Two-Sided Exponential Concentration Bounds for Bayes Error Rate and Shannon Entropy

Jean Honorio JHONORIO@CSAIL.MIT.EDU
CSAIL, MIT, Cambridge, MA 02139, USA

Tommi Jaakkola TOMMI@CSAIL.MIT.EDU
CSAIL, MIT, Cambridge, MA 02139, USA

Abstract

We provide a method that approximates the Bayes error rate and the Shannon entropy with high probability. The Bayes error rate approximation makes possible to build a classifier that polynomially approaches Bayes error rate. The Shannon entropy approximation provides provable performance guarantees for learning trees and Bayesian networks from continuous variables. Our results rely on some reasonable regularity conditions of the unknown probability distributions, and apply to bounded as well as unbounded variables.

1. Introduction

Classification is arguably one of the most well-studied problems in machine learning. This includes the proposal of novel classification algorithms, and the study of generalization bounds (and sample complexity). Intuitively speaking, generalization bounds provide the rate at which the *expected risk* of the best classifier (chosen from some family) for a finite sample, approaches the expected risk of the best classifier (from the same family) for an infinite sample.

Less attention has been given to the approximation of the optimal Bayes error rate, which requires nonparametric methods. Some of the notable exceptions are the study of asymptotic universal Bayes consistency for weighted-average plug-in classifiers (Stone, 1977) (a class that contains for instance k-nearest neighbors) as well as for Parzen windows (Fralick & Scott, 1971).

It is well known that without any further assumption on the probability distributions, no rate-of-convergence results can be obtained (Antos et al., 1999). Given this, several authors have considered regularity conditions in the form of Lipschitz continuity. The Lipschitz continuity assumption upper-bounds the change of a function with respect to a given parameter. Under the assumption of Lipschitz *posterior* probability (i.e. $\mathbb{P}(y = 1|x)$ Lipschitz with respect to x), (Drakopoulos, 1995; Nock & Sebban, 2001; Györfi, 1981; Kulkarni & Posner, 1995) provide generalization bounds for k-nearest neighbors while (Kohler & Krzyżak, 2006) provides generalization bounds average plug-in classifiers. However the generalization bounds in (Drakopoulos, 1995; Nock & Sebban, 2001; Kulkarni & Posner, 1995) do not imply Bayes consistency since the analysis considers "twice Bayes error".

Our first goal is the approximation of Bayes error rate between two unknown distributions, from a given training set. More specifically, we are interested on two-sided exponential concentration bounds. As a byproduct, we obtain a classifier that is Bayes consistent with provable finite-sample rates. Note that while the rates in (Györfi, 1981; Kohler & Krzyżak, 2006) show generalization bounds for Bayes consistent classifiers, they do not provide a recipe for producing the desired bounds. The main reason is that the bounds are one-sided and in *expected risk*. While in practice resampling (Jackknife, bootstrapping and cross-validation) is used to assess classifier performance in finite samples, it is unclear how resampling can produce the bounds we are interested on. In this paper we prove that, given N samples, with probability at least $1-\delta$, we can approximate the Bayes error with a finite-sample rate of $\mathcal{O}(N^{-1/4}\log^{1/2}N,\ \log(1/\delta))$. To the best of our knowledge, this is the first exponential inequality for Bayes error rate estimation.

We assume Lipschitz continuity as the regularity condition for the probability distributions. This is a rea-

Proceedings of the 30

sonable assumption since universal rate-of-convergence results are not possible (Antos et al., 1999).

Our second goal is the approximation of Shannon entropy of an unknown distribution, from a given training set. Again, we are interested on two-sided exponential concentration bounds. Our main motivation is to provide generalization bounds for learning the structure of trees and Bayesian networks from continuous variables. This is due to the fact that a concentration bound for the entropy makes available a concentration bound for related information theoretic measures, such as mutual information and conditional entropy.

Several asymptotic consistency results are available for Shannon entropy estimation. Techniques such as kernel density estimation (Ahmad & Lin, 1976; Eggermont & LaRiccia, 1999; Paninski & Yajima, 2008) and spacings (Van Es, 1992; Tsybakov & Van der Meulen, 1996; Wang et al., 2005) have been previously proposed. The use of k-nearest neighbors was proposed by (Pérez-Cruz, 2008; Póczos & Schneider, 2012) for estimating Shannon entropy and other related information theoretic measures. Convex risk minimization for estimating divergences was analyzed in (Nguyen et al., 2010). The use of nearest-neighbor graphs was proposed by (Pál et al., 2010) for Rényi entropy estimation with finite-sample rates for Lipschitz distributions. We refer the interested reader to the survey articles (Beirlant et al., 1997; Paninski, 2003) for more discussion.

Very recently, a kernel estimator of the Shannon entropy with exponential concentration bounds was proposed by (Liu et al., 2012). The authors focused on distributions of the Hölder class (which is a subset of the Lipschitz class studied here). Moreover, the estimator proposed by (Liu et al., 2012) requires positiveness of the density function, differentiability, vanishing first-order derivative in the boundaries, bounded third-order derivative[1], and prior knowledge of lower/upper bounds of the density function. In contrast, our results also apply to probability distributions with regions of zero-probability, nonsmooth density functions (discontinuous derivative), arbitrary behavior in the boundaries, and we do not require prior knowledge of lower/upper bounds. Additionally, our results apply to both bounded and unbounded variables, unlike (Liu et al., 2012).

As expected, given that our results are relatively more general, our finite-sample rate for the entropy is slower than the rate of (Liu et al., 2012). More specifically, given N samples, with probability at least $1-\delta$, we approximate the Shannon entropy with a finite-sample rate of $\mathcal{O}(N^{-1/4}\log^{3/2}N,\ \log(1/\delta))$. In contrast, (Liu et al., 2012) provides a rate of $\mathcal{O}(N^{-1/2},\ \log(1/\delta))$.

In this paper, we propose the same framework for approximating both the Bayes error rate and the Shannon entropy. Our method is based on splitting the variable domain into bins. Then, we compute empirical probabilities for each bin. Finally, we produce an empirical estimate of the statistical measure (Bayes error and entropy). Interestingly, our method has provable two-sided exponential concentration bounds.

For clarity of exposition, we present our proofs for the one-dimensional case. Given that the extension to several dimensions is trivial, we defer this topic until Section 5.

2. Preliminaries

In this paper, we assume Lipschitz continuity as the regularity condition for the probability distributions. Our Lipschitz assumption upper-bounds the rate of change of the density function. A kernel estimator of the Shannon entropy for the Hölder class was proposed by (Liu et al., 2012). The relationship between the Hölder and the Lipschitz class might not seem immediately obvious. In this section, we show that the Hölder class is a subset of the Lipschitz class.

Next, we present our Lipschitz continuity assumption.

Definition 1. *A probability distribution $\mathcal{P}=p(\cdot)$ is called K-Lipschitz continuous, if its probability density function $p(x)$ is Lipschitz continuous with constant $K \geq 0$, that is:*

$$(\forall x_1, x_2)\ |p(x_1)-p(x_2)| \leq K|x_1-x_2| \qquad (1)$$

or equivalently for a continuous function $p(x)$:

$$(\forall x)\ \left|\frac{\partial p}{\partial x}(x)\right| \leq K \qquad (2)$$

In Table 1 we provide some few examples of very well known parametric distributions that are Lipschitz continuous. This includes the Gaussian, Laplace, Cauchy and Gumbel distributions. Note that by properties of Lipschitz continuity, any mixture of Lipschitz distributions is also Lipschitz.

Next, we show that the Hölder class analyzed in (Liu et al., 2012) is a subset of the Lipschitz class.

Theorem 2. *Let $f(x)$ be a function that belongs to the second-order Holdër class with domain $x \in [0;1]$ and vanishing derivatives in the boundaries. That is, there is constant $L \geq 0$ such that:*

$$(\forall x, u)\ \left|f(x+u)-f(x)-\frac{\partial f}{\partial x}(x)u\right| \leq Lu^2 \qquad (3)$$

[1] See eq.(3.36) of (Liu et al., 2012)

Table 1. Some Lipschitz continuous distributions, their density function $p(x)$ and Lipschitz constant K.

Distrib.	$p(x)$	**Conditions**	K
Gaussian	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}$	$\sigma > 0$	$\frac{1}{\sigma^2\sqrt{2\pi e}}$
Laplace	$\frac{1}{2\lambda}e^{-\frac{\lvert x\rvert}{\lambda}}$	$\lambda > 0$	$\frac{1}{2\lambda^2}$
Cauchy	$\frac{1}{\pi}(\frac{\lambda}{x^2+\lambda^2})$	$\lambda > 0$	$\frac{3\sqrt{3}}{8\pi\lambda^2}$
Gumbel	$\frac{1}{\lambda}e^{-x/\lambda-e^{-x/\lambda}}$	$\lambda > 0$	$\frac{2+\sqrt{5}}{\lambda^2}e^{-\frac{3+\sqrt{5}}{2}}$
Mixture	$\sum_i \alpha_i p_i(x)$	K_i-Lipschitz $p_i(\cdot)$, $\alpha_i > 0$, $\sum_i \alpha_i = 1$	$\sum_i \alpha_i K_i$

and either $\frac{\partial f}{\partial x}(0) = 0$ or $\frac{\partial f}{\partial x}(1) = 0$. Then, the function $f(x)$ is Lipschitz continuous with constant $K = 3L$.

Proof. Assume that $\frac{\partial f}{\partial x}(0) = 0$ holds. By setting $x = 0$ in eq.(3) and by the previous assumption, we have $|f(u) - f(0)| \le Lu^2$. This is a Lipschitz condition at $x = 0$ since $-1 \le u \le 1$ and therefore $Lu^2 \le L|u|$.

Next, we focus at $x > 0$. By the reverse triangle inequality in eq.(3) and by setting $u = -x$, we have:

$$\left|\frac{\partial f}{\partial x}(x)u\right| \le |f(x+u) - f(x)| + Lu^2$$

$$\left|\frac{\partial f}{\partial x}(x)\right| x \le |f(0) - f(x)| + Lx^2 \le Lx^2 + Lx^2 = 2Lx^2$$

Therefore, $\left|\frac{\partial f}{\partial x}(x)\right| \le 2Lx$ for $x > 0$. Recall that $x \le 1$ and $-1 \le u \le 1$. By the reverse triangle inequality in eq.(3), we have:

$$\begin{aligned}|f(x+u) - f(x)| &\le \left|\frac{\partial f}{\partial x}(x)u\right| + Lu^2 \\ &= \left|\frac{\partial f}{\partial x}(x)\right||u| + Lu^2 \\ &\le 2Lx|u| + Lu^2 \\ &\le 2L|u| + L|u| \\ &= 3L|u|\end{aligned}$$

A similar proof can be done for $\frac{\partial f}{\partial x}(1) = 0$ instead. □

3. Bayes error rate

In this section, we present our two-sided exponential concentration bound for Bayes error. We show results for both bounded and unbounded random variables.

3.1. Bounded Domain

In this paper, we approximate statistical measures by splitting the variable domain into bins. To this end, we first provide a general concentration bound for the probability that a variable falls inside a specific bin.

Proposition 3. *Let x be a continuous random variable with domain $\mathcal{D}$. Let d be a subset of $\mathcal{D}$. Let $\mathcal{P} = p(\cdot)$ be a probability distribution with probability density function $p(x)$. Let the true probability $\bar{p}_d = \mathbb{P}_{\mathcal{P}}[x \in d] = \int_{x\in d} p(x)$. Given N i.i.d. samples $x_1, \dots, x_N$ from $\mathcal{P}$. Let the empirical probability $\widehat{p}_d = \widehat{\mathbb{P}}_{\mathcal{P}}[x \in d] = \frac{1}{N}\sum_n 1[x_n \in d]$. The following high probability statement holds:*

$$\mathbb{P}[|\widehat{p}_d - \bar{p}_d| > \varepsilon] \le 2e^{-2N\varepsilon^2} \qquad (4)$$

Proof. Let $z_n \equiv 1[x_n \in d]$. We have:

$$\mathbb{E}_{\mathcal{P}}[z_n] = \mathbb{E}_{\mathcal{P}}[1[x_n \in d]] = \mathbb{P}_{\mathcal{P}}[x_n \in d] = \bar{p}_d$$

Note that $z_n \in \{0, 1\}$ and $\widehat{p}_d = \frac{1}{N}\sum_n z_n$. By Hoeffding's inequality, we prove our claim. □

Next, we provide bounds for the Bayes error rate inside a specific bin. These bounds depend only on the empirical probabilities, the Lipschitz constant K as well as the bin size.

Lemma 4. *Let x be a continuous random variable with compact domain $\mathcal{D}$. Let d be a compact subset of $\mathcal{D}$. Let $\mathcal{P} = p(\cdot)$ and $\mathcal{Q} = q(\cdot)$ be two K-Lipschitz continuous distributions. Let the true probabilities $\bar{p}_d = \mathbb{P}_{\mathcal{P}}[x \in d] = \int_{x\in d} p(x)$ and $\bar{q}_d = \mathbb{P}_{\mathcal{Q}}[x \in d] = \int_{x\in d} q(x)$. The true Bayes error rate on d, given by $\mathcal{B}_d(\mathcal{P}, \mathcal{Q}) = \int_{x\in d} \min(p(x), q(x))$ is bounded as follows:*

$$\min(\bar{p}_d, \bar{q}_d) - \frac{K|d|^2}{2} \le \mathcal{B}_d(\mathcal{P}, \mathcal{Q}) \le \min(\bar{p}_d, \bar{q}_d) \qquad (5)$$

where $|d|$ is the size of d.

Proof. Given that $p(x)$ is Lipschitz continuous with constant K, that we know the integral $\bar{p}_d = \int_{x\in d} p(x)$, and that the maximum change in x is $|d|$. We have $p(x) \ge \frac{\bar{p}_d}{|d|} - Z$ and $q(x) \ge \frac{\bar{q}_d}{|d|} - Z$ where $Z \equiv \frac{K|d|}{2}$.

For finding a lower bound, note that:

$$\begin{aligned}\mathcal{B}_d(\mathcal{P}, \mathcal{Q}) &= \int_{x\in d} \min(p(x), q(x)) \\ &\ge \int_{x\in d} \min\left(\frac{\bar{p}_d}{|d|} - Z, \frac{\bar{q}_d}{|d|} - Z\right) \\ &= \int_{x\in d} \left(\frac{1}{|d|}\min(\bar{p}_d, \bar{q}_d) - Z\right) \\ &= \min(\bar{p}_d, \bar{q}_d) - |d|Z\end{aligned}$$

By replacing Z, we prove that the lower bound holds.

For finding an upper bound, note that given any two functions $p(x)$ and $q(x)$, we have $\min(p(x), q(x)) \le p(x)$. Therefore $\mathcal{B}_d(\mathcal{P}, \mathcal{Q}) = \int_{x\in d} \min(p(x), q(x)) \le \int_{x\in d} p(x) = \bar{p}_d$. Similarly, $\mathcal{B}_d(\mathcal{P}, \mathcal{Q}) \le \bar{q}_d$, and we prove that the upper bound holds. □

Armed with the previous results, we present our first main contribution. That is, we show our two-sided exponential concentration bound for the Bayes error rate of a bounded variable.

Theorem 5. *Let x be a continuous random variable with compact domain $\mathcal{D}$. Let $\mathcal{P} = p(\cdot)$ and $\mathcal{Q} = q(\cdot)$ be two K-Lipschitz continuous distributions. Given N i.i.d. samples $x_1, \dots, x_N$ from $\mathcal{P}$ and $y_1, \dots, y_N$ from $\mathcal{Q}$. We divide $\mathcal{D}$ into $T = N^{1/4}$ compact nonoverlapping equally-sized subsets $d_1, \dots, d_T$. Let the empirical probabilities $\widehat{p}_t = \widehat{\mathbb{P}}_{\mathcal{P}}[x \in d_t] = \frac{1}{N}\sum_n 1[x_n \in d_t]$ and $\widehat{q}_t = \widehat{\mathbb{P}}_{\mathcal{Q}}[x \in d_t] = \frac{1}{N}\sum_n 1[y_n \in d_t]$. Let the empirical Bayes error rate $\widehat{\mathcal{B}}(\mathcal{P}, \mathcal{Q}) = \sum_t \min(\widehat{p}_t, \widehat{q}_t)$. The true Bayes error rate $\mathcal{B}(\mathcal{P}, \mathcal{Q}) = \int_{x \in \mathcal{D}} \min(p(x), q(x))$ is bounded as follows with probability at least $1 - \delta$:*

$$\widehat{\mathcal{B}}(\mathcal{P}, \mathcal{Q}) - \varepsilon^{(1)}_{N\delta} - \varepsilon^{(2)}_{NK\mathcal{D}} \leq \mathcal{B}(\mathcal{P}, \mathcal{Q}) \leq \widehat{\mathcal{B}}(\mathcal{P}, \mathcal{Q}) + \varepsilon^{(1)}_{N\delta} \tag{6}$$

where $\varepsilon^{(1)}_{N\delta} = \frac{1}{N^{1/4}}\sqrt{\frac{1}{8}\log N + \frac{1}{2}\log\frac{4}{\delta}}$, $\varepsilon^{(2)}_{NK\mathcal{D}} = \frac{K|\mathcal{D}|^2}{2N^{1/4}}$ and $|\mathcal{D}|$ is the size of $\mathcal{D}$.

Proof. Let the true probabilities $\bar{p}_t = \mathbb{P}_{\mathcal{P}}[x \in d_t] = \int_{x \in d_t} p(x)$ and $\bar{q}_t = \mathbb{P}_{\mathcal{Q}}[x \in d_t] = \int_{x \in d_t} q(x)$. By Proposition 3 and the union bound, we have $\mathbb{P}[(\exists t)\ |\widehat{p}_t - \bar{p}_t| > \varepsilon \text{ or } |\widehat{q}_t - \bar{q}_t| > \varepsilon] \leq 4Te^{-2N\varepsilon^2} = \delta$. By solving for ε, we have $\varepsilon = \sqrt{\frac{1}{2N}\log\frac{4T}{\delta}}$.

Let $\mathcal{B}_t(\mathcal{P}, \mathcal{Q}) = \int_{x \in d_t} \min(p(x), q(x))$, we have $\mathcal{B}(\mathcal{P}, \mathcal{Q}) = \sum_t \mathcal{B}_t(\mathcal{P}, \mathcal{Q})$. By Lemma 4 and since $|\widehat{p}_t - \bar{p}_t| \leq \varepsilon$ and $|\widehat{q}_t - \bar{q}_t| \leq \varepsilon$, we have:

$$\min(\bar{p}_t, \bar{q}_t) - \frac{K|d_t|^2}{2} \leq \mathcal{B}_t(\mathcal{P}, \mathcal{Q}) \leq \min(\bar{p}_t, \bar{q}_t)$$

$$\min(\widehat{p}_t, \widehat{q}_t) - \varepsilon - \frac{K|d_t|^2}{2} \leq \mathcal{B}_t(\mathcal{P}, \mathcal{Q}) \leq \min(\widehat{p}_t, \widehat{q}_t) + \varepsilon$$

By summing the latter expression for all t and by assuming $|d_t| = \frac{|\mathcal{D}|}{T}$, we have:

$$\widehat{\mathcal{B}}(\mathcal{P}, \mathcal{Q}) - T\varepsilon - \frac{K|\mathcal{D}|^2}{2T} \leq \mathcal{B}(\mathcal{P}, \mathcal{Q}) \leq \widehat{\mathcal{B}}(\mathcal{P}, \mathcal{Q}) + T\varepsilon$$

Finally, we replace ε and set $T = N^{1/4}$. □

3.2. Unbounded Domain

In order to extend our previous result from bounded variables to unbounded variables, we assume a very general concentration inequality. That is, we assume $\mathbb{P}[x \notin \mathcal{D}_\gamma] \leq \gamma$. Such tail bounds are ubiquitous in the machine learning and statistics literature. In Table 2 we provide some few examples: distributions with finite variance, finite m-th moment (both by Chebyshev's inequality), and sub-Gaussian distributions. When several samples are available, we can use the union bound. That is, given N i.i.d. samples $x_1, \dots, x_N$, we have $\mathbb{P}[(\exists n)\ x_n \notin \mathcal{D}_{\gamma/N}] \leq \gamma$.

Table 2. Some distributions for unbounded variables with tail bounds of the form $\mathbb{P}[x \notin \mathcal{D}_\gamma] \leq \gamma$. We include the concentration inequality and the size of the subdomain $|D_\gamma|$.

Distrib.	**Concentration**	**Conditions**	$\lvert\mathcal{D}_\gamma\rvert/2$
Finite variance	$\mathbb{P}[\lvert x\rvert > \varepsilon] \leq \frac{\sigma^2}{\varepsilon^2}$	$\sigma^2 = \mathbb{E}[x^2]$	$\sigma\sqrt{\frac{1}{\gamma}}$
Finite m-th moment	$\mathbb{P}[\lvert x\rvert > \varepsilon] \leq \frac{\lambda^m}{\varepsilon^m}$	$\lambda^m = \mathbb{E}[x^m]$	$\lambda\sqrt[m]{\frac{1}{\gamma}}$
Sub-Gaussian	$\mathbb{P}[\lvert x\rvert > \varepsilon] \leq 2e^{-\frac{\varepsilon^2}{2\sigma^2}}$	$\sigma > 0$	$\sigma\sqrt{2\log\frac{2}{\gamma}}$

Next, we present our two-sided exponential concentration bound for the Bayes error rate of an unbounded variable. The result is given for general tail bounds. For specific distributions, we can plug-in the size of the subdomain $|D_\gamma|$ given in Table 2.

Theorem 6. *Let x be a continuous random variable with domain $\mathbb{R}$. Assume that with high probability, x belongs to a compact set $\mathcal{D}_\gamma$, that is $\mathbb{P}_{\mathcal{P}}[x \notin \mathcal{D}_\gamma] = \int_{x \notin \mathcal{D}_\gamma} p(x) \leq \gamma$ and $\mathbb{P}_{\mathcal{Q}}[x \notin \mathcal{D}_\gamma] = \int_{x \notin \mathcal{D}_\gamma} q(x) \leq \gamma$. Under the same conditions of Theorem 5, the true Bayes error rate $\mathcal{B}(\mathcal{P}, \mathcal{Q}) = \int_{x \in \mathbb{R}} \min(p(x), q(x))$ is bounded as follows with probability at least $1 - \delta$:*

$$\widehat{\mathcal{B}}(\mathcal{P}, \mathcal{Q}) - \varepsilon^{(1)}_{N\delta} - \varepsilon^{(2)}_{NK\mathcal{D}_\gamma} \leq \mathcal{B}(\mathcal{P}, \mathcal{Q}) \leq \widehat{\mathcal{B}}(\mathcal{P}, \mathcal{Q}) + \varepsilon^{(1)}_{N\delta} + \gamma \tag{7}$$

where $\widehat{\mathcal{B}}(\mathcal{P}, \mathcal{Q})$, $\varepsilon^{(1)}_{N\delta}$ and $\varepsilon^{(2)}_{NK\mathcal{D}_\gamma}$ are defined as in Theorem 5.

Proof. Note that $\mathcal{B}(\mathcal{P}, \mathcal{Q}) = \int_{x \in \mathcal{D}_\gamma} \min(p(x), q(x)) + \int_{x \notin \mathcal{D}_\gamma} \min(p(x), q(x))$. Theorem 5 provides bounds for the first term. It suffices to bound the second term.

For finding a lower bound, note that both $p(x) \geq 0$ and $q(x) \geq 0$, therefore $\int_{x \notin \mathcal{D}_\gamma} \min(p(x), q(x)) \geq 0$.

For finding an upper bound, note that given any two functions $p(x)$ and $q(x)$, we have $\min(p(x), q(x)) \leq p(x)$. Therefore $\int_{x \notin \mathcal{D}_\gamma} \min(p(x), q(x)) \leq \int_{x \notin \mathcal{D}_\gamma} p(x) \leq \gamma$. The same bound is obtained from a similar argument with $q(x)$. □

4. Shannon entropy

In this section, we present our two-sided exponential concentration bound for Shannon entropy, for both bounded and unbounded random variables.

4.1. Bounded Domain

First, we provide three general inequalities that will be useful for our purposes. In this paper, we approximate statistical measures by splitting the variable domain into bins. For our specific analysis, we need to bound the change in Shannon entropy at the bin level, as well as at each point. In what follows we present an inequality for bounding the entropy change at the bin level. That is, we consider "interval" probabilities $\bar{p} = \int_{x \in d} p(x) \in [0; 1]$.

Proposition 7. *For $0 \leq \widehat{p}, \bar{p} \leq 1$ such that $|\widehat{p} - \bar{p}| \leq \varepsilon$, we have:*

$$|\widehat{p} \log \widehat{p} - \bar{p} \log \bar{p}| \leq \varepsilon \log \tfrac{e}{\varepsilon} \tag{8}$$

Proof. Note that the bound holds trivially for $\varepsilon = 0$. Next, we focus at $\varepsilon > 0$.

Without loss of generality, assume $0 \leq \bar{p} - \widehat{p} = r \leq 1$. (The same holds for $0 \leq \widehat{p} - \bar{p} = r \leq 1$.) We have:

$$\begin{aligned}
|\widehat{p} \log \widehat{p} - \bar{p} \log \bar{p}| &= |\widehat{p} \log \widehat{p} - \widehat{p} \log \bar{p} + \widehat{p} \log \bar{p} - \bar{p} \log \bar{p}| \\
&\leq |\widehat{p} \log \widehat{p} - \widehat{p} \log \bar{p}| + |\widehat{p} \log \bar{p} - \bar{p} \log \bar{p}| \\
&= \widehat{p} | \log \widehat{p} - \log \bar{p}| - |\widehat{p} - \bar{p}| \log \bar{p} \\
&= \widehat{p} \log \tfrac{\bar{p}}{\widehat{p}} - |\widehat{p} - \bar{p}| \log \bar{p} \\
&\leq \widehat{p} (\tfrac{\bar{p}}{\widehat{p}} - 1) - |\widehat{p} - \bar{p}| \log \bar{p} \\
&= \bar{p} - \widehat{p} - |\widehat{p} - \bar{p}| \log \bar{p} \\
&= r - r \log \bar{p} \\
&\leq r(1 - \log r) \\
&= r \log \tfrac{e}{r}
\end{aligned}$$

where we used $\log z \leq z - 1$ for $z = \frac{\bar{p}}{\widehat{p}}$.

Let $f(r) = r \log \frac{e}{r}$. So far we proved that $|\widehat{p} - \bar{p}| = r \Rightarrow |\widehat{p} \log \widehat{p} - \bar{p} \log \bar{p}| \leq f(r)$. Note that $f(r)$ is increasing in $r \in [0; 1]$. Therefore, for all r such that $0 \leq r \leq \varepsilon \leq 1 \Rightarrow f(r) \leq f(\varepsilon)$, which proves our claim. □

Next, we present an inequality for bounding the entropy change at each point. Recall that continuous distributions allow for point densities greater than one. That is, $p(x) \in [0; +\infty)$.

Proposition 8. *For $p \geq 0$ and $\varepsilon \geq 0$:*

$$-(p - \varepsilon) \log (p - \varepsilon) + p \log p \geq \varepsilon \log \varepsilon \tag{9}$$

$$-(p + \varepsilon) \log (p + \varepsilon) + p \log p \geq \varepsilon \log \tfrac{1}{e(p + \varepsilon)} \tag{10}$$

Proof. For proving the first inequality, by reasonably assuming $p \geq \varepsilon$:

$$\begin{aligned}
-(p - \varepsilon) \log (p - \varepsilon) + p \log p &= \varepsilon \log (p - \varepsilon) + p \log \tfrac{p}{p - \varepsilon} \\
&\geq \varepsilon \log (p - \varepsilon) + \varepsilon \log \tfrac{p}{p - \varepsilon} \\
&= \varepsilon \log p \\
&\geq \varepsilon \log \varepsilon
\end{aligned}$$

For proving the second inequality:

$$\begin{aligned}
-(p + \varepsilon) \log (p + \varepsilon) + p \log p &= \varepsilon \log \tfrac{1}{p + \varepsilon} - p \log (1 + \tfrac{\varepsilon}{p}) \\
&\geq \varepsilon \log \tfrac{1}{p + \varepsilon} - \varepsilon \\
&= \varepsilon \log \tfrac{1}{e(p + \varepsilon)}
\end{aligned}$$

where $-p \log (1 + \frac{\varepsilon}{p}) \geq -\varepsilon$ follows from $\log (1 + z) \leq z$ for $z = \frac{\varepsilon}{p}$. □

In what follows we extend the *log sum inequality* from discrete variables to continuous variables. In this case, straightforward application of the Jensen's inequality is not possible. Therefore, we perform a reparametrization of the probability distribution.

Proposition 9 (Log integral inequality)**.** *Given a nonnegative function $p(x)$ and a positive function $q(x)$, both with domain d. We have:*

$$\int_{x \in d} p(x) \log \frac{p(x)}{q(x)} \geq \int_{x \in d} p(x) \log \frac{\int_{x \in d} p(x)}{\int_{x \in d} q(x)} \tag{11}$$

Proof. Let $\delta(z)$ be the Dirac delta function and $r(x) = p(x)/q(x)$. Let the convex function $f(z) = z \log z$. We have:

$$\begin{aligned}
\textstyle\int_{x \in d} p(x) \log \frac{p(x)}{q(x)} &= \textstyle\int_{x \in d} q(x) f(r(x)) \\
&= \textstyle\int_{x \in d} q(x) \int_{z \in [0; +\infty)} f(z) \delta(z - r(x)) \\
&= \textstyle\int_{z \in [0; +\infty)} f(z) h(z)
\end{aligned}$$

where $h(z) = \int_{x \in d} q(x) \delta(z - r(x))$.

Note that $g(z) = h(z) / \int_{x \in d} q(x)$ is a probability density function. We can write:

$$\begin{aligned}
\mathbb{E}_{\mathcal{G}}[z] \textstyle\int_{x \in d} q(x) &= \textstyle\int_{z \in [0; +\infty)} z h(z) \\
&= \textstyle\int_{z \in [0; +\infty)} z \int_{x \in d} q(x) \delta(z - r(x)) \\
&= \textstyle\int_{x \in d} q(x) \int_{z \in [0; +\infty)} z \delta(z - r(x)) \\
&= \textstyle\int_{x \in d} q(x) r(x) \\
&= \textstyle\int_{x \in d} p(x)
\end{aligned}$$

Finally, we have:

$$\begin{aligned}
\textstyle\int_{x \in d} p(x) \log \frac{p(x)}{q(x)} &= \mathbb{E}_{\mathcal{G}}[f(z)] \textstyle\int_{x \in d} q(x) \\
&\geq f(\mathbb{E}_{\mathcal{G}}[z]) \textstyle\int_{x \in d} q(x)
\end{aligned}$$

By replacing the $\mathbb{E}_{\mathcal{G}}[z]$, we prove our claim. □

Next, we provide bounds for the Shannon entropy inside a specific bin. These bounds depend only on the empirical probability, the Lipschitz constant K as well as the bin and domain size.

Lemma 10. *Let x be a continuous random variable with compact domain $\mathcal{D}$. Let d be a compact subset of $\mathcal{D}$. Let $\mathcal{P} = p(\cdot)$ be a K-Lipschitz continuous distribution. Let the true probability $\bar{p}_d = \mathbb{P}_{\mathcal{P}}[x \in d] = \int_{x \in d} p(x)$. The true Shannon entropy on d, given by $\mathcal{H}_d(\mathcal{P}) = -\int_{x \in d} p(x) \log p(x)$ is bounded as follows:*

$$-\bar{p}_d \log \frac{\bar{p}_d}{|d|} + \varepsilon^{(0)}_{K\mathcal{D}d} \leq \mathcal{H}_d(\mathcal{P}) \leq -\bar{p}_d \log \frac{\bar{p}_d}{|d|} \tag{12}$$

where $\varepsilon^{(0)}_{K\mathcal{D}d} = \frac{K|d|^2}{2} \log \min\left(\frac{K|d|}{2}, \frac{|\mathcal{D}|}{e+eK|\mathcal{D}|^2/2}\right)$ and $|d|$ is the size of d.

Proof. Given that $p(x)$ is Lipschitz continuous with constant K, that the know the integral $\bar{p}_d = \int_{x \in d} p(x)$, and that the maximum change in x is $|d|$. We have $p(x) = \frac{\bar{p}_d}{|d|} + z(x)$ where $|z(x)| \leq \frac{K|d|}{2} \equiv Z$.

For finding a lower bound, by Proposition 8 we have:

$$\begin{aligned}
&\mathcal{H}_d(\mathcal{P}) = \int_{x \in d} -p(x) \log p(x) \\
&\geq \int_{x \in d} \min_{|z(x)| \leq Z} -\left(\frac{\bar{p}_d}{|d|} + z(x)\right) \log \left(\frac{\bar{p}_d}{|d|} + z(x)\right) \\
&= \int_{x \in d} \min_{z(x) \in \{-Z, +Z\}} -\left(\frac{\bar{p}_d}{|d|} + z(x)\right) \log \left(\frac{\bar{p}_d}{|d|} + z(x)\right) \\
&\geq \int_{x \in d} \left(-\frac{\bar{p}_d}{|d|} \log \frac{\bar{p}_d}{|d|} + Z \log \min\left(Z, \frac{|\mathcal{D}|}{e + eK|\mathcal{D}|^2/2}\right)\right) \\
&= -\bar{p}_d \log \frac{\bar{p}_d}{|d|} + |d| Z \log \min\left(Z, \frac{|\mathcal{D}|}{e + eK|\mathcal{D}|^2/2}\right)
\end{aligned}$$

In this derivation, in order to lower-bound eq.(10), we used the fact that $p(x) \leq \frac{1}{|\mathcal{D}|} + \frac{K|\mathcal{D}|}{2} = \frac{1+K|\mathcal{D}|^2/2}{|\mathcal{D}|}$ given that $\int_{x \in \mathcal{D}} p(x) = 1$. By replacing Z, we prove that the lower bound holds.

For finding an upper bound, we apply the *log integral inequality* (Proposition 9) for the given $p(x)$ and a constructed $q(x) = 1$. Thus, $\int_{x \in d} q(x) = |d|$. □

Armed with the previous results, we present our second main contribution. That is, we show our two-sided exponential concentration bound for the Shannon entropy of a bounded variable.

Theorem 11. *Let x be a continuous random variable with compact domain $\mathcal{D}$. Let $\mathcal{P} = p(\cdot)$ be a K-Lipschitz continuous distribution. Given N i.i.d. samples $x_1, \dots, x_N$ from $\mathcal{P}$. We divide $\mathcal{D}$ into $T = N^{1/4}$ compact nonoverlapping equally-sized subsets $d_1, \dots, d_T$. Let the empirical probabilities $\widehat{p}_t = \widehat{\mathbb{P}}_{\mathcal{P}}[x \in d_t] = \frac{1}{N} \sum_n 1[x_n \in d_t]$. Let the empirical Shannon entropy $\widehat{\mathcal{H}}(\mathcal{P}) = -\sum_t \widehat{p}_t \log(\widehat{p}_t / |d_t|)$. The true Shannon entropy $\mathcal{H}(\mathcal{P}) = -\int_{x \in \mathcal{D}} p(x) \log p(x)$ is bounded as follows with probability at least $1 - \delta$:*

$$\widehat{\mathcal{H}}(\mathcal{P}) - \varepsilon^{(1)}_{N\delta\mathcal{D}} - \varepsilon^{(2)}_{NK\mathcal{D}} \leq \mathcal{H}(\mathcal{P}) \leq \widehat{\mathcal{H}}(\mathcal{P}) + \varepsilon^{(1)}_{N\delta\mathcal{D}} \tag{13}$$

where $\varepsilon^{(1)}_{N\delta\mathcal{D}} = \frac{1}{N^{1/4}} \sqrt{\frac{1}{128} \log N + \frac{1}{32} \log \frac{2}{\delta}} \times \left(\log 4(\frac{e}{|\mathcal{D}|})^4 + 3 \log N - 2 \log(\frac{1}{4} \log N + \log \frac{2}{\delta})\right)$, $\varepsilon^{(2)}_{NK\mathcal{D}} = \frac{K|\mathcal{D}|^2}{2N^{1/4}} \log \min\left(\frac{K|\mathcal{D}|}{2N^{1/4}}, \frac{|\mathcal{D}|}{e+eK|\mathcal{D}|^2/2}\right)$ and $|\mathcal{D}|$ is the size of $\mathcal{D}$.

Proof. Let the true probabilities $\bar{p}_t = \mathbb{P}_{\mathcal{P}}[x \in d_t] = \int_{x \in d_t} p(x)$. By Proposition 3 and the union bound, we have $\mathbb{P}[(\exists t)\ |\widehat{p}_t - \bar{p}_t| > \varepsilon] \leq 2Te^{-2N\varepsilon^2} = \delta$. By solving for ε, we have $\varepsilon = \sqrt{\frac{1}{2N} \log \frac{2T}{\delta}}$.

Let $\mathcal{H}_t(\mathcal{P}) = -\int_{x \in d_t} p(x) \log p(x)$, we have $\mathcal{H}(\mathcal{P}) = \sum_t \mathcal{H}_t(\mathcal{P})$. By Lemma 10, Proposition 7, by assuming $|d_t| \leq 1$, and since $|\widehat{p}_t - \bar{p}_t| \leq \varepsilon$, we have:

$$\begin{aligned}
&-\bar{p}_t \log \tfrac{\bar{p}_t}{|d_t|} + \varepsilon^{(0)}_{K\mathcal{D}d_t} \leq \mathcal{H}_t(\mathcal{P}) \leq -\bar{p}_t \log \tfrac{\bar{p}_t}{|d_t|} \\
&-\widehat{p}_t \log \tfrac{\widehat{p}_t}{|d_t|} \dots \\
&-\varepsilon \log \tfrac{e}{\varepsilon|d_t|} + \varepsilon^{(0)}_{K\mathcal{D}d_t} \leq \mathcal{H}_t(\mathcal{P}) \leq -\widehat{p}_t \log \tfrac{\widehat{p}_t}{|d_t|} + \varepsilon \log \tfrac{e}{\varepsilon|d_t|}
\end{aligned}$$

By summing the latter expression for all t and by assuming $|d_t| = \frac{|\mathcal{D}|}{T}$, we have:

$$\begin{aligned}
&\widehat{\mathcal{H}}(\mathcal{P}) - T\varepsilon \log \tfrac{eT}{\varepsilon|\mathcal{D}|} + \tfrac{K|\mathcal{D}|^2}{2T} \log \min\left(\tfrac{K|\mathcal{D}|}{2T}, \tfrac{|\mathcal{D}|}{e+eK|\mathcal{D}|^2/2}\right) \\
&\qquad \dots \leq \mathcal{H}(\mathcal{P}) \leq \widehat{\mathcal{H}}(\mathcal{P}) + T\varepsilon \log \tfrac{eT}{\varepsilon|\mathcal{D}|}
\end{aligned}$$

Finally, we replace ε and set $T = N^{1/4}$. □

4.2. Unbounded Domain

In order to extend our previous result from bounded variables to unbounded variables, we assume a very general concentration inequality as in Section 3.2. That is, we assume $\mathbb{P}[x \notin \mathcal{D}_\gamma] \leq \gamma$.

Next, we present our two-sided exponential concentration bound for the Shannon entropy of an unbounded variable. The result is given for general tail bounds. For specific distributions, we can plug-in the size of the subdomain $|D_\gamma|$ given in Table 2.

Note that a trapezoidal distribution is inside our Lipschitz class. For maximizing the entropy we can reduce the density of the flat region to an infinitesimal value, thus increasing the support. This explains the assumption of bounded variance in the following theorem.

Theorem 12. *Let x be a continuous random variable with domain $\mathbb{R}$, zero mean and bounded variance,* *that is* $\mathbb{E}_{\mathcal{P}}[x] = 0$ *and* $\mathbb{E}_{\mathcal{P}}[x^2] \leq \nu$*. Assume that with high probability,* x *belongs to a compact set* $\mathcal{D}_\gamma$*, that is* $\mathbb{P}_{\mathcal{P}}[x \notin \mathcal{D}_\gamma] = \int_{x \notin \mathcal{D}_\gamma} p(x) \leq \gamma$*. Under the same conditions of Theorem 11, the true Shannon entropy* $\mathcal{H}(\mathcal{P}) = -\int_{x \in \mathbb{R}} p(x) \log p(x)$ *is bounded as follows with probability at least* $1 - \delta$*:*

$$\widehat{\mathcal{H}}(\mathcal{P}) - \varepsilon^{(1)}_{N\delta\mathcal{D}_\gamma} - \varepsilon^{(2)}_{NK\mathcal{D}_\gamma} + \varepsilon^{(3)}_{K\gamma} \leq \mathcal{H}(\mathcal{P}) \leq \widehat{\mathcal{H}}(\mathcal{P}) + \varepsilon^{(1)}_{N\delta\mathcal{D}_\gamma} + \varepsilon^{(4)}_{\nu\gamma} \tag{14}$$

where $\widehat{\mathcal{H}}(\mathcal{P})$*,* $\varepsilon^{(1)}_{N\delta\mathcal{D}_\gamma}$ *and* $\varepsilon^{(2)}_{NK\mathcal{D}_\gamma}$ *are defined as in Theorem 11,* $\varepsilon^{(3)}_{K\gamma} = \min(0, \frac{1}{2}\gamma(1 - \log(2K\gamma)))$ *and* $\varepsilon^{(4)}_{\nu\gamma} = \max(0, \frac{3\gamma}{2}, \frac{1}{2}\gamma(1 - \log(\frac{2\gamma^3}{\pi\nu})))$.

Proof. Note that $\mathcal{H}(\mathcal{P}) = -\int_{x \in \mathcal{D}_\gamma} p(x) \log p(x) - \int_{x \notin \mathcal{D}_\gamma} p(x) \log p(x)$. Theorem 11 provides bounds for the first term. It suffices to bound the second term.

For finding a lower bound, we construct a worst-case scenario in which the density concentrates in a small region. In particular, consider a "half-triangular" distribution $\mathcal{Q}$ with density $q(x) = K'x$ for $0 < K' \leq K$ and $x \in [0; L] \notin \mathcal{D}_\gamma$. That is, the distribution is inside our Lipschitz class and the density vanishes at $q(0) = 0$. (The choice of $x = 0$ as an extreme of the interval is only for clarity of exposition.) By integration we have $\int_{x \in [0;L]} q(x) = \frac{1}{2}K'L^2 = \gamma' \leq \gamma$ and therefore $L = \sqrt{2\gamma'/K'}$. The entropy is given by:

$$\begin{aligned} -\int_{x \in [0;L]} q(x) \log q(x) &= \tfrac{1}{4}K'L^2(1 - 2\log(K'L)) \\ &= \tfrac{1}{2}\gamma'(1 - \log(2K'\gamma')) \\ &\geq \min(0, \tfrac{1}{2}\gamma'(1 - \log(2K'\gamma'))) \end{aligned}$$

Since the latter function is nonincreasing with respect to γ' and K', we prove that the lower bound holds.

For finding an upper bound, we follow a variational calculus argument. Consider a distribution $\mathcal{Q}$ with zero mean and known variance, that is $\mathbb{E}_{\mathcal{Q}}[x] = 0$ and $\mathbb{E}_{\mathcal{Q}}[x^2] \leq \nu$. We are interested on finding the maximum "half-entropy", that is, the entropy in the domain $x \in [0; +\infty) = \mathcal{D}_\gamma$. (The choice of $x = 0$ as an extreme of the interval and the zero-mean assumption are only for clarity of presentation.) Assume $\mathbb{P}_{\mathcal{Q}}[x \notin \mathcal{D}_\gamma] = \int_{x \in [0;+\infty)} q(x) = \gamma' \leq \gamma$. Given the nonnegativity of $x^2 q(x)$, the "half-variance" fulfills $\int_{x \in [0;+\infty)} x^2 q(x) = \nu' \leq \nu$. We need to solve:

$$\begin{aligned} &\max_q - \int_{x \in [0;+\infty)} q(x) \log q(x) \\ &\text{s.t.} \int_{x \in [0;+\infty)} q(x) = \gamma' \\ &\quad\;\; \int_{x \in [0;+\infty)} x^2 q(x) = \nu' \end{aligned}$$

By variational calculus and Lagrange multipliers, the solution of the above problem is $q^*(x) = \sqrt{\frac{2\gamma'^3}{\pi\nu'}} e^{-\frac{\gamma' x^2}{2\nu'}}$. The "half-entropy" of $\mathcal{Q}^*$ is given by:

$$\begin{aligned} h(\nu', \gamma') &= -\int_{x \in [0;+\infty)} q^*(x) \log q^*(x) \\ &= \tfrac{1}{2}\gamma'(1 - \log(\tfrac{2\gamma'^3}{\pi\nu'})) \end{aligned}$$

The function $h(\nu', \gamma')$ is increasing with respect to ν', therefore $h(\nu', \gamma') \leq h(\nu, \gamma')$. The function $h(\nu, \gamma')$ is concave with respect to γ'. In order to obtain an upper bound, we make $\frac{\partial h}{\partial \gamma'}(\nu, \gamma') = 0$ and obtain $\nu^* = \frac{2e^2}{\pi}\gamma'^3$, which produces the maximum value $h(\nu^*, \gamma') = \frac{3\gamma'}{2} \leq \frac{3\gamma}{2}$. By putting everything together, we prove that the upper bound holds. □

5. Discussion

Extension to several dimensions. Our results easily extend to V-dimensional data. Assume, for clarity of exposition that each of the V variables belong to the same domain $\mathcal{D}'$. That is $\mathbf{x} \in \mathcal{D}$ where $\mathcal{D} = \mathcal{D}' \times \cdots \times \mathcal{D}' = \mathcal{D}'^V$. The size of the domain $\mathcal{D}$ is $|\mathcal{D}'|^V$. The term $\frac{|\mathcal{D}|^2}{N^{1/4}}$ in our bounds, explains the exponential dependence of the number of samples N with respect to the dimensionality. That is $N \in \mathcal{O}(|\mathcal{D}'|^{4V})$. For bounded variables where $|\mathcal{D}'| \leq 1$ this is not an issue. For bounded variables where $|\mathcal{D}'| > 1$ and unbounded variables, we have an exponential dependence. However this is expected for a nonparametric method. Note that the results in (Liu et al., 2012) only apply to two variables in the unit square ($\mathcal{D} = [0; 1]^2$). The effect of the domain size and the extension to higher dimensions are not immediately obvious.

In our proofs, for a compact subset d of $\mathcal{D}$, we used the fact that $\int_{x \in d} 1 = |d|$ which is an ℓ_1 measure. Therefore, we extend Lipschitz continuity for several dimensions with respect to the ℓ_1 norm. That is $(\forall \mathbf{x}_1, \mathbf{x}_2)\ |p(\mathbf{x}_1) - p(\mathbf{x}_2)| \leq K\|\mathbf{x}_1 - \mathbf{x}_2\|_1$.

All theorems follow from these assumptions without any modification. Theorem 12 requires a minor change. More specifically, the lower bound produces a term $\frac{\gamma}{V}$ instead of γ. Finally, note that independent variables maximize the entropy, therefore a factor of V is needed in the upper bound.

Implications. To the best of our knowledge, we present the first exponential concentration bounds for Bayes error rate estimation. As a byproduct, we obtain a classifier that is Bayes consistent with provable finite-sample rates.

Regarding the Shannon entropy approximation, we extended the class of distributions with provable finite-sample rates from the Hölder class (Liu et al., 2012) to the Lipschitz class. In contrast to (Liu et al., 2012), our results also apply to probability distributions with regions of zero-probability, nonsmooth density functions, arbitrary behavior in the boundaries, and we do not require prior knowledge of lower/upper bounds neither boundedness of the variable.

The entropy approximation provides provable performance guarantees for learning the structure of trees and Bayesian networks from continuous variables. First, we show a generalization bound for trees (Chow & Liu, 1968).

Theorem 13. *Let N be the number of samples and V the number of continuous variables. Let $\mathcal{T}$ be a tree distribution, that is $\mathcal{T}$ is a collection of $V-1$ edges that form a tree. Define $\mathcal{I}(y,z) = \mathcal{H}(y)+\mathcal{H}(z)-\mathcal{H}(y,z)$ as the mutual information. Let $\widehat{\mathcal{L}}(\mathcal{T}) = \frac{1}{V-1}\sum_{(v,w)\in\mathcal{T}}\widehat{\mathcal{I}}(x_v,x_w)$ be the empirical log-likelihood, and $\mathcal{L}(\mathcal{T}) = \frac{1}{V-1}\sum_{(v,w)\in\mathcal{T}}\mathcal{I}(x_v,x_w)$ the expected log-likelihood. Let $\widehat{\mathcal{T}} = \arg\max_{\mathcal{T}}\widehat{\mathcal{L}}(\mathcal{T})$ be the empirical maximizer, and $\mathcal{T}^* = \arg\max_{\mathcal{T}}\mathcal{L}(\mathcal{T})$ the ground truth model. Under the same conditions of Theorem 11, with probability at least $1-\delta$:*

$$\mathcal{L}(\mathcal{T}^*)-\mathcal{L}(\widehat{\mathcal{T}}) \le \mathcal{O}\left(\tfrac{\log^{3/2} N}{N^{1/4}}, \log^{3/2} V, \log\tfrac{1}{\delta}\right) \qquad (15)$$

Proof. We need to approximate entropies for all nodes and pairs. That is, we need $|\mathcal{H}(x_v)-\widehat{\mathcal{H}}(x_v)| \le \varepsilon$ for every node v, and $|\mathcal{H}(x_v,x_w)-\widehat{\mathcal{H}}(x_v,x_w)| \le \varepsilon$ for every pair (v,w). The number of approximations is $V+\binom{V}{2} \le (V+1)^2$. A minor change in Theorem 11 is required in the union bound ($2(V+1)^2T$ events instead of $2T$ events) in order to obtain ε. Note that a bound in entropy approximation implies a bound for mutual information. That is, we have $|\mathcal{I}(x_v,x_w)-\widehat{\mathcal{I}}(x_v,x_w)| \le 3\varepsilon$ for every pair (v,w). Therefore, we have $|\mathcal{L}(\mathcal{T})-\widehat{\mathcal{L}}(\mathcal{T})| \le 3\varepsilon$ for every tree distribution $\mathcal{T}$. Finally, $\mathcal{L}(\mathcal{T}^*)-\mathcal{L}(\widehat{\mathcal{T}}) \le \widehat{\mathcal{L}}(\mathcal{T}^*)-\widehat{\mathcal{L}}(\widehat{\mathcal{T}})+6\varepsilon \le 6\varepsilon$. □

Next, we show a generalization bound for structure learning of Bayesian networks from continuous variables. This bound complements the results for discrete variables (Friedman & Yakhini, 1997; Höffgen, 1993). In the following theorem we consider maximum likelihood (De Campos, 2006) among models with a prescribed maximum number of parents k.

Theorem 14. *Let N be the number of samples and V the number of continuous variables. Let Π be a Bayesian network where $\pi_v \in \{1,\dots,V\}$ is the parent set for variable x_v. Assume that $(\forall v)\ |\pi_v| \le k$. Define $\mathcal{H}(y|z) = \mathcal{H}(y,z)-\mathcal{H}(z)$ as the conditional entropy. Let $\widehat{\mathcal{L}}(\Pi) = -\frac{1}{V}\sum_v \widehat{\mathcal{H}}(x_v|\mathbf{x}_{\pi_v})$ be the empirical log-likelihood, and $\mathcal{L}(\Pi) = -\frac{1}{V}\sum_v \mathcal{H}(x_v|\mathbf{x}_{\pi_v})$ the expected log-likelihood. Let $\widehat{\Pi} = \arg\max_{\Pi}\widehat{\mathcal{L}}(\Pi)$ be the empirical maximizer, and $\Pi^* = \arg\max_{\Pi}\mathcal{L}(\Pi)$ the ground truth model. Under the same conditions of Theorem 11, with probability at least $1-\delta$:*

$$\mathcal{L}(\Pi^*)-\mathcal{L}(\widehat{\Pi}) \le \mathcal{O}\left(\tfrac{\log^{3/2} N}{N^{1/4}}, ((k+2)\log V)^{3/2}, \log\tfrac{1}{\delta}\right) \qquad (16)$$

Proof. For a maximum number of parents k, we need to approximate entropies from up to $k+1$ variables. That is, we need $|\mathcal{H}(\mathbf{x}_{\mathcal{S}})-\widehat{\mathcal{H}}(\mathbf{x}_{\mathcal{S}})| \le \varepsilon$ for every set $\mathcal{S} \subset \{1,\dots,V\}$ such that $1 \le |\mathcal{S}| \le k+1$. The number of possible sets $\mathcal{S}$ is $\sum_{i=1}^{k+1}\binom{V}{i} \le V^{k+2}$. A minor change in Theorem 11 is required in the union bound ($2V^{k+2}T$ events instead of $2T$ events) in order to obtain ε. Note that a bound in entropy approximation implies a bound for conditional entropy. That is, we have $|\mathcal{H}(x_v|\mathbf{x}_\pi)-\widehat{\mathcal{H}}(x_v|\mathbf{x}_\pi)| \le 2\varepsilon$ for every v and π such that $|\pi| \le k$. Therefore, we have $|\mathcal{L}(\Pi)-\widehat{\mathcal{L}}(\Pi)| \le 2\varepsilon$ for every Bayesian network Π. Finally, $\mathcal{L}(\Pi^*)-\mathcal{L}(\widehat{\Pi}) \le \widehat{\mathcal{L}}(\Pi^*)-\widehat{\mathcal{L}}(\widehat{\Pi})+4\varepsilon \le 4\varepsilon$. □

Algorithmic complexity. For V-dimensional data, instead of building an $\mathcal{O}(2^V)$ matrix with all possible bins, we can perform the following. First, we assign the proper bin to each of the N samples and store these bin-assignments in a $\mathcal{O}(N)$ array. Second, we sort the samples with respect to their bin-assignments, in $\mathcal{O}(N\log N)$-time. Finally, since samples in the same bin are consecutive, we can produce the empirical probabilities in $\mathcal{O}(N)$-time. Thus, our method is $\mathcal{O}(N\log N)$-time and $\mathcal{O}(N)$-space.

Tighter bounds. The Bayes error rate results can be improved to $\mathcal{O}(N^{-1/3}, \log(1/\delta))$ by using a concentration inequality for the ℓ_1 deviation of empirical distributions (Weissman et al., 2003). On the other hand, if we assume a minimum density α by Chernoff bounds we can obtain $\mathcal{O}(1/\alpha, N^{-1/2}\log N, \log(1/\delta))$.

Concluding Remarks. There are several ways of extending this research. Bayes error rate approximation for Lipschitz distributions by using k-nearest neighbors or a more general class of weighted-average plug-in classifiers (Stone, 1977) needs to be analyzed. The extension of kernel methods for Shannon entropy approximation from the Hölder class (Liu et al., 2012) to the Lipschitz class needs to be analyzed. It would be interesting to extend our method to a broader class of probability distributions. Finally, while our method uses equally-sized bins and follows a frequentist approach, more adaptive methods and Bayesian approaches should be analyzed.

References

Ahmad, I. and Lin, P. A nonparametric estimation of the entropy for absolutely continuous distributions. *IEEE Transactions on Information Theory*, 1976.

Antos, A., Devroye, L., and Györfi, L. Lower bounds for Bayes error estimation. *PAMI*, 1999.

Beirlant, J., Dudewicz, E., Györfi, L., and Van der Meulen, E. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 1997.

Chow, C. and Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 1968.

De Campos, L. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *JMLR*, 2006.

Drakopoulos, J. Bounds on the classification error of the nearest neighbor rule. *ICML*, 1995.

Eggermont, P. and LaRiccia, V. Best asymptotic normality of the kernel density entropy estimator for smooth densities. *IEEE Transactions on Information Theory*, 1999.

Fralick, S. and Scott, R. Nonparametric Bayes-risk estimation. *IEEE Transactions on Information Theory*, 1971.

Friedman, N. and Yakhini, Z. On the sample complexity of learning Bayesian networks. *UAI*, 1997.

Györfi, L. The rate of convergence of k_n-NN regression estimates and classification rules. *IEEE Transactions on Information Theory*, 1981.

Höffgen, K. Learning and robust learning of product distributions. *COLT*, 1993.

Kohler, M. and Krzyżak, A. Rate of convergence of local averaging plug-in classication rules under margin condition. *ISIT*, 2006.

Kulkarni, S. and Posner, S. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 1995.

Liu, H., Lafferty, J., and Wasserman, L. Exponential concentration for mutual information estimation with application to forests. *NIPS*, 2012.

Nguyen, X., Wainwright, M., and Jordan, M. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 2010.

Nock, R. and Sebban, M. An improved bound on the finite-sample risk of the nearest neighbor rule. *Pattern Recognition Letters*, 2001.

Pál, D., Póczos, B., and Szepesvári, C. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. *NIPS*, 2010.

Paninski, L. Estimation of entropy and mutual information. *Neural Computation*, 2003.

Paninski, L. and Yajima, M. Undersmoothed kernel entropy estimators. *IEEE Transactions on Information Theory*, 2008.

Pérez-Cruz, F. Estimation of information theoretic measures for continuous random variables. *NIPS*, 2008.

Póczos, B. and Schneider, J. Nonparametric estimation of conditional information and divergences. *AISTATS*, 2012.

Stone, C. Consistent nonparametric regression. *The Annals of Statistics*, 1977.

Tsybakov, A. and Van der Meulen, E. Root-n consistent estimators of entropy for densities with unbounded support. *Scandinavian Journal of Statistics*, 1996.

Van Es, B. Estimating functionals related to a density by a class of statistics based on spacings. *Scandinavian Journal of Statistics*, 1992.

Wang, Q., Kulkarni, S., and Verdú, S. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 2005.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. Inequalities for the ℓ_1 deviation of the empirical distribution. *Technical Report HPL-2003-97*, 2003.