# Analysis of $k$-Nearest Neighbor Distances with Application to Entropy Estimation

**Shashank Singh** SSS1@ANDREW.CMU.EDU
**Barnabás Póczos** BAPOCZOS@CS.CMU.EDU
Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213 USA

## Abstract

Estimating entropy and mutual information consistently is important for many machine learning applications. The Kozachenko-Leonenko (KL) estimator (Kozachenko & Leonenko, 1987) is a widely used nonparametric estimator for the entropy of multivariate continuous random variables, as well as the basis of the mutual information estimator of Kraskov et al. (2004), perhaps the most widely used estimator of mutual information in this setting. Despite the practical importance of these estimators, major theoretical questions regarding their finite-sample behavior remain open. This paper proves finite-sample bounds on the bias and variance of the KL estimator, showing that it achieves the minimax convergence rate for certain classes of smooth functions. In proving these bounds, we analyze finite-sample behavior of $k$-nearest neighbors ($k$-NN) distance statistics (on which the KL estimator is based). We derive concentration inequalities for $k$-NN distances and a general expectation bound for statistics of $k$-NN distances, which may be useful for other analyses of $k$-NN methods.

## 1. Introduction

Estimating entropy and mutual information in a consistent manner is of importance in a number problems in machine learning. For example, entropy estimators have applications in goodness-of-fit testing (Goria et al., 2005), parameter estimation in semi-parametric models (Wolsztynski et al., 2005), studying fractal random walks (Alemany & Zanette, 1994), and texture classification (Hero et al., 2002a;b). Mutual information estimators have applications in feature selection (Peng & Dind, 2005), clustering (Aghagolzadeh et al., 2007), causality detection (Hlaváckova-Schindler et al., 2007), optimal experimental design (Lewi et al., 2007; Póczos & Lőrincz, 2009), fMRI data processing (Chai et al., 2009), prediction of protein structures (Adami, 2004), and boosting and facial expression recognition (Shan et al., 2005). Both entropy estimators and mutual information estimators have been used for independent component and subspace analysis (Learned-Miller & Fisher, 2003; Szabó et al., 2007; Póczos & Lőrincz, 2005; Hulle, 2008), as well as for image registration (Kybic, 2006; Hero et al., 2002a;b). For further applications, see (Leonenko et al., 2008).

In this paper, we focus on the problem of estimating the Shannon entropy of a continuous random variable given samples from its distribution. All of our results extend to the estimation of mutual information, since the latter can be written as a sum of entropies. [1] In our setting, we assume we are given $n$ IID samples from an unknown probability measure $P$. Under nonparametric assumptions (on the smoothness and tail behavior of $P$), our task is then to estimate the differential Shannon entropy of $P$.

Estimators of entropy and mutual information come in many forms (as reviewed in Section 2), but one common approach is based on statistics of $k$-nearest neighbor ($k$-NN) distances (i.e., the distance from a sample to its $k^{th}$ nearest neighbor amongst the samples, in some metric on the space). These nearest-neighbor estimates are largely based on initial work by Kozachenko & Leonenko (1987), who proposed an estimate for differential Shannon entropy and showed its weak consistency. Henceforth, we refer to this historic estimator as the 'KL estimator', after its discoverers. Although there has been much work on the problem of entropy estimation in the nearly three decades since the KL estimator was proposed, there are still major open questions about the finite-sample behavior of the KL estimator. The goal of this paper is to address some of these questions in the form of finite-sample bounds on the bias

[1] Specifically, for random variables $X$ and $Y$, $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

and variance of the estimator.

Specifically, our **main contributions** are the following:

1. We derive $O\left((k/n)^{\beta/D}\right)$ bounds on the bias of the KL estimate, where $\beta$ is a measure of the smoothness (i.e., Hölder continuity) of the sampling density, $D$ is the intrinsic dimension of the support of the distribution, and $n$ is the sample size.

2. We derive $O\left(n^{-1}\right)$ bounds on the variance of the KL estimator.

3. We derive concentration inequalities for $k$-NN distances, as well as general bounds on expectations of $k$-NN distance statistics, with important special cases:

   (a) We bound the moments of $k$-NN distances, which play a role in analysis of many applications of $k$-NN methods, including both the bias and variance of the KL estimator. In particular, we significantly relax strong assumptions underlying previous results by Evans et al. (2002), such as compact support and smoothness of the sampling density. Our results are also the first which apply to negative moments (i.e., $\mathbb{E}\left[X^\alpha\right]$ with $\alpha < 0$); these are important for bounding the variance of the KL estimator.

   (b) We give upper and lower bounds on the logarithms of $k$-NN distances. These are important for bounding the variance of the KL estimator, as well as $k$-NN estimators for divergences and mutual informations.

We present our results in the general setting of a set equipped with a metric, a base measure, a probability density, and an appropriate definition of dimension. This setting subsumes Euclidean spaces, in which $k$-NN methods have traditionally been analyzed, [2] but also includes, for instance, Riemannian manifolds, and perhaps other spaces of interest. We also strive to weaken some of the restrictive assumptions, such as compact support and boundedness of the density, on which most related work depends.

We anticipate that the some of the tools developed here may be used to derive error bounds for $k$-NN estimators of mutual information, divergences (Wang et al., 2009), their generalizations (e.g., Rényi and Tsallis quantities (Leonenko et al., 2008)), norms, and other functionals of probability densities. We leave such bounds to future work.

**Organization**

Section 2 discusses related work. Section 3 gives theoretical context and assumptions underlying our work. In Section 4, we prove concentration boundss for $k$-NN distances, and we use these in Section 5 to derive bounds on the expectations of $k$-NN distance statistics. Section 6 describes the KL estimator, for which we prove bounds on the bias and variance in Sections 7 and 8, respectively.

## 2. Related Work

Here, we review previous work on the analysis of $k$-nearest neighbor statistics and their role in estimating information theoretic functionals, as well as other approaches to estimating information theoretic functionals.

### 2.1. The Kozachenko-Leonenko Estimator of Entropy

In general contexts, only weak consistency of the KL estimator is known (Kozachenko & Leonenko, 1987). Biau & Devroye (2015) recently reviewed finite-sample results known for the KL estimator. They show (Theorem 7.1) that, if the density $p$ has compact support, then the variance of the KL estimator decays as $O(n^{-1})$. They also claim (Theorem 7.2) to bound the bias of the KL estimator by $O(n^{-\beta})$, under the assumptions that $p$ is $\beta$-Hölder continuous ($\beta \in (0, 1]$), bounded away from 0, and supported on the interval $[0, 1]$. However, in their proof Biau & Devroye (2015) neglect the additional bias incurred at the boundaries of $[0, 1]$, where the density cannot simultaneously be bounded away from 0 and continuous. In fact, because the KL estimator does not attempt to correct for boundary bias, for densities bounded away from 0, the estimator may suffer bias worse than $O(n^{-\beta})$.

The KL estimator is also important for its role in the mutual information estimator proposed by Kraskov et al. (2004), which we refer to as the KSG estimator. The KSG estimator expands the mutual information as a sum of entropies, which it estimates via the KL estimator with a particular *random* (i.e., data-dependent) choice of the nearest-neighbor parameter $k$. The KSG estimator is perhaps the most widely used estimator for the mutual information between continuous random variables, despite the fact that it currently appears to have no theoretical guarantees, even asymptotically. In fact, one of the few theoretical results, due to Gao et al. (2015b), concerning the KSG estimator is a negative result: when estimating the mutual information between strongly dependent variables, the KSG estimator tends to systematically underestimate mutual information, due to increased boundary bias. [3] Nevertheless, the

[2] A recent exception in the context of classification, is Chaudhuri & Dasgupta (2014) which considers general metric spaces.

[3] To alleviate this, Gao et al. (2015b) provide a heuristic correction based on using local PCA to estimate the support of the distribution. Gao et al. (2015a) provide and prove asymptotic un-

widespread use of the KSG estimator motivates study of its behavior. We hope that our analysis of the KL estimator, in terms of which the KSG estimator can be written, will lead to a better understanding of the latter.

### 2.2. Analysis of nearest-neighbor distance statistics

Evans (2008) derives a law of large numbers for $k$-NN statistics with uniformly bounded (central) kurtosis as the sample size $n \to \infty$. Although it is not obvious that the kurtosis of $\log$-$k$-NN distances is uniformly bounded (indeed, each $\log$-$k$-NN distance approaches $-\infty$ almost surely), we show in Section 8 that this is indeed the case, and we apply the results of Evans (2008) to bound the variance of the KL estimator.

Evans et al. (2002) derives asymptotic limits and convergence rates for moments of $k$-NN distances, for sampling densities with bounded derivatives and compact domain. In contrast, we use weaker assumptions to simply prove bounds on the moments of $k$-NN distances. Importantly, whereas the results of Evans et al. (2002) apply only to non-negative moments (i.e., $\mathbb{E}\left[|X|^\alpha\right]$ with $\alpha \geq 0$), our results also hold for certain negative moments, which is crucial for our bounds on the variance of the KL estimator.

### 2.3. Other Approaches to Estimating Information Theoretic Functionals

**Analysis of convergence rates:** For densities over $\mathbb{R}^D$ satisfying a Hölder smoothness condition parametrized by $\beta \in (0, \infty)$, the minimax rate for estimating entropy has been known since Birge & Massart (1995) to be $O\left(n^{-\min\left\{\frac{8\beta}{4\beta+D},1\right\}}\right)$ in mean squared error, where $n$ is the sample size.

Quite recently, there has been much work on analyzing new estimators for entropy, mutual information, divergences, and other functionals of densities. Most of this work has been along one of three approaches. One series of papers (Liu et al., 2012; Singh & Poczos, 2014b;a) studied boundary-corrected plug-in approach based on under-smoothed kernel density estimation. This approach has strong finite sample guarantees, but requires prior knowledge of the support of the density and can necessitate computationally demanding numerical integration. A second approach (Krishnamurthy et al., 2014; Kandasamy et al., 2015) uses von Mises expansion to correct the bias of optimally smoothed density estimates. This approach shares the difficulties of the previous approach, but is statistically more efficient. Finally, a long line of work (Pérez-Cruz, 2008; Pál et al., 2010; Sricharan et al., 2012; Sricharan et al., 2010; Moon & Hero, 2014) has studied entropy estimation based on continuum limits of certain properties of graphs (including $k$-NN graphs, spanning trees, and other sample-based graphs).

Most of these estimators achieve rates of $O\left(n^{-\min\left\{\frac{2\beta}{\beta+D},1\right\}}\right)$ or $O\left(n^{-\min\left\{\frac{4\beta}{2\beta+D},1\right\}}\right)$. Only the von Mises approach of Krishnamurthy et al. (2014) is known to achieve the minimax rate for general $\beta$ and $D$, but due to its high computational demand ($O(2^D n^3)$), the authors suggest the use of other statistically less efficient estimators for moderately sized datasets. In this paper, we prove that, for $\beta \in (0, 2]$, the KL estimator converges at the rate $O\left(n^{-\min\left\{\frac{4\beta}{2\beta+D},1\right\}}\right)$. It is also worth noting the relative computational efficiency of the KL estimator ($O\left(Dn^2\right)$, or $O\left(2^D n \log n\right)$ using $k$-d trees for small $D$).

**Boundedness of the density:** For all of the above approaches, theoretical finite-sample results known so far assume that the sampling density is lower and upper bounded by positive constants. This also excludes most distributions with unbounded support, and hence, many distributions of practical relevance. A distinctive feature of our results is that they hold for a variety of densities that approach 0 and $\infty$ on their domain, which may be unbounded. Our bias bounds apply, for example, to densities that decay exponentially, such as Gaussian distributions. To our knowledge, the only previous results that apply to unbounded densities are those of Tsybakov & van der Meulen (1996), who show $\sqrt{n}$-consistency of a truncated modification of the KL estimate for a class of functions with exponentially decaying tails. In fact, components of our analysis are inspired by Tsybakov & van der Meulen (1996), and some of our assumptions are closely related. Their analysis only applies to the case $\beta = 2$ and $D = 1$, for which our results also imply $\sqrt{n}$-consistency, so our results can be seen in some respects as a generalization of this work.

## 3. Setup and Assumptions

While most prior work on $k$-NN estimators has been restricted to $\mathbb{R}^D$, we present our results in a more general setting. This includes, for example, Riemannian manifolds embedded in higher dimensional spaces, in which case we note that our results depend on the *intrinsic*, rather than *extrinsic*, dimension. Such data can be better behaved in their native space than when embedded in a lower dimensional Euclidean space (e.g., working directly on the unit circle avoids boundary bias caused by mapping data to the interval $[0, 2\pi]$).

**Definition 1. (Metric Measure Space):** *A quadruple* $(\mathbb{X}, d, \Sigma, \mu)$ *is called a* metric measure space *if* $\mathbb{X}$ *is a set,* $d : \mathbb{X} \times \mathbb{X} \to [0, \infty)$ *is a metric on* $\mathbb{X}$*,* $\Sigma$ *is a* $\sigma$*-algebra on* $\mathcal{X}$ *containing the Borel* $\sigma$*-algebra induced by* $d$*, and* $\mu : \Sigma \to [0, \infty]$ *is a* $\sigma$*-finite measure on the measurable*

biasedness of another estimator, based on local Gaussian density estimation, that directly adapts to the boundary.

*space* $(\mathbb{X}, \Sigma)$.

**Definition 2. (Dimension):** *A metric measure space* $(\mathbb{X}, d, \Sigma, \mu)$ *is said to have* dimension $D \in [0, \infty)$ *if there exist constants* $c_D, \rho > 0$ *such that,* $\forall r \in [0, \rho]$, $x \in \mathcal{X}$, $\mu(B(x, r)) = c_D r^D$. [4]

**Definition 3. (Full Dimension):** *Given a metric measure space* $(\mathbb{X}, d, \Sigma, \mu)$ *of dimension* $D$*, a measure* $P$ *on* $(\mathbb{X}, \Sigma)$ *is said to have* full dimension *on a set* $\mathcal{X} \subseteq \mathbb{X}$ *if there exist functions* $\gamma_*, \gamma^* : \mathcal{X} \to (0, \infty)$ *such that, for all* $r \in [0, \rho]$ *and* $\mu$*-almost all* $x \in \mathcal{X}$,

$$\gamma_*(x) r^D \leq P(B(x, r)) \leq \gamma^*(x) r^D.$$

**Remark 4.** *If* $\mathbb{X} = \mathbb{R}^D$, $d$ *is the Euclidean metric, and* $\mu$ *is the Lebesgue measure, then the dimension of the metric measure space is* $D$*. However, if* $\mathbb{X}$ *is a lower dimensional subspace of* $\mathbb{R}^D$*, then the dimension may be less than* $D$*. For example, if* $\mathbb{X} = \mathbb{S}_{D-1} := \{x \in \mathbb{R}^D : \|x\|_2 = 1\}$*),* $d$ *is the geodesic distance on* $\mathbb{S}_{D-1}$*, and* $\mu$ *is the* $(D-1)$*-dimensional surface measure, then the dimension is* $D - 1$.

**Remark 5.** *In previous work on* $k$*-NN statistics (Evans et al., 2002; Biau & Devroye, 2015) and estimation of information theoretic functionals (Sricharan et al., 2010; Krishnamurthy et al., 2014; Singh & Poczos, 2014b; Moon & Hero, 2014), it has been common to make the assumption that the sampling distribution has full dimension with* constant $\gamma_*$ *and* $\gamma^*$ *(or, equivalently, that the density is lower and upper bounded by positive constants). This excludes distributions with densities approaching* $0$ *or* $\infty$ *on their domain, and hence also densities with unbounded support. By letting* $\gamma_*$ *and* $\gamma^*$ *be functions, our results extend to unbounded densities that instead satisfy certain tail bounds.*

In order to ensure that entropy is well defined, we assume that $P$ is a probability measure absolutely continuous with respect to $\mu$, and that its probability density function $p : \mathcal{X} \to [0, \infty)$ satisfies [5]

$$H(p) := \mathop{\mathbb{E}}_{X \sim P} [\log p(X)] = \int_{\mathcal{X}} p(x) \log p(x) \, d\mu(x) \in \mathbb{R}. \tag{1}$$

Finally, we assume we have $n + 1$ samples $X, X_1, ..., X_n$ drawn IID from $P$. We would like to use these samples to estimate the entropy $H(p)$ as defined in Equation (1).

Our analysis and methods relate to the $k$-nearest neighbor distance $\varepsilon_k(x)$, defined for any $x \in \mathcal{X}$ by $\varepsilon_k(x) = d(x, X_i)$, where $X_i$ is the $k^{th}$-nearest neighbor of $x$ in the set $\{X_1, ..., X_n\}$. Note that, since the definition of dimension used precludes the existence of atoms (i.e., for all $x \in \mathcal{X}$, $p(x) = \mu(\{x\}) = 0$), $\varepsilon_k(x) > 0$, $\mu$-almost everywhere. This is important, since we will study $\log \varepsilon_k(x)$.

Initially (i.e., in Sections 4 and 5), we will study $\log \varepsilon_k(x)$ with fixed $x \in \mathcal{X}$, for which we will derive bounds in terms of $\gamma_*(x)$ and $\gamma^*(x)$. When we apply these results to analyze the KL estimator in Section 7 and 8, we will need to take expectations such as $\mathbb{E}[\log \varepsilon_k(X)]$ (for which we reserve the extra sample $X$), leading to 'tail bounds' on $p$ in terms of the functions $\gamma_*$ and $\gamma^*$.

## 4. Concentration of $k$-NN Distances

We begin with a consequence of the multiplicative Chernoff bound, asserting a sort of concentration of the distance of any point in $\mathcal{X}$ from its $k^{th}$-nearest neighbor in $\{X_1, \ldots, X_n\}$. Since the results of this section are concerned with fixed $x \in \mathcal{X}$, for notational simplicity, we suppress the dependence of $\gamma_*$ and $\gamma^*$ on $x$.

**Lemma 6.** *Let* $(\mathbb{X}, d, \Sigma, \mu)$ *be a metric measure space of dimension* $D$*. Suppose* $P$ *is an absolutely continuous probability measure with full dimension on* $\mathcal{X} \subseteq \mathbb{X}$ *and density function* $p : \mathcal{X} \to [0, \infty)$*. For* $x \in \mathcal{X}$*, if* $r \in \left[\left(\frac{k}{\gamma_* n}\right)^{1/D}, \rho\right]$*, then*

$$\mathbb{P}[\varepsilon_k(x) > r] \leq e^{-\gamma_* r^D n} \left(e \frac{\gamma_* r^D n}{k}\right)^k.$$

*and, if* $r \in \left[0, \min\left\{\left(\frac{k}{\gamma^* n}\right)^{1/D}, \rho\right\}\right]$*, then*

$$\mathbb{P}[\varepsilon_k(x) \leq r] \leq \left(\frac{e \gamma^* r^D n}{k}\right)^{k\gamma_*/\gamma^*}.$$

## 5. Bounds on Expectations of KNN Statistics

Here, we use the concentration bounds of Section 4 to bound expectations of functions of $k$-nearest neighbor distances. Specifically, we give a simple formula for deriving bounds that applies to many functions of interest, including logarithms and (positive and negative) moments. As in the previous section, the results apply to a fixed $x \in \mathcal{X}$, and we continue to suppress the dependence of $\gamma_*$ and $\gamma^*$ on $x$.

**Theorem 7.** *Let* $(\mathcal{X}, d, \Sigma, \mu)$ *be a metric measure space of dimension* $D$*. Suppose* $P$ *is an absolutely continuous probability measure with full dimension and density function* $p : \mathcal{X} \to [0, \infty)$ *that satisfies the tail condition* [6]

$$\mathop{\mathbb{E}}_{X \sim P}\left[\int_{\rho}^{\infty} \left[1 - P(B(X, f^{-1}(r)))\right]^n\right] \leq \frac{C_T}{n} \tag{2}$$

[4] Here and in what follows, $B(x, r) := \{y \in \mathcal{X} : d(x, y) < r\}$ denotes the open ball of radius $r$ centered at $x$.

[5] See (Baccetti & Visser, 2013) for discussion of sufficient conditions for $H(p) < \infty$.

[6] Since $f$ need not be surjective, we use the generalized inverse $f^{-1} : \mathbb{R} \to [0, \infty]$ defined by $f^{-1}(\varepsilon) := \inf\{x \in (0, \infty) : f(x) \geq \varepsilon\}$.

*for some constant $C_T > 0$. Suppose $f : (0, \infty) \to \mathbb{R}$ is continuously differentiable, with $f' > 0$. Fix $x \in \mathcal{X}$. Then, we have the upper bound*

$$\mathbb{E}\left[f_+(\varepsilon_k(x))\right] \leq f_+\left(\left(\frac{k}{\gamma_* n}\right)^{\frac{1}{D}}\right) + \frac{C_T}{n} \tag{3}$$

$$+ \frac{(e/k)^k}{D(n\gamma_*)^{\frac{1}{D}}} \int_k^\infty e^{-y} y^{k+\frac{1}{D}-1} f'\left(\left(\frac{y}{n\gamma_*}\right)^{\frac{1}{D}}\right) dy$$

*and the lower bound*

$$\mathbb{E}\left[f_-(\varepsilon_k(x))\right] \leq f_-\left(\left(\frac{k}{\gamma^* n}\right)^{1/D}\right) + \frac{C_T}{n}$$

$$+ \left(\frac{en\gamma^*}{k}\right)^{\frac{k\gamma_*}{\gamma^*}} \int_0^{\left(\frac{k}{\gamma^* n}\right)^{\frac{1}{D}}} y^{Dk\gamma_*/\gamma^*} f'(y)\, dy \tag{4}$$

*($f_+(x) = \max\{0, f(x)\}$ and $f_-(x) = -\min\{0, f(x)\}$ denote the positive and negative parts of $f$, respectively).*

**Remark 8.** *If $f : (0, \infty) \to \mathbb{R}$ is continuously differentiable with $f' < 0$, we can apply Theorem 7 to $-f$. Also, similar techniques can be used to prove analogous lower bounds (i.e., lower bounds on the positive part and upper bounds on the negative part).*

**Remark 9.** *The tail condition (2) is difficult to validate directly for many distributions. Clearly, it is satisfied when the support of $p$ is bounded. However, (Tsybakov & van der Meulen, 1996) show that, for the functions $f$ we are interested in (i.e., logarithms and power functions), when $\mathcal{X} = \mathbb{R}^D$, $d$ is the Euclidean metric, and $\mu$ is the Lebesgue measure, (2) is also satisfied by upper-bounded densities with exponentially decreasing tails. More precisely, that is when there exist $a, b, \alpha, \delta > 0$ and $\beta > 1$ such that, whenever $\|x\|_2 > \delta$,*

$$ae^{-\alpha\|x\|^\beta} \leq p(x) \leq be^{-\alpha\|x\|^\beta},$$

*which permits, for example, Gaussian distributions. It should be noted that the constant $C_T$ depends only on the metric measure space, the distribution $P$, and the function $f$, and, in particular, not on $k$.*

### 5.1. Applications of Theorem 7

We can apply Theorem 7 to several functions $f$ of interest. Here, we demonstrate the cases $f(x) = \log x$ and $f(x) = x^\alpha$ for certain $\alpha$, as we will use these bounds when analyzing the KL estimator.

When $f(x) = \log(x)$, (3) gives

$$\mathbb{E}\left[\log_+(\varepsilon_k(x))\right] \leq \frac{1}{D}\log_+\left(\frac{k}{\gamma_* n}\right) + \left(\frac{e}{k}\right)^k \frac{\Gamma(k,k)}{D}$$

$$\leq \frac{1}{D}\left(1 + \log_+\left(\frac{k}{\gamma_* n}\right)\right) \tag{5}$$

(where $\Gamma(s,x) := \int_x^\infty t^{s-1}e^{-t}\,dt$ denotes the upper incomplete Gamma function, and we used the bound $\Gamma(s,x) \leq x^{s-1}e^{-x}$), and (4) gives

$$\mathbb{E}\left[\log_-(\varepsilon_k(x))\right] \leq \frac{1}{D}\log_-\left(\frac{k}{\gamma^* n}\right) + C_1, \tag{6}$$

for $C_1 = \frac{\gamma^* e^{k\gamma_*/\gamma^*}}{Dk\gamma_*}$. For $\alpha > 0$, $f(x) = x^\alpha$, (3) gives

$$\mathbb{E}\left[\varepsilon_k^\alpha(x)\right] \leq \left(\frac{k}{\gamma_* n}\right)^{\frac{\alpha}{D}} + \left(\frac{e}{k}\right)^k \frac{\alpha\Gamma\left(k+\alpha/D, k\right)}{D(n\gamma_*)^{\alpha/D}}$$

$$\leq C_2\left(\frac{k}{\gamma_* n}\right)^{\frac{\alpha}{D}}, \tag{7}$$

where $C_2 = 1 + 2\frac{\alpha}{D}$. For any $\alpha \in [-Dk\gamma_*/\gamma^*, 0]$, when $f(x) = -x^\alpha$, (4) gives

$$\mathbb{E}\left[\varepsilon_k^\alpha(x)\right] \leq C_3\left(\frac{k}{\gamma^* n}\right)^{\frac{\alpha}{D}}, \tag{8}$$

where $C_3 = 1 + \frac{\alpha\gamma^* e^{k\gamma_*/\gamma^*}}{Dk\gamma_* + \alpha\gamma^*}$.

## 6. The KL Estimator for Entropy

Recall that, for a random variable $X$ sampled from a probability density $p$ with respect to a base measure $\mu$, the Shannon entropy is defined as

$$H(X) = -\int_{\mathcal{X}} p(x)\log p(x)\,dx.$$

As discussed in Section 1, many applications call for estimate of $H(X)$ given $n$ IID samples $X_1, \ldots, X_n \sim p$. For a positive integer $k$, the KL estimator is typically written as

$$\hat{H}_k(X) = \psi(n) - \psi(k) + \log c_D + \frac{D}{n}\sum_{i=1}^n \log\varepsilon_k(X_i),$$

where $\psi : \mathbb{N} \to \mathbb{R}$ denotes the digamma function. The motivating insight is the observation that, independent of the sampling distribution, [7]

$$\mathbb{E}\left[\log P(B(X_i, \varepsilon_k(X_i)))\right] = \psi(k) - \psi(n),$$

Hence,

$$\mathbb{E}\left[\hat{H}_k(X)\right]$$

$$= \mathbb{E}\left[-\log P(B(X_i, \varepsilon_k(X_i))) + \log c_D + \frac{D}{n}\sum_{i=1}^n \log\varepsilon_k(X_i)\right]$$

$$= -\,\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \log\left(\frac{P(B(x_i, \varepsilon_k(X_i)))}{c_D\varepsilon_k^D(X_i)}\right)\right]$$

[7] See (Kraskov et al., 2004) for a concise proof of this fact.

$$= -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log p_{\varepsilon_k(i)}(X_i)\right] = -\mathbb{E}\left[\log p_{\varepsilon_k(X_1)}(X_1)\right],$$

where, for any $x \in \mathcal{X}$, $\varepsilon > 0$,

$$p_\varepsilon(x) = \frac{1}{c_D\varepsilon^D}\int_{B(x,\varepsilon)} p(y)\,d\mu(y) = \frac{P(B(x,\varepsilon))}{c_D\varepsilon^D}$$

denotes the local average of $p$ in a ball of radius $\varepsilon$ around $x$. Since $p_\varepsilon$ is a smoothed approximation of $p$ (with smoothness increasing with $\varepsilon$), the KL estimate can be intuitively thought of as a plug-in estimator for $H(X)$, using a density estimate with an adaptive smoothing parameter.

In the next two sections, we utilize the bounds derived in Section 5 to bound the bias and variance of the KL estimator. We note that, for densities in the $\beta$-Hölder smoothness class ($\beta \in (0, 2]$), our results imply a mean-squared error of $O(n^{-2\beta/D})$ when $\beta < D/2$ and $O(n^{-1})$ when $\beta \geq D/2$.

## 7. Bias Bound

In this section, we prove bounds on the bias of the KL estimator, first in a relatively general setting, and then, as a corollary, in a more specific but better understood setting.

**Theorem 10.** *Suppose* $(\mathbb{X}, d, \Sigma, \mu)$ *and* $P$ *satisfy the conditions of Theorem 7, and there exist* $C, \beta \in (0, \infty)$ *with*

$$\sup_{x\in\mathcal{X}} |p(x) - p_\varepsilon(x)| \leq C_\beta\varepsilon^\beta,$$

*and suppose* $p$ *satisfies a 'tail bound'*

$$\Gamma_B := \mathop{\mathbb{E}}_{X\sim P}\left[(\gamma_*(X))^{-\frac{\beta+D}{D}}\right] < \infty. \tag{9}$$

*Then,*

$$\left|\mathbb{E}\left[H(X) - \hat{H}_k(X)\right]\right| \leq C_B\left(\frac{k}{n}\right)^{\frac{\beta}{D}},$$

*where* $C_B = (1 + c_D)C_2C_\beta\Gamma_B$.

We now show that the conditions of Theorem 10 are satisfied by densities in the commonly used nonparametric class of $\beta$-Hölder continuous densities on $\mathbb{R}^D$.

**Definition 11.** *Given a constant* $\beta > 0$ *and an open set* $\mathcal{X} \subseteq \mathbb{R}^D$*, a function* $f : \mathcal{X} \to \mathbb{R}$ *is called* $\beta$-Hölder continuous *if* $f$ *is* $\ell$ *times differentiable and there exists* $L > 0$ *such that, for any multi-index* $\alpha \in \mathbb{N}^D$ *with* $|\alpha| < \beta$,

$$\sup_{x\neq y\in\mathcal{X}} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{\|x - y\|^{\beta-\ell}} \leq L,$$

*where* $\ell := \lfloor\beta\rfloor$ *is the greatest integer* strictly *less than* $\beta$.

**Definition 12.** *Given an open set* $\mathcal{X} \subseteq \mathbb{R}^D$ *and a function* $f : \mathcal{X} \to \mathbb{R}$, $f$ *is said to* vanish on the boundary $\partial\mathcal{X}$ *of* $\mathcal{X}$ *if, for any sequence* $\{x_i\}_{i=1}^\infty$ *in* $\mathcal{X}$ *with* $\inf_{x'\in\partial\mathcal{X}} \|x - x'\|_2 \to 0$ *as* $i \to \infty$, $f(x) \to 0$ *as* $i \to \infty$. *Here,*

$$\partial\mathcal{X} := \{x \in \mathbb{R}^D : \forall\delta > 0, B(x,\delta) \nsubseteq \mathcal{X} \text{ and } B(x,\delta) \nsubseteq \mathcal{X}^c\},$$

*denotes the boundary of* $\mathcal{X}$.

**Corollary 13.** *Consider the metric measure space* $(\mathbb{R}^D, d, \Sigma, \mu)$*, where* $d$ *is Euclidean and* $\mu$ *is the Lebesgue measure. Let* $P$ *be an absolute continuous probability measure with full dimension and density* $p$ *supported on an open set* $\mathcal{X} \subseteq \mathbb{R}^D$. *Suppose* $p$ *satisfies (9) and the conditions of Theorem 7 and is* $\beta$*-Hölder continuous (*$\beta \in (0, 2]$*) with constant* $L$. *Assume* $p$ *vanishes on* $\partial\mathcal{X}$. *If* $\beta > 1$, *assume* $\|\nabla p\|_2$ *vanishes on* $\partial\mathcal{X}$. *Then,*

$$\left|\mathbb{E}\left[\hat{H}_k(X) - H(X)\right]\right| \leq C_H\left(\frac{n}{k}\right)^{-\frac{\beta}{D}},$$

*where* $C_H = (1 + c_D)C_2\Gamma\frac{LD}{D+\beta}$.

**Remark 14.** *The assumption that* $p$ *(and perhaps* $\|\nabla p\|$*) vanish on the boundary of* $\mathcal{X}$ *can be thought of as ensuring that the trivial continuation* $q : \mathbb{R}^D \to [0, \infty)$

$$q(x) = \begin{cases} p(x) & x \in \mathcal{X} \\ 0 & x \in \mathbb{R}^D\backslash\mathcal{X} \end{cases}$$

*of* $p$ *to* $\mathbb{R}^D$ *is* $\beta$*-Hölder continuous. This reduces boundary bias, for which the KL estimator does not correct.* [8]

## 8. Variance Bound

We first use the bounds proven in Section 5 to prove uniform (in $n$) bounds on the moments of $\mathbb{E}[\log\varepsilon_k(X)]$. We the for any fixed $x \in \mathcal{X}$, although $\log\varepsilon_k(x) \to -\infty$ almost surely as $n \to \infty$, $\mathbb{V}[\log\varepsilon_k(x)]$, and indeed all higher central moments of $\log\varepsilon_k(x)$, are bounded, uniformly in $n$. In fact, there exist exponential bounds, independent of $n$, on the density of $\log\varepsilon_k(x) - \mathbb{E}[\log\varepsilon_k(x)]$.

### 8.1. Moment Bounds on Logarithmic $k$-NN distances

**Lemma 15.** *Suppose* $(\mathbb{X}, d, \Sigma, \mu)$ *and* $P$ *satisfy the conditions of Theorem 7. Suppose also that* $\Gamma_0 := \sup_{x\in\mathcal{X}}\frac{\gamma^*(x)}{\gamma_*(x)} < \infty$. *Let* $\lambda \in \left(0, \frac{Dk}{\Gamma_0}\right)$ *and assume the following expectations are finite:*

$$\Gamma := \mathop{\mathbb{E}}_{X\sim P}\left[\frac{\gamma^*(X)}{\gamma_*(X)}\right] < \infty. \tag{10}$$

$$\Gamma_*(\lambda) := \mathop{\mathbb{E}}_{X\sim P}\left[(\gamma_*(X))^{-\lambda/D}\right] < \infty. \tag{11}$$

[8] Several estimators controlling for boundary bias have been proposed (e.g., Sricharan et al. (2010) give a modified $k$-NN estimator that accomplishes this *without* prior knowledge of $\mathcal{X}$.

$$\Gamma^*(\lambda) := \mathop{\mathbb{E}}_{X \sim P}\left[(\gamma^*(X))^{\lambda/D}\right] < \infty. \tag{12}$$

*Then, for any integer $\ell > 1$, the $\ell^{th}$ central moment*

$$M_\ell := \mathbb{E}\left[(\log \varepsilon_k(X) - \mathbb{E}\left[\log \varepsilon_k(X)\right])^\ell\right]$$

*satisfies*

$$M_\ell \leq C_M \ell!/\lambda^\ell, \tag{13}$$

*where $C_M > 0$ is a constant independent of $n$, $\ell$, and $\lambda$.*

**Remark 16.** *The conditions (10), (11), and (12) are mild. For example, when $\mathcal{X} = \mathbb{R}^D$, $d$ is the Euclidean metric, and $\mu$ is the Lebesgue measure, it suffices that $p$ is Lipschitz continuous* [9] *and there exist $c, r > 0, p > \frac{D^2}{D-\alpha}$ such that $p(x) \leq c\|x\|^{-p}$ whenever $\|x\|_2 > r$. The condition $\Gamma_0 < \infty$ is more prohibitive, but still permits many (possibly unbounded) distributions of interest.*

**Remark 17.** *If the terms $\log \varepsilon_k(X_i)$ were independent, a Bernstein inequality, together with the moment bound (13) would imply a sub-Gaussian concentration bound on the KL estimator about its expectation. This may follow from one of several more refined concentration results relaxing the independence assumption that have been proposed.*

### 8.2. Bound on the Variance of the KL Estimate

Bounds on the variance of the KL estimator now follow from the law of large numbers in Evans (2008) (itself an application of the Efron-Stein inequality to $k$-NN statistics).

**Theorem 18.** *Suppose $(\mathbb{X}, d, \Sigma, \mu)$ and $P$ satisfy the conditions of Lemma 15, and that that there exists a constant $N_k \in \mathbb{N}$ such that, for any finite $F \subseteq \mathcal{X}$, any $x \in F$ can be among the $k$-NN of at most $N_k$ other points in that set. Then, $\hat{H}_k(X) \to \mathbb{E}\left[\hat{H}_k(X)\right]$ almost surely (as $n \to \infty$), and, for $n \geq 16k$ and $M_4$ satisfying (13).*

$$\mathbb{V}\left[\hat{H}_k(X)\right] \leq \frac{5(3 + kN_k)(3 + 64k)M_4}{n} \in O\left(\frac{1}{nk}\right),$$

**Remark 19.** *$N_k$ depends only on $k$ and the geometry of the metric space $(\mathcal{X}, d)$. For example, Corollary A.2 of Evans (2008) shows that, when $\mathcal{X} = \mathbb{R}^D$ and $d$ is the Euclidean metric, then $N_k \leq kK(D)$, where $K(D)$ is the kissing number of $\mathbb{R}^d$.*

## 9. Bounds on the Mean Squared Error

The bias and variance bounds (Theorems 10 and 18) imply a bound on the mean squared error of the KL estimator:

**Corollary 20.** *Suppose $p$*

1. *is $\beta$-Hölder continuous with $\beta \in (0, 2]$.*
2. *vanishes on $\partial\mathcal{X}$. If $\beta > 1$, then also suppose $\|\nabla p\|_2$ vanishes on $\partial\mathcal{X}$.*
3.

*[TODO: Other assumptions.] satisfies the assumptions of Theorems 10 and 18. Then,*

$$\mathbb{E}\left[\left(\hat{H}_k(X) - H(X)\right)^2\right] \leq C_B^2 \left(\frac{k}{n}\right)^{2\beta/D} + \frac{C_V}{nk}. \tag{14}$$

*If we let $k$ scale as $k \asymp n^{\max\left\{0, \frac{2\beta - D}{2\beta + D}\right\}}$ this gives an overall convergence rate of*

$$\mathbb{E}\left[\left(\hat{H}_k(X) - H(X)\right)^2\right] \leq C_B^2 \left(\frac{k}{n}\right)^{2\beta/D} + \frac{C_V}{nk}. \tag{15}$$

## 10. Conclusions and Future Work

This paper derives finite sample bounds on the bias and variance of the KL estimator under general conditions, including for certain classes of unbounded distributions. As intermediate results, we proved concentration inequalities for $k$-NN distances and bounds on the expectations of statistics of $k$-NN distances. We hope these results and methods may lead to convergence rates for the widely used KSG mutual information estimator, or to generalize convergence rates for other estimators of entropy and related functionals to unbounded distributions.

## Acknowledgements

This material is based upon work supported by a National Science Foundation Graduate Research Fellowship to the first author under Grant No. DGE-1252522.

## References

Adami, C. Information theory in molecular biology. *Physics of Life Reviews*, 1:3–22, 2004.

Aghagolzadeh, M., Soltanian-Zadeh, H., Araabi, B., and Aghagolzadeh, A. A hierarchical clustering based on mutual information maximization. In *in Proc. of IEEE International Conference on Image Processing*, pp. 277–280, 2007.

Alemany, P. A. and Zanette, D. H. Fractal random walks from a variational formalism for Tsallis entropies. *Phys. Rev. E*, 49(2):R956–R958, Feb 1994. doi: 10.1103/PhysRevE.49.R956.

[9] Significantly milder conditions than Lipschitz continuity suffice, but are difficult to state here due to space limitations.

Baccetti, Valentina and Visser, Matt. Infinite shannon entropy. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(04):P04010, 2013.

Biau, Gérard and Devroye, Luc. Entropy estimation. In *Lectures on the Nearest Neighbor Method*, pp. 75–91. Springer, 2015.

Birge, L. and Massart, P. Estimation of integral functions of a density. *A. Statistics*, 23:11–29, 1995.

Chai, B., Walther, D. B., Beck, D. M., and Fei-Fei, L. Exploring functional connectivity of the human brain using multivariate information analysis. In *NIPS*, 2009.

Chaudhuri, Kamalika and Dasgupta, Sanjoy. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pp. 3437–3445, 2014.

Evans, D. A law of large numbers for nearest neighbor statistics. In *Proceedings of the Royal Society*, volume 464, pp. 3175–3192, 2008.

Evans, Dafydd, Jones, Antonia J, and Schmidt, Wolfgang M. Asymptotic moments of near–neighbour distance distributions. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 458, pp. 2839–2849. The Royal Society, 2002.

Gao, Shuyang, Steeg, Greg Ver, and Galstyan, Aram. Estimating mutual information by local gaussian approximation. *arXiv preprint arXiv:1508.00536*, 2015a.

Gao, Shuyang, Ver Steeg, Greg, and Galstyan, Aram. Efficient estimation of mutual information for strongly dependent variables. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 277–286, 2015b.

Goria, M. N., Leonenko, N. N., Mergel, V. V., and Inverardi, P. L. Novi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparametric Statistics*, 17:277–297, 2005.

Hero, A. O., Ma, B., Michel, O., and Gorman, J. Alpha-divergence for classification, indexing and retrieval, 2002a. Communications and Signal Processing Laboratory Technical Report CSPL-328.

Hero, A. O., Ma, B., Michel, O. J. J., and Gorman, J. Applications of entropic spanning graphs. *IEEE Signal Processing Magazine*, 19(5):85–95, 2002b.

Hlaváckova-Schindler, K., Paluŝb, M., Vejmelkab, M., and Bhattacharya, J. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441:1–46, 2007.

Hulle, M. M. Van. Constrained subspace ICA based on mutual information optimization directly. *Neural Computation*, 20:964–973, 2008.

Kandasamy, Kirthevasan, Krishnamurthy, Akshay, Poczos, Barnabas, Wasserman, Larry, et al. Nonparametric von mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems*, pp. 397–405, 2015.

Kozachenko, L. F. and Leonenko, N. N. A statistical estimate for the entropy of a random vector. *Problems of Information Transmission*, 23:9–16, 1987.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Phys. Rev. E*, 69:066138, 2004.

Krishnamurthy, A., Kandasamy, K., Poczos, B., and Wasserman, L. Nonparametric estimation of renyi divergence and friends. In *International Conference on Machine Learning (ICML)*, 2014.

Kybic, J. Incremental updating of nearest neighbor-based high-dimensional entropy estimation. In *Proc. Acoustics, Speech and Signal Processing*, 2006.

Learned-Miller, E. G. and Fisher, J. W. ICA using spacings estimates of entropy. *J. Machine Learning Research*, 4: 1271–1295, 2003.

Leonenko, N., Pronzato, L., and Savani, V. A class of Rényi information estimators for multidimensional densities. *Annals of Statistics*, 36(5):2153–2182, 2008.

Lewi, J., Butera, R., and Paninski, L. Real-time adaptive information-theoretic optimization of neurophysiology experiments. In *Advances in Neural Information Processing Systems*, volume 19, 2007.

Liu, H., Lafferty, J., and Wasserman, L. Exponential concentration inequality for mutual information estimation. In *Neural Information Processing Systems (NIPS)*, 2012.

Moon, Kevin R and Hero, Alfred O. Ensemble estimation of multivariate f-divergence. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pp. 356–360. IEEE, 2014.

Pál, D., Póczos, B., and Szepesvári, Cs. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Proceedings of the Neural Information Processing Systems*, 2010.

Peng, H. and Dind, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans On Pattern Analysis and Machine Intelligence*, 27, 2005.

Pérez-Cruz, F. Estimation of information theoretic measures for continuous random variables. In *Advances in Neural Information Processing Systems 21*, 2008.

Póczos, B. and Lőrincz, A. Independent subspace analysis using geodesic spanning trees. In *ICML*, pp. 673–680, 2005.

Póczos, B. and Lőrincz, A. Identification of recurrent neural networks by Bayesian interrogation techniques. *J. Machine Learning Research*, 10:515–554, 2009.

Shan, C., Gong, S., and Mcowan, P. W. Conditional mutual information based boosting for facial expression recognition. In *British Machine Vision Conference (BMVC)*, 2005.

Singh, S. and Poczos, B. Exponential concentration of a density functional estimator. In *Neural Information Processing Systems (NIPS)*, 2014a.

Singh, S. and Poczos, B. Generalized exponential concentration inequality for Rényi divergence estimation. In *International Conference on Machine Learning (ICML)*, 2014b.

Sricharan, K., Raich, R., and Hero, A. Empirical estimation of entropy functionals with confidence. Technical Report, http://arxiv.org/abs/1012.4188, 2010.

Sricharan, K., Wei, D., and Hero, A. Ensemble estimators for multivariate entropy estimation, 2012. http://arxiv.org/abs/1203.5829.

Szabó, Z., Póczos, B., and Lőrincz, A. Undercomplete blind subspace deconvolution. *J. Machine Learning Research*, 8:1063–1095, 2007.

Tsybakov, A. B. and van der Meulen, E. C. Root-$n$ consistent estimators of entropy for densities with unbounded support. *Scandinavian J. Statistics*, 23:75–83, 1996.

Wang, Q., Kulkarni, S.R., and Verdú, S. Divergence estimation for multidimensional densities via $k$-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5), 2009.

Wolsztynski, E., Thierry, E., and Pronzato, L. Minimum-entropy estimation in semi-parametric models. *Signal Process.*, 85(5):937–949, 2005. ISSN 0165-1684. doi: http://dx.doi.org/10.1016/j.sigpro.2004.11.028.