathcal{S}} dZ}{\int_{\mathbf{S}_k(X)} dZ}$$

be the fraction of the volume of the $k$-NN ball inside the boundary of the support. Also define $\mathbf{V}_{k,M}(X)$ to be the $k$-NN ball volume in a sample of size $M$. For interior points $X \in \mathcal{S}'$, $\alpha_k(X) = 1$, while for boundary points $X \in \mathcal{S} - \mathcal{S}'$, $\alpha_k(X)$ is closer to $0$ when the points are closer to the boundary. For boundary points we then have

$$\mathbb{E}[\hat{\mathbf{f}}_k(X)] - f(X) = (1 - \alpha_k(X))f(X) + o(1). \tag{B.1}$$

Therefore the bias is much higher at the boundary of the support ($O(1)$) as compared to its interior ($O((k/M)^{2/d})$) (A.24). Furthermore, the bias at the support boundary does not decay to $0$ as $k/M \to 0$.

In the next section, we detect interior points $\mathcal{I}_\mathcal{N}$ which lie in $\mathcal{S}'$ with high probability $O(N\mathcal{C}(k))$. The results on bias, variance and cross-moments derived in the previous Appendix for points $X \in \mathcal{S}'$ therefore carry over to the points $\mathcal{I}_\mathcal{N}$. A density estimate at points $\mathcal{B}_\mathcal{N}$ is then proposed that will reduce the bias of density estimates close to the boundary.

### *B. Boundary point detection*

Define $V_{k,M}(X) := \frac{k}{M\alpha_k(X)f(X)}$. Let $p(k, M)$ be any positive function satisfying $p(k, M) = \Theta((k/M)^{2/d}) + (\sqrt{6}/k^{\delta/2})$. From the concentration inequality (A.1) and Taylor series expansion of the coverage function (A.13), for small values of $k/M$, we have

$$1 - Pr\left(\left|\frac{\mathbf{V}_{k,M}(X)}{V_{k,M}(X)} - 1\right| \leq p(k, M)\right) = O(\mathcal{C}(k)).$$

To determine $\mathcal{I}_\mathcal{N}$ and $\mathcal{B}_\mathcal{N}$, we first construct a $K$-NN graph on the samples $\mathcal{X}_N$ where $K = \lfloor k \times (N/M) \rfloor$. For any $X \in \mathcal{X}_\mathcal{N}$, from the concentration inequality (A.1)

$$1 - Pr\left(\left|\frac{\mathbf{V}_{K,N}(X)}{V_{K,N}(X)} - 1\right| \leq p(K, N)\right) = O(\mathcal{C}(K)) = O(\mathcal{C}(k)), \tag{B.2}$$

where $\mathcal{C}(K) = O(\mathcal{C}(k))$ because by $(\mathcal{A}.0)$, $K = \theta(k)$. This implies that, with high probability, the radius of the $K$-NN ball at $X$ concentrates around $(V_{K,N}(X)/c_d)^{1/d}$. By this concentration inequality (B.2), this choice of $K$ guarantees that the size of the $k$-NN ball in the partitioned sample is the same as the the size of the $K$-NN ball in the pooled sample with high probability $1 - \mathcal{C}(k)$. By the union bound and (B.2), the probability that

$$\left| \frac{\mathbf{V}_{K,N}(X)}{V_{K,N}(X)} - 1 \right| \leq p(K, N)$$

is satisfied by every $X_i \in \mathcal{X}_N$ is lower bounded by $1 - O(N\mathcal{C}(k))$.

Using the $K$-NN graph, for each sample $\mathbf{X} \in \mathcal{X}_\mathcal{N}$, we compute the number of points in $\mathcal{X}_\mathcal{N}$ that have $\mathbf{X}$ as a $l$-th nearest neighbor ($l$-NN), $l = \{1, \ldots, K\}$. Denote this count as $count(\mathbf{X})$. Let $Y$ be the $l$-nearest neighbor of $X$, $l = \{1, \ldots, K\}$. Then $Y$ can be represented as $Y = X + R_K(X)u$ where $u$ is an arbitrary vector with $||u|| \leq 1$.

For $X$ to be one of the $K$-NN of $Y$ it is necessary that $R_K(Y) \geq ||Y - X||$ or equivalently, $R_K(Y)/R_K(X) \geq ||u||$. Using the concentration inequality (B.2) for $R_K(X)$ and $R_K(Y)$, a sufficient condition for this is

$$\frac{\alpha_K(X) f(X)}{\alpha_K(Y) f(Y)} (1 - 2p(K, N)) \geq ||u||. \tag{B.3}$$

Because $f$ is differentiable and has a finite support, $f$ is Lipschitz continuous. Denote the Lipschitz constant by $\mathbb{L}$. Then, we have $|f(Y) - f(X)| \leq \mathbb{L}(K/c_d N\epsilon_0)^{1/d}$. Define $q(K, N) = (\mathbb{L}/\epsilon_0)(K/c_d N\epsilon_0)^{1/d} + 2\sqrt{6}/k^{\delta/2}$. Then (B.3) is satisfied if

$$\frac{\alpha_K(X)}{\alpha_K(Y)} (1 - q(K, N)) \geq ||u||.$$

For points $X \in \mathcal{S}'$, $\alpha_K(X) = 1$ with probability $1 - \mathcal{C}(k)$. This implies that $X$ will be one of the $K$-NN of $Y$ if $||u|| \leq 1 - q(K, N)$. This implies that, with probability $1 - O(N\mathcal{C}(k))$, $count(\mathbf{X}) \geq K(1 - q(K, N))$ whenever $X \in \mathcal{S}'$. On the other hand, for $X \in \mathcal{S} - \mathcal{S}'$, $\alpha_K(X) < 1$ with probability $1 - \mathcal{C}(k)$. It is also clear that for small values of $K/N$, $\alpha_K(X) < \alpha_K(Y)$ for at least $K/2$ $l$-NN $Y$ of $X$. This then implies that $count(\mathbf{X}) < K(1 - q(K, N))$ for $X \in \mathcal{S} - \mathcal{S}'$ with probability $1 - O(N\mathcal{C}(k))$. We therefore can apply the threshold $K(1 - q(K, N))$ to detect interior points $\mathcal{I}_\mathcal{N} = \mathcal{X}_\mathcal{N} \cap \mathcal{S}'$ and boundary points $\mathcal{B}_N = \mathcal{X}_\mathcal{N} - \mathcal{I}_\mathcal{N} = \mathcal{X}_\mathcal{N} \cap (\mathcal{S} - \mathcal{S}')$ with high probability $1 - O(N\mathcal{C}(k))$. Algorithm 1, shown below, codifies this into a precise procedure.

### *C. Boundary corrected density estimator*

Here the boundary corrected $k$-NN density estimator is defined and its asymptotic rates are computed. The proposed density estimator corrects the $k$-NN ball volumes for points that are close to the boundary. To estimate the density at a boundary point $\mathbf{X} \in \mathcal{B}_\mathcal{N}$, we find a point $\mathbf{Y} \in \mathcal{I}_\mathcal{N}$ that is close to $\mathbf{X}$. Because of the proximity of $\mathbf{X}$ and $\mathbf{Y}$, $f(\mathbf{X}) \approx f(\mathbf{Y})$. We can then estimate the density at $\mathbf{Y}$ instead and use this as an estimate of $f(\mathbf{Y})$. This informal argument is made more precise in what follows.

Consider the corrected density estimator $\tilde{\mathbf{f}}_k$ defined in (II.3). This estimator has bias of order $O((k/M)^{1/d})$, which can be shown as follows. Let $\mathbf{X}$ denote $\mathbf{X}_i$ for some fixed $i \in \{1, .., N\}$. Also, let $\mathbf{X}_{-1} = \arg\min_{x \in \mathcal{S}'} d(x, \mathbf{X})$.

**Algorithm 1** Detect boundary points $\mathcal{B}_{\mathcal{N}}$

1. Construct $K$-NN tree on $\mathcal{X}_{\mathcal{N}}$
2. Compute $count(\mathbf{X})$ for each $\mathbf{X} \in \mathcal{X}_{\mathcal{N}}$
3. Detect boundary points $\mathcal{B}_{\mathcal{N}}$:

**for** each $\mathbf{X} \in \mathcal{X}_{\mathcal{N}}$ **do**
  **if** $count(\mathbf{X}) < (1 - q(K, N))K$ **then**
    $\mathcal{B}_N \leftarrow \mathbf{X}$
  **else**
    $\mathcal{I}_N \leftarrow \mathbf{X}$
  **end if**
**end for**

Given $\mathcal{X}_N$, if $X \in \mathcal{I}_N$, then by (A.24),

$$\mathbb{E}[\tilde{\mathbf{f}}_k(X)] \quad = \quad \mathbb{E}[\hat{\mathbf{f}}_k(X)] = f(X) + O((k/M)^{2/d}) + O(\mathcal{C}(k)).$$

Next consider the alternative case $X \in \mathcal{B}_N$. Let $X_n \in \mathcal{I}_N$ be the closest interior point to $X$. Define $h = X - X_n$. $h$ can be rewritten as $h = h_1 + h_2$, where $h_1 = X - X_{-1}$ and $h_2 = X_{-1} - X_n$. Since $X \in \mathcal{B}_N$ implies that $X \in \mathcal{S} - \mathcal{S}'$ with probability $1 - O(N\mathcal{C}(k))$, consequently $||h_1|| = ||X - X_{-1}|| = O((k/M)^{1/d})$ with probability $1 - O(N\mathcal{C}(k))$. Again with probability $1 - O(N\mathcal{C}(k))$, $X_n \in \mathcal{S}'$ and consequently $||h_2|| = ||X_{-1} - X_n|| = o((k/M)^{1/d})$ [31]. This implies that $||h|| = O((k/M)^{1/d})$. Now,

$$f(X) = f(X_n) + O(||h||).$$

If $X_n$ is located in the interior $\mathcal{S}'$, by (A.24),

$$\mathbb{E}[\hat{\mathbf{f}}_k(X_n)] \quad = \quad f(X_n) + O((k/M)^{2/d}) + O(\mathcal{C}(k)), \tag{B.4}$$

and therefore

$$\begin{aligned}
\mathbb{E}[\tilde{\mathbf{f}}_k(X)] &= \mathbb{E}[\hat{f}_k(\mathbf{X}_n)] + O(N\mathcal{C}(k)) \\
&= f(X_n) + O((k/M)^{2/d}) + O(N\mathcal{C}(k)) \\
&= f(X) + O(||h||) + O((k/M)^{2/d}) + O(N\mathcal{C}(k)) \\
&= f(X) + O((k/M)^{1/d}) + O(N\mathcal{C}(k)),
\end{aligned} \tag{B.5}$$

where the $O(N\mathcal{C}(k))$ accounts for error in the case of the event that $X_{n(i)} \notin \mathcal{S}'$. This implies that the corrected density estimate has lower bias as compared to the standard $k$-NN density estimate (compare to (A.24) and (B.1)). In particular, boundary compensation has reduced the bias of the estimator at points near the boundary from $O(1)$ to $O((k/M)^{1/d}) + O(N\mathcal{C}(k))$.

*D. Properties of boundary corrected density estimator*

By section B-B, $\mathcal{I}_{\mathcal{N}} \in \mathcal{S}'$ with probability $1 - N\mathcal{C}(k)$. The results on bias, variance and cross-moments of the standard $k$-NN density estimator $\hat{\mathbf{f}}_k$ derived in the previous Appendix for points $X \in \mathcal{S}'$ therefore carry over to the corrected density estimator $\tilde{\mathbf{f}}_k$ for points $\mathcal{I}_{\mathcal{N}}$ with error of order $O(N\mathcal{C}(k))$.

In the definition of the corrected estimator $\tilde{\mathbf{f}}_k$ in (II.3), $\hat{\mathbf{f}}_k(\mathbf{X}_{n(i)})$ is the standard $k$-NN density estimates and $\mathbf{X}_{n(i)} \in \mathcal{S}'$ . It therefore follows that the variance and other central and cross moments of the corrected density estimator $\tilde{\mathbf{f}}_k$ will continue to decay at the same rate as the standard $k$-NN density estimator in the interior, as given by (A.59) and (A.60).

Given these identical rates and that the probability of a point being in the boundary region $\mathcal{S} - \mathcal{S}'$ is $O((k/M)^{1/d}) = o(1)$, the contribution of the boundary region to the overall variance and other cross moments of the boundary corrected density estimator $\tilde{\mathbf{f}}_k$ are asymptotically negligible compared to the contribution from the interior. As a result we can now generalize the results from Appendix A on the central moments and cross moments to include the boundary regions as follows. Denote $\tilde{\mathbf{f}}_k(X) - \mathbb{E}_X[\tilde{\mathbf{f}}_k(X) \mid X]$ by $\mathbf{e}(X)$.

*1) Central and cross moments:* For positive integers $q, r < k$

$$\mathbb{E}[\gamma(\mathbf{X})\mathbf{e}^q(\mathbf{X})] = 1_{\{q=2\}}\mathbb{E}\big[\gamma(\mathbf{X})f^2(\mathbf{X})\big]\left(\frac{1}{k}\right) + o\left(\frac{1}{k}\right) + O(N\mathcal{C}(k)), \tag{B.6}$$

$$\begin{aligned} &Cov[\gamma_1(\mathbf{X})\mathbf{e}^q(\mathbf{X}), \gamma_2(\mathbf{Y})\mathbf{e}^r(\mathbf{Y})] \\ &= 1_{\{q,r=1\}}Cov[1_{\{\mathbf{X}\in\mathcal{S}'\}}\gamma_1(\mathbf{X})f(\mathbf{X}), 1_{\{\mathbf{Y}\in\mathcal{S}'\}}\gamma_2(\mathbf{Y})f(\mathbf{Y})]\left(\frac{1}{M} + o(1/M)\right) \\ &+ 1_{\{q+r>2\}}\left(O\left(\frac{1}{k^{((q+r)\delta/2-1)}M}\right) + O(k_M^{2/d}/M) + O(1/M^2)\right) + O(N\mathcal{C}(k)). \end{aligned} \tag{B.7}$$

Next, we derive the following result on the bias of boundary corrected estimators.

*2) Bias:* For $k > 2$,

$$\begin{aligned} &\mathbb{E}[\gamma(\mathbb{E}[\tilde{\mathbf{f}}_k(\mathbf{X}) \mid \mathbf{X}]) - \gamma(f(\mathbf{X}))] = \mathbb{E}\left[\mathbb{E}\left[(\gamma(\tilde{\mathbf{f}}_k(\mathbf{X})) - \gamma(f(\mathbf{X}))) \mid \mathcal{X}_N\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[1_{\{X\in\mathcal{I}_N\}}(\gamma(\mathbb{E}[\tilde{\mathbf{f}}_k(X)]) - \gamma(f(X))) \mid \mathcal{X}_N\right]\right] + \mathbb{E}\left[\mathbb{E}\left[1_{\{X\in\mathcal{B}_N\}}(\gamma(\mathbb{E}[\tilde{\mathbf{f}}_k(X)]) - \gamma(f(X))) \mid \mathcal{X}_N\right]\right] \\ &= I + II. \end{aligned} \tag{B.8}$$

From (A.24), and $Pr(\mathbf{X} \in \mathcal{B}_N) = O((k/M)^{1/d})$, we have

$$I = \mathbb{E}\left[\gamma'(f(\mathbf{X}))h(\mathbf{X})\right]\left(\frac{k}{M}\right)^{2/d} + o\left(\frac{k}{M}\right)^{2/d} + O(N\mathcal{C}(k)). \tag{B.9}$$

Next, we will now derive $II$.

$$\begin{aligned} II &= \mathbb{E}\left[\mathbb{E}\left[1_{\{X\in\mathcal{B}_N\}}(\gamma(\mathbb{E}[\tilde{\mathbf{f}}_k(X)]) - \gamma(f(X))) \mid \mathcal{X}_N\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[1_{\{X\in\mathcal{B}_N\}}(\gamma(f(X_n)) - \gamma(f(X))) + O\left(\frac{k}{M}\right)^{2/d} \mid \mathcal{X}_N\right]\right] + O(N\mathcal{C}(k)), \end{aligned} \tag{B.10}$$

where the last step follows by (B.4). Let us concentrate on the inner expectation now. By section B-B, we know that with probability $1-O(N\mathcal{C}(k))$, if $X \in \mathcal{B}_N$, then $X \in \mathcal{S}-\mathcal{S}'$ and if $X_n \in \mathcal{I}_N$, then $X_n \in \mathcal{S}'$. Furthermore, $||X-X_{-1}|| = O(k/M)^{1/d}$ and $||X_{-1}-X_n|| = o(k/M)^{1/d}$ with probability $1-O(N\mathcal{C}(k))$. This implies that

$$\begin{aligned}&\mathbb{E}\left[1_{\{X\in\mathcal{B}_N\}}(\gamma(f(X_n))-\gamma(f(X)))+O\left(\frac{k}{M}\right)^{2/d} \mid \mathcal{X}_N\right]\\ &= \mathbb{E}\left[1_{\{X\in\mathcal{S}-\mathcal{S}'\}}(\gamma(f(X_{-1}))-\gamma(f(X))) \mid \mathcal{X}_N\right]+o\left(\frac{k}{M}\right)^{1/d}+O(N\mathcal{C}(k)).\end{aligned}$$

Since $Pr(\mathbf{X}\in\mathcal{S}-\mathcal{S}') = O((k/M)^{1/d})$, this in turn implies that

$$\begin{aligned}II &= \mathbb{E}\left[\mathbb{E}\left[1_{\{X\in\mathcal{B}_N\}}(\gamma(\mathbb{E}[\tilde{\mathbf{f}}_k(X)])-\gamma(f(X))) \mid \mathcal{X}_N\right]\right]\\ &= \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}-\mathcal{S}'\}}(\gamma(f(\mathbf{X}_{-1}))-\gamma(f(\mathbf{X})))]+o\left(\frac{k}{M}\right)^{2/d}+O(N\mathcal{C}(k)). \qquad \text{(B.11)}\end{aligned}$$

We therefore finally get,

$$\begin{aligned}&\mathbb{E}[\gamma(\mathbb{E}[\tilde{\mathbf{f}}_k(\mathbf{X}) \mid \mathbf{X}])-\gamma(f(\mathbf{X})))] = I+II\\ &= \mathbb{E}\left[\gamma'(f(\mathbf{X}))h(\mathbf{X})\right]\left(\frac{k}{M}\right)^{2/d}+\mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}-\mathcal{S}'\}}(\gamma(f(\mathbf{X}_{-1}))-\gamma(f(\mathbf{X})))]+o\left(\frac{k}{M}\right)^{2/d}+O(N\mathcal{C}(k)) \qquad \text{(B.12)}\end{aligned}$$

Note that $||\mathbf{X}-\mathbf{X}_{-1}|| = O((k/M)^{1/d})$ with probability $1-O(N\mathcal{C}(k))$. This therefore implies that

$$c_3 = \mathbb{E}[1_{\{\mathbf{X}\in\mathcal{S}-\mathcal{S}'\}}(\gamma(f(\mathbf{X}_{-1}))-\gamma(f(\mathbf{X})))] = O((k/M)^{1/d})\times O((k/M)^{1/d})+O(N\mathcal{C}(k)) = O((k/M)^{2/d})+O(N\mathcal{C}(k)).$$

*3) Optimality of boundary correction:* Comparing (B.12), (B.6) and (B.7) with (A.24), (A.59) and (A.60) respectively, oracle rates of convergence of bias, and central and cross moments for the boundary corrected density estimate are attained. The oracle rates are defined as the rates of MSE convergence attainable by the *oracle* density estimate that knows the boundary of $\mathcal{S}$

$$\tilde{\mathbf{f}}_{k,o} = \frac{k-1}{M\mathbf{V}_{k,o}(X)},$$

where $\mathbf{V}_{k,o}(X)$ is the volume of the region $\mathbf{S}_k(X)\cap\mathcal{S}$. It follows that the boundary compensated BPI estimator is adaptive in the sense that it's asymptotic MSE rate of convergence is identical to that of a $k$-NN plug-in estimator that knows the true boundary.

## APPENDIX C
## PROOF FOR BIAS AND VARIANCE OF PLUG-IN ESTIMATORS

**Lemma C.1.** *Assume that $U(x,y)$ is any arbitrary functional which satisfies*

$$\begin{aligned}(i)\ &\sup_{x\in(\epsilon_0,\epsilon_1)}|U(x,y)| = G_0 < \infty,\\ (ii)\ &\sup_{x\in(q_l,q_u)}|U(x,y)|\mathcal{C}(k) = G_1 < \infty,\\ (iii)\mathbb{E}[&\sup_{x\in(p_l,p_u)}|U(x/\mathbf{p},y)|] = G_2 < \infty.\end{aligned}$$

*Let* $\mathbf{Z}$ *denote* $\mathbf{X}_i$ *for some fixed* $i \in \{1,..,N\}$. *Let* $\zeta_{\mathbf{Z}}$ *be any random variable which almost surely lies in the range* $(f(\mathbf{Z}), \tilde{\mathbf{f}}_k(\mathbf{Z}))$. *Then,*

$$\mathbb{E}[|U(\zeta_{\mathbf{Z}}, \mathbf{Z})|] < \infty.$$

*Proof:*

We will show that the conditional expectation $\mathbb{E}[|U(\zeta_Z, Z)| \mid \mathcal{X}_N] < \infty$. Because $0 < \epsilon_0 < f(X) < \epsilon_\infty < \infty$ by $(\mathcal{A}.1)$, it immediately follows that

$$\mathbb{E}[|U(\zeta_{\mathbf{Z}}, \mathbf{Z})|] = \mathbb{E}[\mathbb{E}[|U(\zeta_Z, Z)| \mid \mathcal{X}_N]] < \infty.$$

For fixed $\mathcal{X}_N$, $Z \in \mathcal{I}_N$ or $Z \in \mathcal{B}_N$. These two cases are handled seperately.

*Case 1:* $Z \in \mathcal{I}_N$*:* In this case, $\tilde{\mathbf{f}}_k(Z) = \hat{\mathbf{f}}_k(Z)$. By (A.19) and $(\mathcal{A}.1)$, we know that if $\natural(Z)$ holds, $p_l/\mathbf{P}(Z) < \hat{\mathbf{f}}_k(Z) < p_u/\mathbf{P}(Z)$. On the other hand, if $\natural^c(Z)$ holds, by (A.21) and $(\mathcal{A}.1)$, $q_l < \hat{\mathbf{f}}_k(Z) < q_u$. This therefore implies that if $\natural(Z)$ holds, $\min\{\epsilon_0, p_l/\mathbf{P}(Z)\} < \zeta_Z < \max\{\epsilon_\infty, p_u/\mathbf{P}(Z)\}$ and if $\natural^c(Z)$ holds, $\min\{\epsilon_0, q_l\} < \zeta_Z < \max\{\epsilon_\infty, q_u\}$. Then,

$$\begin{aligned}
\mathbb{E}[|U(\zeta_Z, Z)| \mid \mathcal{X}_N] &= \mathbb{E}[1_{\natural(Z)}|U(\zeta_Z, Z)| \mid \mathcal{X}_N] + \mathbb{E}[1_{\natural^c(Z)}|U(\zeta_Z, Z)| \mid \mathcal{X}_N] \\
&\leq G_0 + \mathbb{E}[1_{\natural(Z)} \sup_{x \in (p_l, p_u)} |U(x/\mathbf{P}(Z), Z)|] + \max\{G_0, G_1/\mathcal{C}(k)\}(1 - Pr(\natural(Z))) \\
&\leq G_0 + \mathbb{E}[\sup_{x \in (p_l, p_u)} |U(x/\mathbf{P}(Z), Z)|] + \max\{G_0, G_1/\mathcal{C}(k)\}(1 - Pr(\natural(Z))) \\
&= G_0 + G_2 + \max\{G_1/\mathcal{C}(M), G_0\}\mathcal{C}(k) \\
&= G_0 + G_2 + \max\{G_1, G_0\mathcal{C}(k)\} < \infty \qquad \text{(C.1)}
\end{aligned}$$

where the final step follows from the fact that $\mathcal{C}(k) = o(1)$.

*Case 2:* $Z \in \mathcal{B}_N$*:* If $Z \in \mathcal{B}_N$, let $Y_n$ be the nearest neighbor of $Z$ in the set $\mathcal{I}_N$. Then,

$$\tilde{\mathbf{f}}_k(Z) = \hat{\mathbf{f}}_k(Y_n) \qquad \text{(C.2)}$$

This implies that we can now condition on the event $\natural(Y_n)$, and follow the exact procedure as in case 1 to obtain

$$\begin{aligned}
\mathbb{E}[|U(\zeta_Z, Z)| \mid \mathcal{X}_N] &= \mathbb{E}[1_{\natural(Y_n)}|U(\zeta_Z, Z)| \mid \mathcal{X}_N] + \mathbb{E}[1_{\natural^c(Y_n)}|U(1/\zeta_Z, Z)| \mid \mathcal{X}_N] \\
&\leq G_0 + G_2 + \max\{G_1, G_0\mathcal{C}(k)\} < \infty \qquad \text{(C.3)}
\end{aligned}$$

where the final step follows from the fact that $\mathcal{C}(k) = o(1)$. This concludes the proof.

■

**Proof of Theorem III.1**.

*Proof:*

Using the continuity of $g'''(x,y)$, construct the following third order Taylor series of $g(\tilde{\mathbf{f}}_k(\mathbf{Z}),\mathbf{Z})$ around the conditional expected value $\mathbb{E}_Z[\tilde{\mathbf{f}}_k(\mathbf{Z})] = \mathbb{E}[\tilde{\mathbf{f}}_k(\mathbf{Z}) \mid \mathbf{Z}]$.

$$g(\tilde{\mathbf{f}}_k(\mathbf{Z}),\mathbf{Z}) = g(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z}) + g'(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z})\mathbf{e}(\mathbf{Z}) + \frac{1}{2}g''(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z})\mathbf{e}^2(\mathbf{Z}) + \frac{1}{6}g^{(3)}(\zeta_{\mathbf{Z}},\mathbf{Z})\mathbf{e}^3(\mathbf{Z}),$$

where $\zeta_{\mathbf{Z}} \in (\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})], \tilde{\mathbf{f}}_k(\mathbf{Z}))$ is defined by the mean value theorem. This gives

$$\begin{aligned}&\mathbb{E}[(g(\tilde{\mathbf{f}}_k(\mathbf{Z}),\mathbf{Z}) - g(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z}))]\\ &= \mathbb{E}\left[\frac{1}{2}g''(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z})\mathbf{e}^2(\mathbf{Z})\right] + \mathbb{E}\left[\frac{1}{6}g^{(3)}(\zeta_{\mathbf{Z}},\mathbf{Z})\mathbf{e}^3(\mathbf{Z})\right]\end{aligned}$$

Let $\Delta(\mathbf{Z}) = \frac{1}{6}g^{(3)}(\zeta_{\mathbf{Z}},\mathbf{Z})$. Direct application of Lemma C.1 in conjunction with assumptions $(\mathcal{A}.5)$ , $(\mathcal{A}.6)$ implies that $\mathbb{E}[\Delta^2(\mathbf{Z})] = O(1)$. By Cauchy-Schwarz and assumption $(\mathcal{A}.4)$ applied to (B.6) for the choice $q = 6$,

$$\left|\mathbb{E}\left[\frac{1}{6}\Delta(\mathbf{Z})\mathbf{e}^3(\mathbf{Z})\right]\right| \leq \sqrt{\mathbb{E}\left[\frac{1}{36}\Delta^2(\mathbf{Z})\right]\mathbb{E}\left[\mathbf{e}^6(\mathbf{Z})\right]} = o\left(\frac{1}{k}\right) + O(N\mathcal{C}(k)).$$

By observing that the density estimates $\{\tilde{\mathbf{f}}_k(\mathbf{X}_i)\}, i = 1,\ldots,N$ are identical, we therefore have

$$\begin{aligned}&\mathbb{E}[\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)] - G(f) = \mathbb{E}[g(\tilde{\mathbf{f}}_k(\mathbf{Z}),\mathbf{Z}) - g(f(\mathbf{Z}),\mathbf{Z})]\\ &= \mathbb{E}[g(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z}) - g(f(\mathbf{Z}),\mathbf{Z})] + \mathbb{E}\left[\frac{1}{2}g''(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z})\mathbf{e}^2(\mathbf{Z})\right] + o(1/k) + O(N\mathcal{C}(k)).\end{aligned}$$

By (B.12) and (B.6) for the choice $q = 2$, in conjunction with assumption $(\mathcal{A}.4)$,this implies that

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)] - G(f) &= \mathbb{E}[g'(f(\mathbf{Z}),\mathbf{Z})h(\mathbf{Z})]\left(\frac{k}{M}\right)^{2/d} + \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}-\mathcal{S}_I\}}(g(f(\mathbf{Z}_{-1}),\mathbf{Z}_{-1}) - g(f(\mathbf{Z}),\mathbf{Z}))]\\ &\quad + \mathbb{E}[f^2(\mathbf{Z})g''(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z})/2]\left(\frac{1}{k}\right) + O(N\mathcal{C}(k)) + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{2/d}\right)\\ &= \mathbb{E}[g'(f(\mathbf{Z}),\mathbf{Z})h(\mathbf{Z})]\left(\frac{k}{M}\right)^{2/d} + \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}-\mathcal{S}_I\}}(g(f(\mathbf{Z}_{-1}),\mathbf{Z}_{-1}) - g(f(\mathbf{Z}),\mathbf{Z}))]\\ &\quad + \mathbb{E}[f^2(\mathbf{Z})g''(f(\mathbf{Z}),\mathbf{Z})/2]\left(\frac{1}{k}\right) + O(N\mathcal{C}(k)) + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{2/d}\right)\\ &= c_1\left(\frac{k}{M}\right)^{2/d} + c_2\left(\frac{1}{k}\right) + c_3 + O(N\mathcal{C}(k)) + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{2/d}\right),\end{aligned}$$

where the last but one step follows because, by (A.24) and (B.5), we know $\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})] = f(\mathbf{Z}) + o(1)$. This in turn implies $\mathbb{E}[f^2(\mathbf{Z})g''(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z})/2] = \mathbb{E}[f^2(\mathbf{Y})g''(f(\mathbf{Y}),\mathbf{Y})/2]$. Finally, by assumption $(\mathcal{A}.5)$ and $(\mathcal{A}.2)$, the leading constants $c_1$ and $c_2$ are bounded. We have also shown in equation (B.11) that $c_3 = O((k/M)^{2/d})$. This concludes the proof.

■

**Proof of Theorem IV.1**

*Proof:* Let $\mathbf{X}$ denote $\mathbf{X}_i$ for some fixed $i \in \{1,..,N\}$. Also, let $\mathbf{X}_{-1} = \arg\min_{x \in \mathcal{S}_I} d(x, \mathbf{X})$. Using (A.25), we can derive the following in an identical manner to (B.12):

$$\begin{aligned}
\mathbb{B}(\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)) &= \mathbb{E}[\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)] - \int g(f(x),x)f(x)dx \\
&= (\mathbb{E}[g(\tilde{\mathbf{f}}_k(\mathbf{Z}),\mathbf{Z})] - g_2(k,M))/g_1(k,M) - \int g(f(x),x)f(x)dx \\
&= \mathbb{E}[\mathbb{E}[(g(\tilde{\mathbf{f}}_k(\mathbf{Z}),\mathbf{X}) - g_2(k,M))/g_1(k,M) \mid \mathcal{X}_N]] - \int g(f(x),x)f(x)dx \\
&= \mathbb{E}[\mathbb{E}[(g(\tilde{\mathbf{f}}_k(\mathbf{X}),\mathbf{X}) - g_2(k,M))/g_1(k,M) \mid \mathcal{X}_N], X \in \mathcal{I}_N] \\
&\quad + \mathbb{E}[\mathbb{E}[(g(\tilde{\mathbf{f}}_k(\mathbf{X}),\mathbf{X}) - g_2(k,M))/g_1(k,M) \mid \mathcal{X}_N], X \in \mathcal{B}_N] \\
&\quad - \int g(f(x),x)f(x)dx \\
&= \mathbb{E}[g(f(\mathbf{X}),\mathbf{X}) + \frac{g'(f(\mathbf{X}),\mathbf{X})h(\mathbf{X})}{g_1(k,M)}(k/M)^{2/d} \\
&\quad + \frac{1_{\{\mathbf{X} \in \mathcal{S} - \mathcal{S}'\}}}{g_1(k,M)}(g(f(\mathbf{X}_{-1}),\mathbf{X}_{-1}) - g(f(\mathbf{X}),\mathbf{X})) \\
&\quad + o((k/M)^{2/d}) + O(N\mathcal{C}(k))] - \int g(f(x),x)f(x)dx \\
&= \frac{c_1}{g_1(k,M)}\left(\frac{k}{M}\right)^{2/d} + \frac{c_3}{g_1(k,M)} + o\left(\left(\frac{k}{M}\right)^{2/d}\right) + O(N\mathcal{C}(k)).
\end{aligned}$$

Because we assume the logarithmic growth condition $k = O((\log(M))^{2/(1-\delta)})$, it follows that $O(N\mathcal{C}(k)) = O(N/M^3) = o(1/T)$. Also, by (IV.3), $g_1(k,M) = 1 + o(1)$. This implies that

$$\mathbb{B}(\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)) = c_1\left(\frac{k}{M}\right)^{2/d} + c_3 + o\left(\left(\frac{k}{M}\right)^{2/d}\right). \tag{C.4}$$

■

**Proof of Theorem III.2 and Theorem IV.2**.

*Proof:* By the continuity of $g^{(\lambda)}(x,y)$, we can construct the following Taylor series of $g(\tilde{\mathbf{f}}_k(\mathbf{Z}),\mathbf{Z})$ around the conditional expected value $\mathbb{E}_Z[\tilde{\mathbf{f}}_k(\mathbf{Z})]$.

$$\begin{aligned}
g(\tilde{\mathbf{f}}_k(\mathbf{Z}),\mathbf{Z}) &= g(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z}) + g'(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z})\mathbf{e}(\mathbf{Z}) \\
&+ \left(\sum_{i=2}^{\lambda-1}\frac{g^{(i)}(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})],\mathbf{Z})}{i!}\mathbf{e}^i(\mathbf{Z})\right) + \frac{g^{(\lambda)}(\xi_{\mathbf{Z}},\mathbf{Z})}{\lambda!}\mathbf{e}^\lambda(\mathbf{Z}),
\end{aligned}$$

where $\xi_{\mathbf{Z}} \in (g(\mathbb{E}_Z[\tilde{\mathbf{f}}_k(\mathbf{Z})], g(\tilde{\mathbf{f}}_k(\mathbf{Z}))))$. Denote $(g^\lambda(\xi_{\mathbf{Z}},\mathbf{Z}))/\lambda!$ by $\Psi(\mathbf{Z})$. Further define the operator $\mathcal{M}(\mathbf{Z}) = \mathbf{Z} - \mathbb{E}[\mathbf{Z}]$ and

$$\begin{aligned}
p_i &= \mathcal{M}(g(\mathbb{E}_{\mathbf{X}_i}[\tilde{\mathbf{f}}_k(\mathbf{X}_i)],\mathbf{X_i})), \\
q_i &= \mathcal{M}(g'(\mathbb{E}_{\mathbf{X}_i}[\tilde{\mathbf{f}}_k(\mathbf{X}_i)],\mathbf{X_i})\mathbf{e}(\mathbf{X_i})), \\
r_i &= \mathcal{M}\left(\sum_{i=2}^{\lambda}\frac{g^{(i)}(\mathbb{E}_{\mathbf{X}_i}[\tilde{\mathbf{f}}_k(\mathbf{X}_i)],\mathbf{X_i})}{i!}\mathbf{e}^i(\mathbf{X_i})\right) \\
s_i &= \mathcal{M}\left(\Psi(\mathbf{X_i})\mathbf{e}^\lambda(\mathbf{X_i})\right)
\end{aligned}$$

The variance of the estimator $\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)$ is given by

$$\begin{aligned}\mathbb{V}[\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)] &= \mathbb{E}[(\hat{\mathbf{G}}(f) - \mathbb{E}[\hat{\mathbf{G}}(f)])^2] \\ &= \frac{1}{N}\mathbb{E}\big[(p_1+q_1+r_1+s_1)^2\big] \\ &+ \frac{N-1}{N}\mathbb{E}[(p_1+q_1+r_1+s_1)(p_2+q_2+r_2+s_2)].\end{aligned}$$

Because $\mathbf{X}_1$, $\mathbf{X}_2$ are independent, we have $\mathbb{E}[(p_1)(p_2+q_2+r_2+s_2)] = 0$. Furthermore,

$$\mathbb{E}\big[(p_1+q_1+r_1+s_1)^2\big] \quad = \quad \mathbb{E}[{p_1}^2] + o(1) = \mathbb{V}[g(\mathbb{E}_{\mathbf{Z}}[\hat{\mathbf{f}}(\mathbf{Z})], \mathbf{Z})] + o(1).$$

From assumption $(\mathcal{A}.4)$ applied to (B.6) and (B.7), in conjunction with assumption $(\mathcal{A}.3)$, it follows that

- $\mathbb{E}[{p_1}^2] = \mathbb{V}[g(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z})]$
- $\mathbb{E}[q_1 q_2] = \mathbb{V}[g'(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z}) f(\mathbf{Z})] \left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right) + O(N\mathcal{C}(k))$
- $\mathbb{E}[q_1 r_2] = \sum_{i=2}^{\lambda-1} O\left(\frac{1}{k^{((1+i)\delta/2-1)}M}\right) + O\left(\frac{\lambda(k_M^{2/d}+1/M)}{M}\right) + O(N\mathcal{C}(k)) = o\left(\frac{1}{M}\right) + O(N\mathcal{C}(k))$
- $\mathbb{E}[r_1 r_2] = \sum_{i_1=2}^{\lambda-1}\sum_{i_2=2}^{\lambda-1} O\left(\frac{1}{k^{((i_1+i_2)\delta/2-1)}M}\right) + O\left(\frac{\lambda^2(k_M^{2/d}+1/M)}{M}\right) + O(N\mathcal{C}(k)) = o\left(\frac{1}{M}\right) + O(N\mathcal{C}(k))$

Since $q_1$ and $s_2$ are $0$ mean random variables

$$\begin{aligned}\mathbb{E}[q_1 s_2] &= \mathbb{E}\left[q_1 \Psi(\mathbf{X_2})(\hat{\mathbf{f}}(\mathbf{X_2}) - \mathbb{E}_{\mathbf{X}_2}[\tilde{\mathbf{f}}_k(\mathbf{X}_2)])^{\lambda}\right] \\ &= \mathbb{E}\left[q_1 \Psi(\mathbf{X_2})(\hat{\mathbf{f}}(\mathbf{X_2}) - \mathbb{E}_{\mathbf{X}_2}[\tilde{\mathbf{f}}_k(\mathbf{X}_2)])^{\lambda}\right] \\ &\leq \sqrt{\mathbb{E}\left[\Psi^2(\mathbf{X_2})\right]\mathbb{E}\left[q_1^2(\hat{\mathbf{f}}(\mathbf{X_2}) - \mathbb{E}_{\mathbf{X}_2}[\tilde{\mathbf{f}}_k(\mathbf{X}_2)])^{2\lambda}\right]} \\ &= \sqrt{\mathbb{E}\left[\Psi^2(\mathbf{Z})\right]}\left(o\left(\frac{1}{k^{\lambda}}\right) + O(N\mathcal{C}(k))\right)\end{aligned}$$

Direct application of Lemma C.1 in conjunction with assumptions $(\mathcal{A}.5)$, $(\mathcal{A}.6)$ implies that $\mathbb{E}\left[\Psi^2(\mathbf{Z})\right] = O(1)$. Note that from assumption $(\mathcal{A}.3)$, $o\left(\frac{1}{k^{\lambda}}\right) = o(1/M)$ . In a similar manner, it can be shown that $\mathbb{E}[r_1 s_2] = o\left(\frac{1}{M}\right) + O(N\mathcal{C}(k))$ and $\mathbb{E}[s_1 s_2] = o\left(\frac{1}{M}\right) + O(N\mathcal{C}(k))$. Finally, by (A.24) and (B.5), we know $\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})] = \mathbb{E}[\tilde{\mathbf{f}}_k(\mathbf{Z})] = f(\mathbf{Z}) + o(1)$. This implies that

$$\begin{aligned}\mathbb{V}[\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)] &= \frac{1}{N}\mathbb{E}\big[{p_1}^2\big] + \frac{(N-1)}{N}\mathbb{E}[q_1 q_2] + O(N\mathcal{C}(k)) + o\left(\frac{1}{M}+\frac{1}{N}\right) \\ &= \mathbb{V}[g(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z})]\left(\frac{1}{N}\right) + \mathbb{V}[g'(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z}) f(\mathbf{Z})]\left(\frac{1}{M}\right) + O(N\mathcal{C}(k)) + o\left(\frac{1}{M}+\frac{1}{N}\right) \\ &= \mathbb{V}[g(f(\mathbf{Z}), \mathbf{Z})]\left(\frac{1}{N}\right) + \mathbb{V}[g'(f(\mathbf{Z}), \mathbf{Z}) f(\mathbf{Z})]\left(\frac{1}{M}\right) + O(N\mathcal{C}(k)) + o\left(\frac{1}{M}+\frac{1}{N}\right) \\ &= c_4\left(\frac{1}{N}\right) + c_5\left(\frac{1}{M}\right) + O(N\mathcal{C}(k)) + o\left(\frac{1}{M}+\frac{1}{N}\right),\end{aligned}$$

where the last but one step follows because, by (A.24) and (B.5), we know $\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})] = f(\mathbf{Z}) + o(1)$. This in turn implies $\mathbb{V}[g(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z})] = \mathbb{V}[g(f(\mathbf{Z}), \mathbf{Z})]$ and $\mathbb{V}[g'(\mathbb{E}_{\mathbf{Z}}[\tilde{\mathbf{f}}_k(\mathbf{Z})], \mathbf{Z}) f(\mathbf{Z})] = \mathbb{V}[g'(f(\mathbf{Z}), \mathbf{Z}) f(\mathbf{Z})]$. Finally, by assumptions $(\mathcal{A}.5)$ and $(\mathcal{A}.2)$, the leading constants $c_4$ and $c_5$ are bounded. This concludes the proof of Theorem III.2.

Under the logarithmic growth condition $k = O((\log(M))^{2/(1-\delta)})$, $g_2(k,M) = o(1)$ and $g_1(k,M) = 1 + o(1)$ by assumption (IV.3). Theorem IV.2 follows by observing that $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k) = (\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) - g_1(k,M))/g_2(k,M)$ ■

**Bias of Baryshnikov's estimator: Proof of equation (III.2)**

*Proof:* We will first prove that

$$\mathbb{B}(\tilde{\mathbf{G}}_N(\hat{\mathbf{f}}_k)) = \Theta((k/M)^{1/d} + 1/k), \tag{C.5}$$

Because the standard $k$-NN density estimate $\hat{\mathbf{f}}_{kS}(\mathbf{X}_i)$ is identical to the partitioned $k$-NN density estimate $\hat{\mathbf{f}}_k(\mathbf{X}_i)$ defined on the partition $\{\mathbf{X}_i\}$ and $\{\mathbf{X}_1, .., \mathbf{X}_T\} - \{\mathbf{X}_i\}$, it follows that

$$\mathbb{B}(\tilde{\mathbf{G}}_N(\hat{\mathbf{f}}_{kS})) = \Theta((k/T)^{1/d} + 1/k). \tag{C.6}$$

From the definition of set $\mathcal{S}'$ in section A-A2, we can choose the set $\mathcal{S}'$, such that $Pr(\mathbf{Z} \notin \mathcal{S}') = O((k/M)^{1/d})$.

$$\begin{aligned}
&\mathbb{E}[\hat{\mathbf{G}}_N(\hat{\mathbf{f}}_k)] - G(f) = \mathbb{E}[g(\hat{\mathbf{f}}_k(\mathbf{Z}), \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z})] \\
&= \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}'\}} g(\hat{\mathbf{f}}_k(\mathbf{Z}), \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z})] + \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}-\mathcal{S}'\}} g(\hat{\mathbf{f}}_k(\mathbf{Z}), \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z})] \\
&= I + II
\end{aligned} \tag{C.7}$$

Using the exact same method as in the Proof of Theorem III.1, using (A.24) and (A.59), and the fact that $Pr(\mathbf{Z} \notin \mathcal{S}') = O((k/M)^{1/d}) = o(1)$, we have

$$I = \mathbb{E}[g'(f(\mathbf{Z}), \mathbf{Z}) h(\mathbf{Z})] \left(\frac{k}{M}\right)^{2/d} + \mathbb{E}[f^2(\mathbf{Z}) g''(f(\mathbf{Z}), \mathbf{Z})/2] \left(\frac{1}{k}\right) + O(\mathcal{C}(k)) + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{2/d}\right),$$

Because we assume that $g$ satisfies assumption $(\mathcal{A}.6)$, from the proof of Lemma C.1, for $Z \in \mathcal{S} - \mathcal{S}'$, we have $\mathbb{E}[g(\hat{\mathbf{f}}_k(Z), Z) - g(f(Z), Z)] = O(1)$. This implies that,

$$\begin{aligned}
II &= \mathbb{E}[1_{\{\mathbf{Z}\in\mathcal{S}-\mathcal{S}'\}} g(\hat{\mathbf{f}}_k(\mathbf{Z}), \mathbf{Z}) - g(f(\mathbf{Z}), \mathbf{Z})] \\
&= \mathbb{E}\left[\mathbb{E}[g(\hat{\mathbf{f}}_k(Z), Z) - g(f(Z), Z)] \mid 1_{\{\mathbf{Z}\in\mathcal{S}-\mathcal{S}'\}}\right] \times Pr(\mathbf{Z} \notin \mathcal{S}') \\
&= O(1) \times O((k/M)^{1/d}) = O((k/M)^{1/d}).
\end{aligned} \tag{C.8}$$

This concludes the proof.

■

## APPENDIX D
## CLT FOR INTERCHANGEABLE PROCESSES

Define the random variables $\{\mathbf{Y}_{M,i}; i = 1, \ldots, N\}$ for any fixed $M$

$$\mathbf{Y}_{M,i} = \frac{g(\tilde{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i) - \mathbb{E}[g(\tilde{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i)]}{\sqrt{\mathbb{V}[g(\tilde{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i)]}},$$

and define the sum $\mathbf{S}_{\mathbf{N},\mathbf{M}}$

$$\mathbf{S}_{N,M} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{Y}_{M,i},$$

where the indices $N$ and $M$ explicitly stress the dependence of the sum $\mathbf{S}_{N,M}$ on the number of random variables $N+M$. Observe that the random variables $\{\mathbf{Y}_{M,i}; i = 1, \ldots, N\}$ belong to an $0$ mean, unit variance, interchangeable process [36] for all values of $M$. To establish the CLT for $\mathbf{S}_{N,M}$, we will exploit the fact the random variables $\{\mathbf{Y}_{M,i}; i = 1, \ldots, N\}$ are interchangeable by appealing to DeFinetti's theorem, which we describe below.

### *A. De Finetti's Theorem*

Let $\mathcal{F}$ be the class of one dimensional distribution functions and for each pair of real numbers $x$ and $y$ define $\mathcal{F}(x, y) = \{F \in \mathcal{F} | F(x) \leq y\}$. Let $\mathcal{B}$ be the Borel field of subsets of $\mathcal{F}$ generated by the class of sets $\mathcal{F}(x, y)$. Then De Finetti's theorem asserts that for any interchangeable process $\{\mathbf{Z}_i\}$ there exists a probability measure $\mu$ defined on $\mathcal{B}$ such that

$$Pr\{\mathbf{B}\} = \int_{\mathcal{F}} Pr_F\{\mathbf{B}\} d\mu(F), \tag{D.1}$$

for any Borel measurable set defined on the sample space of the sequence $\{\mathbf{Z}_i\}$. Here $Pr\{\mathbf{B}\}$ is the probability of the event $\mathbf{B}$ and $Pr_F\{\mathbf{B}\}$ is the probability of the event $B$ under the assumption that component random variables $\mathbf{X}_i$ of the interchangeable process are independent and identically distributed with distribution $F$.

### *B. Necessary and Sufficient conditions for CLT*

For each $F \in \mathcal{F}$ define $m(F)$ and $\sigma^2(F)$ as $m(F) = \int_{-\infty}^{\infty} x dF(x)$, $\sigma(F) = \int_{-\infty}^{\infty} x^2 dF(x) - 1$ and for all real numbers $m$ and non-negative real numbers $\sigma^2$ let $\mathcal{F}_{m,\sigma^2}$ be the set of $F \in \mathcal{F}$ for which $m(F) = m$ and $\sigma^2(F) = \sigma^2$.

Let $\{\mathbf{Z}_i; i = 1, 2, \ldots\}$ be an interchangeable stochastic process with $0$ mean and variance $1$. Blum *etal* [36] showed that the random variable $\mathbf{S}_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{Z}_i$ converges in distribution to $N(0, 1)$ if and only if $\mu(\mathcal{F}_{0,0}) = 1$. Furthermore, they show that the condition $\mu(\mathcal{F}_{0,0}) = 1$ is equivalent to the condition that $Cov(\mathbf{Z}_1, \mathbf{Z}_2) = 0$ and $Cov(\mathbf{Z}_1^2, \mathbf{Z}_2^2) = 0$. We will *e*xtend Blum *etal*'s results to interchangeable processes where $Cov(\mathbf{Z}_1, \mathbf{Z}_2) = o(1)$ and $Cov(\mathbf{Z}_1^2, \mathbf{Z}_2^2) = o(1)$.

In particular, we will show that $Cov(\mathbf{Y}_{M,1}, \mathbf{Y}_{M,2})$ and $Cov(\mathbf{Y}_{M,1}^2, \mathbf{Y}_{M,2}^2)$ are $O(1/M)$. Subsequently we will show that the random variable $\mathbf{S}_{N,M} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{Y}_{M,i}$ converges in distribution to $N(0, 1)$ and conclude that Theorem III.3 holds.

### C. CLT for Asymptotically Uncorrelated processes

Let $\mathbf{X}$ be a random variable with density $f$. In the proof of Theorem III.2, we showed that

$$\begin{aligned} Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}) &= \frac{Cov(g(\tilde{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i), g(\tilde{\mathbf{f}}_k(\mathbf{X}_j), \mathbf{X}_j))}{\sqrt{\mathbb{V}[g(\tilde{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i)]\mathbb{V}[g(\tilde{\mathbf{f}}_k(\mathbf{X}_j), \mathbf{X}_j)]}} \\ &= \frac{Cov(p_i + q_i + r_i + s_i, p_j + q_j + r_j + s_j)}{\sqrt{\mathbb{V}[g(\tilde{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i)]\mathbb{V}[g(\tilde{\mathbf{f}}_k(\mathbf{X}_j), \mathbf{X}_j)]}} \\ &= \frac{Cov(p_i + q_i + r_i + s_i, p_j + q_j + r_j + s_j)}{\sqrt{\mathbb{V}[g(\tilde{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i)]\mathbb{V}[g(\tilde{\mathbf{f}}_k(\mathbf{X}_j), \mathbf{X}_j)]}} \\ &= \frac{\mathbb{V}(g'(f(\mathbf{X}), \mathbf{X})f(\mathbf{X}))}{\mathbb{V}[g(f(\mathbf{X}_i), \mathbf{X}_i)]}\left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right) + O(N\mathcal{C}(k)) \\ &= \frac{\mathbb{V}(g'(f(\mathbf{X}), \mathbf{X})f(\mathbf{X}))}{\mathbb{V}[g(f(\mathbf{X}_i), \mathbf{X}_i)]}\left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right), \end{aligned} \tag{D.2}$$

where the last but one step follows by observing that $N\mathcal{C}(k)/M \to 0$ under the logarithmic growth condition $k = O((\log(M))^{2/(1-\delta)})$. Define the function $d(x, y) = g(x, y)(g(x, y)-c)$, where the constant $c = \mathbb{E}[g(\tilde{\mathbf{f}}_k(\mathbf{X}), \mathbf{X})]$. Then, similar to the derivation of (D.2), we have,

$$\begin{aligned} Cov(\mathbf{Y}^2_{M,i}, \mathbf{Y}^2_{M,j}) &= \frac{Cov(d(\tilde{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i), d(\tilde{\mathbf{f}}_k(\mathbf{X}_j), \mathbf{X}_j))}{\sqrt{\mathbb{V}[d(\tilde{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i)]\mathbb{V}[d(\tilde{\mathbf{f}}_k(\mathbf{X}_j), \mathbf{X}_j)]}} \\ &= \frac{\mathbb{V}(d'(f(\mathbf{X}), \mathbf{X})f(\mathbf{X}))}{\mathbb{V}[d(f(\mathbf{X}_i), \mathbf{X}_i)]}\left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right). \end{aligned} \tag{D.3}$$

**Proof of Theorem III.3 and Theorem IV.3**.

*Proof:*

Let $\delta_\mu(M)$ and $\delta_\sigma(M)$ be a strictly positive functions parameterized by $M$ such that $\delta_\mu(M) = o(1)$; $\frac{1}{M\delta_\mu(M)} = o(1)$, $\delta_\sigma(M) = o(1)$; $\frac{1}{M\delta_\sigma(M)} = o(1)$. Denote the set of $F \in \mathcal{F}$ with $\mathcal{F}_{m,\delta,M} := \{m^2(F) \geq \delta_\mu(M)\}$; $\mathcal{F}_{\sigma,\delta,M} := \{\sigma^2(F) \geq \delta_\sigma(M)\}$; $\mathcal{F}^*_{m,\delta,M} := \{m^2(F) \in (0, \delta_\mu(M))\}$ and $\mathcal{F}^*_{\sigma,\delta,M} := \{\sigma^2(F) \in (0, \delta_\sigma(M))\}$. Denote the measures of these sets by $\mu_{m,\delta,M}$, $\mu_{\sigma,\delta,M}$, $\mu^*_{m,\delta,M}$ and $\mu^*_{\sigma,\delta,M}$ respectively. We have from (D.1) that

$$\begin{aligned} \int_{\mathcal{F}} m^2(F)d\mu(F) &= Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}) \\ \int_{\mathcal{F}} \sigma^2(F)d\mu(F) &= \int_{\mathcal{F}} [\mathbb{E}_F[\mathbf{Z}^2 - 1]]^2 d\mu(F) = Cov(\mathbf{Y}^2_{M,i}, \mathbf{Y}^2_{M,j}). \end{aligned} \tag{D.4}$$

Applying the Chebyshev inequality, we get

$$\delta_\mu(M)\mu_{m,\delta,M} \leq Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}),$$

$$\delta_\sigma(M)\mu_{\sigma,\delta,M} \leq Cov(\mathbf{Y}^2_{M,i}, \mathbf{Y}^2_{M,j}).$$

Because the covariances decay at $O(1/M)$, $\mu_{m,\delta,M}$ and $\mu_{\sigma,\delta,M} \to 0$ as $M \to \infty$. From the definition of $\mathcal{F}^*_{m,\delta,M}$ and $\mathcal{F}^*_{\sigma,\delta,M}$, we also have that $\mu^*_{m,\delta,M}$ and $\mu^*_{\sigma,\delta,M} \to 0$ as $M \to \infty$. We also have

$$1 - (\mu_{m,\delta,M} + \mu_{\sigma,\delta,M} + \mu^*_{m,\delta,M} + \mu^*_{\sigma,\delta,M}) \leq \mu(\mathcal{F}_{0,0}) \leq 1,$$

and therefore

$$\lim_{M\to\infty} \mu(\mathcal{F}_{0,0}) = 1. \tag{D.5}$$

We will now show that $\tilde{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) = (\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) - \mathbb{E}[\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)])/(\sqrt{\mathbb{V}[\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)]})$ converges weakly to $\mathbb{N}(0,1)$. Denote $g(\tilde{\mathbf{f}}_k(\mathbf{X}_i), \mathbf{X}_i)$ by $\mathbf{g}_i$. Observe that

$$\begin{aligned}
&\lim_{\Delta\to 0} Pr\{\tilde{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) \le \alpha\} = \lim_{\Delta\to 0}\int_{\mathcal{F}} Pr_F\{\tilde{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) \le \alpha\} d\mu(F) \\
&= \lim_{\Delta\to 0}\int_{\mathcal{F}_{0,0}} Pr_F\{\tilde{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) \le \alpha\} d\mu(F) + \lim_{\Delta\to 0}\int_{\mathcal{F}} 1_{\{F\in\mathcal{F}-\mathcal{F}_{0,0}\}} Pr_F\{\tilde{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) \le \alpha\} d\mu(F) \\
&= \lim_{\Delta\to 0}\int_{\mathcal{F}_{0,0}} Pr_F\{\tilde{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) \le \alpha\} d\mu(F) + \int_{\mathcal{F}} \lim_{\Delta\to 0}\left(1_{\{F\in\mathcal{F}-\mathcal{F}_{0,0}\}} Pr_F\{\tilde{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) \le \alpha\}\right) d\mu(F) \qquad \text{(D.6)} \\
&= \lim_{\Delta\to 0}\int_{\mathcal{F}_{0,0}} Pr_F\{\tilde{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) \le \alpha\} d\mu(F) \qquad \text{(D.7)} \\
&= \lim_{\Delta\to 0}\int_{\mathcal{F}_{0,0}} Pr_F\left\{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{g(\tilde{\mathbf{f}}_k(\mathbf{X}_i),\mathbf{X}_i) - \mathbb{E}[g(\tilde{\mathbf{f}}_k(\mathbf{X}_i),\mathbf{X}_i)]}{\sqrt{\mathbb{V}[\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)]}}\right) \le \alpha\right\} d\mu(F) \\
&= \lim_{\Delta\to 0}\int_{\mathcal{F}_{0,0}} Pr_F\left\{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{g(\tilde{\mathbf{f}}_k(\mathbf{X}_i),\mathbf{X}_i) - \mathbb{E}[g(\tilde{\mathbf{f}}_k(\mathbf{X}_i),\mathbf{X}_i)]}{\sqrt{\mathbb{V}[\mathbf{g}_i]/N + ((N-1)/N)Cov[\mathbf{g}_i,\mathbf{g}_j]}}\right) \le \alpha\right\} \int_{\mathcal{F}_{0,0}} d\mu(F) \\
&= \lim_{\Delta\to 0}\int_{\mathcal{F}_{0,0}} Pr_F\left\{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{g(\tilde{\mathbf{f}}_k(\mathbf{X}_i),\mathbf{X}_i) - \mathbb{E}[g(\tilde{\mathbf{f}}_k(\mathbf{X}_i),\mathbf{X}_i)]}{\sqrt{\mathbb{V}[\mathbf{g}_i]/N + ((N-1)/N)\sqrt{\mathbb{V}[\mathbf{g}_i]\mathbb{V}[\mathbf{g}_j]}Cov[\mathbf{Y}_{M,i},\mathbf{Y}_{M,j}]}}\right) \le \alpha\right\} d\mu(F) \\
&= \lim_{\Delta\to 0}\int_{\mathcal{F}_{0,0}} Pr_F\left\{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{g(\tilde{\mathbf{f}}_k(\mathbf{X}_i),\mathbf{X}_i) - \mathbb{E}[g(\tilde{\mathbf{f}}_k(\mathbf{X}_i),\mathbf{X}_i)]}{\sqrt{\mathbb{V}[\mathbf{g}_i]/N}}\right) \le \alpha\right\} d\mu(F) \qquad \text{(D.8)} \\
&= \lim_{\Delta\to 0}\int_{\mathcal{F}_{0,0}} Pr_F\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{Y}_{M,i} \le \alpha\right\} d\mu(F) \\
&= \int_{\mathcal{F}} \lim_{\Delta\to 0}\left(1_{\{F\in\mathcal{F}_{0,0}\}} Pr_F\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{Y}_{M,i} \le \alpha\right\}\right) d\mu(F) \\
&= \int_{\mathcal{F}} \phi(\alpha) d\mu(F) = \phi(\alpha), \qquad \text{(D.9)}
\end{aligned}$$

where $\phi(.)$ is the distribution function of a Gaussian random variable with mean $0$ and variance $1$. Step (D.6) follows from the Dominated Convergence theorem. By (D.5), $\lim_{\Delta\to 0} 1_{\{F\in\mathcal{F}-\mathcal{F}_{0,0}\}} = 0$ almost surely. This gives Step (D.7). Step (D.8) is obtained by observing that, by (D.4), $Cov[\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}] = 0$ when $F \in \mathcal{F}_{0,0}$. The last step (D.9) follows from the CLT for sums of $0$ mean, unit variance, i.i.d random variables and (D.5). This concludes the proof of Theorem III.3.

To show Theorem IV.3, observe that under the logarithmic growth condition $k = O((\log(M))^{2/(1-\delta)})$, $g_2(k,M) = o(1)$ and $g_1(k,M) = 1 + o(1)$ by assumption (IV.3). Since $\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k) = (\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) - g_1(k,M))/g_2(k,M)$, it follows that the asymptotic distribution of

$$\frac{\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k) - \mathbb{E}[\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)]}{\sqrt{\mathbb{V}[\hat{\mathbf{G}}_{N,BC}(\tilde{\mathbf{f}}_k)]}}$$

is equal to the asymptotic distribution of $\tilde{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) = (\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k) - \mathbb{E}[\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)])/(\sqrt{\mathbb{V}[\hat{\mathbf{G}}_N(\tilde{\mathbf{f}}_k)]})$.

■

## ACKNOWLEDGMENT

We thank the reviewers for their helpful suggestions and comments. This work is partially funded by the Air Force Office of Scientific Research, grant number FA9550-09-1-0471.

## REFERENCES

[1] A. O. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *Signal Processing Magazine, IEEE*, vol. 19, no. 5, pp. 85 – 95, sep 2002.

[2] H. Neemuchwala and A. O. Hero, "Image registration in high dimensional feature space," *Proc. of SPIE Conference on Electronic Imaging, San Jose*, January 2005.

[3] E. G. Miller and J. W. Fisher III, "ICA using spacings estimates of entropy," *Proc. 4th Intl. Symp. on ICA and BSS*, pp. 1047–1052, 2003.

[4] A. Lakhina, M. Crovella, and C. Diot, "Mining anomalies using traffic feature distributions," in *In ACM SIGCOMM*, 2005, pp. 217–228.

[5] A. Jain, "Image data compression: A review," *Proceedings of the IEEE*, vol. 69, no. 3, pp. 349 – 389, March 1981.

[6] O. Vasicek, "A test for normality based on sample entropy." *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 38, pp. 54–59, 1976.

[7] E. J. Dudewicz and E. C. van der Meulen, "Entropy-based tests of uniformity." *Journal of the American Statistical Association*, vol. 76, pp. 967–974, 1981.

[8] R. C. H. Cheng and N. A. K. Amin, "Estimating parameters in continuous univariate distributions with a shifted origin." *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 11, pp. 394–403, 1983.

[9] B. Ranneby, "The maximum spacing method. an estimation method related to the maximum likelihood method." *Scandinavian Journal of Statistics*, vol. 11, pp. 93–112, 1984.

[10] N. Leonenko, L. Prozanto, and V. Savani, "A class of rényi information estimators for multidimensional densities," *Annals of Statistics*, vol. 36, pp. 2153–2182, 2008.

[11] A. O. Hero, J. Costa, and B. Ma, "Asymptotic relations between minimal graphs and alpha-entropy," *Technical Report CSPL-334 Communications and Signal Processing Laboratory, The University of Michigan*, March 2003.

[12] D. Pál, B. Póczos, and C. Szepesvári, "Estimation of R\'enyi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs," *ArXiv e-prints*, Mar. 2010.

[13] B. van Es, "Estimating functionals related to a density by class of statistics based on spacing," *Scandinavian Journal of Statistics*, 1992.

[14] M. Goria, N. Leonenko, V. Mergel, and P. L. N. Inverardi, "A new class of random vector entropy estimators and its applications in testing statistical hypotheses," *Nonparametric Statistics*, 2004.

[15] E. Liitiäinen, A. Lendasse, and F. Corona, "On the statistical estimation of rényi entropies," in *Proceedings of IEEE/MLSP 2009 International Workshop on Machine Learning for Signal Processing, Grenoble (France)*, September 2-4 2009.

[16] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *Information Theory, IEEE Transactions on*, vol. 51, no. 9, pp. 3064–3074, 2005.

[17] I. Ahmad and Pi-Erh Lin, "A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.)," *Information Theory, IEEE Transactions on*, vol. 22, no. 3, pp. 372 – 375, may 1976.

[18] P. B. Eggermont and V. N. LaRiccia, "Best asymptotic normality of the kernel density entropy estimator for smooth densities," *Information Theory, IEEE Transactions on*, vol. 45, no. 4, pp. 1321 –1326, May 1999.

[19] P. J. Bickel and Y. Ritov, "Estimating integrated squared density derivatives: Sharp best order of convergence estimates," *Sankhya: The Indian Journal of Statistics*, vol. 50, pp. 381–393, October 1988.

[20] P. Hall and J. S. Marron, "Estimation of integrated squared density derivatives," *Stat. Prob. Lett*, pp. 109–115, 1987.

[21] L. Birge and P. Massart, "Estimation of integral functions of a density," *The Annals of Statistics*, vol. 23, no. 1, pp. 11–29, 1995.

[22] E. Giné and D. Mason, "Uniform in bandwidth estimation of integral functionals of the density function," *Scandinavian Journal of Statistics*, vol. 35, p. 739761, 2008.

[23] M. M. V. Hulle, "Edgeworth approximation of multivariate differential entropy," *Neural Computation*, vol. 17, no. 9, pp. 1903–1910, 2005.

[24] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *Information Theory, IEEE Transactions on*, vol. 56, no. 11, pp. 5847 –5861, November 2010.

[25] B. Laurent, "Efficient estimation of integral functionals of a density," *The Annals of Statistics*, vol. 24, no. 2, pp. 659–681, 1996.

[26] H. Singh, N. Misra, and V. Hnizdo, "Nearest neighbor estimators of entropy," *The Annals of Statistics*, 2005.

[27] D. Evans, A. Jones, and W. M. Schmidt, "Asymptotic moments of nearest neighbor distance distributions," *Proceedings of the Royal Society A*, vol. 458, pp. 2839–2849, 2008.

[28] Y. Baryshnikov, M. D. Penrose, and J. Yukich, "Gaussian limits for generalized spacings," *Ann. Appl. Probab.*, vol. 19, no. 1, pp. 158–185, 2009.

[29] D. Evans, "A law of large numbers for nearest neighbor statistics," *Proceedings of the Royal Society A*, vol. 464, pp. 3175–3192, 2008.

[30] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Ann. Math. Statist.*, 1965.

[31] K. Sricharan, R. Raich, and A. O. Hero, "Empirical estimation of entropy functionals with confidence," *ArXiv e-prints*, February 2012.

[32] Y. P. Mack and M. Rosenblatt, "Multivariate k-nearest neighbor density estimates," *Journal of Multivariate Analysis*, vol. 9, no. 1, pp. 1 – 15, 1979.

[33] K. Fukunaga and L. D. Hostetler, "Optimization of k-nearest-neighbor density estimates," *IEEE Transactions on Information Theory*, 1973.

[34] X. S. Raymond, *Elementary Introduction to the Theory of Pseudodifferential Operators*. CRC Press, 1991.

[35] D. S. Moore and J. W. Yackel, "Consistency properties of nearest neighbor density function estimators," *The Annals of Statistics*, 1977.

[36] J. Blum, H. Chernoff, M. Rosenblatt, and H. Teicher, "Central limit theorems for interchangeable processes," *Canadian Journal of Mathematics*, June 1957.

[37] N. Leonenko, L. Prozanto, and V. Savani, "A class of rényi information estimators for multidimensional densities," *Annals of Statistics*, vol. 38, pp. 3837–3838, 2010.

[38] K. Sricharan and A. Hero, "Weighted k-nn graphs for rényi entropy estimation in high dimensions," in *Statistical Signal Processing Workshop (SSP), 2011 IEEE*, june 2011, pp. 773 –776.

[39] V. C. Raykar and R. Duraiswami, "Fast optimal bandwidth selection for kernel density estimation," in *Proceedings of the sixth SIAM International Conference on Data Mining*, J. Ghosh, D. Lambert, D. Skillicorn, and J. Srivastava, Eds., 2006, pp. 524–528.