Exponential Concentration for Mutual Information Estimation with Application to Forests

Han Liu  
Department of Operations Research  
and Financial Engineering  
Princeton University, NJ 08544  
hanliu@princeton.edu

John Lafferty  
Department of Computer Science  
Department of Statistics  
University of Chicago, IL 60637  
lafferty@galton.uchicago.edu

Larry Wasserman  
Department of Statistics  
Machine Learning Department  
Carnegie Mellon University, PA 15231  
larry@stat.cmu.edu

Abstract

We prove a new exponential concentration inequality for a plug-in estimator of the Shannon mutual information. Previous results on mutual information estimation only bounded expected error. The advantage of having the exponential inequality is that, combined with the union bound, we can guarantee accurate estimators of the mutual information for many pairs of random variables simultaneously. As an application, we show how to use such a result to optimally estimate the density function and graph of a distribution which is Markov to a forest graph.

1 Introduction

We consider the problem of nonparametrically estimating the Shannon mutual information between two random variables. Let $X_1 \in \mathcal{X}_1$ and $X_2 \in \mathcal{X}_2$ be two random variables with domains $\mathcal{X}_1$ and $\mathcal{X}_2$ and joint density $p(x_1, x_2)$. The mutual information between $X_1$ and $X_2$ is

$$I(X_1; X_2) := \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} p(x_1, x_2) \log \left( \frac{p(x_1, x_2)}{p(x_1)p(x_2)} \right) dx_1 dx_2 = H(X_1) + H(X_2) - H(X_1, X_2),$$

where $H(X_1, X_2) = -\int_{\mathcal{X}_1} \int_{\mathcal{X}_2} p(x_1, x_2) \log p(x_1, x_2) dx_1 dx_2$ (and similarly for $H(X_1)$ and $H(X_2)$) are the corresponding Shannon entropies [4]. The mutual information is a measure of dependence between $X_1$ and $X_2$. To estimate $I(X_1; X_2)$ well, it suffices to estimate $H(X_1, X_2) := H(p)$.

A simple way to estimate the Shannon entropy is to use a kernel density estimator (KDE) [22, 1, 9, 5, 20, 7], i.e., the densities $p(x, y)$, $p(x)$, and $p(y)$ are separately estimated from samples and the estimated densities are used to calculate the entropy. Alternative methods involve estimation of the entropies using spacings [25, 26, 23], k-nearest neighbors [11, 12], the Edgeworth expansion [24], and convex optimization [17]. More discussions can be found in the survey articles [2, 19]. There have been many recent developments in the problem of estimating Shannon entropy and related quantities as well as application of these results to machine learning problems [18, 21, 8, 6]. Under weak conditions, it has been shown that there are estimators that achieve the parametric $\sqrt{n}/n$-rate of convergence in mean squared error (MSE), where $n$ is the sample size.

In this paper, we construct an estimator with this rate, but we also prove an exponential concentration inequality for the estimator. More specifically, we show that our estimator $\hat{H}$ of $H(p)$ satisfies

$$\sup_{p \in \mathcal{P}} \mathbb{P} \left( |\hat{H} - H(p)| > \epsilon \right) \leq 2 \exp \left( -\frac{n \epsilon^2}{36 \kappa^2} \right)$$ (1.1)
where $\Sigma$ is a nonparametric class of distributions defined in Section 2 and $\kappa$ is a constant. To the best of our knowledge, this is the first such exponential inequality for nonparametric Shannon entropy and mutual information estimation. The advantage of this result, over the usual results which state that $\mathbb{E}(|\hat{H} - H(p)|^2) = O(n^{-1})$, is that we can apply the union bound and thus guarantee accurate mutual information estimation for many pairs of random variables simultaneously. As an application, we consider forest density estimation [15], which, in a $d$-dimensional problem, requires estimating $d(d+1)/2$ mutual informations in order to apply the Chow-Liu algorithm. As long as $\frac{\log d}{n} \to 0$ as $n \to \infty$, we can estimate the forest graph well, even if $d = d(n)$ increases with $n$ exponentially fast.

The rest of this paper is organized as follows. The assumptions and estimator are given in Section 2. The main theoretical analysis is in Section 3. In Section 4 we show how to apply the result to forest density estimation. Some discussion and possible extensions are provided in the last section.

2 Estimator and Main Result

Let $X = (X_1, X_2) \in \mathbb{R}^2$ be a random vector with density $p(x) := p(x_1, x_2)$ and let $x^1, \ldots, x^n \in X \subset \mathbb{R}^2$ be a random sample from $p$. In this paper, we only consider the case of bounded domain $X = [0, 1]^2$. We want to estimate the Shannon entropy

$$H(p) = -\int_X p(x) \log p(x) dx.$$  \hspace{1cm} (2.1)

We start with some assumptions on the density function $p(x_1, x_2)$.

**Assumption 2.1 (Density assumption).** We assume the density $p(x_1, x_2)$ belongs to a 2nd-order H"older class $\Sigma_2(2, L)$ and is bounded away from zero and infinity. In particular, there exist constants $\kappa_1, \kappa_2$

$$0 < \kappa_1 \leq \min_{x \in X} p(x) \leq \max_{x \in X} p(x) \leq \kappa_2 < \infty,$$  \hspace{1cm} (2.2)

and for any $(x_1, x_2)^T \in X$, there exists a constant $L$ such that, for any $(u, v)^T \in X$

$$|p(x_1 + u, x_2 + v) - p(x_1, x_2) - \frac{\partial p(x_1, x_2)}{\partial x_1} u - \frac{\partial p(x_1, x_2)}{\partial x_2} v| \leq L(u^2 + v^2).$$  \hspace{1cm} (2.3)

**Assumption 2.2 (Boundary assumption).** If $\{x_n\} \in X$ is any sequence converging to a boundary point $x^\ast$, we require the density $p(x)$ has vanishing first order partial derivatives:

$$\lim_{n \to \infty} \frac{\partial p(x_n)}{\partial x_1} = \lim_{n \to \infty} \frac{\partial p(x_n)}{\partial x_2} = 0.$$  \hspace{1cm} (2.4)

To efficiently estimate the entropy in (2.1), we use a KDE based “plug-in” estimator. Bias at the boundaries turns out to be very important in this problem; see [10] for a discussion of boundary bias. To correct the boundary effects, we use the following “mirror image” kernel density estimator:

$$\tilde{p}_h(x_1, x_2) := \frac{1}{nh^2} \sum_{i=1}^{n} \left\{ K \left( \frac{x_1 - x_i^1}{h} \right) K \left( \frac{x_2 - x_i^2}{h} \right) + K \left( \frac{x_1 + x_i^1}{h} \right) K \left( \frac{x_2 - x_i^2}{h} \right) \right. \right.$$  \hspace{1cm} (2.5)

$$+ K \left( \frac{x_1 - x_i^1}{h} \right) K \left( \frac{x_2 + x_i^2}{h} \right) + K \left( \frac{x_1 + x_i^1}{h} \right) K \left( \frac{x_2 + x_i^2}{h} \right) \right.$$  \hspace{1cm} (2.5)

$$+ K \left( \frac{x_1 - 2 + x_i^1}{h} \right) K \left( \frac{x_2 - x_i^2}{h} \right) + K \left( \frac{x_1 + x_i^1}{h} \right) K \left( \frac{x_2 + x_i^2}{h} \right) \right.$$  \hspace{1cm} (2.5)

$$+ K \left( \frac{x_1 - 2 + x_i^1}{h} \right) K \left( \frac{x_2 + x_i^2}{h} \right) \right\}.$$

Here $h$ is the bandwidth and $K(\cdot)$ is a univariate kernel function. We denote by $K_2(u, v) := K(u)K(v)$ the bivariate product kernel. This estimator has nine terms; one corresponds to the original data in the unit square $[0, 1]^2$, and each of the remaining terms corresponds to reflecting the data across one of the four sides or four corners of the square.
Assumption 2.3 (Kernel assumption). The kernel \( K(\cdot) \) is nonnegative and has a bounded support \([-1,1]\) with \( \int_{-1}^{1} K(u)du = 1 \) and \( \int_{-1}^{1} uK(u)du = 0 \).

By Assumption 2.1, the values of the true density lie in the interval \([\kappa_1, \kappa_2]\). We propose a clipped KDE estimator
\[
\hat{p}_h(x) = T_{\kappa_1, \kappa_2}(\tilde{p}_h(x)),
\]
where \( T_{\kappa_1, \kappa_2}(a) = \kappa_1 \cdot I(a < \kappa_1) + a \cdot I(\kappa_1 \leq a \leq \kappa_2) + \kappa_2 \cdot I(a > \kappa_2) \), so that the estimated density also has this property. Letting \( g(u) = u \log u \), we propose the following plug-in entropy estimator:
\[
H(\hat{p}_h) := -\int_X g(\hat{p}_h(x)) \, dx = -\int_X \hat{p}_h(x) \log \hat{p}_h(x) \, dx.
\]

Remark 2.1. The clipped estimator \( \hat{p}_h \) requires the knowledge of \( \kappa_1 \) and \( \kappa_2 \). In applications, we do not need to know the exact values of \( \kappa_1 \) and \( \kappa_2 \); lower and upper bounds are sufficient.

Our main technical result is the following exponential concentration inequality on \( H(\hat{p}_h) \) around the population quantity \( H(p) \). Our proof is given in Section 3.

Theorem 2.1. Under Assumptions 2.1, 2.2, and 2.3, if we choose the bandwidth according to \( h \asymp n^{-1/4} \), then there exists a constant \( N_0 \) such that for all \( n > N_0 \),
\[
\sup_{p \in \Sigma_2(2, \kappa)} \mathbb{P}(\left| H(\hat{p}_h) - H(p) \right| > \epsilon) \leq 2 \exp\left(-\frac{n \epsilon^2}{36 \kappa^2}\right),
\]
where \( \kappa = \max\{\log \kappa_1, \log \kappa_2\} + 1 \).

To the best of our knowledge, this is the first time an exponential inequality like (2.8) has been established for Shannon entropy estimation over the Hölder class. It is easy to see that (2.8) implies the parametric \( \sqrt{n} \)-rate of convergence in mean squared error, \( \mathbb{E}(\|\hat{H} - H(p)\|) = O(n^{-1/2}) \). The bandwidth \( h \asymp n^{-1/4} \) in the above theorem is different from the usual choice for optimal bivariate density estimation, which is \( h_P \asymp n^{-1/6} \) for the 2nd-order Hölder class. By using \( h \asymp n^{-1/4} \), we undersmooth the density estimate. As we show in the next section, such a bandwidth choice is important for achieving the optimal rate for entropy estimation.

Let \( I(p) := I(X_1; X_2) \) be the Shannon mutual information, and define
\[
I(\hat{p}_h) := \int_{X_1} \int_{X_2} \hat{p}_h(x_1, x_2) \log \frac{\hat{p}_h(x_1, x_2)}{\hat{p}_h(x_1)\hat{p}_h(x_2)} \, dx_1 \, dx_2.
\]

The next corollary provides an exponential inequality for Shannon mutual information estimation.

Corollary 2.1. Under the same conditions as in Theorem 2.1, if we choose \( h \asymp n^{-1/4} \), then there exists a constant \( N_1 \) such that for all \( n > N_1 \),
\[
\sup_{p \in \Sigma_2(2, \kappa)} \mathbb{P}(\left| I(\hat{p}_h) - I(p) \right| > \epsilon) \leq 6 \exp\left(-\frac{n \epsilon^2}{324 \kappa^2}\right),
\]
where \( \kappa = \max\{\log \kappa_1, \log \kappa_2\} + 1 \).

Proof. Using the same proof for Theorem 2.1, we can show that (2.8) also holds for estimating univariate entropies \( H(X_1) \) and \( H(X_2) \). The desired result then follows from the union bound since \( I(p) := I(X_1; X_2) = H(X_1) + H(X_2) - H(X_1, X_2) \).

Remark 2.2. We use the same bandwidth \( h \asymp n^{-1/4} \) to estimate the bivariate density \( p(x_1, x_2) \) and univariate densities \( p(x_1), p(x_2) \). A related result is presented in [15]. They consider the same problem setting as ours and also use a KDE based plug-in estimator to estimate the mutual information. However, unlike our proposal, they advocate the use of different bandwidths for bivariate and univariate entropy estimations. For bivariate case they use \( h_2 \asymp n^{-1/6} \); for univariate case they use \( h_1 \asymp n^{-1/5} \). Such bandwidths \( h_1 \) and \( h_2 \) are useful for optimally estimating the density functions. However, such a choice achieves a suboptimal rate in terms of mutual information estimation:
\[
\sup_{p \in \Sigma_2(2, \kappa)} \mathbb{P}(\left| I(\hat{p}_h) - I(p) \right| > \epsilon) \leq c_1 \exp\left(-c_2 n^{2/3} \epsilon^2\right),
\]
where \( c_1 \) and \( c_2 \) are two constants. Our method achieves the faster parametric rate.
3 Theoretical Analysis

Here we present the detailed proof of Theorem 2.1. To analyze the error $|H(\hat{p}_h) - H(p)|$, we first decompose it into a bias or approximation error term, and a “variance” or estimation error term:

$$|H(\hat{p}_h) - H(p)| \leq \left| \frac{H(\hat{p}_h) - \mathbb{E}H(\hat{p}_h)}{\mathbb{E}H(\hat{p}_h)} \right| + \left| \frac{\mathbb{E}H(\hat{p}_h) - H(p)}{H(p)} \right|. \tag{3.1}$$

We are going to show that

$$\sup_{p \in \Sigma_n(2, L)} \mathbb{P}\left( \left| \frac{H(\hat{p}_h) - \mathbb{E}H(\hat{p}_h)}{\mathbb{E}H(\hat{p}_h)} \right| > \epsilon \right) \leq 2 \exp\left( -\frac{\epsilon^2}{2\sigma^2} \right), \tag{3.2}$$

$$\sup_{p \in \Sigma_n(2, L)} \left| \frac{\mathbb{E}H(\hat{p}_h) - H(p)}{H(p)} \right| \leq c_1 h^2 + \frac{c_3}{nh^2}, \tag{3.3}$$

where $c_1$ and $c_3$ are two constants. Since the bound on the variance in (3.2) does not depend on $h$, to optimize the rate, we only need to choose $h$ to minimize the right-hand side of (3.3). Therefore $h \approx n^{-1/4}$ achieves the optimal rate. In the rest of this section, we bound the bias and variance terms separately.

3.1 Analyzing the Bias Term

Here we prove (3.3). Let $u$ be a vector. We denote the sup norm by $\|u\|_\infty$. The next lemma bounds the integrated squared bias of the kernel density estimator over the support $X := [0, 1]^2$.

**Lemma 3.1.** Under Assumptions 2.1, 2.2, and 2.3, there exists a constant $c > 0$ such that

$$\sup_{p \in \Sigma_n(2, L)} \int_X (\mathbb{E}\tilde{p}_h(x) - p(x))^2 \, dx \leq ch^4. \tag{3.4}$$

**Proof.** We partition the support $X := [0, 1]^2$ into three regions $X = B \cup C \cup I$, the boundary area $B$, the corner area $C$, and the interior area $I$:

$$C = \{x : \|x - u\|_\infty \leq h \text{ for } u = (0, 0)^T, \text{ or } (0, 1)^T, \text{ or } (1, 0)^T, \text{ or } (1, 1)^T\},$$

$$B = \{x : x \text{ is within distance } h \text{ to an edge of } X, \text{ but does not belong to } C\},$$

$$I = X \setminus (C \cup B).$$

We have the following decomposition:

$$\int_X (\mathbb{E}\tilde{p}_h(x) - p(x))^2 \, dx = \int_I + \int_C + \int_B (\mathbb{E}\tilde{p}_h(x) - p(x))^2 \, dx = T_I + T_C + T_B.$$

From standard results on kernel density estimation, we know that $\sup_{p \in \Sigma(2, L)} T_I \leq ch^4$. In the next two subsections, we bound $T_B := \int_B (\mathbb{E}\tilde{p}_h(x) - p(x))^2 \, dx$ and $T_C := \int_C (\mathbb{E}\tilde{p}_h(x) - p(x))^2 \, dx$.

3.1.1 Analyzing $T_B$

Let $A := \{x : 0 \leq x_1 \leq h \text{ and } h \leq x_2 \leq 1 - h\}$. We have

$$T_B = \int_B (\mathbb{E}\tilde{p}_h(x) - p(x))^2 \, dx \leq c \int_A (\mathbb{E}\tilde{p}_h(x) - p(x))^2 \, dx. \tag{3.8}$$

For $x \in A$, we have

$$\tilde{p}_h(x) = \frac{1}{nh^2} \sum_{i=1}^n K\left( \frac{x_1 - x_1^i}{h} \right) K\left( \frac{x_2 - x_2^i}{h} \right) + K\left( \frac{x_1 + x_1^i}{h} \right) K\left( \frac{x_2 - x_2^i}{h} \right). \tag{3.9}$$

Therefore, for $x \in A$ we have

$$\mathbb{E}\tilde{p}_h(x) = \frac{1}{h^2} \int_0^1 \int_0^1 K\left( \frac{x_1 - t_1}{h} \right) K\left( \frac{x_2 - t_2}{h} \right) p(t_1, t_2) dt_1 dt_2$$
\[ + \frac{1}{h^2} \int_0^1 \int_0^1 K \left( \frac{x_1 + t_1}{h} \right) K \left( \frac{x_2 - t_2}{h} \right) p(t_1, t_2) dt_1 dt_2 \]
\[ = \int_{-1}^1 \int_{-1}^1 K(u_1)K(u_2)p(x_1 + u_1h, x_2 + u_2h)du_1 du_2 \]
\[ + \int_{-1}^1 \int_{-1}^1 \left[ K(u_1)K(u_2)p(x_1 - u_1h, x_2 - u_2h)du_1 du_2. \right. \]

(3.10)

Since \( p \in \Sigma_k(2, L) \) and \( 0 < x_1 \leq h \), we have
\[ |p(x_1 + u_1h, x_2 + u_2h) - p(x_1, x_2) - \langle \nabla p(x), u \rangle h| \leq L\|u\|^2 h^2, \]
\[ |p(x_1 - u_1h, x_2 - u_2h) - p(x_1, x_2) + \frac{\partial p(x)}{\partial x_1}(2x_1 + u_1h) + \frac{\partial p(x)}{\partial x_2}(u_2h)| \leq L[(2 + u_1^2 + u_2^2)]^2 h^2. \]

Since \( |u_1|, |u_2| \leq 1 \), we have \( |p(x_1 + u_1h, x_2 + u_2h) - p(x_1, x_2)| \leq |\frac{\partial p(x)}{\partial x_1}| h + |\frac{\partial p(x)}{\partial x_2}| h + 10Lh^2 \).

For any \( x \in A \), we can bound the bias term
\[ |\mathbb{E}\tilde{p}_h(x) - p(x)| \leq 10 \left| \frac{\partial p(x)}{\partial x_1} \right| h + 2 \left| \frac{\partial p(x)}{\partial x_2} \right| h + 12Lh^2 \]
\[ \leq 12Lh^2 + 12Lh^2 \]
\[ = 24Lh^2, \]

where the last inequality follows from the fact that \( \left| \frac{\partial p(x)}{\partial x_1} \right|, \left| \frac{\partial p(x)}{\partial x_2} \right| \leq Lh \), by the Hölder condition and the assumption that the density \( p(x) \) has vanishing partial derivatives on the boundary points. Therefore, we have \( T_B \leq ch^5 \).

3.1.2 Analyzing \( T_C \)

Let \( A_1 := \{ x : 0 \leq x_1, x_2 \leq h \} \). We now analyze the term \( T_C \):
\[ T_C = \int_{\mathcal{C}} (\mathbb{E}\tilde{p}_h(x) - p(x))^2 \, dx \leq c \int_{A_1} (\mathbb{E}\tilde{p}_h(x) - p(x))^2 \, dx. \]

(3.15)

For notational simplicity, we write
\[ U_{x,h}(a,b) = K \left( \frac{x_1 - a}{h} \right) K \left( \frac{x_2 - b}{h} \right). \]

(3.16)

For \( x \in A_1 \), we have
\[ \tilde{p}_h(x) = \frac{1}{nh^2} \sum_{i=1}^n \left[ U_{x,h}(x_1^i, x_2^i) + U_{x,h}(-x_1^i, x_2^i) + U_{x,h}(x_1^i, -x_2^i) + U_{x,h}(-x_1^i, -x_2^i) \right]. \]

(3.17)

Therefore, for \( x \in A_1 \) we have
\[ \mathbb{E}\tilde{p}_h(x) \]
\[ = \frac{1}{h^2} \int_0^1 \int_0^1 \left[ U_{x,h}(t_1, t_2) + U_{x,h}(-t_1, t_2) + U_{x,h}(t_1, -t_2) + U_{x,h}(-t_1, -t_2) \right] p(t_1, t_2) dt_1 dt_2. \]
Under Assumptions 2.1, 2.2, and 2.3, there exists a universal constant 3.1.3 Analyzing the Bias of the Entropy Estimator

\[ x \] is free of boundary bias. Thus the desired result of Lemma 3.1 is proved.

For \( x = (x_1, x_2) \in A_1 \),

\[ p(x_1, x_2) = \int_{-\frac{x_2}{2}}^{\frac{x_2}{2}} \int_{-\frac{x_1}{2}}^{\frac{x_1}{2}} p(x_1, x_2)K(u_1, u_2)du_1du_2. \]

Using the fact that \( p \in \Sigma_s(2, L) \), \( 0 \leq x_1, x_2 \leq h \), and \(-1 \leq u_1, u_2 \leq 1\), we have

\[ |p(x_1 + u_1, x_2 + u_2h) - p(x_1, x_2)| \leq 4Lh^2, \]

\[ |p(u_1, x_1, x_2 + u_2h) - p(x_1, x_2)| \leq 20Lh^2, \]

\[ |p(u_1, x_1, x_2 + u_2h) - p(x_1, x_2)| \leq 20Lh^2, \]

\[ |p(u_1, x_1 - u_1, u_2h - u_2) - p(x_1, x_2)| \leq 36Lh^2. \]

For \( x \in A_1 \), we can then bound the bias term as

\[ |E\tilde{p}_h(x) - p(x)| \]

\[ = \left| E\tilde{p}_h(x) - \int_{-1}^{1} \int_{-1}^{1} K(u_1, u_2)p(t_1, t_2)du_1du_2 \right| \]

\[ \leq \int_{-\frac{x_2}{2}}^{\frac{x_2}{2}} \int_{-\frac{x_1}{2}}^{\frac{x_1}{2}} K(u_1)K(u_2)|p(x_1 + u_1, x_2 + u_2h) - p(x_1, x_2)|du_1du_2 \]

\[ + \int_{-\frac{x_2}{2}}^{\frac{x_2}{2}} \int_{-\frac{x_1}{2}}^{\frac{x_1}{2}} K(u_1)K(u_2)|p(u_1, x_1, x_2 + u_2h) - p(x_1, x_2)|du_1du_2 \]

\[ + \int_{-\frac{x_2}{2}}^{\frac{x_2}{2}} \int_{-\frac{x_1}{2}}^{\frac{x_1}{2}} K(u_1)K(u_2)|p(u_1, x_1 + u_1, u_2h - u_2) - p(x_1, x_2)|du_1du_2 \]

\[ + \int_{-\frac{x_2}{2}}^{\frac{x_2}{2}} \int_{-\frac{x_1}{2}}^{\frac{x_1}{2}} K(u_1)K(u_2)|p(u_1 - u_1, u_2h - u_2) - p(x_1, x_2)|du_1du_2 \]

\[ \leq 80Lh^2. \]

Therefore, we have \( T_C \leq c'h^6 \).

Combining the analysis of \( T_B, T_C, \) and \( T_L \), we show that the mirror image kernel density estimator is free of boundary bias. Thus the desired result of Lemma 3.1 is proved.

3.1.3 Analyzing the Bias of the Entropy Estimator

**Lemma 3.2.** Under Assumptions 2.1, 2.2, and 2.3, there exists a universal constant \( C^* \) that does not depend on the true density \( p \), such that

\[ \sup_{p \in \Sigma_s(2, L)} \left| E\tilde{H}(\tilde{p}_h) - H(p) \right| \leq \frac{C^*}{\sqrt{n}} \]
Proof. Recalling that \( g(u) = u \log u \), by Taylor’s theorem we have

\[
g(\hat{p}_h(x)) - g(p(x)) = (\log(p(x)) + 1) \cdot \hat{p}_h(x) - p(x) + \frac{1}{2\xi(x)} \cdot \left(\hat{p}_h(x) - p(x)\right)^2, \tag{3.36}
\]

where \( \xi(x) \) lies in between \( \hat{p}_h(x) \) and \( p(x) \). It is obvious that \( \kappa_1 \leq \xi(x) \leq \kappa_2 \).

Let \( \kappa \) be as defined in the statement of the theorem. Using Fubini’s theorem, Hölder’s inequality and the fact that the Lebesgue measure of \( \mathcal{X} \) is 1, we have

\[
|\mathbb{E} H(\hat{p}_h) - H(p)| \leq |\mathbb{E} [g(\hat{p}_h(x)) - g(p(x))] dx| \leq |\int_X (\log(p(x)) + 1) \cdot \mathbb{E} [\hat{p}_h(x) - p(x)] dx| + \left| \int_X \frac{1}{2\xi(x)} \cdot \mathbb{E} [\hat{p}_h(x) - p(x)]^2 dx \right| \leq \kappa \int_X \left( \mathbb{E} [\hat{p}_h(x) - p(x)]^2 + \frac{1}{2\kappa} \cdot \mathbb{E} [\hat{p}_h(x) - p(x)]^2 dx \right) \leq \kappa \int_X \left( \mathbb{E} [\hat{p}_h(x) - p(x)]^2 + \frac{1}{2\kappa} \cdot \mathbb{E} [\hat{p}_h(x) - p(x)]^2 dx \right) \leq c_1h^2 + c_2h^4 + c_3\frac{\kappa^2}{nh^2}. \tag{3.42}
\]

The last inequality follows from standard results of kernel density estimation and Lemma 3.1, where \( c_1, c_2, c_3 \) are three constants. We get the desired result by setting \( h \approx n^{-1/4} \).

3.2 Analyzing the Variance Term

Lemma 3.3. Under Assumptions 2.1, 2.2, and 2.3, we have,

\[
\sup_{p \in \Sigma_{n,(2,L)}} \mathbb{P}\left( |H(\hat{p}_h) - \mathbb{E} H(\hat{p}_h)| > \epsilon \right) \leq 2 \exp\left( -\frac{n\epsilon^2}{32\kappa^2} \right). \tag{3.43}
\]

Proof. Let \( \hat{p}_h(x) \) be the kernel density estimator defined as in (2.6) but with the \( j^{th} \) data point \( x^j \) replaced by an arbitrary value \( (x^j)' \). Since \( g'(u) = \log u + 1 \), by Assumption 2.1, we have \( \max \| g'(\hat{p}_h(x)) \|, \| g'(\hat{p}_h(x)) \| \leq \kappa \).

For notational simplicity, we write the product kernel as \( K_2 = K \cdot K \). Using the mean-value theorem and the fact that \( T_{\kappa_1,\kappa_2}(\cdot) \) is a contraction, we have

\[
\sup_{x^1,\ldots,x^n,(x^j)'} |H(\hat{p}_h) - \mathbb{E} H(\hat{p}_h)| \leq \kappa \sup_{x^1,\ldots,x^n,(x^j)'} \left| \int_{\mathcal{X}} \left( g(\hat{p}_h(x)) - g(\hat{p}_h'(x)) \right) dx \right| \leq 4\kappa \sup_{x^1,\ldots,x^n,(x^j)'} \int_{\mathcal{X}} \left| \frac{1}{nh^2} K_2 \left( \frac{x^j - x}{h} \right) - \frac{1}{nh^2} K_2 \left( \frac{(x^j)'}{h} \right) \right| dx \leq 8\kappa \sup_{y} \int_{\mathcal{X}} \frac{1}{nh^2} K_2 \left( \frac{y - x}{h} \right) dx \leq 8\kappa \int_{\mathcal{X}} K_2(u) du = \frac{8\kappa}{n}. \tag{3.50}
\]

Therefore, using McDiarmid’s inequality [16], we get the desired inequality (3.43). The uniformity result holds since the constant does not depend on the true density \( p \). \( \square \)
Algorithm. Under the same condition as Theorem 12 in [15], if we choose is minimax optimal over the class $P$. Theorem 4.1 allows for both graph recovery and density estimation. Their proofs are provided in a technical report [14].

4 Application to Forest Density estimation

Let $x_1, \ldots, x^n \in \mathbb{R}^d$ from a $d$-dimensional random vector $X$. We have two learning tasks: (i) we want to estimate an acyclic undirected graph $F = (V, E)$, where $V$ is the vertex set containing all the random variables and $E$ is the edge set such that an edge $(j, k) \in E$ if and only if the corresponding random variables $X_j$ and $X_k$ are conditionally independent given the other variables $X_{\setminus \{j,k\}}$; (ii) once we have an estimated graph $\hat{F}$, we want to estimate the density function $p(x)$.

Using the negative log-likelihood loss, Liu et al. [15] show that the graph estimation problem can be recast as the problem of finding the maximum weight spanning forest for a weighted graph, where the weight of the edge connecting nodes $j$ and $k$ is $I(X_j; X_k)$, the mutual information between these two variables. Empirically, we replace $I(X_j; X_k)$ by its estimate $\hat{I}(X_j; X_k)$ from (2.9). The forest graph can be obtained by the Chow-Liu algorithm [3, 13], which is an iterative algorithm. At each iteration the algorithm adds an edge connecting that pair of variables with maximum mutual information among all pairs not yet visited by the algorithm, if doing so does not form a cycle. When stopped early, after $s < d - 1$ edges have been added, it yields the best $s$-edge weighted forest. Once a forest graph $\hat{F} = (V, \hat{E})$ is estimated, we propose to estimate the forest density as

$$\hat{p}_F(x) = \prod_{(j,k) \in \hat{E}} \frac{\hat{p}_{h_2}(x_j, x_k)}{\hat{p}_{h_2}(x_j) \hat{p}_{h_2}(x_k)} \cdot \prod_{u \in \hat{U}} \hat{p}_{h_2}(x_u) \cdot \prod_{\ell \in V \setminus \hat{U}} \hat{p}_{h_1}(x_{\ell}),$$

where $\hat{U}$ is the set of isolated vertices in the estimated forest $\hat{F}$. Our estimator is different from the estimator proposed by [15]—once the graph $\hat{F}$ is given, we treat the isolated variables differently than the connected variables. As will be shown in Theorem 4.2, such a choice leads to minimax optimal forest density estimation, while the obtained rate from [15] is suboptimal.

Let $F'_d$ denote the set of forest graphs with $d$ nodes and no more than $s$ edges. Let $D(\cdot \| \cdot)$ be the Kullback-Leibler divergence. We define the $s$-oracle forest $F^*_s := (V, E^*_s)$ and its corresponding oracle density estimator $p_{F^*_s}$ to be

$$F^*_s = \arg \min_{F \in F'_d} D(p \| p_F) \quad \text{and} \quad p_{F^*_s} := \prod_{(j,k) \in E^*_s} \frac{p(x_j, x_k)}{p(x_j) p(x_k)} \prod_{\ell \in V} p(x_{\ell}).$$

Let $\Sigma_n(2, L)$ be defined as in Assumption (2.1). We define a density class $\mathcal{P}_n$ as

$$\mathcal{P}_n := \{ p : p \text{ is a } d\text{-dimensional density with } p(x_j, x_k) \in \Sigma_n(2, L) \text{ for any } j \neq k \}. \quad (4.3)$$

The next two theorems show that the above forest density estimation procedure is minimax optimal for both graph recovery and density estimation. Their proofs are provided in a technical report [14].

Theorem 4.1 (Graph Recovery). Let $\hat{F}$ be the estimated $s$-edge forest graph using the Chow-Liu algorithm. Under the same condition as Theorem 12 in [15], if we choose $h \asymp n^{-1/4}$ for the mutual information estimator in (2.9), then

$$\sup_{p \in \mathcal{P}_n} \mathbb{P} \left( \hat{F} \neq F^*_s \right) = O \left( \sqrt{\frac{s}{n}} \right) \quad \text{whenever} \quad \frac{\log d}{n} \rightarrow 0. \quad (4.4)$$

Theorem 4.2 (Density Estimation). Once the $s$-edge forest graph $\hat{F}$ as in Theorem 4.1 has been obtained, we calculate the density estimator (B.1) by choosing $h_1 \asymp n^{-1/5}$ and $h_2 \asymp n^{-1/6}$. Then,

$$\sup_{p \in \mathcal{P}_n} \mathbb{E} \int_X |\hat{p}_F(x) - p_{F^*_s}(x)| \, dx \leq C \cdot \sqrt{\frac{s}{n^{2/3}}} + \frac{d - s}{n^{4/5}}. \quad (4.5)$$

5 Discussions and Conclusions

Theorem 4.1 allows $d$ to increase exponentially fast as $n$ increases and still guarantees graph recovery consistency. Theorem 4.2 provides the rate of convergence for the $L_1$-risk. The obtained rate is minimax optimal over the class $\mathcal{P}_n$. The term $sn^{-2/3}$ corresponds to the price paid to estimate bivariate densities; while the term $(d - s)n^{-4/5}$ corresponds to the price paid to estimate univariate densities. In this way, we see that the exponential concentration inequality for Shannon mutual information leads to significantly improved theoretical analysis of the forest density estimation, in terms of both graph estimation and density estimation. This research was supported by NSF grant IIS-1116730 and AFOSR contract FA9550-09-1-0373.
References


