INFORMATION THEORY MEETS BIG DATA: THEORY, ALGORITHMS AND APPLICATIONS TO DEEP LEARNING

BY

WEIHAO GAO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Pramod Viswanath, Chair
Associate Professor Maxim Raginsky
Assistant Professor Sewoong Oh
Assistant Professor Sreeram Kannan, University of Washington

ABSTRACT

As the era of big data arises, people get access to numerous amounts of multi-view data. Measuring, discovering and understanding the underlying relationship among different aspects of data is the core problem in information theory. However, traditional information theory research focuses on solving this problem in an abstract population-level way. In order to apply information-theoretic tools to real-world problems, it is necessary to revisit information theory from sample-level.

One important bridge between traditional information theory and real-world problems is the information-theoretic quantity estimators. These estimators enable computing of traditional information-theoretic quantities from big data and understanding hidden relationships in data. Information-theoretic tools can also be utilized to improve modern machine learning techniques. In this dissertation, several problems of information-theoretic quantity estimators and their applications are investigated.

This dissertation consists of the following topics: (1) theoretical study of the fundamental limit of information-theoretic quantity estimators, especially k-nearest neighbor estimators of differential entropy and mutual information; (2) designing novel algorithms of differential entropy and mutual information estimators for some special and challenging practical scenarios, as well as new information-theoretic measures to discover complex relationships among data which cannot be found by traditional measures; (3) applying information-theoretic tools to improve training algorithms and model compression algorithms in deep learning.

To my parents, for their love and support.

ACKNOWLEDGMENTS

I was extremely fortunate to work with my advisors Sewoong Oh and Pramod Viswanath during the past years. When I entered the campus, I was a young man with no experience and understanding of research at all. Sewoong and Pramod were always patient to discuss ideas with me, revise my ill-written papers and give me guidance on the road of research. They taught me how to do solid and "first-order" research, rather than just following the hot topics. I also learned a lot from their vision, hardwork and passionate attitude. I would give Sewoong and Pramod most of the credit for becoming the kind of researcher I am today. I am always indebted to them.

I would like to thank Maxim Raginsky for serving as both my doctoral committee member and my qualifying exam committee member, Sreeram Kannan for serving as my doctoral committee member, and Yuliy Baryshnikov and Yihong Wu for serving as my qualifying exam committee members. Their guidance helped me a lot during my journey. I would like to thank all the wonderful professors who taught me at UIUC, including Chandra Chekuri, Jiawei Han, Xiaochun Li, Pierre Moulin, Sewoong Oh, Maxim Raginsky, Zhongjin Ruan, Renming Song, Rayadurgam Srikant and Venugopal V. Veeravalli. Their classes provide me important knowledge useful for research. I also want to thank all the TAs in these courses.

I would like to thank all the research collaborators other than my two advisors, including Yanjun Han, Jiantao Jiao, Sreeram Kannan, Hyeji Kim, Ashok V. Makkuva and Chong Wang. I benefit considerably from discussing and collabrating with them. This dissertation would not have been possible without their contributions.

I would like to thank all the managers and colleagues during my internships at Facebook and Google, especially Chong Wang and Dengyong (Denny) Zhou. Their guidance was crucial for me to learn how I could incorporate my research into industry.

I would like to thank all my labmates, roommates and friends at UIUC, including but not limited to Yuheng Bu, Jiyang Chen, Yuchen Fan, Giulia Fanti, Hongyu Gong, Peter Kairouz, Ashish K. Khetan, Pan Li, Ruochen Lu, Jiaqi Mu, Ranvir Rana, Tarek Sakakini, Du Su, Kiran K. Thekumparampil, Shaileshh B. Vankatakrishnan, Gerui Wang, Xuechao Wang, Qiaomin Xie, Kaiqing Zhang, Zhenzhe Zheng and Shaofeng Zou. I will always remember the happy moments I spent with them on this beautiful campus.

Finally, I would like to thank my parents and family for providing me with support and encouragement throughout my years of study. Without their support from across the Pacific Ocean, this dissertation would not have been possible. To them I dedicate this dissertation.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Information theory, originally proposed by Claude Shannon in 1948 [1], studies the fundamental limit on quantization, storage and communication of data. During the past decades, information theory has gained great success in the area of wireless communication, data compression, statistic inference, natural language processing and numerous other fields.

Since the development of Internet in the beginning of the 21st century, people get access to a huge amount data from different aspects. How to understand, analyze and utilize the big data is of great interest to both theorists and practitioners. Information theory — the fundamental mathematical tool of understanding big data — can be widely applied in the big data era.

However, the focus of traditional information theory is mostly on the population level. Several information-theoretic quantities such as information entropy, mutual information and Kullback-Leibler divergence are defined and studied based on the probability distribution of data. In the big data era, we usually have sampled data but not the distribution, hence bridging this gap between theoretical information theory and practical application is an important problem.

In this dissertation, we aim to build this bridge between information theory and big data, by taking the following steps from theory toward practice.

- Analyze and understand the theory of the fundamental limit for estimating information-theoretic quantities from data samples.
- Design novel algorithms to compute traditional information-theoretic quantities for some challenging but practically useful scenarios, and new information-theoretic measures capturing complicated relationships from different aspects of multi-view data.
- Use information-theoretic tools to understand and improve modern machine learning models, especially deep learning models.

1.1 Understanding fundamental limit

Before applying information-theoretic tools to machine learning and data science, we need to estimate information-theoretic quantities, such as differential entropy and mutual information, from high-dimensional samples. That brings a fundamental question — how well can we estimate these quantities given finite data. The geometry of Euclidean space and the dimensionality of the domain bring difficulty to the problem. In this dissertation, we studied the estimation of differential entropy and mutual information separately, and provide a breakthrough on the understanding of fundamental limit of information-theoretic quantity estimators.

We analyze the Kozachenko–Leonenko (KL) fixed k-nearest neighbor estimator for the differential entropy [2] in Chapter 2. We obtain the first uniform upper bound on its performance for any fixed k over Hölder balls on a torus without assuming any conditions on how close the density could be from zero. Accompanying a recent minimax lower bound over the Hölder ball, we show that the KL estimator for any fixed k is achieving the minimax rates up to logarithmic factors without cognizance of the smoothness parameter s of the Hölder ball for $s \in (0, 2]$ and arbitrary dimension d, rendering it the first estimator that provably satisfies this property [3].

For the problem of estimating mutual information, the most popular estimator is one proposed by Kraskov and Stögbauer and Grassberger (KSG) [4], and is nonparametric and based on fixed k-nearest neighbor distances as well. Despite its widespread use, theoretical properties of this estimator have been largely unexplored. In Chapter 3, we demonstrate that the estimator is consistent and also identify an upper bound on the rate of convergence of the ℓ_2 error as a function of number of samples. We argue that the performance benefits of the KSG estimator stems from a curious "correlation boosting" effect and build on this intuition to modify the KSG estimator in novel ways to construct a superior estimator [5].

1.2 Designing novel algorithms and measures

Given our understanding of the theoretical fundamental limit for estimating information theoretical quantities, we want to bring it into application. How-

ever in practice, data usually has certain special structure, which requires us to design novel algorithms for these specific practical cases.

Following the theoretical understanding, we notice that the dimensionality affects the performance of differential entropy and mutual information estimators dramatically. In the big data era, data is usually high-dimensional but have relatively low intrinsic dimension. The basic issue of k-NN entropy/mutual information estimators is that they are unable to take advantage of the small intrinsic dimension.

In Chapter 4, we propose an estimator that can take this advantage. State-of-the-art approaches have been either geometric (nearest neighbor (NN) based) or kernel based (with a globally chosen bandwidth). In this chapter, we combine both these approaches to design new estimators of entropy and mutual information that outperform state-of-the-art methods. Our estimator borrows the idea from Local Likelihood Densisty Estimator (LLDE) [6, 7] and uses local bandwidth choices of k-NN distances with a finite k, independent of the sample size. Such a local and data dependent choice ameliorates boundary bias and improves performance in practice, but the bandwidth is vanishing at a fast rate, leading to a non-vanishing bias. We show that the asymptotic bias of the proposed estimator is *universal*; it is independent of the underlying distribution. Hence, it can be pre-computed and subtracted from the estimate. As a byproduct, we obtain a unified way of obtaining *both* kernel and NN estimators. The corresponding theoretical contribution relating the asymptotic geometry of nearest neighbors to order statistics is of independent mathematical interest [8].

Previous research on mutual information estimators focus on either of two cases — the data is either purely discrete or purely continuous, whereas mutual information is a well-defined quantity in general probability spaces. But in practical downstream applications, we often have to deal with a mixture of continuous and discrete data. The data can be mixed in several ways: (i) one dimension of data is continuous and another dimension is discrete; (ii) a single-dimensional data can be a mixture of discrete and continuous components; (iii) any dimension of the data can be a mixture. In the aforementioned cases, mutual information is well-defined, but no algorithms have been studied.

In Chapter 5, we designed an algorithm that estimates mutual information from data for all the aforementioned cases. The algorithm is based on KSG

mutual information estimator, but automatically detects which case of mixture the data is by examining the k-nearest neighbor distances. We prove that the estimator is ℓ_2 consistent and demonstrate its excellent practical performance through several experiments [9].

Beyond understanding and designing algorithms for computing traditional information-theoretic quantities, we also want to use these algorithms to extract useful information from data. A common task in machine learning is to discover the underlying complicated relationship among various aspects of data. To solve this problem, we need to develop appropriate information-theoretic measures for different scenarios and develop efficient algorithms for the new measures.

While existing correlation measures such as mutual information and Shannon capacity studied before are suitable for discovering average correlation, they fail to discover hidden or potential correlations.

In Chapter 6, we postulate a set of natural axioms that we expect a measure of potential correlation to satisfy and show that the rate of information bottleneck, i.e., the hypercontractivity coefficient [10], satisfies all the proposed axioms. Then we design a novel estimator to estimate the hypercontractivity coefficient from samples and provide numerical experiments demonstrating that this proposed estimator discovers potential correlations among various indicators of WHO datasets [11], is robust in discovering gene interactions from gene expression time series data, and is statistically more powerful than the estimators for other correlation measures in binary hypothesis testing of canonical potential correlations [12].

1.3 Applications to deep learning

Since the success of AlexNet [13], deep learning, or artificial neural networks has been widely used in practise and thoroughly studied in theory. Armed with our understanding of information-theoretic tools, we could help deep learners understand the training of deep neural networks well.

Training of neural networks is a challenging problem due to its well-known non-convex landscape. Significant advances have been made recently on training neural networks, where the main challenge is in solving an optimization problem with abundant critical points. However, existing ap-

proaches [14] to address this issue crucially rely on a restrictive assumption: the training data is drawn from a Gaussian distribution.

In Chapter 7, we provide a novel unified framework to design loss functions with desirable landscape properties for a wide range of general input distributions. On these loss functions, remarkably, stochastic gradient descent theoretically recovers the true parameters with global initialization and empirically outperforms the existing approaches. Our loss function design bridges the notion of score functions [15] with the topic of neural network optimization.

Central to our approach is the task of estimating the score function from samples, which is of basic and independent interest to theoretical statistics. Traditional estimation methods fail right at the outset. We bring statistical methods of local likelihood to design a novel estimator of score functions, that provably adapts to the local geometry of the unknown density [16].

Besides providing deeper understanding of the training process of neural nets, information theory can also help improving the efficiency of neural networks. The enormous size of modern deep neural nets makes it challenging to deploy those models in memory and communication limited scenarios. Thus, compressing a trained model without a significant loss in performance has become an increasingly important task. Tremendous advances [17, 18] have been made recently, where the main technical building blocks are pruning, quantization, and low-rank factorization.

In Chapter 8, we propose principled approaches to improve upon the common heuristics used in those building blocks, by studying the fundamental limit for model compression via the rate distortion theory [19]. We prove a lower bound for the rate distortion function for model compression and prove its achievability for linear models. Although this achievable compression scheme is intractable in practice, this analysis motivates a novel objective function for model compression, which can be used to improve classes of the model compressor such as pruning or quantization. Theoretically, we prove that the proposed scheme is optimal for compressing one-hidden-layer ReLU neural networks. Empirically, we show that the proposed scheme improves upon the baseline in the compression-accuracy tradeoff [20].

CHAPTER 2

ANALYSIS OF NEAREST NEIGHBOR DIFFERENTIAL ENTROPY ESTIMATOR

Information-theoretic measures such as entropy, Kullback-Leibler divergence and mutual information quantify the amount of information among random variables. They have many applications in modern machine learning tasks, such as classification [21], clustering [22, 23, 24, 25] and feature selection [26, 27]. Information-theoretic measures and their variants can also be applied in several data science domains such as causal inference [28], sociology [11] and computational biology [29]. Estimating information-theoretic measures from data is a crucial sub-routine in the aforementioned applications and has attracted much interest in statistics community. In this chapter, we study the problem of estimating Shannon differential entropy, which is the basis of estimating other information-theoretic measures for continuous random variables.

Suppose we observe n independent identically distributed random vectors $\mathbf{X} = \{X_1, \ldots, X_n\}$ drawn from density function f where $X_i \in \mathbb{R}^d$. We consider the problem of estimating the differential entropy

$$h(f) \quad = -\int f(x) \ln f(x) dx, \tag{2.1}$$

from the empirical observations $\mathbf{X}$. The fundamental limit of estimating the differential entropy is given by the minimax risk

$$\inf_{\hat{h}} \sup_{f \in \mathcal{F}} \left(\mathbb{E}(\hat{h}(\mathbf{X}) - h(f))^2 \right)^{1/2}, \tag{2.2}$$

where the infimum is taken over all estimators $\hat{h}$ that is a function of the empirical data $\mathbf{X}$. Here $\mathcal{F}$ denotes a (nonparametric) class of density functions.

The problem of differential entropy estimation has been investigated extensively in the literature. As discussed in [30], there exist two main approaches, where one is based on kernel density estimators [31], and the other is based

on the nearest neighbor methods [32, 33, 34, 35, 36], which is pioneered by the work of [2].

The problem of differential entropy estimation lies in the general problem of estimating nonparametric functionals. Unlike the parametric counterparts, the problem of estimating nonparametric functionals is challenging even for smooth functionals. Initial efforts have focused on inference of linear, quadratic, and cubic functionals in Gaussian white noise and density models and have laid the foundation for the ensuing research. We do not attempt to survey the extensive literature in this area, but instead refer to the interested reader to, e.g., [37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47] and the references therein. For non-smooth functionals such as entropy, there is some recent progress [48, 49, 50] on designing theoretically minimax optimal estimators, while these estimators typically require the knowledge of the smoothness parameters, and the practical performances of these estimators are not yet known.

The k-nearest-neighbor differential entropy estimator, or the Kozachenko-Leonenko (KL) estimator is computed in the following way. Let $R_{i,k}$ be the distance between X_i and its k-nearest neighbor among the remaining samples $\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$. Precisely, $R_{i,k}$ equals the k-th smallest number in the list $\{\|X_i - X_j\| : j \neq i, j \in [n]\}$, here $[n] = \{1, 2, \ldots, n\}$. Let $B(x, \rho)$ denote the closed ℓ_2 ball centered at x of radius ρ and λ be the Lebesgue measure on $\mathbb{R}^d$. The KL differential entropy estimator is defined as

$$\hat{h}_{n,k}(\mathbf{X}) \quad = \ln k - \psi(k) + \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \ln \left(\tfrac{n}{k} \lambda(B(X_i, R_{i,k})) \right), \tag{2.3}$$

where $\psi(x)$ is the digamma function with $\psi(1) = -\gamma$, $\gamma = -\int_0^\infty e^{-t} \ln t dt = 0.5772156\ldots$ is the Euler–Mascheroni constant.

There exists an intuitive explanation behind the construction of the KL differential entropy estimator. Writing informally, we have

$$h(f) = \mathbb{E}_f[-\ln f(X)] \approx \frac{1}{n} \sum_{i=1}^{n} - \ln f(X_i) \approx \frac{1}{n} \sum_{i=1}^{n} - \ln \hat{f}(X_i), \tag{2.4}$$

where the first approximation is based on the law of large numbers, and in the second approximation we have replaced f by a nearest neighbor density estimator $\hat{f}$. The nearest neighbor density estimator $\hat{f}(X_i)$ follows from the

"intuition"[1] that

$$\hat{f}(X_i)\lambda(B(X_i, R_{i,k})) \approx \frac{k}{n}. \tag{2.5}$$

Here the final additive bias correction term $\ln k - \psi(k)$ follows from a detailed analysis of the bias of the KL estimator, which will become apparent later.

We focus on the regime where k is a fixed constant: in other words, it does not grow as the number of samples n increases. The fixed k version of the KL estimator is widely applied in practice and enjoys smaller computational complexity, see [34].

There exists extensive literature on the analysis of the KL differential entropy estimator, which we refer to [51] for a recent survey. One of the major difficulties in analyzing the KL estimator is that the nearest neighbor density estimator exhibits a huge bias when the density is small. Indeed, it was shown in [52] that the bias of the nearest neighbor density estimator in fact does not vanish even when $n \to \infty$ and deteriorates as $f(x)$ gets close to zero. In the literature, a large collection of work assume that the density is uniformly bounded away from zero [53, 54, 55, 31, 33], while others put various assumptions quantifying on average how close the density is to zero [56, 57, 32, 58, 8, 34, 35]. In this chapter, we focus on removing assumptions on how close the density is to zero.

Main contributions of Chapter 2:

Let $\mathcal{H}_d^s(L; [0,1]^d)$ be the Hölder ball in the unit cube (torus) (formally defined later in Section 2.1) and $s \in (0, 2]$ is the Hölder smoothness parameter. Then, the worst-case risk of the fixed k-nearest neighbor differential entropy estimator over $\mathcal{H}_d^s(L; [0,1]^d)$ is controlled by the following theorem.

Theorem 1. *Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be i.i.d. samples from density function f. Then, for $0 < s \leq 2$, the fixed k-nearest neighbor KL differential entropy estimator $\hat{h}_{n,k}$ in (2.3) satisfies*

$$\begin{aligned} &\left(\sup_{f \in \mathcal{H}_d^s(L;[0,1]^d)} \mathbb{E}_f \left(\hat{h}_{n,k}(\mathbf{X}) - h(f) \right)^2 \right)^{\frac{1}{2}} \\ \leq \;& C \left(n^{-\frac{s}{s+d}} \ln(n+1) + n^{-\frac{1}{2}} \right), \end{aligned} \tag{2.6}$$

[1] Precisely, we have $\int_{B(X_i, R_{i,k})} f(u)du \sim \mathsf{Beta}(k, n-k)$ [51, Chap. 1.2]. A $\mathsf{Beta}(k, n-k)$ distributed random variable has mean $\frac{k}{n}$.

where C is a constant depends only on s, L, k and d.

The KL estimator is in fact nearly minimax up to logarithmic factors, as shown in the following result from [49].

Theorem 2. *[49] Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be i.i.d. samples from density function f. Then, there exists a constant L_0 depending on s, d only such that for all $L \geq L_0, s > 0$,*

$$\begin{aligned} &\left(\inf_{\hat{h}} \sup_{f \in \mathcal{H}_d^s(L;[0,1]^d)} \mathbb{E}_f \left(\hat{h}(\mathbf{X}) - h(f) \right)^2 \right)^{\frac{1}{2}} \\ \geq \;& c \left(n^{-\frac{s}{s+d}} (\ln(n+1))^{-\frac{s+2d}{s+d}} + n^{-\frac{1}{2}} \right), \end{aligned} \tag{2.7}$$

where c is a constant depends only on s, L and d.

Remark 1. *We emphasize that one cannot remove the condition $L \geq L_0$ in Theorem 2. Indeed, if the Hölder ball has a too small width, then the density itself is bounded away from zero, which makes the differential entropy a smooth functional, with minimax rates $n^{-\frac{4s}{4s+d}} + n^{-1/2}$ [59, 60, 61].*

Theorems 1 and 2 imply that for any fixed k, the KL estimator achieves the minimax rates up to logarithmic factors without knowing s for all $s \in (0, 2]$, which implies that it is near minimax rate-optimal (within logarithmic factors) when the dimension $d \leq 2$. We cannot expect the vanilla version of the KL estimator to adapt to higher order of smoothness since the nearest neighbor density estimator can be viewed as a variable width kernel density estimator with the box kernel, and it is well known in the literature (see, e.g., [62, Chapter 1]) that any positive kernel cannot exploit the smoothness $s > 2$. We refer to [49] for a more detailed discussion on this difficulty and potential solutions. The Jackknife idea, such as the one presented in [35, 36] might be useful for adapting to $s > 2$.

The significance of our work is multi-folded:

- We obtain the first uniform upper bound on the performance of the fixed k-nearest neighbor KL differential entropy estimator over Hölder balls without assuming how close the density could be from zero. We emphasize that assuming conditions of this type, such as the density is bounded away from zero, could make the problem significantly easier.

For example, if the density f is assumed to satisfy $f(x) \geq c$ for some constant $c > 0$, then the differential entropy becomes a *smooth* functional and consequently, the general technique for estimating smooth nonparametric functionals [59, 60, 61] can be directly applied here to achieve the minimax rates $n^{-\frac{4s}{4s+d}} + n^{-1/2}$. The main technical tools that enabled us to remove the conditions on how close the density could be from zero are the Besicovitch covering lemma (Lemma. 4) and the generalized Hardy–Littlewood maximal inequality.

- We show that, for any fixed k, the k-nearest neighbor KL entropy estimator nearly achieves the minimax rates without knowing the smoothness parameter s. In the functional estimation literature, designing estimators that can be theoretically proved to adapt to unknown levels of smoothness is usually achieved using the Lepski method [63, 64, 65, 66, 50], which is not known to be performing well in general in practice. On the other hand, a simple plug-in approach can achieves the rate of $n^{-s/(s+d)}$, but only when s is known [49]. The KL estimator is well known to exhibit excellent empirical performance, but existing theory has not yet demonstrated its near-"optimality" when the smoothness parameter s is not known. Recent works [36, 34, 35] analyzed the performance of the KL estimator under various assumptions on how close the density could be to zero, with no matching lower bound up to logarithmic factors in general. Our work makes a step towards closing this gap and provides a theoretical explanation for the wide usage of the KL estimator in practice.

The rest of the chapter is organized as follows. In Section 2.1 we formally discuss the definition of Hölder balls. Section 2.2 and Section 2.3 are dedicated to the proof of bias and variance in Theorem 1. Section 2.4 provides proof to the lemmas.

Notations. For positive sequences a_γ, b_γ, we use the notation $a_\gamma \lesssim_\alpha b_\gamma$ to denote that there exists a universal constant C that only depends on α such that $\sup_\gamma \frac{a_\gamma}{b_\gamma} \leq C$, and $a_\gamma \gtrsim_\alpha b_\gamma$ is equivalent to $b_\gamma \lesssim_\alpha a_\gamma$. Notation $a_\gamma \asymp_\alpha b_\gamma$ is equivalent to $a_\gamma \lesssim_\alpha b_\gamma$ and $b_\gamma \lesssim_\alpha a_\gamma$. We write $a_\gamma \lesssim b_\gamma$ if the constant is universal and does not depend on any parameters. Notation $a_\gamma \gg b_\gamma$ means that $\liminf_\gamma \frac{a_\gamma}{b_\gamma} = \infty$, and $a_\gamma \ll b_\gamma$ is equivalent to $b_\gamma \gg a_\gamma$.

We write $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

2.1 Definition of Hölder ball

In order to define the Hölder ball in the unit cube $[0,1]^d$, we first review the definition of Hölder ball in $\mathbb{R}^d$.

Definition 1 (Hölder ball in $\mathbb{R}^d$)**.** *The Hölder ball $\mathcal{H}_d^s(L; \mathbb{R}^d)$ is specified by the parameters $s > 0$ (order of smoothness), $d \in \mathbb{Z}_+$ (dimension of the argument) and $L > 0$ (smoothness constant) and is as follows. A positive real s can be uniquely represented as*

$$s = m + \alpha, \tag{2.8}$$

where m is a non-negative integer and $0 < \alpha \leq 1$. By definition, $\mathcal{H}_d^s(L; \mathbb{R}^d)$ is comprised of all m times continuously differentiable functions

$$f : \mathbb{R}^d \mapsto \mathbb{R}, \tag{2.9}$$

with Hölder continuous, with exponent α and constant L, derivatives of order m:

$$\begin{aligned} & |D^m f(x)[\delta_1, \ldots, \delta_m] - D^m f(x')[\delta_1, \ldots, \delta_m]| \\ \leq \; & L\|x - x'\|^\alpha \|\delta\|^m, \quad \forall x, x' \in \mathbb{R}^d, \delta \in \mathbb{R}^d. \end{aligned} \tag{2.10}$$

Here $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^d$, and $D^m f(x)[\delta_1, \ldots, \delta_m]$ is the m-th differential of f taken as a point x along the directions $\delta_1, \ldots, \delta_m$:

$$\begin{aligned} & D^m f(x)[\delta_1, \ldots, \delta_m] \\ = \; & \left. \frac{\partial^m}{\partial t_1 \ldots \partial t_m} \right|_{t_1 = \ldots = t_m = 0} f(x + t_1\delta_1 + \ldots + t_m\delta_m). \end{aligned} \tag{2.11}$$

In this chapter, we consider functions that lie in Hölder balls in $[0,1]^d$. The Hölder ball in the compact set $[0,1]^d$ is defined as follows.

Definition 2 (Hölder ball in the unit cube)**.** *A function $f : [0,1]^d \mapsto \mathbb{R}$ is said to belong to the Hölder ball $\mathcal{H}_d^s(L; [0,1]^d)$ if and only if there exists*

another function $f_1 \in \mathcal{H}_d^s(L;\mathbb{R}^d)$ *such that*

$$f(x) \;\; = f_1(x), \quad x \in [0,1], \tag{2.12}$$

and $f_1(x)$ *is a* 1*-periodic function in each variable. Here* $\mathcal{H}_d^s(L;[0,1]^d)$ *is introduced in Definition 1. In other words,*

$$f_1(x+e_j) \;\; = f_1(x), \quad \forall x \in \mathbb{R}^d, 1 \le j \le d, \tag{2.13}$$

where $\{e_j : 1 \le j \le d\}$ *is the standard basis in* $\mathbb{R}^d$*.*

Definition 2 has appeared in the literature [67]. It is motivated by the observations that sliding window kernel methods usually cannot deal with the boundary effects without additional assumptions [68]. Indeed, near the boundary the sliding window kernel density estimator may have a significantly larger bias than that of the interior points. In the nonparametric statistics literature, it is usually assumed that the density has its value and all the derivatives vanishing at the boundary, which is stronger than our assumptions.

2.2 Upper bound of bias

In this and the following section, we will prove that

$$\left(\mathbb{E}\left(\hat{h}_{n,k}(\mathbf{X}) - h(f)\right)^2\right)^{\frac{1}{2}} \lesssim_{s,L,d,k} n^{-\frac{s}{s+d}} \ln(n+1) + n^{-\frac{1}{2}}, \tag{2.14}$$

for any $f \in \mathcal{H}_d^s(L;[0,1]^d)$ and $s \in (0,2]$. The proof consists two parts: (i) the upper bound of the bias in the form of $O_{s,L,d,k}(n^{-s/(s+d)}\ln(n+1))$; (ii) the upper bound of the variance is $O_{s,L,d,k}(n^{-1})$. Below we show the bias proof and the variance proof is in the next section.

First, we introduce the following notation

$$f_t(x) \;\; = \;\; \frac{\mu(B(x,t))}{\lambda(B(x,t))} = \frac{1}{V_d t^d}\int_{u:|u-x|\le t} f(u)du. \tag{2.15}$$

Here μ is the probability measure specified by density function f on the torus, λ is the Lebesgue measure on $\mathbb{R}^d$, and $V_d = \pi^{d/2}/\Gamma(1+d/2)$ is the Lebesgue

measure of the unit ball in d-dimensional Euclidean space. Hence $f_t(x)$ is the average density of a neighborhood near x. We first state two main lemmas about $f_t(x)$ which will be used later in the proof.

Lemma 1. *If $f \in \mathcal{H}_d^s(L;[0,1]^d)$ for some $0 < s \le 2$, then for any $x \in [0,1]^d$ and $t > 0$, we have*

$$| f_t(x) - f(x) | \;\le\; \frac{dLt^s}{s+d}. \tag{2.16}$$

Lemma 2. *If $f \in \mathcal{H}_d^s(L;[0,1]^d)$ for some $0 < s \le 2$ and $f(x) \ge 0$ for all $x \in [0,1]^d$, then for any x and any $t > 0$, we have*

$$f(x) \;\lesssim_{s,L,d}\; \max\Big\{ f_t(x), \big(f_t(x) V_d t^d \big)^{s/(s+d)} \Big\}. \tag{2.17}$$

Furthermore, $f(x) \lesssim_{s,L,d} 1$.

Now we investigate the bias of $\hat{h}_{n,k}(\mathbf{X})$. The following argument reduces the bias analysis of $\hat{h}_{n,k}(\mathbf{X})$ to a function analytic problem. For notation simplicity, we introduce a new random variable $X \sim f$ independent of $\{X_1,\dots,X_n\}$ and study $\hat{h}_{n+1,k}(\{X_1,\dots,X_n,X\})$. For every $x \in \mathbb{R}^d$, denote $R_k(x)$ by the k-nearest neighbor distance from x to $\{X_1,X_2,\dots,X_n\}$ under distance $d(x,y) = \min_{m\in\mathbb{Z}^d} \|m+x-y\|$, i.e., the k-nearest neighbor distance on the torus. Then,

$$\begin{aligned}
& \mathbb{E}[\hat{h}_{n+1,k}(\{X_1,\dots,X_n,X\})] - h(f) \\
= \;& -\psi(k) + \mathbb{E}\left[\ln\big((n+1)\lambda(B(X,R_k(X))) \big)\right] + \mathbb{E}\left[\ln f(X)\right] \\
= \;& \mathbb{E}\left[\ln\left(\frac{f(X)\lambda(B(X,R_k(X)))}{\mu(B(X,R_k(X)))} \right) \right] \\
& + \mathbb{E}\left[\ln\big((n+1)\mu(B(X,R_k(X)))\big)\right] - \psi(k) \\
= \;& \mathbb{E}\left[\ln \frac{f(X)}{f_{R_k(X)}(X)}\right] + \left(\mathbb{E}\left[\ln\big((n+1)\mu(B(X,R_k(X)))\big)\right] - \psi(k) \right).
\end{aligned} \tag{2.18}$$

We first show that the second term $\mathbb{E}\left[\ln\big((n+1)\mu(B(X,R_k(X)))\big)\right] - \psi(k)$ can be universally controlled regardless of the smoothness of f. Indeed, the random variable $\mu(B(X,R_k(X))) \sim \mathsf{Beta}(k,n+1-k)$ [51, Chap. 1.2] and it was shown in [51, Theorem 7.2] that there exists a universal constant $C > 0$

such that

$$\left| \mathbb{E}\left[\ln\left((n+1)\mu(B(X, R_k(X)))\right)\right] - \psi(k) \right| \;\leq\; \frac{C}{n}. \tag{2.19}$$

Hence, it suffices to show that for $0 < s \leq 2$,

$$\left| \mathbb{E}\left[\ln \frac{f(X)}{f_{R_k(X)}(X)}\right] \right| \;\lesssim_{s,L,d,k}\; n^{-\frac{s}{s+d}} \ln(n+1). \tag{2.20}$$

We split our analysis into two parts. In Section 2.2.1, we will show that $\mathbb{E}\left[\ln \frac{f_{R_k(X)}(X)}{f(X)}\right] \lesssim_{s,L,d,k} n^{-\frac{s}{s+d}}$. In Section 2.2.2, we will show that $\mathbb{E}\left[\ln \frac{f(X)}{f_{R_k(X)}(X)}\right] \lesssim_{s,L,d,k} n^{-\frac{s}{s+d}} \ln(n+1)$, which completes the proof.

2.2.1 Upper bound on $\mathbb{E}\left[\ln \frac{f_{R_k(X)}(X)}{f(X)}\right]$

By the fact that $\ln y \leq y - 1$ for any $y > 0$, we have

$$\begin{aligned} & \mathbb{E}\left[\ln \frac{f_{R_k(X)}(X)}{f(X)}\right] \leq \mathbb{E}\left[\frac{f_{R_k(X)}(X) - f(X)}{f(X)}\right] \\ = & \int_{[0,1]^d \cap \{x: f(x) \neq 0\}} \left(\mathbb{E}[f_{R_k(x)}(x)] - f(x)\right) dx. \end{aligned} \tag{2.21}$$

Here the expectation is taken with respect to the randomness in $R_k(x) = \min_{1 \leq i \leq n, m \in \mathbb{Z}^d} \|m + X_i - x\|, x \in \mathbb{R}^d$. Define function $g(x; f, n)$ as

$$g(x; f, n) \;=\; \sup\left\{ u \geq 0 : V_d u^d f_u(x) \leq \frac{1}{n} \right\}, \tag{2.22}$$

where $g(x; f, n)$ intuitively means the distance R such that the probability mass $\mu(B(x, R))$ within R is $1/n$. Then for any $x \in [0,1]^d$, we can split $\mathbb{E}[f_{R_k(x)}(x)] - f(x)$ into three terms as

$$\begin{aligned} & \mathbb{E}[f_{R_k(x)}(x)] - f(x) \\ = \; & \mathbb{E}[(f_{R_k(x)}(x) - f(x))\mathbb{I}(R_k(x) \leq n^{-1/(s+d)})] \\ + \; & \mathbb{E}[(f_{R_k(x)}(x) - f(x))\mathbb{I}(n^{-1/(s+d)} < R_k(x) \leq g(x; f, n))] \\ + \; & \mathbb{E}[(f_{R_k(x)}(x) - f(x))\mathbb{I}(R_k(x) > g(x; f, n) \vee n^{-1/(s+d)})] \\ = \; & C_1 + C_2 + C_3. \end{aligned} \tag{2.23}$$

Now we handle three terms separately. Our goal is to show that for every $x \in [0,1]$, $C_i \lesssim_{s,L,d} n^{-s/(s+d)}$ for $i \in \{1,2,3\}$. Then, taking the integral with respect to x leads to the desired bound.

1. Term C_1: whenever $R_k(x) \leq n^{-1/(s+d)}$, by Lemma 1, we have

$$|f_{R_k(x)}(x) - f(x)| \leq \frac{dLR_k(x)^s}{s+d} \lesssim_{s,L,d} n^{-s/(s+d)}, \tag{2.24}$$

which implies that

$$\begin{aligned} C_1 &\leq \mathbb{E}\left[\left|f_{R_k(x)}(x) - f(x)\right| \mathbb{I}(R_k(x) \leq n^{-1/(s+d)})\right] \\ &\lesssim_{s,L,d} n^{-s/(s+d)}. \end{aligned} \tag{2.25}$$

2. Term C_2: whenever $R_k(x)$ satisfies that $n^{-1/(s+d)} < R_k(x) \leq g(x; f, n)$, by definition of $g(x; f, n)$, we have $V_d R_k(x)^d f_{R_k(x)}(x) \leq \frac{1}{n}$, which implies that

$$f_{R_k(x)}(x) \leq \frac{1}{nV_d R_k(x)^d} \leq \frac{1}{nV_d n^{-d/(s+d)}} \lesssim_{s,L,d} n^{-s/(s+d)}. \tag{2.26}$$

It follows from Lemma 2 that in this case

$$\begin{aligned} & f(x) \lesssim_{s,L,d} f_{R_k(x)}(x) \vee \left(f_{R_k(x)}(x) V_d R_k(x)^d \right)^{s/(s+d)} \\ \leq\ & n^{-s/(s+d)} \vee n^{-s/(s+d)} = n^{-s/(s+d)}. \end{aligned} \tag{2.27}$$

Hence, we have

$$\begin{aligned} C_2 &= \mathbb{E}\left[(f_{R_k(x)}(x) - f(x))\mathbb{I}\left(n^{-1/(s+d)} < R_k(x) \leq g(x; f, n)\right)\right] \\ &\leq \mathbb{E}\left[(f_{R_k(x)}(x) + f(x))\mathbb{I}\left(n^{-1/(s+d)} < R_k(x) \leq g(x; f, n)\right)\right] \\ &\lesssim_{s,L,d} n^{-s/(s+d)}. \end{aligned} \tag{2.28}$$

3. Term C_3: we have

$$C_3 \leq \mathbb{E}\left[(f_{R_k(x)}(x) + f(x))\mathbb{I}\left(R_k(x) > g(x; f, n) \vee n^{-1/(s+d)}\right)\right]. \tag{2.29}$$

For any x such that $R_k(x) > n^{-1/(s+d)}$, we have

$$f_{R_k(x)}(x) \lesssim_{s,L,d} V_d R_k(x)^d f_{R_k(x)}(x) n^{d/(s+d)}, \tag{2.30}$$

and by Lemma 2,

$$\begin{aligned} & f(x) \lesssim_{s,L,d} f_{R_k(x)}(x) \vee (V_d R_k(x)^d f_{R_k(x)}(x))^{s/(s+d)} \\ \leq \; & f_{R_k(x)}(x) + (V_d R_k(x)^d f_{R_k(x)}(x))^{s/(s+d)}. \end{aligned} \tag{2.31}$$

Hence,

$$\begin{aligned} & f(x) + f_{R_k(x)}(x) \\ \lesssim_{s,L,d} \; & 2 f_{R_k(x)}(x) + (V_d R_k(x)^d f_{R_k(x)}(x))^{s/(s+d)} \\ \lesssim_{s,L,d} \; & V_d R_k(x)^d f_{R_k(x)}(x) n^{d/(s+d)} + (V_d R_k(x)^d f_{R_k(x)}(x))^{s/(s+d)} \\ \lesssim_{s,L,d} \; & V_d R_k(x)^d f_{R_k(x)}(x) n^{d/(s+d)}, \end{aligned} \tag{2.32}$$

where in the last step we have used the fact that $V_d R_k(x)^d f_{R_k(x)}(x) > n^{-1}$ since $R_k(x) > g(x; f, n)$. Finally, we have

$$\begin{aligned} & C_3 \lesssim_{s,L,d} n^{d/(s+d)} \mathbb{E}[(V_d R_k(x)^d f_{R_k(x)}(x)) \mathbb{I}(R_k(x) > g(x; f, n))] \\ = \; & n^{d/(s+d)} \mathbb{E}\left[(V_d R_k(x)^d f_{R_k(x)}(x)) \mathbb{I}\left(V_d R_k(x)^d f_{R_k(x)}(x) > 1/n\right)\right]. \end{aligned} \tag{2.33}$$

Note that $V_d R_k(x)^d f_{R_k(x)}(x) \sim \mathsf{Beta}(k, n+1-k)$, and if $Y \sim \mathsf{Beta}(k, n+1-k)$, we have

$$\mathbb{E}[Y^2] = \left(\frac{k}{n+1}\right)^2 + \frac{k(n+1-k)}{(n+1)^2(n+2)} \lesssim_k \frac{1}{n^2}. \tag{2.34}$$

Notice that $\mathbb{E}[Y \mathbb{I}(Y > 1/n)] \leq n \mathbb{E}[Y^2]$. Hence, we have

$$\begin{aligned} C_3 \quad & \lesssim_{s,L,d} \quad n^{d/(s+d)}\, n\, \mathbb{E}\left[(V_d R_k(x)^d f_{R_k(x)}(x))^2\right] \\ & \lesssim_{s,L,d,k} \quad \frac{n^{d/(s+d)} n}{n^2} = n^{-s/(s+d)}. \end{aligned} \tag{2.35}$$

2.2.2 Upper bound on $\mathbb{E}\left[\ln \frac{f(X)}{f_{R_k(X)}(X)}\right]$

By splitting the term into two parts, we have

$$\begin{aligned}
\mathbb{E}\left[\ln \frac{f(X)}{f_{R_k(X)}(X)}\right] &= \mathbb{E}\left[\int_{[0,1]^d \cap \{x: f(x)\neq 0\}} f(x) \ln \frac{f(x)}{f_{R_k(x)}(x)} dx\right] \\
= \quad & \mathbb{E}\left[\int_A f(x) \ln \frac{f(x)}{f_{R_k(x)}(x)} \mathbb{I}(f_{R_k(x)}(x) > n^{-s/(s+d)}) dx\right] \\
+ \quad & \mathbb{E}\left[\int_A f(x) \ln \frac{f(x)}{f_{R_k(x)}(x)} \mathbb{I}(f_{R_k(x)}(x) \leq n^{-s/(s+d)}) dx\right] \\
= \quad & C_4 + C_5. \qquad (2.36)
\end{aligned}$$

Here we denote $A = [0,1]^d \cap \{x : f(x) \neq 0\}$ for simplicity of notation. For the term C_4, we have

$$\begin{aligned}
C_4 &\leq \mathbb{E}\left[\int_A f(x) \left(\frac{f(x) - f_{R_k(x)}(x)}{f_{R_k(x)}(x)}\right) \mathbb{I}(f_{R_k(x)}(x) > n^{-s/(s+d)}) dx\right] \\
= \quad & \mathbb{E}\left[\int_A \frac{(f(x) - f_{R_k(x)}(x))^2}{f_{R_k(x)}(x)} \mathbb{I}(f_{R_k(x)}(x) > n^{-s/(s+d)}) dx\right] \\
+ \quad & \mathbb{E}\left[\int_A \left(f(x) - f_{R_k(x)}(x)\right) \mathbb{I}(f_{R_k(x)}(x) > n^{-s/(s+d)}) dx\right] \\
\leq \quad & n^{s/(s+d)} \mathbb{E}\left[\int_A \left(f(x) - f_{R_k(x)}(x)\right)^2 dx\right] + \mathbb{E}\left[\int_A \left(f(x) - f_{R_k(x)}(x)\right) dx\right]. \qquad (2.37)
\end{aligned}$$

In the proof of upper bound of $\mathbb{E}\left[\ln \frac{f_{R_k(X)}(X)}{f(X)}\right]$, we show that $\mathbb{E}[f_{R_k(x)}(x) - f(x)] \lesssim_{s,L,d,k} n^{-s/(s+d)}$ for any $x \in A$. Similarly as in the proof of upper bound of $\mathbb{E}\left[\ln \frac{f_{R_k(X)}(X)}{f(X)}\right]$, we have $\mathbb{E}\left[(f_{R_k(x)}(x) - f(x))^2\right] \lesssim_{s,L,d,k} n^{-2s/(s+d)}$ for every $x \in A$. Therefore, we have

$$C_4 \quad \lesssim_{s,L,d,k} \quad n^{s/(s+d)} n^{-2s/(s+d)} + n^{-s/(s+d)} \lesssim_{s,L,d,k} n^{-s/(s+d)}. \qquad (2.38)$$

Now we consider C_5. We conjecture that $C_5 \lesssim_{s,L,d,k} n^{-s/(s+d)}$ in this case, but we were not able to prove it. Below we prove that $C_5 \lesssim_{s,L,d,k} n^{-s/(s+d)} \ln(n+1)$. Define the function

$$M(x) \quad = \quad \sup_{t>0} \frac{1}{f_t(x)}. \qquad (2.39)$$

Since $f_{R_k(x)}(x) \le n^{-s/(s+d)}$, we have $M(x) = \sup_{t>0}(1/f_t(x)) \ge 1/f_{R_k(x)}(x) \ge n^{s/(s+d)}$. Denote $\ln^+(y) = \max\{\ln(y), 0\}$ for any $y > 0$, therefore, we have that

$$\begin{aligned}
& C_5 \le \mathbb{E}\left[\int_A f(x) \ln^+\left(\frac{f(x)}{f_{R_k(x)}(x)}\right) \mathbb{I}(f_{R_k(x)}(x) \le n^{-s/(s+d)})dx\right] \\
\le\ & \mathbb{E}\left[\int_A f(x) \ln^+\left(\frac{f(x)}{f_{R_k(x)}(x)}\right) \mathbb{I}(M(x) \ge n^{s/(s+d)})dx\right] \\
\le\ & \int_A f(x)\mathbb{E}\left[\ln^+\left(\frac{1}{(n+1)V_d R_k(x)^d f_{R_k(x)}(x)}\right)\right] \mathbb{I}(M(x) \ge n^{s/(s+d)})dx \\
+\ & \int_A f(x)\mathbb{E}\left[\ln^+\left((n+1)V_d R_k(x)^d f(x)\right)\right] \mathbb{I}(M(x) \ge n^{s/(s+d)})dx \\
=\ & C_{51} + C_{52}, \qquad (2.40)
\end{aligned}$$

where the last inequality uses the fact $\ln^+(xy) \le \ln^+ x + \ln^+ y$ for all $x, y > 0$. As for C_{51}, since $V_d R_k(x)^d f_{R_k(x)}(x) \sim \mathsf{Beta}(k, n+1-k)$, and for $Y \sim \mathsf{Beta}(k, n+1-k)$, we have

$$\begin{aligned}
& \mathbb{E}\left[\ln^+\left(\frac{1}{(n+1)Y}\right)\right] = \int_0^{\frac{1}{n+1}} \ln\left(\frac{1}{(n+1)x}\right) p_Y(x)dx \\
=\ & \mathbb{E}\left[\ln\left(\frac{1}{(n+1)Y}\right)\right] + \int_{\frac{1}{n+1}}^1 \ln\left((n+1)x\right) p_Y(x)dx \\
\le\ & \mathbb{E}\left[\ln\left(\frac{1}{(n+1)Y}\right)\right] + \ln(n+1)\int_{\frac{1}{n+1}}^1 p_Y(x)dx \\
\le\ & \mathbb{E}\left[\ln\left(\frac{1}{(n+1)Y}\right)\right] + \ln(n+1) \\
\le\ & \ln(n+1), \qquad (2.41)
\end{aligned}$$

where in the last inequality we used the fact that $\mathbb{E}\left[\ln\left(\frac{1}{(n+1)Y}\right)\right] = \psi(n+1) - \psi(k) - \ln(n+1) \le 0$ for any $k \ge 1$. Hence,

$$C_{51} \ \lesssim_{s,L,d}\ \ln(n+1)\int_A f(x)\mathbb{I}(M(x) \ge n^{s/(s+d)})dx. \qquad (2.42)$$

Now we introduce the following lemma, which is proved in Section 2.4.

Lemma 3. *Let μ_1, μ_2 be two Borel measures that are finite on the bounded*

Borel sets of $\mathbb{R}^d$. *Then, for all* $t > 0$ *and any Borel set* $A \subset \mathbb{R}^d$,

$$\mu_1\left(\left\{x \in A : \sup_{0<\rho\le D}\left(\frac{\mu_2(B(x,\rho))}{\mu_1(B(x,\rho))}\right) > t\right\}\right) \;\le\; \frac{C_d}{t}\mu_2(A_D). \quad (2.43)$$

Here $C_d > 0$ *is a constant that depends only on the dimension* d *and*

$$A_D = \{x : \exists y \in A, |y - x| \le D\}. \quad (2.44)$$

Applying the second part of Lemma 3 with μ_2 being the Lebesgue measure and μ_1 being the measure specified by $f(x)$ on the torus, we can view the function $M(x)$ as

$$M(x) = \sup_{0<\rho\le 1/2} \frac{\mu_2(B(x,\rho))}{\mu_1(B(x,\rho))}. \quad (2.45)$$

Taking $A = [0,1]^d \cap \{x : f(x) \neq 0\}, t = n^{s/(s+d)}$, then $\mu_2(A_{\frac{1}{2}}) \le 2^d$, so we know that

$$\begin{aligned} C_{51} \;&\lesssim_{s,L,d}\; \ln(n+1)\cdot\int_A f(x)\mathbb{I}(M(x) \ge n^{s/(s+d)})dx \\ &=\; \ln(n+1)\cdot\mu_1\left(x \in [0,1]^d, f(x) \neq 0, M(x) \ge n^{s/(s+d)}\right) \\ &\le\; \ln(n+1)\cdot C_d n^{-s/(s+d)}\mu_2(A_{\frac{1}{2}}) \lesssim_{s,L,d} n^{-s/(s+d)}\ln(n+1). \end{aligned} \quad (2.46)$$

Now we deal with C_{52}. Recall that in Lemma 2, we know that $f(x) \lesssim_{s,L,d} 1$ for any x, and $R_k(x) \le 1$, so $\ln^+((n+1)V_d R_k(x)^d f(x)) \lesssim_{s,L,d} \ln(n+1)$. Therefore,

$$\begin{aligned} C_{52} \;&\lesssim_{s,L,d}\; \ln(n+1)\cdot\int_A f(x)\mathbb{I}(M(x) \ge n^{s/(s+d)})dx \\ &\lesssim_{s,L,d}\; n^{-s/(s+d)}\ln(n+1). \end{aligned} \quad (2.47)$$

Therefore, we have proved that $C_5 \le C_{51} + C_{52} \lesssim_{s,L,d} n^{-s/(s+d)}\ln(n+1)$, which completes the proof of the upper bound on $\mathbb{E}\left[\ln \frac{f(X)}{f_{R_k(X)}(X)}\right]$.

2.3 Upper bound of variance

Our goal is to prove

$$\mathrm{Var}\left(\hat{h}_{n,k}(\mathbf{X})\right) \lesssim_{d,k} \frac{1}{n}. \tag{2.48}$$

The proof is based on the analysis in [51, Section 7.2] which utilizes the Efron–Stein inequality. Let $\mathbf{X}^{(i)} = \{X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n\}$ be a set of sample where only X_i is replaced by X_i'. Then Efron–Stein inequality [69] states

$$\mathrm{Var}\left(\hat{h}_{n,k}(\mathbf{X})\right) \;\leq\; \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{h}_{n,k}(\mathbf{X}) - \hat{h}_{n,k}(\mathbf{X}^{(i)})\right)^2\right]. \tag{2.49}$$

Note that KL estimator is symmetric of sample indices, so $\hat{h}_{n,k}(\mathbf{X}) - \hat{h}_{n,k}(\mathbf{X}^{(i)})$ has the same distribution for any i. Furthermore, we bridge $\hat{h}_{n,k}(\mathbf{X})$ and $\hat{h}_{n,k}(\mathbf{X}^{(i)})$ by introducing an estimator from $n-1$ samples. Precisely, for any $i = 2, \ldots, n$, define $R_{i,k}'$ be the k-nearest neighbor distance from X_i to $\{X_2, \ldots, X_n\}$ (note that X_1 is removed), under the distance $d(x,y) = \min_{m\in\mathbb{Z}^d} \|x - y - m\|$. Define

$$\hat{h}_{n-1,k}(\mathbf{X}) = -\psi(k) + \frac{1}{n}\sum_{i=2}^{n}\ln(n\lambda(B(X_i, R_{i,k}'))). \tag{2.50}$$

Notice that $\hat{h}_{n,k}(\mathbf{X}) - \hat{h}_{n-1,k}(\mathbf{X})$ has the same distribution as $\hat{h}_{n,k}(\mathbf{X}^{(1)}) - \hat{h}_{n-1,k}(\mathbf{X})$. Therefore, the variance is bounded by

$$\begin{aligned}
&\mathrm{Var}\left(\hat{h}_{n.k}(\mathbf{X})\right) \leq \frac{n}{2}\mathbb{E}\left[\left(\hat{h}_{n,k}(\mathbf{X}) - \hat{h}_{n,k}(\mathbf{X}^{(1)})\right)^2\right] \\
=\;& 2n\mathbb{E}\left[\left(\hat{h}_{n,k}(\mathbf{X}) - \hat{h}_{n-1,k}(\mathbf{X})\right)^2\right].
\end{aligned} \tag{2.51}$$

Now we deal with the term $\mathbb{E}\left[\left(\hat{h}_{n,k}(\mathbf{X}) - \hat{h}_{n-1,k}(\mathbf{X})\right)^2\right]$. Define the indicator function $E_i^{(k)} = \mathbb{I}\{X_1 \text{ is in the } k-\text{nearest neighbor of } X_i\}$ for $i \neq 1$. Note that $R_{i,k}' = R_{i,k}$ if $E_i^{(k)} \neq 1$ and $i \neq 1$. As shown in [9, Lemma B.1], the set $S = \{i : E_i^{(k)} = 1\}$ has cardinality at most $k\beta_d$ for a constant β_d only

depends on d. Therefore, we have

$$\begin{aligned}
&\mathrm{Var}\left(\hat{h}_{n,k}(\mathbf{X})\right) \leq 2n\mathbb{E}\left[\left(\hat{h}_{n,k}(\mathbf{X}) - \hat{h}_{n-1,k}(\mathbf{X})\right)^2\right] \\
= \quad & 2n\mathbb{E}\left[\frac{1}{n^2}\left(\sum_{i\in S\cup\{1\}} \ln(n\lambda(B(X_i, R_{i,k}))) - \sum_{i\in S} \ln(n\lambda(B(X_i, R'_{i,k})))\right)^2\right] \\
\leq \quad & \frac{2+4|S|}{n}\mathbb{E}\left[\sum_{i\in S\cup\{1\}} \ln^2(n\lambda(B(X_i, R_{i,k}))) + \sum_{i\in S} \ln^2(n\lambda(B(X_i, R'_{i,k})))\right] \\
\lesssim_{d,k} \quad & \frac{1}{n}\left(\mathbb{E}\left[\ln^2(n\lambda(B(X_1, R_{1,k})))\right] + \mathbb{E}\left[\ln^2(n\lambda(B(X_1, R'_{1,k})))\right]\right). \qquad (2.52)
\end{aligned}$$

Since $\mathbb{E}\left[\ln^2(n\lambda(B(X_1, R_{1,k})))\right] \lesssim_{d,k} 1$ and $\mathbb{E}\left[\ln^2(n\lambda(B(X_1, R'_{1,k})))\right] \lesssim_{d,k} 1$. Using Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\ln^2(n\lambda(B(X_1, R_{1,k})))\right] \\
\leq \quad & 2\left(\mathbb{E}\left[\ln^2(\frac{\lambda(B(X_1, R_{1,k}))}{\mu(B(X_1, R_{1,k}))})\right] + \mathbb{E}\left[\ln^2(n\mu(B(X_1, R_{1,k})))\right]\right), \qquad (2.53)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\ln^2(n\lambda(B(X_1, R'_{1,k})))\right] \\
\leq \quad & 3\Big(\mathbb{E}\left[\ln^2(\frac{\lambda(B(X_1, R'_{1,k}))}{\mu(B(X_1, R'_{1,k}))})\right] + \mathbb{E}\left[\ln^2((n-1)\mu(B(X_1, R'_{1,k})))\right] \\
& + \ln^2(\frac{n}{n-1})\Big). \qquad (2.54)
\end{aligned}$$

Since $\mu(B(X_1, R_{1,k})) \sim \mathsf{Beta}(k, n+1-k)$ and $\mu(B(X_1, R'_{1,k})) \sim \mathsf{Beta}(k, n-k)$, therefore we know that both the quantities $\mathbb{E}\left[\ln^2(n\mu(B(X_1, R_{1,k})))\right]$ and $\mathbb{E}\left[\ln^2((n-1)\mu(B(X_1, R'_{1,k})))\right]$ equal to certain constants that only depends on k. $\ln^2(n/(n-1))$ is smaller than $\ln^2 2$ for $n \geq 2$. So we only need to prove that $\mathbb{E}\left[\ln^2(\frac{\lambda(B(X_1, R_{1,k}))}{\mu(B(X_1, R_{1,k}))})\right] \lesssim_{d,k} 1$ and $\mathbb{E}\left[\ln^2(\frac{\lambda(B(X_1, R'_{1,k}))}{\mu(B(X_1, R'_{1,k}))})\right] \lesssim_{d,k} 1$. Recall that we have defined the maximal function as follows,

$$M(x) = \sup_{0\leq r\leq 1/2} \frac{\lambda(B(x,r))}{\mu(B(x,r))}. \qquad (2.55)$$

Similarly, we define

$$m(x) = \sup_{0 \le r \le 1/2} \frac{\mu(B(x,r))}{\lambda(B(x,r))}. \tag{2.56}$$

Therefore,

$$\begin{aligned} & \mathbb{E}\left[\ln^2(\frac{\lambda(B(X_1,R_{1,k}))}{\mu(B(X_1,R_{1,k}))})\right] \\ \le\ & \mathbb{E}\left[\max\{\ln^2(M(x)), \ln^2(m(x))\}\right] \\ \le\ & \mathbb{E}\left[\ln^2(M(x)+1) + \ln^2(m(x)+1)\right] \\ =\ & \mathbb{E}\left[\ln^2(M(x)+1)\right] + \mathbb{E}\left[\ln^2(m(x)+1)\right]. \end{aligned} \tag{2.57}$$

Similarly this inequality holds if we replace $R_{1,k}$ by $R'_{1,k}$. By Lemma 3, we have

$$\begin{aligned} & \mathbb{E}\left[\ln^2(M(x)+1)\right] = \int_{[0,1]^d} \ln^2(M(x)+1) d\mu(x) \\ =\ & \int_{t=0}^{\infty} \mu\left(\left\{x \in [0,1]^d : \ln^2(M(x)+1) > t\right\}\right) dt \\ =\ & \int_{t=0}^{\infty} \mu\left(\left\{x \in [0,1]^d : M(x) > e^{\sqrt{t}} - 1\right\}\right) dt \\ \lesssim_d\ & \int_{t=0}^{\infty} \frac{1}{e^{\sqrt{t}} - 1} dt \lesssim_d 1. \end{aligned} \tag{2.58}$$

For $\mathbb{E}[\ln^2(m(x)+1)]$, we rewrite the term as

$$\begin{aligned} & \mathbb{E}\left[\ln^2(m(x)+1)\right] = \int_{[0,1]^d} f(x) \ln^2(m(x)+1) d\lambda(x) \\ =\ & \int_{t=0}^{\infty} \lambda\left(\left\{x \in [0,1]^d : f(x)\ln^2(m(x)+1) > t\right\}\right) dt. \end{aligned} \tag{2.59}$$

For a sufficiently large T_0 and $t > T_0$, $f(x)\ln^2(m(x)+1) > t$ implies either $m(x) > t^2$ or $f(x) > t/\ln^2(t^2+1)$. For $t \le T_0$, simply we use $\lambda\left(\left\{x \in [0,1]^d : f(x)\ln^2(m(x)+1) > t\right\}\right) \le 1$. Moreover, if $f(x) > t/\ln^2(t^2+1)$ then

$$f(x)\ln^2 f(x) > \frac{t(\ln t - 2\ln\ln(t^2+1))^2}{\ln^2(t^2+1)} > \frac{t}{T_0^2}, \tag{2.60}$$

since $(\ln t - 2\ln\ln(t^2+1))^2/\ln^2(t^2+1) > 1/T_0^2$ for any $t > T_0$. So for $t > T_0$,

$$\begin{aligned} & \lambda\left(\left\{x \in [0,1]^d : f(x)\ln^2(m(x)+1) > t\right\}\right) \\ \leq\ & \lambda\left(\left\{x \in [0,1]^d : m(x) > t^2\right\}\right) + \lambda\left(\left\{x \in [0,1]^d : f(x)\ln^2 f(x) > t/T_0^2\right\}\right). \end{aligned} \tag{2.61}$$

Therefore,

$$\begin{aligned} & \int_{t=0}^{\infty} \lambda\left(\left\{x \in [0,1]^d : f(x)\ln^2(m(x)+1) > t\right\}\right) dt \\ \leq\ & \int_{t=0}^{T_0} 1\, dt + \int_{t=T_0}^{\infty} \lambda\left(\left\{x \in [0,1]^d : m(x) > t^2\right\}\right) dt \\ & + \int_{t=T_0}^{\infty} \lambda\left(\left\{x \in [0,1]^d : f(x)\ln^2 f(x) > t/T_0^2\right\}\right) dt \\ \lesssim_d\ & T_0 + \int_{t=T_0}^{\infty} \frac{1}{t^2} dt + T_0^2 \int_{[0,1]^d} f(x)\ln^2 f(x) dx \\ \lesssim\ & 1. \end{aligned} \tag{2.62}$$

Hence, the proof is completed.

2.4 Proof of lemmas in Chapter 2

2.4.1 Proof of Lemma 1

We consider the cases $s \in (0,1]$ and $s \in (1,2]$ separately. For $s \in (0,1]$, following the definition of Hölder smoothness, we have,

$$\begin{aligned} & |\, f_t(x) - f(x)\,| = \Big|\, \frac{1}{V_d t^d} \int_{u:||u-x||\leq t} f(u)du - f(x)\,\Big| \\ \leq\ & \frac{1}{V_d t^d} \int_{u:\|u-x\|\leq t} |f(u) - f(x)| du \\ \leq\ & \frac{1}{V_d t^d} \int_{u:\|u-x\|\leq t} L\|u-x\|^s du. \end{aligned} \tag{2.63}$$

By denoting $\rho = \|u-x\|$ and considering $\theta \in S^{d-1}$ on the unit d-dimensional sphere, we rewrite the above integral using polar coordinate system and

obtain,

$$
\begin{aligned}
&| f_t(x) - f(x) | \leq \frac{1}{V_d t^d} \int_{\rho=0}^{t} \int_{\theta \in S^{d-1}} L\rho^s \rho^{d-1} d\rho d\theta \\
= \quad & \frac{1}{V_d t^d} \int_{\rho=0}^{t} dV_d L \rho^{s+d-1} d\rho = \frac{dV_d L t^{s+d}}{(s+d)V_d t^d} = \frac{dLt^s}{s+d}. \qquad (2.64)
\end{aligned}
$$

Now we consider the case $s \in (1, 2]$. Now we rewrite the difference as

$$
\begin{aligned}
&| f_t(x) - f(x) | = \Big| \frac{1}{V_d t^d} \int_{u:\|u-x\| \leq t} f(u) du - f(x) \Big| \\
= \quad & \Big| \frac{1}{2V_d t^d} \int_{v:\|v\| \leq t} (f(x+v) + f(x-v)) \, dv - f(x) \Big| \\
\leq \quad & \frac{1}{2V_d t^d} \int_{v:\|v\| \leq t} \Big| f(x+v) + f(x-v) - 2f(x) \Big| dv. \qquad (2.65)
\end{aligned}
$$

For fixed v, we bound $|f(x+v)+f(x-v)-2f(x)|$ using the gradient theorem and the definition of Hölder smoothness as follows,

$$
\begin{aligned}
& |f(x+v) + f(x-v) - 2f(x)| \\
= \quad & \Big| (f(x+v) - f(x)) + (f(x-v) - f(x)) \Big| \\
= \quad & \Big| \int_{\alpha=0}^{1} \nabla f(x+\alpha v) \cdot d(x+\alpha v) + \int_{\alpha=0}^{-1} \nabla f(x+\alpha v) \cdot d(x+\alpha v) \Big| \\
= \quad & \Big| \int_{\alpha=0}^{1} (\nabla f(x+\alpha v) \cdot v) \, d\alpha - \int_{\alpha=0}^{1} (\nabla f(x-\alpha v) \cdot v) \, d\alpha \Big| \\
= \quad & \Big| \int_{\alpha=0}^{1} (\nabla f(x+\alpha v) - \nabla f(x-\alpha v)) \cdot v d\alpha \Big| \\
\leq \quad & \int_{\alpha=0}^{1} \| \nabla f(x+\alpha v) - \nabla f(x-\alpha v) \| \|v\| d\alpha \\
\leq \quad & \int_{\alpha=0}^{1} L \|2\alpha v\|^{s-1} \|v\| d\alpha \\
= \quad & L\|v\|^s \int_0^1 (2\alpha)^{s-1} d\alpha = \frac{L\|v\|^s 2^{s-1}}{s}. \qquad (2.66)
\end{aligned}
$$

Plug it into (2.65) and using the similar method in the $s \in (0, 1]$ case, we have

$$\begin{aligned} & | f_t(x) - f(x) | \leq \frac{1}{2V_d t^d} \int_{v:\|v\| \leq t} \frac{L\|v\|^s 2^{s-1}}{s} dv \\ = & \frac{1}{2V_d t^d} \int_{\rho=0}^{t} \int_{\theta \in S^{d-1}} \frac{L\rho^s 2^{s-1}}{s} \rho^{d-1} d\rho d\theta = \frac{1}{2V_d t^d} \int_{\rho=0}^{t} \frac{dV_d L \rho^{s+d-1} 2^{s-1}}{s} d\rho \\ = & \frac{1}{2V_d t^d} \frac{dV_d L 2^{s-1}}{s} \frac{t^{s+d}}{s+d} \leq \frac{dLt^s}{s+d}, \end{aligned} \tag{2.67}$$

where the last inequality uses the fact that $s \in (1, 2]$.

2.4.2 Proof of Lemma 2

We consider the following two cases. If $f(x) \geq 2dLt^s/(s+d)$, then by Lemma 1, we have

$$f(x) \leq f_t(x) + \frac{dLt^s}{s+d} \leq f_t(x) + \frac{f(x)}{2}. \tag{2.68}$$

Hence, $f(x) \leq 2f_t(x)$ in this case. If $f(x) < 2dLt^s/(s+d)$, then define $t_0 = (f(x)(s+d)/2dL)^{1/s} < t$. By the non-negativity of f, we have

$$\begin{aligned} & f_t(x) V_d t^d = \int_{B(x,t)} f(x) dx \geq \int_{B(x,t_0)} f(x) dx \\ = & f_{t_0}(x) V_d t_0^d \geq \left(f(x) - \frac{dLt_0^s}{s+d} \right) V_d t_0^d \\ = & f(x) V_d \left(\frac{f(x)(s+d)}{2dL} \right)^{d/s} - \frac{dL}{s+d} V_d \left(\frac{f(x)(s+d)}{2dL} \right)^{(s+d)/s} \\ = & f(x)^{(s+d)/s} V_d \left(\frac{s+d}{dL} \right)^{d/s} \left(2^{-d/s} - 2^{-(s+d)/s} \right). \end{aligned} \tag{2.69}$$

Therefore, we have $f(x) \lesssim_{s,L,d} (f_t(x) V_d t^d)^{s/(s+d)}$ in this case. We obtain the desired statement by combining the two cases. Furthermore, by taking $t = 1/2$, we have $V_d t^d f_t(x) < 1$, so $f_t(x) \lesssim_{s,L,d} 1$. By applying this lemma immediately we obtain $f(x) \lesssim_{s,L,d} 1$.

2.4.3 Proof of Lemma 3

We first introduce the Besicovitch covering lemma, which plays a crucial role in the analysis of nearest neighbor methods.

Lemma 4. *[70, Theorem 1.27][Besicovitch covering lemma] Let $A \subset \mathbb{R}^d$, and suppose that $\{B_x\}_{x\in A}$ is a collection of balls such that $B_x = B(x, r_x), r_x > 0$. Assume that A is bounded or that $\sup_{x\in A} r_x < \infty$. Then there exist an at most countable collection of balls $\{B_j\}$ and a constant C_d depending only on the dimension d such that*

$$A \subset \bigcup_j B_j \ , \quad \text{and} \quad \sum_j \chi_{B_j}(x) \leq C_d. \tag{2.70}$$

Here $\chi_B(x) = \mathbb{I}(x \in B)$.

Now we are ready to prove the lemma. Let

$$M(x) = \sup_{0<\rho\leq D} \left(\frac{\mu_2(B(x,\rho))}{\mu_1(B(x,\rho))} \right). \tag{2.71}$$

Let $O_t = \{x \in A : M(x) > t\}$. Hence, for all $x \in O_t$, there exists $B_x = B(x, r_x)$ such that $\mu_2(B_x) > t\mu_1(B_x), 0 < r_x \leq D$. It follows from the Besicovitch lemma applying to the set O_t that there exists a set $E \subset O_t$, which has at most countable cardinality, such that

$$O_t \subset \bigcup_{j\in E} B_j \ , \quad \text{and} \quad \sum_{j\in E} \chi_{B_j}(x) \leq C_d. \tag{2.72}$$

Let $A_D = \{x : \exists y \in A, |y - x| \leq D\}$, therefore $B_j \subset A_D$ for every j. Then,

$$\begin{aligned} \mu_1(O_t) &\leq \sum_{j\in E} \mu_1(B_j) < \frac{1}{t} \sum_{j\in E} \mu_2(B_j) \\ = \ & \frac{1}{t} \sum_{j\in E} \int_{A_D} \chi_{B_j} d\mu_2 = \frac{1}{t} \int_{A_D} \sum_{j\in E} \chi_{B_j} d\mu_2 \leq \frac{C_d}{t} \mu_2(A_D). \end{aligned} \tag{2.73}$$

CHAPTER 3

ANALYSIS OF KSG MUTUAL INFORMATION ESTIMATORS

Information-theoretic quantities such as mutual information measure *relations* between random variables. A key property of these measures is that they are invariant to one-to-one transformations of the random variables and obey the data processing inequality [71, 72]. These properties combine to make information-theoretic quantities attractive in several data science applications involving clustering [22, 23, 24], classification [21] and more generally as a basic feature that can be used in several downstream applications [27, 26, 73, 74]. A canonical question in all these applications is to estimate the information-theoretic quantities from *samples*, typically supposed to be drawn i.i.d. from an unknown distribution. This fundamental question has been of longstanding interest in the theoretical statistics community where it is a canonical question of estimating a functional of the (unknown) density [41] but also in the information theory [75, 76, 77, 78], machine learning [79, 80] and theoretical computer science [81, 82, 83] communities, with significant renewed interest of late, summarized in detail in Section 3.4. The most fundamental information-theoretic quantity of interest is the mutual information between a pair of random variables, which is also the primary focus of this chapter, in the context of real valued random variables (in potentially high dimensions).

The basic estimation question takes a different hue depending on whether the underlying distribution is discrete or continuous. In the discrete setting, significant understanding of the minimax rate-optimal estimation of functionals, including entropy and mutual information, of an unknown probability mass function is attained via recent works [76, 84, 81, 85, 77]. The continuous setting is significantly different, bringing to fore the interplay of geometry of the Euclidean space as well as the role of dimensionality of the domain in terms of estimating the information-theoretic quantities; this setting is the focus of this chapter. Among the various estimation methods, of

great theoretical interest and high practical relevance, are the *nearest neighbor* (NN) methods: the quantities of interest are estimated based on distances (in an appropriate norm) of the samples to their k-nearest neighbors (k-NN). Of particular practical interest is the situation when k is a small *fixed* integer – typically in the range of 4∼8 – and the estimators based on fixed k-NN statistics typically perform significantly better than alternative approaches, discussed in detail in Section 3.4, both in simulations and when tested in the wild; this is especially true when the random variables are in high dimensions.

The exemplar fixed k-NN estimator is that of differential entropy from i.i.d. samples proposed in 1987 by Kozachenko and Leonenko [2] which involved a novel bias correction term, and we refer to as the KL estimator (of differential entropy). Since the mutual information between two random variables is the sum and difference of three differential entropy terms, any estimator of differential entropy naturally lends itself into an estimator of mutual information, which we christen as the 3KL estimator (of mutual information). In an inspired work in 2004, Kraskov and Stögbauer and Grassberger [4], proposed a different fixed k-NN estimator of the mutual information, which we name the KSG estimator, that involved subtle (sample dependent) alterations to the 3KL estimator. The authors of [4, 86] empirically demonstrated that the KSG estimator consistently improves over the 3KL estimator in a variety of settings. Indeed, the simplicity of the KSG estimator, combined with its superior performance, has made it a very popular estimator of mutual information in practice.

Despite its widespread use, even basic theoretical properties of the KSG estimator are unknown – it is not even clear if the estimator has vanishing bias (i.e., consistent) as the number of samples grows, much less any understanding of the asymptotic behavior of the bias as a function of the number of samples. As observed elsewhere [87], characterizing the theoretical properties of the KSG estimator is of first order importance – this study could shed light on why the sample-dependent modifications lead to improved performance and perhaps this understanding could lead to the design of even better mutual information estimators. Such are the goals of this chapter.

Main contribution of Chapter 3:

- Our main result is to show that the KSG estimator is consistent. We also show upper bounds to the rate of convergence of the bias as a

function of the dimensions of the two random variables involved: in the special case when the dimensions of the two random variables are equal and no more than one, the rate of convergence of the ℓ_2 error is $1/\sqrt{N}$, which is the parametric rate of convergence.

- We argue that the improvement of the KSG estimator over the 3KL estimator comes from a "correlation boosting" effect, which can be further amplified by a suitable modification to the KSG estimator. This leads to a novel mutual information estimator, which we call the bias-improved-KSG estimator (BI-KSG). The asymptotic theoretical guarantees we show of the BI-KSG estimator are the same as the KSG estimator, but the improved performance can be seen empirically – especially for moderate values of N.

- We extend the idea of "correlation boosting" to multivariate mutual information and general functional of entropies, propose an estimator of MMI, and demonstrate its empirical performance.

- We demonstrate sharp bounds on the ℓ_2 rate of convergence of the KL estimator of (differential) entropy for arbitrary k and arbitrary dimensions d, showing that the parametric rate of convergence of $1/\sqrt{N}$ is achievable when $d \leq 2$.

In the rest of the chapter, we mathematically summarize these main results, following up with empirical evidence.

Outline of this chapter. In Section 3.1, we show the consistency and the convergence rate of KSG estimator of mutual information, also providing brief sketches of, and intuitions behind, the corresponding proofs. In Section 3.2 we discuss the insights behind the KSG estimator: the correlation boosting effect and how this understanding leads to the BI-KSG estimator with improved empirical performance. In Section 3.3 we discuss generalization of the KSG estimator to multivariate mutual information estimators. Section 3.4 puts our results in context of the vast literature on entropy (and mutual information) estimators. Finally, the proofs of the main results are in Sections 3.5 through 3.7.

3.1 KSG estimator: Consistency and convergence rate

A detailed understanding of the KL estimator sets the stage for the main results of this chapter: deriving theoretical properties of the KSG estimator of mutual information. Our main result is that the KSG estimator is consistent, as is our proposed modification, the so-called bias-improved KSG estimator (BI-KSG); these results are under some (fairly standard) assumptions on the joint pdf of (X, Y).

Consider two random variables X in $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and Y in $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$. Given N i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^N$ from the underlying joint probability density function $f_{X,Y}(x, y)$, we want to estimate the mutual information $I(X;Y)$. Mutual information between two random variables X and Y is the sum and difference of differential entropy terms: $I(X;Y) = H(X) + H(Y) - H(X,Y)$. Thus given KL entropy estimator, there is a straightforward and consistent estimation of the mutual information:

$$\widehat{I}_{\mathrm{3KL}}(X;Y) \;=\; \widehat{H}_{\mathrm{KL}}(X) + \widehat{H}_{\mathrm{KL}}(Y) - \widehat{H}_{\mathrm{KL}}(X,Y). \tag{3.1}$$

While this estimator performs fairly well in practice, the authors of [4] introduced a simple, but inspired, modification of the 3KL estimator that does even better. Let $n_{x,i,p} \equiv \sum_{j \neq i} \mathbb{I}\{\|X_j - X_i\|_p \leq \rho_{k,i,p}\}$, which can be interpreted as the number of samples that are within a X-dimensions-only distance of $\rho_{k,i,p}$ with respect to sample i. Since $\rho_{k,i,p}$ is the k-NN distance (in terms of both the dimensions of X and Y) of the sample i it must be that $n_{x,i,p} \geq k$. Finally, $n_{y,i,p}$ is defined analogously. The KSG estimator measures distances using the ℓ_∞ norm, so $p = \infty$ in the notation above.

The KSG mutual information estimator introduced in [4] is given by:

$$\widehat{I}_{\mathrm{KSG}}(X;Y) \;\equiv\; \psi(k) + \log N - \frac{1}{N} \sum_{i=1}^{N} \left(\psi(n_{x,i,\infty} + 1) + \psi(n_{y,i,\infty} + 1) \right), \tag{3.2}$$

where $\psi(x) = \Gamma^{-1}(x) d\Gamma(x)/dx$ is the digamma function. Observe that the estimate of the joint differential entropy $H(X,Y)$ is done exactly as in the KL estimator using fixed k-NN distances, but the KL estimates of $H(X)$ and $H(Y)$ are done using $n_{x,\cdot,\infty}$ and $n_{y,\cdot,\infty}$ NN distances, respectively, which are *sample dependent*. The point is that by this choice, the k-NN distance

terms are canceled away exactly, although it is not clear why this would be a good idea. In fact, it is not even clear if the estimator is consistent. On the other hand, the authors of [4] showed empirically that the KSG estimator is uniformly superior to the 3KL estimator in many synthetic experiments. A theoretical understanding of the KSG estimator, including a mathematical justification for the improved performance, has been missing in the literature. Our main results fill this gap.

One of our main results is to show that the KSG estimator is indeed consistent. We prove this result by deriving a vanishingly small upper bound on the bias, subject to regularity conditions on the Radon-Nikodym derivatives of X and Y and standard smoothness conditions on the joint pdf which includes both bounded and unbounded supports.

3.1.1 Consistency

We make the following assumptions on the joint pdf of (X, Y). The first assumption is essentially needed to define the joint differential entropy of (X, Y), the second assumption makes some regularity conditions on the Radon-Nikodym derivatives of X and Y, and the third assumption is regarding standard smoothness conditions on the joint pdf. We note that these conditions are readily met by most popular pdfs, including multivariate Gaussians, and no assumption is made on the boundedness of the support.

Assumption 1. (*a*) $\int f(x,y) \left|\log f(x,y)\right| dxdy < \infty$.

(*b*) *There exists a finite constant* C' *such that the conditional pdf* $f_{Y|X}(y|x) < C'$ *and* $f_{X|Y}(x|y) < C'$ *almost everywhere.*

(*c*) $f(x,y)$ *is twice continuously differentiable and the Hessian matrix* H_f *satisfy* $\|H_f(x,y)\|_2 < C$ *almost everywhere.*

The following theorem states that under these assumptions, the KSG estimator is consistent in probability.

Theorem 3. *Under the Assumption 1 and for finite* $k > \max\{d_x/d_y, d_y/d_x\}$, $d_x, d_y = O(1)$, *and for all* $\varepsilon > 0$,

$$\lim_{N\to\infty} \Pr\left(\left| \widehat{I}_{\mathrm{KSG}}(X;Y) - I(X;Y) \right| > \varepsilon \right) = 0. \tag{3.3}$$

Also, consider the following BI(biased-improved)-KSG estimator

$$\begin{aligned} & \widehat{I}_{\mathrm{BI-KSG}}(X;Y) \\ \equiv \;& \psi(k) + \log N + \log\left(\frac{c_{d_x,2}c_{d_y,2}}{c_{d_x+d_y,2}}\right) - \frac{1}{N}\sum_{i=1}^{N}\left(\log(n_{x,i,2}) + \log(n_{y,i,2})\right), \end{aligned} \tag{3.4}$$

which will be further discussed in Section 3.2. Under the same assumption 1 and for finite $k > \max\{d_x/d_y, d_y/d_x\}$, $d_x, d_y = O(1)$, *and for all* $\varepsilon > 0$,

$$\lim_{N\to\infty} \Pr\left(\left|\widehat{I}_{\mathrm{BI-KSG}}(X;Y) - I(X;Y)\right| > \varepsilon\right) = 0. \tag{3.5}$$

3.1.2 Convergence rate

To understand the rate of convergence of the bias of the KSG and BI-KSG estimators, we first truncate the k-NN distance $\rho_{k,\cdot,\cdot}$ by a certain threshold. For any $\delta > 0$, let the truncation threshold be:

$$a_N = \left(\frac{(\log N)^{1+\delta}}{N}\right)^{1/(d_x+d_y)}, \tag{3.6}$$

where d_x and d_y are the dimensions of the random variables X and Y respectively. We define local information estimates $\iota_{k,i,\infty}$ by:

$$\iota_{k,i,\infty} \;=\; \psi(k) + \log N - \psi(n_{x,i,\infty}+1) - \psi(n_{y,i,\infty}+1), \tag{3.7}$$

if $\rho_{k,i,\infty} \le a_N$ and $\iota_{k,i,\infty} = 0$ if $\rho_{k,i,\infty} > a_N$. Similarly, we define $\iota_{k,i,2}$ as, and

$$\iota_{k,i,2} \;=\; \psi(k) + \log N + \log(\frac{c_{d_x,2}c_{d_y,2}}{c_{d_x+d_y,2}}) - \log(n_{x,i,2}) - \log(n_{y,i,2}), \tag{3.8}$$

if $\rho_{k,i,2} \leq a_N$ and $\iota_{k,i,2} = 0$ if $\rho_{k,i,2} > a_N$. The modified (via truncation) KSG and BI-KSG estimators (compare with (3.2) and (3.4)) are:

$$\widehat{I}_{tKSG}(X;Y) \equiv \frac{1}{N}\sum_{i=1}^{N} \iota_{k,i,\infty}, \tag{3.9}$$

$$\widehat{I}_{tBI-KSG}(X;Y) \equiv \frac{1}{N}\sum_{i=1}^{N} \iota_{k,i,2}. \tag{3.10}$$

The following theorem provides an upper bound on the rate of convergence of the bias and variance, under the conditions in Assumption 2 below, and holds for any k and $\delta > 0$ (parameter in the truncation threshold, cf. (3.6)).

Assumption 2. *We make the following assumptions: there exist finite constants C_a,C_b,C_c,C_d,C_e,C_f,C_g,C_h and C_0 such that*

(a) $f(x,y) \leq C_a < \infty$ *almost everywhere.*

(b) *There exists $\gamma > 0$ such that $\int f(x,y)\,(\log f(x,y)\,)^{1+\gamma}\,dxdy \leq C_b < \infty$.*

(c) $\int f(x,y)\exp\{-bf(x,y)\}dxdy \leq C_c e^{-C_0 b}$ *for all $b > 1$.*

(d) $f(x,y)$ *is twice continuously differentiable and the Hessian matrix H_f satisfy $\|H_f(x,y)\|_2 < C_d$ almost everywhere.*

(e) *The conditional pdf $f_{Y|X}(y|x) < C_e$ and $f_{X|Y}(x|y) < C_e$ almost everywhere.*

(f) *The marginal pdf $f_X(x) < C_f$ and $f_Y(y) < C_f$ almost everywhere.*

(g) *The set of points violating (d) has finite $d_x + d_y - 1$-dimensional Hausdorff measure, i.e.,*

$$H^{d_x+d_y-1}\left(\{(x,y) : \|H_f(x,y)\| \geq C_d\}\right) \leq C_g.$$

(h) *The set of points such that $H_{f_X}(x)$ or $H_{f_Y}(y)$ is larger than C_d also has finite $d_x - 1$ (or $d_y - 1$)-dimensional Hausdorff measure, i.e.,*

$$H^{d_x-1}\left(\{x : \|H_{f_X}(x)\| \geq C_d\}\right) \leq C_h, \tag{3.11}$$

$$H^{d_y-1}\left(\{y : \|H_{f_Y}(y)\| \geq C_d\}\right) \leq C_h. \tag{3.12}$$

The assumptions (a)-(d) come from the assumptions of the KL estimator in [32] and are slightly stronger than those in [32], where assumption (a) is not required (and with some technical finesse might be eliminated here as well), assumption (b) was weaker requiring only $\int f(x,y)|\log f(x,y)|dxdy < \infty$, and assumption (c) was weaker requiring only $\int f(x,y)\exp\{-bf(x,y)\}dxdy \leq O(1/b)$. The assumption (c) is satisfied for any distribution with bounded support and pdf bounded away from zero. This assumption provides a sufficient condition to bound the average effect of the truncation. Our analysis can be generalized to relax this assumption on the smoothness, requiring only $\int f(x,y)\exp\{-bf(x,y)\}dxdy \leq C_c b^{-\beta}$ for all $b > 1$, in which case the resulting guarantees will also depend on β. This recovers the result of [32] with $\beta = 1$ which holds for $d = 1$, and we assume stronger conditions here since we seek sharp convergence rates in higher dimensions. The assumption (d) assumes that the pdf is reasonably smooth, and it is essential for NN-based methods. More general families of smoothness conditions have been assumed for other approaches, such as the Hölder condition,and we have made formal comparisons in Section 3.4.

Assumption 2.(e) makes sure that the marginal entropy estimator converges at certain rate. Compared to Assumption 1, we need an upper bound for the joint entropy (a). The condition (b) is slightly stronger than Assumption 1 by changing the power from 1 to $1+\gamma$. The condition (c) is the tail bound which ensures the convergence rate of truncated KL joint entropy estimator.

Note that there exist (families of) distributions, satisfying the assumptions (a)–(d), where the convergence rates of k-NN estimators can be made arbitrarily slow. Consider a family of distributions in two-dimensional rectangle with uniform measure parametrized by ℓ, such that one side has a length ℓ and the other $1/\ell$. This family of distributions has differential entropy zero. However, for any sample size N, there exists ℓ large enough such that the k-NN distances are arbitrarily large and the estimated entropy is also large. To provide a sharp convergence rate for k-NN estimators, we need to restrict the space of distributions by adding appropriate assumptions that captures this phenomenon.

The challenge in the above example has been addressed under the notion of *boundary bias*. The k-NN distances are larger near the boundaries, which results in underestimating the density at boundaries. This effect is

prominent for those distributions that (i) have non-smooth boundaries such as a uniform distribution on a compact support, and (ii) have large surface area at the boundary. There are two solutions; either we strengthen Assumption 2.(d) and require twice continuously differentiability *everywhere* including the boundaries or we can add another assumption on the surface area of the boundaries. In this chapter, we take the second route. The reason is that the first option conflicts with the current Assumption 2.(c) where the only examples we know have lower bounded densities, which implies non-smooth boundaries. It is an interesting future research direction to relax Assumption 2.(c) as suggested above, and capture the trade-off between the lightness of the tail in β and also the smoothness in the boundaries.

Instead, we assume in 2.(h) that the surface area of the boundaries is finite. Recall that the Hausdorff measure of a set S is defined as [88]

$$H^{d-1}(S) \;=\; \lim_{\delta \to 0} \inf_{\{U_i\}_{i=1}^{\infty}} \Big\{ \sum_{i=1}^{\infty} (\operatorname{diam} U_i)^{d-1} \;:\; \bigcup_{i=1}^{\infty} U_i \supseteq S \,,\, \operatorname{diam} U_i < \delta \Big\}. \tag{3.13}$$

Here the *diameter* of the set U is defined as

$$\operatorname{diam} U = \sup\{\|x - y\| | x, y \in U\}. \tag{3.14}$$

The Hausdorff measure of a set is a measure of its *surface area*. Note that this could be unbounded for the boundary of a family of distributions, as is the case for the uniform rectangle example above. Assumption 1.(h) restricts it to be finite, allowing us to limit the boundary bias to $\tilde{O}(N^{-1/d})$. Since in the (smooth) interior of the support, the bias is $\tilde{O}(N^{-2/d})$, the boundary bias dominates the error for the proposed k-NN estimator. We note that truncated multivariate Gaussians and uniform random variables meet these constraints.

Theorem 4. *Under Assumption 2, and for finite* $k > \max\{d_x/d_y, d_y/d_x\}$,

$d_x, d_y = O(1)$,

$$\begin{aligned} & \mathbb{E}\left[\widehat{I}_{\mathrm{tKSG}}(X;Y)\right] - I(X;Y) \\ = \; & O\left(\frac{(\log N)^{(1+\delta)(1+\frac{1}{d_x+d_y})}}{N^{\frac{1}{d_x+d_y}}}\right). \end{aligned} \tag{3.15}$$

$$\begin{aligned} & \mathbb{E}\left[\widehat{I}_{\mathrm{tBI-KSG}}(X;Y)\right] - I(X;Y) \\ = \; & O\left(\frac{(\log N)^{(1+\delta)(1+\frac{1}{d_x+d_y})}}{N^{\frac{1}{d_x+d_y}}}\right). \end{aligned} \tag{3.16}$$

The following theorem establishes an upper bound for the variance of truncated KSG and BI-KSG estimators.

Theorem 5. *Under Assumption 2, and for finite $k \geq 2$,*

$$\mathrm{Var}\left[\widehat{\mathrm{I}}_{\mathrm{tKSG}}(\mathrm{X};\mathrm{Y})\right] = \mathrm{O}\left(\frac{(\log \mathrm{N})^2}{\mathrm{N}}\right). \tag{3.17}$$

$$\mathrm{Var}\left[\widehat{\mathrm{I}}_{\mathrm{tBI-KSG}}(\mathrm{X};\mathrm{Y})\right] = \mathrm{O}\left(\frac{(\log \mathrm{N})^2}{\mathrm{N}}\right). \tag{3.18}$$

Combining Theorem 4 and Theorem 5, we obtain the following upper bound on the MSE of truncated KSG or BI-KSG estimator.

Corollary 1. *Under the Assumption 2 and for finite $k = O(1)$ and $d = O(1)$, the MSE of the truncated KSG or BI-KSG mutual information estimator using N i.i.d. samples is bounded by:*

$$\begin{aligned} & \mathbb{E}\left[\left(\widehat{I}_{\mathrm{tKSG}}(X;Y) - I(X;Y)\right)^2\right] \\ = \; & O\left(\frac{(\log N)^{2(1+\delta)(1+\frac{1}{d_x+d_y})}}{N^{\frac{2}{d_x+d_y}}} + \frac{(\log N)^2}{N}\right). \end{aligned} \tag{3.19}$$

$$\begin{aligned} & \mathbb{E}\left[\left(\widehat{I}_{\mathrm{tBI-KSG}}(X;Y) - I(X;Y)\right)^2\right] \\ = \; & O\left(\frac{(\log N)^{2(1+\delta)(1+\frac{1}{d_x+d_y})}}{N^{\frac{2}{d_x+d_y}}} + \frac{(\log N)^2}{N}\right). \end{aligned} \tag{3.20}$$

Corollary 2. *If $d_x = d_y = 1$, we obtain:*

$$\begin{aligned} & \mathbb{E}\left[\left(\widehat{I}_{\mathrm{tKSG}}(X;Y) - I(X;Y)\right)^2\right] \\ = \; & O\left(\frac{(\log N)^{(2k+2)(1+\delta)}}{N}\right). \qquad (3.21) \\ & \mathbb{E}\left[\left(\widehat{I}_{\mathrm{tBI-KSG}}(X;Y) - I(X;Y)\right)^2\right] \\ = \; & O\left(\frac{(\log N)^{(2k+2)(1+\delta)}}{N}\right). \qquad (3.22) \end{aligned}$$

This establishes the $1/N$ convergence rate of the MSE of the KSG and BI-KSG and 3KL estimators up to a poly-logarithmic factor; this (parametric) convergence rate cannot be improved upon.

We compare the upper bound exponent from theory and experiment to see whether the upper bound should be improved or not. For each $N \in \{100, 200, 400, 800, 1500, 3000\}$ and $d \in \{1, 2, \ldots, 8\}$, we choose N i.i.d. samples $\{X_i\}_{i=1}^N$ from $\mathrm{Unif}[0,1]^d$ and let $Y_i = X_i + \mathrm{Unif}[0,1]^d$, and compute $\widehat{I}_{\mathrm{KSG}}(X;Y)$ averaged over 500 trails. We use standard linear regression to compute $\log(\mathrm{MSE})/\log N$, which is the experimental exponent. We compare the exponent with the theoretical upper bound 3.22 and lower bound from [41] (also with the exponents for other estimators: resubstitution [54] and von Mises expansion estimators [31]). From Figure 3.1 we conclude that the exponent from simulation is quite closed to the upper bound. We conjecture that the lower bound can be further improved to close the gap.

3.2 Correlation boosting

The goal of this section is to build some intuition toward a deeper theoretical understanding of the KSG estimator, where we see a curious *correlation boosting* effect which explains the superior performance of the KSG estimator and allows us to derive an even better estimator of mutual information. A related intuitive explanation is provided in [89].

Correlation boosting effect. We begin by rewriting the KSG estimator,

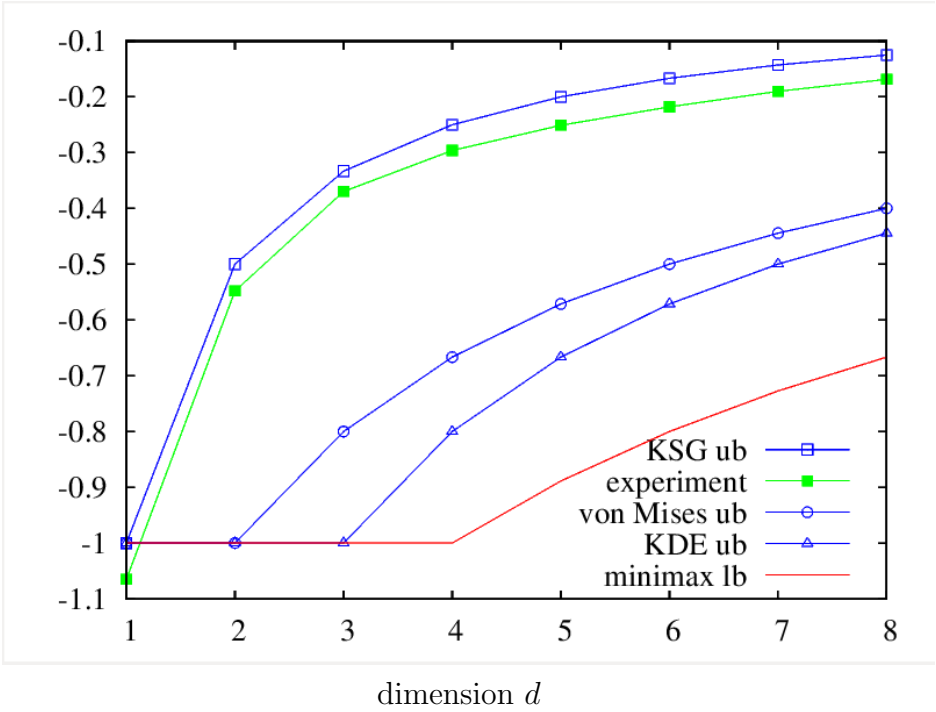

Figure 3.1: MSE for mutual information versus sample size in log-log scale.

cf. (3.2), as:

$$\begin{aligned}\widehat{I}_{KSG}(X;Y) &= \frac{1}{N}\sum_{i=1}^{N}\iota_{k,i,\infty} \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\,\xi_{k,i,\infty}(X)+\xi_{k,i,\infty}(Y)-\xi_{k,i,\infty}(X,Y)\,\right), \end{aligned} \tag{3.23}$$

where

$$\begin{aligned}\xi_{k,i,\infty}(X,Y) &\equiv -\psi(k)+\log N+\log c_{d_x,\infty}c_{d_y,\infty}+(d_x+d_y)\log\rho_{k,i,\infty}, \\ \xi_{k,i,\infty}(X) &\equiv -\psi(n_{x,i,\infty}+1)+\log N+\log c_{d_x,\infty}+d_x\log\rho_{k,i,\infty}, \\ \xi_{k,i,\infty}(Y) &\equiv -\psi(n_{y,i,\infty}+1)+\log N+\log c_{d_y,\infty}+d_y\log\rho_{k,i,\infty}. \end{aligned} \tag{3.24}$$

Here $\xi_{k,i,\infty}(X,Y), \xi_{k,i,\infty}(X)$ and $\xi_{k,i,\infty}(Y)$ are local estimates of the differential entropies $H(X,Y), H(X)$ and $H(Y)$, respectively, at the i^{th} sample. We will show that the local bias of joint entropy estimate $b_{k,i,\infty}(X,Y) = \xi_{k,i,\infty}(X,Y) - H(X,Y)$ is positively correlated to the local bias of marginal entropy estimates $b_{k,i,\infty}(X) = \xi_{k,i,\infty}(X) - H(X)$ and $b_{k,i,\infty}(Y) = \xi_{k,i,\infty}(Y) - H(Y)$. Formally, we can see this effect in the context of an example.

Theorem 6. *X and Y be independently and uniformly distributed in $[0,1]$. Then:*

$$\begin{aligned}&\mathbb{E}\left[b_{k,i,\infty}(X)|b_{k,i,\infty}(X,Y)\right] \\ =\ &\sqrt{\frac{1}{N}}\big(a_k^{(1)}+b_{k,i,\infty}(X,Y)a_k^{(2)}+O(b_{k,i,\infty}(X,Y)^2)\big)+O(\frac{1}{N}), \end{aligned} \tag{3.25}$$

where $a_k^{(1)}$ and $a_k^{(2)} > 0$ are constants that only depend on k.

We observe that the local biases are positively correlated, although the correlation decreases with the sample size. We simulate 300 i.i.d. samples uniformly from $[0,1]^2$, draw the scatter plot of local biases $b_{k,i,\infty}(X)$ versus $b_{k,i,\infty}(X,Y)$ in Figure. 3.2 and conclude that the scatter plot matches prediction from Theorem 6 reasonably well.

Since the global bias of the KSG estimator is simply equal to

$$\frac{1}{N}\sum_{i=1}^{N} b_{k,i,\infty}(X,Y) - b_{k,i,\infty}(X) - b_{k,i,\infty}(Y), \tag{3.26}$$

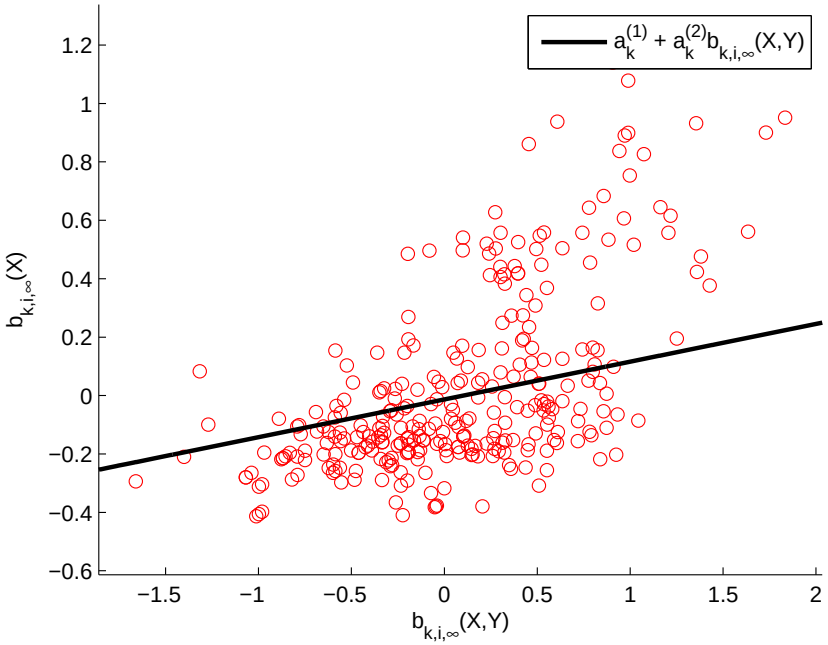

Figure 3.2: Scatter plot of the local biases $b_{k,i,\infty}(X,Y)$ and $b_{k,i,\infty}(X)$. The Pearson correlation is 0.4745.

the global bias is reduced if the joint bias $b(X,Y) = \frac{1}{N}\sum_{i=1}^{N} b_{k,i,\infty}(X,Y)$ is *positively* correlated with marginal bias $b(X) = \frac{1}{N}\sum_{i=1}^{N} b_{k,i,\infty}(X)$ and $b(Y) = \frac{1}{N}\sum_{i=1}^{N} b_{k,i,\infty}(Y)$. The same effect is true for the 3KL estimator, which is already based on estimating the three differential entropy terms separately. We tabulate the Pearson correlation coefficients of the global biases in Table 3.1 for two exemplar pdfs (independent uniforms and Gaussians). The main empirical observation is that the correlation is positive even for the 3KL estimator but is significantly higher for the KSG estimator (and at times even higher for the BI-KSG estimator which we introduce below).

Table 3.1: Pearson correlation coefficient $\rho\,(\,b(X,Y), b(X))$ for different mutual information estimators.

	$(X,Y) \sim \text{Unif}([0,1]^2)$			$(X,Y) \sim \mathcal{N}(0, I_2)$		
N	1024	2048	4096	1024	2048	4096
3KL	0.1276	0.1259	0.0930	0.4602	0.4471	0.3717
KSG	**0.9312**	**0.9328**	**0.9085**	0.6750	0.7151	0.6687
BI-KSG	0.9253	0.9251	0.8880	**0.6823**	**0.7330**	**0.6939**

We hypothesize that this correlation *boosting* effect is the main reason for

the KSG estimator having smaller mean-square error than the 3KL one. To get a feel for the effect for finite sample sizes, we simulate 100 i.i.d. samples uniformly from $[0,1]^2$ and map the scatter-plot of the biases $b(X,Y)$ and $b(X)$ in Figure 3.3, where the boosted correlation for the KSG estimator is visibly significant.

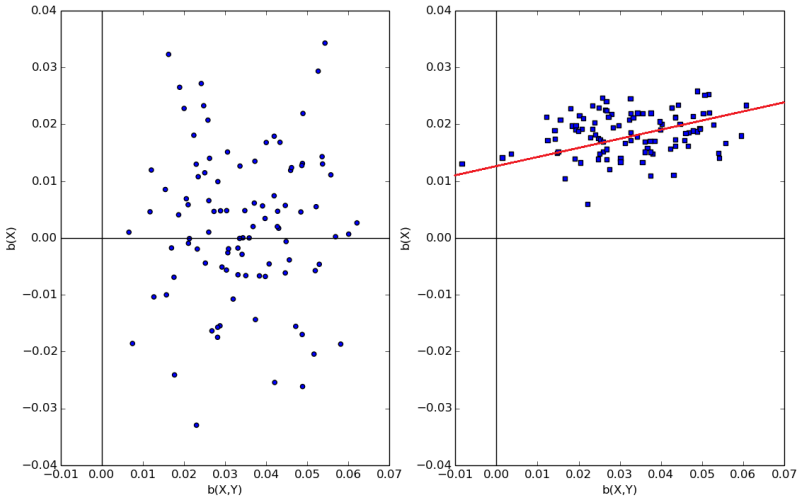

Figure 3.3: Scatter plot of the biases $b(X,Y)$ and $b(X)$ to illustrate the correlation boosting effect. Left: 3KL. Right: KSG. The solid red line is a regression line.

New estimator of mutual information. Given the understanding of the correlation boosting effect, it is natural to ask if this can lead to a new estimator that furthers the improvement in MSE. This goal is achieved below, where we discuss potential areas of improvement of the KSG estimator and conclude with our proposal: Bias Improved KSG (BI-KSG) estimator of mutual information. One of the key differences comes from using ℓ_2 norm to measure k-NN distances, while KSG uses ℓ_∞ distance. Next, BI-KSG uses $\log(n_{x,i,2})$ and $\log(n_{y,i,2})$ instead of $\psi(n_{x,i,\infty}+1)$ and $\psi(n_{y,i,\infty}+1)$, respectively. We briefly discuss the intuitions behind these changes below. We begin by noting that the KSG estimator can be written as:

$$\widehat{I}_{\mathrm{KSG}}(X;Y) \;=\; \widehat{H}_{\mathrm{KSG}}(X) + \widehat{H}_{\mathrm{KSG}}(Y) - \widehat{H}_{\mathrm{KL}}(X,Y), \tag{3.27}$$

where $\widehat{H}_{\mathrm{KL}}(X;Y)$ is the KL entropy estimator (and already known to be

consistent). The marginal entropy estimator is

$$\begin{aligned}\widehat{H}_{\mathrm{KSG}}(X) &= \frac{1}{N}\sum_{i=1}^{N}\big(-\psi(n_{x,i,\infty}+1) \\ &\quad +\psi(N)+\log c_{d_x,\infty}+d_x\log\rho_{k,i,\infty}\big), \end{aligned} \tag{3.28}$$

and we note that this has a form similar to that of the KL entropy estimator, except that k is replaced by $n_{x,i,\infty}+1$, which is sample dependent. Suppose $(X_i^{(k)}, Y_i^{(k)})$ be the k-NN of (X_i, Y_i) with distance $\rho_{k,i,\infty}$, then the "KSG entropy estimator" in (3.28) implicitly assumes that $\rho_{k,i,\infty}$ is *both* the $(n_{x,i,\infty}+1)$-NN distance of X_i on X-space *and* the $(n_{y,i,\infty}+1)$-NN of Y_i on Y-space. But since ℓ_∞-distance is used, $(X_i^{(k)}, Y_i^{(k)})$ either lies on the X-boundary of the hypercube $S_{(X,Y,\rho_{k,i,\infty})} = \{\,(x,y) : \max\{\|x-X_i\|_\infty, \|y-y_i\|_\infty\} \le \rho_{k,i,\infty}\,\}$, or on the Y-boundary of $S_{(X,Y,\rho_{k,i,\infty})}$ (the chance of lying on a corner, and thus on both the boundaries, has zero probability). If the k-NN lies on the X-boundary, i.e. $\|X_i^{(k)} - X_i\| = \rho_{k,i,\infty}$ and $\|Y_i^{(k)} - Y_i\|_\infty < \rho_{k,i,\infty}$, then $\rho_{k,i,\infty}$ is the $(n_{x,i,\infty}+1)$-NN distance of X_i, but *not* the $(n_{y,i,\infty}+1)$-NN distance of Y_i. Thus, while the estimate of entropy of X is correct, the entropy of Y is over-estimated. Since $\rho_{k,i,\infty}$ is between the $n_{y,i,\infty}$-th and $(n_{y,i,\infty}+1)$-th NN distance, the "KSG entropy estimator" in (3.28) introduces a bias of order $1/n_{y,i,\infty}$. Similarly, a $1/n_{x,i,\infty}$-bias if $(X_i^{(k)}, Y_i^{(k)})$ is introduced if the k-NN sample lies on the Y-boundary.

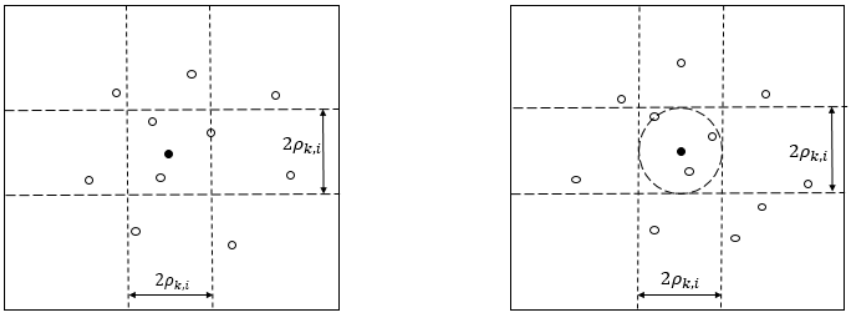

Figure 3.4: Illustration of choice of $\rho_{k,i}$ for $k=3$. Left: use ℓ_∞-distance. Right: use ℓ_2-distance.

This discussion suggests that we use an ℓ_2 ball, instead of an ℓ_∞ ball to find the k-NN. This would ensure that $\rho_{k,i,2}$ is neither the $(n_{x,i,2}+1)$-NN distance

of X_i on X-space nor the $(n_{y,i,2}+1)$-NN distance of Y_i on Y-space. But then, we are unable to directly use the KL estimator for $H(X)$ and $H(Y)$ with this distance. The following theorem sheds some light on this conundrum.

Theorem 7. *Given $(X_i, Y_i) = (x, y)$ such that the density f is twice continuously differentiable at (x,y) and $\rho_{k,i,2} = r < r_N$ for some deterministic sequence of r_N such that $\lim_{N\to\infty} r_N = 0$, the number of neighbors $n_{x,i,2} - k$ is distributed as $\sum_{l=k+1}^{N-1} U_l$, where U_l are i.i.d. Bernoulli random variables with mean p, and there exists a positive constant C_1 such that for sufficiently large N. $r^{-d_x} \left| p - f_X(x) c_{d_x,2} r^{d_x} \right| \leq C_1 \left(r^2 + r^{d_y} \right)$.*

Intuitively, the theorem says that $E[n_{x,i,2}] \approx N f_X(x) c_{d_x,2} \rho_{k,i,2}^{d_x}$. This suggests that we estimate the log of the density $(\log \widehat{f_X}(x))$ by $\log(n_{x,i,2}) - \log N - \log c_{d_x,2} - d_x \log \rho_{k,i,2}$. The resubstitution estimate of the marginal entropy $H(X)$ is now:

$$\begin{aligned} \widehat{H}_{\mathrm{BI-KSG}}(X) &= \log N + \log c_{d_x,2} \\ &\quad + \frac{1}{N} \sum_{i=1}^{N} \left(-\log(n_{x,i,2}) + d_x \log \rho_{k,i,2} \right), \end{aligned} \tag{3.29}$$

which is different from the KL estimate only via replacing the digamma function by the logarithm. This technique kills the $O(1/n_{x,i,2} + 1/n_{y,i,2})$ bias of the "KSG entropy estimator" and leads to the new estimator of mutual information that we christen *bias-improved KSG estimator*:

$$\begin{aligned} \widehat{I}_{\mathrm{BI-KSG}}(X;Y) &\equiv \psi(k) + \log N + \log \left(\frac{c_{d_x,2} c_{d_y,2}}{c_{d_x+d_y,2}} \right) \\ &\quad - \frac{1}{N} \sum_{i=1}^{N} \log(n_{x,i,2}) + \log(n_{y,i,2}), \end{aligned} \tag{3.30}$$

where $c_{d,2} = \pi^{\frac{d}{2}} / \Gamma(\frac{d}{2} + 1)$ be the volume of d-dimensional unit ℓ_2 ball.

The theoretical performance of this new estimator, which mimics our result on the KSG estimator have been analyzed in Section 3.1. The consistency of BI-KSG estimator is proved in Theorem 3. The bias of the BI-KSG estimator is $\tilde{O}(N^{-\frac{1}{d_x+d_y}})$ shown in Theorem 4 and the variance is $\tilde{O}(1/N)$ shown in Theorem 5. Thus the ℓ_2 error of the BI-KSG estimator is $\tilde{O}(\frac{1}{\sqrt{N}} + N^{-\frac{1}{d_x+d_y}})$.

Indeed, when N gets large, so do $n_{x,i,2}$ and $n_{y,i,2}$, and hence the KSG and BI-KSG estimators asymptotically perform similarly. But when k is small

and N is moderate and X and Y are not independent, then $n_{x,i,2}$ and $n_{y,i,2}$ are *expected* to be small. In such cases, BI-KSG should outperform KSG. We demonstrate this empirically in Table 3.2 where we choose $k = 1$ and X and Y are joint Gaussian with mean 0 and covariance $\Sigma = [1, 0.9; 0.9, 1]$. We can see that all the estimators converge to the ground truth as N goes to infinity, but BI-KSG has the best sample complexity for moderate values of N. Overall, the empirical gains of correlation boosting are most seen in moderate sample sizes.

Our current theoretical understanding leads to the same upper bounds on the asymptotic rates of convergence for the KSG and BI-KSG and 3KL estimators (cf. Corollary 1), and fails to explain the correlation boosting effects. We suspect that the gains of correlation boosting are not in the first order terms in the rates of convergence (of bias and variance) but in the multiplicative constants. A theoretical understanding of these constant terms is an interesting future direction of research; such an effort has been successfully conducted for entropy estimators based on kernel density estimators [54].

Table 3.2: Comparison of bias for different mutual information estimators.

N	100	200	400	800	1600
3KL	0.0590	0.1025	0.0313	0.0053	0.0097
KSG	0.0240	0.0100	0.0217	0.0024	0.0087
BI-KSG	**0.0096**	**-0.0035**	**0.0133**	**-0.0012**	**0.0071**

3.3 Multivariate mutual information

Generalizations of the standard mutual information that measure the relation *among a sequence* of random variables are routinely used in various applications of machine learning. We discuss two such multivariate versions of mutual information below and show how the correlation boosting ideas from the previous section can be used to construct sample-efficient estimators. The first version is a straightforward generalization and routinely used in unsupervised clustering and correlation extraction, cf. [90, 23, 91, 92] for

a few recent applications:

$$I(X_1; X_2; X_3; \cdots; X_L) = \sum_{\ell=1}^{L} H(X_\ell) - H(X_1, X_2, \ldots, X_L). \quad (3.31)$$

This definition of multivariate mutual information originated in [93] as an upper bound for secrecy capacity in a multiterminal source model. This divergence expression has also been termed "shared information" by its originators in [94].

One natural way to estimate this multivariate mutual information (MMI) is to use the sum and differences of the basic entropy estimators. In particular, one can use the fixed k-NN based KL entropy estimator to estimate MMI from i.i.d. samples (we can christen such a method as the $L+1$-KL estimator, generalizing from the 3KL estimator). Alternatively, one can use the correlation boosting ideas of KSG and BI-KSG to construct superior MMI estimators. Generalizing from (3.2) and (3.4) we construct the estimators:

$$\begin{aligned} & I_{\mathrm{KSG}}(X_1; X_2; X_3; \cdots; X_L) \\ = \; & \psi(k) + \log N - \frac{1}{N} \sum_{i=1}^{N} \sum_{\ell=1}^{L} \psi(n_{x_\ell, i, \infty}), \quad (3.32) \\ & I_{\mathrm{BI-KSG}}(X_1; X_2; X_3; \cdots; X_L) \\ = \; & \psi(k) + \log N + \log \Big(\frac{\prod_{\ell=1}^{L} c_{d_\ell, 2}}{c_{\sum_{\ell=1}^{L} d_\ell, 2}} \Big) - \frac{1}{N} \sum_{i=1}^{N} \sum_{\ell=1}^{L} \log(n_{x_\ell, i, 2}). \quad (3.33) \end{aligned}$$

Here d_ℓ is the dimension of X_ℓ. The key property we used in constructing these estimators is that the definition of MMI is *balanced* with respect to each of the L random variables: for every entropy term with a positive coefficient featuring a random variable X_ℓ there is a corresponding entropy term with a negative coefficient featuring the same random variable X_ℓ. From a theoretical perspective, the balance property ensures that the theoretical properties (including consistency) proved in the (pairwise) mutual information setting in Section 3.1 carry over to this MMI setting as well. From an empirical perspective, we see that the correlation boosting estimators perform significantly better than the simpler $(L+1)$-KL estimator defined as $\widehat{I}_{(L+1)-\mathrm{KL}} = \sum_{j=1}^{L} \widehat{H}_{\mathrm{KL}}(X_j) - \widehat{H}_{\mathrm{KL}}(X_1, \ldots, X_L)$ in Figure 3.5 where $N = 100 \sim 3000$ and $L = 3$ and the random variables are jointly

Gaussian with covariance matrix [1 1/2 1/4; 1/2 1 1/2; 1/4 1/2 1].

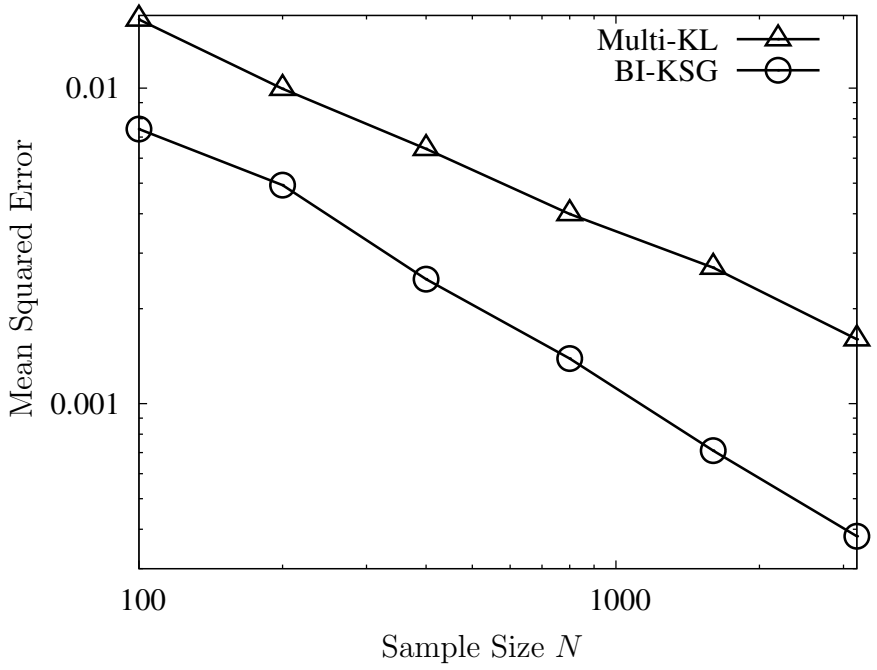

Figure 3.5: Plot of MSE with sample size. BI-KSG performs marginally better than KSG.

As another application of our ideas, we consider a more general form of multivariate mutual information:

$$\mathrm{MMI}(X_1; X_2; X_3; \cdots; X_L) \;=\; \sum_{S \subset \{1,\ldots,L\}} a_S \;\cdot\; H(X_S), \qquad (3.34)$$

for some *balanced* real valued set function a_S, i.e., for every $\ell = 1 \ldots L$ we have $\sum_{S \ni \ell \in S} a_S = 0$. Such a metric was posited recently in the context of causal influence measurement on probabilistic graphical models (cf. Equation (9) in [95]) and widely studied in the information theory community due to its invariance to scaling (cf. [96] for a recent example). The definition in (3.31) is a special case with the set function equal to 1 for singletons and -1 for the whole set and 0, otherwise (and can be viewed as arising out of a graphical model with a single latent variable). Such MMI can be estimated from samples using the correlation boosting ideas presented in this chapter: we briefly describe the procedure in the context of an example (which can be viewed as a certain causal strength measurement [95] with respect to

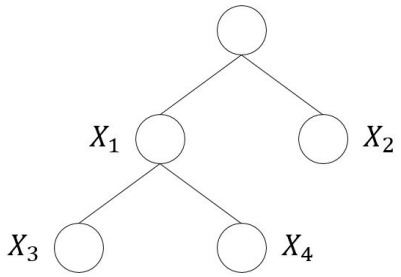

Figure 3.6: Causal influence on a specific graphical model.

the graphical model in Figure 3.6): $\mathrm{MMI}(X_1, X_2, X_3, X_4) = H(X_1X_3) + H(X_1X_4) - H(X_1) + H(X_2) - H(X_1X_2X_3X_4)$. For each sample i, we first find the k-NN distance ρ in the joint space (of four random variables) and use it estimate the joint entropy using the KL estimator. Then we use this distance to calculate the number of neighbors in each of the other subset of random variables (in this case two pairwise ones ((X_1X_3) and (X_1X_4)), and two marginal ones (X_1 and X_2), and use these to estimate the corresponding entropies. The balanced nature of the metric ensures that the actual distance ρ is *precisely canceled out* when all the entropy estimators are put together. In this case, the full estimator (in the spirit of the KSG estimator) is the following, and directly inherits the theoretical and empirical flavor of results from those in Section 3.1:

$$\begin{aligned} \mathrm{MMI}_{\mathrm{KSG}} &= \psi(k) + \log N - \frac{1}{N}\sum_{i=1}^{N}\big(\psi(n_{x_1x_3,i,\infty}) + \psi(n_{x_1x_4,i,\infty}) \\ &\quad -\psi(n_{x_1,i,\infty}) + \psi(n_{x_2,i,\infty})\big). \end{aligned} \tag{3.35}$$

3.4 Discussion and related work

In this section we address the question of estimating functionals, including the entropy and the mutual information, of an unknown distribution. Significant understanding of the minimax rate-optimal estimators has been attained via [76, 84, 81, 85, 77]. In the case of continuous random vec-

tors, this fundamental question of estimating a functional of the (unknown) density has been of longstanding interest in the statistics community [41]. Further, this has been investigated in the machine learning [79, 80], information theory [75, 76, 77, 78], and theoretical computer science [81, 82, 83] communities. The popularity of mutual information and other information theoretic quantities comes from their wide use as basic features in several downstream applications [27, 26, 73, 74].

A conceptually straightforward way to estimate the differential entropy and mutual information is to use a kernel density estimator (KDE) [97, 54, 98, 99, 100, 56]: the densities $f_{X,Y}, f_X, f_Y$ are separately estimated from samples and the estimated densities are then used to calculate the entropy and mutual information via the resubstitution estimator. A typical approach to avoid overfitting is to conduct data splitting (DS): split the samples and use one part for KDE and the other for the resubstitution.

In some cases, the parametric rate of convergence of $\sqrt{N}$ of ℓ_2 error is achieved: of particular interest is the result of [54] where the parametric rate is achieved for differential entropy estimation via KDE of density followed by the resubstitution estimator when the dimension is no more than 6. Numerical evidence suggests the hypothesis that the lower bounds could perhaps be improved when the dimension is more than 4 and estimators constrained to only use fixed k-NN distances. Under certain very strong conditions on the density class (that are relevant in certain applications on graphical model selection [101]), exponential rate of convergence can be demonstrated [102, 103]. Recent works [67, 31] have studied the performance of the leave-one-out (LOO) approach where all but the sample of resubstitution are used for KDE, involving techniques such as von Mises expansion methods.

Alternative methods involve estimation of the entropies using spacings [104, 55], the Edgeworth expansion [105], and convex optimization [106]. Among the k-NN methods, there are two variants: either k is chosen to grow with the sample size N or k is fixed. There is a large literature on the former, where the classical result is the possibility of consistent estimation of the density from k-NN distances [107, 108], including recent sharper consistency characterizations [109, 110]. Several works have applied this basic insight towards the estimation of the specific case of information theoretic quantities [111, 112], general nonlinear functions of densities [33] and extensions to generalized NN graphs [113]. For fixed k-NN methods, apart from the works

referred to in the main text, detailed experimental comparisons are in [114] and local Gaussian approaches studied in [79, 87, 115] bringing together local likelihood density estimation methods [7, 6] with k-NN driven choices of kernel bandwidth.

In this chapter we have considered the smoothness of the class of pdfs studied via bounded Hessians. In nonparametric estimation, a standard feature is to consider whole families of smooth pdfs as defined by how the differences of derivatives relate to the differences of the samples. Of specific interest is the Hölder family: $\Sigma(s, C)$, i.e., for any tuple $r = (r_1, \ldots, r_d)$, define $D^r = \frac{\partial^{r_1+\cdots+r_d}}{\partial x_1^{r_1} \ldots \partial x_d^{r_d}}$. Then for any r such that $\sum_j r_j = \lfloor s \rfloor$, where $\lfloor s \rfloor$ is the largest integer smaller than s, we have:

$$\|D^r f(x) - D^r f(y)\| \leq C \|x - y\|^{s - \sum_j r_j}, \tag{3.36}$$

for any x, y. The rate of convergence of various nonparametric estimators depends on the parameter s of the Hölder family under consideration, cf. [67, 31] for recent work on convergence rate characterization of information theoretic quantities via KDE and resubstitution estimators as a function of the smoothness parameter s. Recent work [34] showed convergence rate for fixed k-NN entropy estimator under such smoothness assumptions. It is natural to ask if such smoothness considerations could lead to a refined understanding of the rates of convergence of KSG estimators studied here.

For small enough r, defining $P(x, r)(u) = \Pr\{\|X - x\| < r\}$, we seek to understand how this probability can be approximated by the density at x. With bounded Hessian norms, we assert the following:

$$\left| P(x, r) - f(x) c_d r^d \right| \;\; \leq \;\; C r^{d+2}, \tag{3.37}$$

which is crucial in deriving the rate of convergence upper bounds on the KL estimator. A fairly straightforward calculation shows that this condition does not change even if we allow for smoother class of families of pdfs, as defined via the Hölder class – we conclude that refined rates of convergence for fixed k-NN estimators do not materialize by standard approaches such as the Hölder class.

Although our analysis technique is inspired by that of [32], several subtle differences emerge while generalizing it to higher dimensions, and [32] does

not imply our result even for $d = 1$: hence, we complement the understanding of k-NN estimators even for univariate random variables. For example, random variables with strictly positive densities over a bounded support are covered by our analysis, whereas random variables with unbounded support that are smooth everywhere are covered by the results of [32]. The reason is that non-smooth boundaries are not handled in [32] and densities approaching zero are not handled by our analysis. We believe it is possible to extend our analysis to have a theorem that includes both types of random variables, which is an interesting future research direction.

In this chapter, k is assumed to be a finite constant, and we do not keep track of how the convergence rate depends on k. Analyses on fixed ρ estimators [116], where instead of fixing k and using the distance ρ_k, one fixes the distance ρ and uses the number of neighbors k_ρ within that distance, we expect the convergence rate of the variance to be independent of k, and the convergence rate of bias to be of order $O((k/N)^{1/d})$. Recently, the idea of using an ensemble of k-NN entropy estimators to achieve a faster convergence rate has been introduced in [116, 117], and applied in [118, 119, 120]. If the first-order terms in the convergence rate is known, then it is possible to achieve the parametric rate of $O(1/N)$ by taking a (weighted) linear combination of multiple estimators with varying k, whose weight depends on the convergence rate. Applying this idea together with KSG (and KL) estimators have the potential to improve the convergence rate we provide in this chapter.

3.5 Proof of Theorem 3 on the consistency of KSG estimator

Note that

$$\begin{aligned} & \widehat{I}_{KSG}(X;Y) \\ = \; & \widehat{H}_{KSG}(X) + \widehat{H}_{KSG}(Y) - \widehat{H}_{KL,\infty}(X,Y), \qquad (3.38) \\ & \widehat{I}_{BI-KSG}(X;Y) \\ = \; & \widehat{H}_{BI-KSG}(X) + \widehat{H}_{BI-KSG}(Y) - \widehat{H}_{KL,2}(X,Y), \qquad (3.39) \end{aligned}$$

where

$$\begin{aligned}
&\widehat{H}_{KL,\infty}(X,Y) \\
\equiv\ & -\psi(k) + \log N + \log c_{d_x,\infty} c_{d_y,\infty} + (d_x + d_y) \log \rho_{k,i,\infty}, \qquad (3.40) \\
&\widehat{H}_{KSG}(X) \\
\equiv\ & -\frac{1}{N} \sum_{i=1}^{N} \psi(n_{x,i,\infty} + 1) + \log N + \log c_{d_x,\infty} + d_x \log \rho_{k,i,\infty}, \qquad (3.41) \\
&\widehat{H}_{KSG}(Y) \\
\equiv\ & -\frac{1}{N} \sum_{i=1}^{N} \psi(n_{y,i,\infty} + 1) + \log N + \log c_{d_y,\infty} + d_y \log \rho_{k,i,\infty}, \qquad (3.42)
\end{aligned}$$

and

$$\begin{aligned}
&\widehat{H}_{KL,2}(X,Y) \\
\equiv\ & -\psi(k) + \log N + \log c_{d_x+d_y,2} + (d_x + d_y) \log \rho_{k,i,2}, \qquad (3.43) \\
&\widehat{H}_{BI-KSG}(X) \\
\equiv\ & -\frac{1}{N} \sum_{i=1}^{N} \log n_{x,i,2} + \log N + \log c_{d_x,2} + d_x \log \rho_{k,i,2}, \qquad (3.44) \\
&\widehat{H}_{BI-KSG}(Y) \\
\equiv\ & -\frac{1}{N} \sum_{i=1}^{N} \log n_{y,i,2} + \log N + \log c_{d_y,2} + d_y \log \rho_{k,i,2}. \qquad (3.45)
\end{aligned}$$

We prove the following technical lemma that shows the convergence of the marginal entropy estimate (3.41) and (3.44) . The convergence of (3.42) and (3.45) is immediate by interchanging X and Y. The convergence in probability of the joint entropy estimate (3.40) and (3.43) are known from [4]. This proves the desired claim.

Lemma 5. *Under the hypotheses of Theorem 3, the estimated marginal entropy converges to the true entropy, i.e. for all* $\varepsilon > 0$

$$\lim_{N \to \infty} \Pr\left(\left| \widehat{H}_{KSG}(X) - H(X) \right| > \varepsilon \right) = 0, \qquad (3.46)$$

$$\lim_{N \to \infty} \Pr\left(\left| \widehat{H}_{BI-KSG}(X) - H(X) \right| > \varepsilon \right) = 0. \qquad (3.47)$$

3.5.1 Proof of Lemma 5

Define the following quantities

$$\widehat{f}_X^{KSG}(X_i) \equiv \frac{\exp\{\psi(n_{x,i,\infty}+1)\}}{Nc_{d_x,\infty}\rho_{k,i,\infty}^{d_x}}, \tag{3.48}$$

$$\widehat{f}_X^{BI-KSG}(X_i) \equiv \frac{n_{x,i,2}}{Nc_{d_x,2}\rho_{k,i,2}^{d_x}}. \tag{3.49}$$

Then we have the following equations $\widehat{H}_{KSG}(X) = -\frac{1}{N}\sum_{i=1}^{N}\log \widehat{f}_X^{KSG}(X_i)$ and $\widehat{H}_{BI-KSG}(X) = -\frac{1}{N}\sum_{i=1}^{N}\log \widehat{f}_X^{KSG}(X_i)$. From now on we will skip the subscript KSG or BI-KSG and the subscript 2 or ∞ if the formula holds for both. We will specify it when necessary. Now we write $|\widehat{H}(X) - H(X)|$ as:

$$\begin{aligned}
&\left|\widehat{H}(X) - H(X)\right| = \left| -\frac{1}{N}\sum_{i=1}^{N}\log \widehat{f}_X(X_i) - \left(-\int f_X(x)\log f_X(x)dx\right)\right| \\
\leq\ & \left|\frac{1}{N}\sum_{i=1}^{N}\log f_X(X_i) - \int f_X(x)\log f_X(x)dx\right| \\
&+\frac{1}{N}\sum_{i=1}^{N}\left|\log \widehat{f}_X(X_i) - \log f_X(X_i)\right|.
\end{aligned} \tag{3.50}$$

The first term is the error from the empirical mean. Notice that $\log f_X(X_i)$ are i.i.d. random variables, satisfying

$$\mathbb{E}\,|\log f_X(X_i)\,| = \int f_X(x)|\log f_X(x)|dx < +\infty, \tag{3.51}$$

where the mean is given by:

$$\mathbb{E}\,(\log f_X(X_i)\,) = \int f_X(x)\log f_X(x)dx. \tag{3.52}$$

Therefore, by weak law of large numbers, we have:

$$\begin{aligned}
&\lim_{N\to\infty}\Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N}\log f_X(X_i) - \int f_X(x)\log f_X(x)dx\right| > \varepsilon\right) \\
=\ & 0,
\end{aligned} \tag{3.53}$$

for any $\varepsilon > 0$.

The second term comes from density estimation. We denote $Z = (X, Y)$

and $f(z) = f(x, y)$ for short, then for any fixed $\varepsilon > 0$, we obtain:

$$\begin{aligned}
& \Pr\left(\frac{1}{N} \sum_{i=1}^{N} \left| \log \widehat{f}_X(X_i) - \log f_X(X_i) \right| > \varepsilon \right) \\
\leq\ & \Pr\left(\bigcup_{i=1}^{N} \left\{ \left| \log \widehat{f}_X(X_i) - \log f_X(X_i) \right| > \varepsilon \right\} \right) \\
\leq\ & N \Pr\left(\left| \log \widehat{f}_X(X_i) - \log f_X(X_i) \right| > \varepsilon \right) \\
=\ & N \int \underbrace{\Pr\left(\left| \log \widehat{f}_X(X_i) - \log f_X(X_i) \right| > \varepsilon \middle| Z_i = z \right)}_{=I_1(z)+I_2(z)+I_3(z)} f(z) dz, \quad (3.54)
\end{aligned}$$

where

$$I_1(z) = \Pr\left(\rho_{k,i} > \frac{\log N}{(N f(z) c_{d_x+d_y})^{\frac{1}{d_x+d_y}}} \middle| Z_i = z \right), \quad (3.55)$$

$$I_2(z) = \Pr\left(\rho_{k,i} < \frac{(\log N)^2}{(N f_X(x) c_{d_x})^{\frac{1}{d_x}}} \middle| Z_i = z = (x, y) \right), \quad (3.56)$$

$$\begin{aligned}
I_3(z) =\ & \int_{r=(\log N)^2 (N f_X(x) c_{d_x})^{-\frac{1}{d_x}}}^{\log N (N f(z) c_{d_x+d_y})^{-\frac{1}{d_x+d_y}}} \Pr\Big(\left| \log \widehat{f}_X(X_i) - \log f_X(X_i) \right| > \varepsilon \\
& \left| \rho_{k,i} = r, Z_i = z \right) f_{\rho_{k,i}}(r) dr, \quad (3.57)
\end{aligned}$$

where $f_{\rho_{k,i}}(r)$ is the pdf of $\rho_{k,i}$ given $Z_i = z$. We will consider the three terms separately, and show that each is bounded by $o(N^{-1})$.

I_1: Let $B_Z(z, r) = \{Z : \|Z - z\| < r\}$ be the $(d_x + d_y)$-dimensional ball centered at z with radius r. Since the Hessian matrix of $H(f)$ exists and $\|H(f)\|_2 < C$ almost everywhere, then for sufficiently small r, there exists z' such that

$$\begin{aligned}
& \left| \Pr(\, u \in B_Z(z, r)\,) - f(z) c_{d_x+d_y} r^{d_x+d_y} \right| \\
=\ & \int_{\|u-z\| \leq r} (\, f(u) - f(z)\,)\, du \\
=\ & \int_{\|u-z\| \leq r} \Big((u - z)^T \nabla f(z) + (u - z)^T H_f(z')(u - z) du \Big) \\
\leq\ & C f(z) c_{d_x+d_y} r^{d_x+d_y+2}. \quad (3.58)
\end{aligned}$$

Then for sufficiently large N such that $C(\log N/(N f(z) c_{d_x+d_y})^{\frac{1}{d_x+d_y}})^2 < 1/2$,

we have

$$\begin{aligned} & p_1 = \Pr\left(u \in B_Z(z, \log N (N f(z) c_{d_x+d_y})^{-\frac{1}{d_x+d_y}}) \right) \\ \geq & \frac{1}{2} f(z) c_{d_x+d_y} \left(\frac{\log N}{(N f(z) c_{d_x+d_y})^{\frac{1}{d_x+d_y}}} \right)^{d_x+d_y} \\ \geq & \frac{(\log N)^{d_x+d_y}}{2N}. \end{aligned} \tag{3.59}$$

Therefore, for any $d_x, d_y \geq 1$, $I_1(z)$ is upper bounded by:

$$\begin{aligned} & I_1(z) = \Pr\left(\rho_{k,i} > \frac{\log N}{(N f(z) c_{d_x+d_y})^{\frac{1}{d_x+d_y}}} \big| Z_i = z \right) \\ = & \sum_{m=0}^{k-1} \tbinom{N}{m} p_1^m (1-p_1)^{N-1-m} \leq \sum_{m=0}^{k-1} N^m (1-p_1)^{N-1-m} \\ \leq & k N^{k-1} (1 - \frac{(\log N)^{d_x+d_y}}{2N})^{N-k-1} \\ \leq & k N^{k-1} \exp\{-\frac{(\log N)^{d_x+d_y}(N-k-1)}{2N}\} \\ \leq & k N^{k-1} \exp\{-\frac{(\log N)^{d_x+d_y}}{4}\}. \end{aligned} \tag{3.60}$$

I_2: For sufficiently large N such that $C((\log N)^2)/(N f_X(x) c_{d_x})^{\frac{1}{d_x}})^2 < 1$, we have

$$\begin{aligned} & p_2 = \Pr\left(u \in B_Z(z, \frac{(\log N)^2}{(N f_X(x) c_{d_x})^{\frac{1}{d_x}}}) \right) \\ \leq & 2 f(z) c_{d_x+d_y} \left(\frac{(\log N)^2}{(N f_X(x) c_{d_x})^{\frac{1}{d_x}}} \right)^{d_x+d_y} \\ \leq & \frac{2 f(z) c_{d_x+d_y}}{(f(x) c_{d_x})^{\frac{d_x+d_y}{d_x}}} (\log N)^{2(d_x+d_y)} N^{-\frac{d_x+d_y}{d_x}} \\ \leq & 2 f_{Y|X}(y|x) \frac{c_{d_x+d_y}}{c_{d_x}} (\log N)^{2(d_x+d_y)} N^{-\frac{d_x+d_y}{d_x}} \\ \leq & 2 C_e \frac{c_{d_x+d_y}}{c_{d_x}} (\log N)^{2(d_x+d_y)} N^{-\frac{d_x+d_y}{d_x}}. \end{aligned} \tag{3.61}$$

For any $d_x, d_y \geq 1$ and $k \geq 1$, I_2 is upper bounded by:

$$\begin{aligned} & I_2(z) = \Pr\left(\rho_{k,i} < \frac{(\log N)^2}{(N f_X(x) c_{d_x})^{\frac{1}{d_x}}} \big| Z_i = z \right) \\ = & \sum_{m=k}^{N-1} \binom{N-1}{m} p_2^m (1-p_2)^{N-1-m} \leq \sum_{m=k}^{N-1} N^m p_2^m \\ \leq & \sum_{m=k}^{N-1} (2C_e \frac{c_{d_x+d_y}}{c_{d_x}} (\log N)^{2(d_x+d_y)} N^{-\frac{d_y}{d_x}})^m \\ \leq & (4C_e \frac{c_{d_x+d_y}}{c_{d_x}})^k (\log N)^{2K(d_x+d_y)} N^{-\frac{k d_y}{d_x}}. \end{aligned} \tag{3.62}$$

I_3**:** Now we will consider KSG and BI-KSG separately. Also we need to specify whether we are considering ℓ_2 or ℓ_∞ norm. For KSG, given that $Z_i = z = (x, y)$ and $\rho_{k,i,\infty} = r$, we have:

$$\begin{aligned} & \Pr_{r,z}\Big(\Big| \log \widehat{f}_X^{KSG}(X_i) - \log f_X(X_i) \Big| > \varepsilon \Big) \\ = & \Pr_{r,z}\Big(\big| \psi(n_{x,i,\infty}+1) \log\big(N c_{d_x,\infty} \rho_{k,i,\infty}^{d_x} f_X(x)\big) \big| > \varepsilon \Big). \end{aligned} \tag{3.63}$$

Here $\Pr_{r,z}$ denotes the probability given $\rho_{k,i,\cdot} = r$ and $Z = z$ for notation simplicity. Notice that for any integer $x \geq 2$, we have $\log(x-1) < \psi(x) < \log(x)$. Therefore

$$\begin{aligned} & \Pr_{r,z}\Big(\psi(n_{x,i,\infty}+1) - \log\big(N c_{d_x,\infty} \rho_{k,i,\infty}^{d_x} f_X(x)\big) < -\varepsilon \Big) \\ \leq & \Pr_{r,z}\Big(\log n_{x,i,\infty} - \log\big(N c_{d_x,\infty} \rho_{k,i,\infty}^{d_x} f_X(x)\big) < -\varepsilon \Big) \\ = & \Pr_{r,z}\Big(n_{x,i,\infty} < N c_{d_x,\infty} r^{d_x} f_X(x) e^{-\varepsilon} \Big). \end{aligned} \tag{3.64}$$

In the other direction,

$$\begin{aligned} & \Pr_{r,z}\Big(\psi(n_{x,i,\infty}+1) - \log\big(N c_{d_x,\infty} \rho_{k,i,\infty}^{d_x} f_X(x)\big) > \varepsilon \Big) \\ \leq & \Pr_{r,z}\Big(\log(n_{x,i,\infty}+1) - \log\big(N c_{d_x,\infty} \rho_{k,i,\infty}^{d_x} f_X(x)\big) > \varepsilon \Big) \\ = & \Pr_{r,z}\Big(n_{x,i,\infty} > N c_{d_x,\infty} r^{d_x} f_X(x) e^{\varepsilon} - 1 \Big). \end{aligned} \tag{3.65}$$

For BI-KSG, we have:

$$\begin{aligned}
&\Pr_{r,z}\Big(\Big|\log \widehat{f}_X^{BI-KSG}(X_i) - \log f_X(X_i)\Big| > \varepsilon \Big) \\
=\ &\Pr_{r,z}\Big(\big|\log n_{x,i,2} - \log\big(Nc_{d_x,2}\rho_{k,i,2}^{d_x} f_X(x)\big)\big| > \varepsilon, \Big) \\
=\ &\Pr_{r,z}\Big(n_{x,i,2} > Nc_{d_x,2} r^{d_x} f_X(x) e^{\varepsilon} \Big) + \Pr_{r,z}\Big(n_{x,i,2} < Nc_{d_x,2} r^{d_x} f_X(x) e^{-\varepsilon} \Big).
\end{aligned} \tag{3.66}$$

Combine them together, we have:

$$\begin{aligned}
&\Pr_{r,z}\Big(\Big|\log \widehat{f}_X(X_i) - \log f_X(X_i)\Big| > \varepsilon \Big) \\
\le\ &\Pr_{r,z}\Big(n_{x,i} < Nc_{d_x} r^{d_x} f_X(x) e^{-\varepsilon} \Big) + \Pr_{r,z}\Big(n_{x,i} > Nc_{d_x} r^{d_x} f_X(x) e^{\varepsilon} - 1 \Big),
\end{aligned} \tag{3.67}$$

holds for both KSG and BI-KSG estimates. Recall that in Theorem 7, given that $\rho_{k,i} = r$ and $Z_i = z$, $n_{x,i} - k$ is distributed as $\sum_{l=k+1}^{N-1} U_l$, where U_l are i.i.d Bernoulli random variables with mean p satisfying

$$r^{-d_x}\,\big|\, p - f_X(x) c_{d_x} r^{d_x} \,\big| \le C_1(r^2 + r^{d_y}). \tag{3.68}$$

For small enough r such that $C_1(r^2 + r^{d_y}) \le \varepsilon/2$, we obtain

$$\begin{aligned}
&\Pr_{r,z}\big(n_{x,i} > (N-1) c_{d_x} r^{d_x} f_X(x) e^{\varepsilon} - 1 \big) \\
=\ &\Pr\left(\sum_{l=k+1}^{N-1} U_l > (N-1) c_{d_x} r^{d_x} f_X(x) e^{\varepsilon} - k - 1 \right) \\
=\ &\Pr\Big(\sum_{l=k+1}^{N-1} (U_l - \mathbb{E}[U_l]) > (N-1) c_{d_x} r^{d_x} f_X(x) e^{\varepsilon} - k - 1 \\
&\quad -(N-k-1)\mathbb{E}[U_l] \Big),
\end{aligned} \tag{3.69}$$

and the right-hand side in the probability is lower bounded by

$$\begin{aligned}
& Nc_{d_x} r^{d_x} f_X(x) e^{\varepsilon} - k - 1 - (N-k-1)\mathbb{E}[U_l] \\
\geq\ & Nc_{d_x} r^{d_x} f_X(x) e^{\varepsilon} - k - 1 - (N-k-1) f_X(x) c_{d_x} r^{d_x} (1+\varepsilon/2) \\
\geq\ & (N-k-1) c_{d_x} r^{d_x} f_X(x)(e^{\varepsilon} - 1 - \varepsilon/2) - k - 1 \\
\geq\ & (N-k-1) c_{d_x} r^{d_x} f_X(x) \varepsilon/4, \qquad (3.70)
\end{aligned}$$

for sufficiently large N such that $(N-k-1)c_{d_x} r^{d_x} f_X(x)(e^{\varepsilon} - 1 - \varepsilon/4) > k+1$. Since U_l is Bernoulli, we have $\mathbb{E}[U_l^2] = \mathbb{E}[U_l]$. Now applying Bernstein's inequality, (3.69) is upper bounded by:

$$\begin{aligned}
& \Pr\Big(\sum_{l=k+1}^{N-1} (U_l - \mathbb{E}[U_l]) > (N-1) c_{d_x} r^{d_x} f_X(x) e^{\varepsilon} - k - (N-k-1)\mathbb{E}[U_l] \Big) \\
\leq\ & \Pr\Big(\sum_{l=k+1}^{N-1} (U_l - \mathbb{E}[U_l]) > (N-k-1) c_{d_x} r^{d_x} f_X(x) \varepsilon/4 \Big) \\
\leq\ & \exp\Big\{ - \frac{((N-k-1) c_{d_x} r^{d_x} f_X(x) \varepsilon/4)^2}{2(N-k-1)\mathbb{E}[U_l^2] + \frac{2}{3}((N-k-1) c_{d_x} r^{d_x} f_X(x) \varepsilon/4)} \Big\} \\
=\ & \exp\Big\{ - \frac{\varepsilon^2}{32(1+\frac{7\varepsilon}{12})} (N-k-1) c_{d_x} r^{d_x} f_X(x) \Big\}. \qquad (3.71)
\end{aligned}$$

Similarly, the tail bound on the other way is given by:

$$\begin{aligned}
& \Pr_{r,z}\big(n_{x,i} < Nc_{d_x} r^{d_x} f_X(x) e^{-\varepsilon} \big) \\
=\ & \Pr\left(\sum_{l=k+1}^{N-1} U_l < Nc_{d_x} r^{d_x} f_X(x) e^{-\varepsilon} - k \right) \\
=\ & \Pr\Big(\sum_{l=k+1}^{N-1} (U_l - \mathbb{E}[U_l]) < Nc_{d_x} r^{d_x} f_X(x) e^{-\varepsilon} - k - (N-k-1)\mathbb{E}[U_l] \Big), \\
& \qquad (3.72)
\end{aligned}$$

and the right-hand side in the probability is upper bounded by

$$\begin{aligned}
& Nc_{d_x} r^{d_x} f_X(x) e^{-\varepsilon} - k - (N-k-1)\mathbb{E}[U_l] \\
\leq\ & Nc_{d_x} r^{d_x} f_X(x) e^{-\varepsilon} - k - (N-k-1) f_X(x) c_{d_x} r^{d_x} (1-\varepsilon/2) \\
\leq\ & (N-k-1) c_{d_x} r^{d_x} f_X(x) e^{-\varepsilon} - (N-k-1) f_X(x) c_{d_x} r^{d_x} (1-\varepsilon/2) \\
=\ & (N-k-1) c_{d_x} r^{d_x} f_X(x) \left(e^{-\varepsilon} - 1 + \varepsilon/2 \right) \\
\leq\ & -(N-k-1) c_{d_x} r^{d_x} f_X(x) \varepsilon/4, \qquad (3.73)
\end{aligned}$$

for sufficiently small r such that $(k+1)c_{d_x} r^{d_x} f_X(x) e^{-\varepsilon} < k$ and sufficiently small ε such that $e^{-\varepsilon} - 1 + \varepsilon/2 \leq -\varepsilon/4$. Similarly, by applying Bernstein's inequality, (3.72) is upper bounded by:

$$\begin{aligned}
& \Pr\Big(\sum_{l=k+1}^{N-1} (U_l - \mathbb{E}[U_l]) < Nc_{d_x} r^{d_x} f_X(x) e^{-\varepsilon} - k - (N-k-1)\mathbb{E}[U_l] \Big) \\
\leq\ & \Pr\Big(\sum_{l=k+1}^{N-1} (U_l - \mathbb{E}[U_l]) < -(N-k-1) c_{d_x} r^{d_x} f_X(x) \varepsilon/4 \Big) \\
\leq\ & \exp\Big\{ - \frac{((N-k-1) c_{d_x} r^{d_x} f_X(x) \varepsilon/4)^2}{2(N-k-1)\mathbb{E}[U_l^2] + \frac{2}{3}((N-k-1) c_{d_x} r^{d_x} f_X(x) \varepsilon/4)} \Big\} \\
=\ & \exp\Big\{ - \frac{\varepsilon^2}{32(1+\frac{7\varepsilon}{12})} (N-k-1) c_{d_x} r^{d_x} f_X(x) \Big\}. \qquad (3.74)
\end{aligned}$$

Therefore, $I_3(z)$ is upper bounded by:

$$\begin{aligned}
& I_3(z) = \int_{r=(\log N)^2 (N f_X(x) c_{d_x})^{-\frac{1}{d_x}}}^{\log N (N f(z) c_{d_x+d_y})^{-\frac{1}{d_x+d_y}}} \Pr\Big(\Big| \log \widehat{f}_X(X_i) - \log f_X(X_i) \Big| > \varepsilon \\
& \Big| \rho_{k,i} = r, Z_i = z \Big) f_{\rho_{k,i}}(r) dr \\
\leq\ & \int_{r=(\log N)^2 (N f_X(x) c_{d_x})^{-\frac{1}{d_x}}}^{\log N (N f(z) c_{d_x+d_y})^{-\frac{1}{d_x+d_y}}} 2 \exp\Big\{ - \frac{\varepsilon^2 (N-k-1)}{32(1+\frac{7\varepsilon}{12})} c_{d_x} r^{d_x} f_X(x) \Big\} f_{\rho_{k,i}}(r) dr \\
\leq\ & 2 \exp\Big\{ - \frac{\varepsilon^2}{64} N c_{d_x} f_X(x) \Big(\frac{(\log N)^2}{(N f_X(x) c_{d_x})^{1/d_x}} \Big)^{d_x} \Big\} \\
\leq\ & 2 \exp\Big\{ - \frac{\varepsilon^2}{64} (\log N)^{2d_x} \Big\}. \qquad (3.75)
\end{aligned}$$

for sufficiently large N such that $(N-k-1)/(1+\frac{7}{12}\varepsilon) > N/2$ and any

$d_x \geq 1$. The upper bounds of $I_1(z)$, $I_2(z)$ and $I_3(z)$ are all independent of z. Therefore, combine the upper bounds of $I_1(z)$, $I_2(z)$ and $I_3(z)$, we obtain

$$\begin{aligned}
& \Pr\left(\frac{1}{N}\sum_{i=1}^{N} \left| \log \widehat{f}_X(X_i) - \log f_X(X_i) \right| > \varepsilon \right) \\
\leq\ & N \int (I_1(z) + I_2(z) + I_3(z)) f(z) dz \\
=\ & kN^k \exp\Big\{ -\frac{(\log N)^{d_x+d_y}}{4} \Big\} + \Big(4C' \frac{c_{d_x+d_y}}{c_{d_x}} \Big)^k (\log N)^{2k(d_x+d_y)} N^{1-\frac{k\,d_y}{d_x}} \\
& + 2N \exp\Big\{ -\frac{\varepsilon^2}{64} (\log N)^{2d_x} \Big\}. \qquad (3.76)
\end{aligned}$$

If $k > d_y/d_x$ as per our assumption, each of the three terms goes to 0 as $N \to \infty$.

Therefore, by combining the convergence of error from sampling and error from density estimation, we obtain that $\widehat{H}(X)$ converges to $H(X)$ in probability.

3.6 Proof of Theorem 4 on the bias of KSG estimator

We will introduce some notations first. Let $Z = (X, Y)$, $f(x) = f(x, y)$ and $d = d_x + d_y$ for short. Let $B(z, r)$ denote the d-dimensional ball centered at z with radius r, $B_X(x, r)$ denote the d_x-dimensional ball (on X space) centered at x with radius r. $P(z, r)$ denotes the probability mass inside $B(z, r)$, i.e., $P(z, r) = \int_{B(z,r)} f(t)dt$. Similarly, $P_X(x, r) = \int_{B_X(x,r)} f_X(t)dt$ denotes the probability mass inside $B_X(z, r)$. Now note that if $\rho_{k,i,\cdot} \leq a_N$, we can write $\iota_{k,i,2}$ and $\iota_{k,i,\infty}$ as:

$$\begin{aligned}
\iota_{k,i,\infty} &= \xi_{k,i,\infty}(X) + \xi_{k,i,\infty}(Y) - \xi_{k,i,\infty}(Z), && (3.77) \\
\iota_{k,i,2} &= \xi_{k,i,2}(X) + \xi_{k,i,2}(Y) - \xi_{k,i,2}(Z), && (3.78)
\end{aligned}$$

where

$$\begin{aligned}
\xi_{k,i,\infty}(Z) &\equiv -\psi(k) + \log N + \log c_{d_x,\infty} c_{d_y,\infty} + d \log \rho_{k,i,\infty}, && (3.79) \\
\xi_{k,i,\infty}(X) &\equiv -\psi(n_{x,i,\infty} + 1) + \log N + \log c_{d_x,\infty} + d_x \log \rho_{k,i,\infty}, && (3.80) \\
\xi_{k,i,\infty}(Y) &\equiv -\psi(n_{y,i,\infty} + 1) + \log N + \log c_{d_y,\infty} + d_y \log \rho_{k,i,\infty}, && (3.81)
\end{aligned}$$

and

$$\begin{aligned}
\xi_{k,i,2}(Z) &\equiv -\psi(k) + \log N + \log c_{d,2} + d\,\log \rho_{k,i,2}, && (3.82)\\
\xi_{k,i,2}(X) &\equiv -\log(n_{x,i,2}) + \log N + \log c_{d_x,2} + d_x \log \rho_{k,i,2}, && (3.83)\\
\xi_{k,i,2}(Y) &\equiv -\log(n_{y,i,2}) + \log N + \log c_{d_y,2} + d_y \log \rho_{k,i,2}. && (3.84)
\end{aligned}$$

If $\rho_{k,i,\cdot} > a_N$, just define $\xi_{k,i,\cdot}(X) = \xi_{k,i,\cdot}(Y) = \xi_{k,i,\cdot}(Z) = 0$. Similar as the proof of Theorem 3, we drop the superscript KSG or BI-KSG and subscript 2 and ∞ for statements that holds for both. Since $\iota_{k,i}$'s are identically distributed, we have $\mathbb{E}[\widehat{I}(X;Y)] = \mathbb{E}[\iota_{k,1}]$. By triangular inequality, the bias of $\widehat{I}(X;Y)$ can be written as:

$$\begin{aligned}
&\mathbb{E}\left[\widehat{I}(X;Y)\right] - I(X;Y) = \mathbb{E}\left[\iota_{k,1}\right] - I(X;Y)\\
\leq\ & |\,\mathbb{E}\left[\xi_{k,1}(X)\right] - H(X)\,| + |\,\mathbb{E}\left[\xi_{k,1}(Y)\right] - H(Y)\,| + |\,\mathbb{E}\left[\xi_{k,1}(Z)\right] - H(Z)\,|\\
\leq\ & |\,\mathbb{E}\left[\,(\,\xi_{k,1}(X) - H(X)\,)\cdot \mathbb{I}\{\rho_{k,1} \leq a_N\}\,\right]\,|\\
&+ |\,\mathbb{E}\left[\,(\,\xi_{k,1}(X) - H(X)\,)\cdot \mathbb{I}\{\rho_{k,1} > a_N\}\,\right]\,|\\
&+ |\,\mathbb{E}\left[\,(\,\xi_{k,1}(Y) - H(Y)\,)\cdot \mathbb{I}\{\rho_{k,1} \leq a_N\}\,\right]\,|\\
&+ |\,\mathbb{E}\left[\,(\,\xi_{k,1}(Y) - H(Y)\,)\cdot \mathbb{I}\{\rho_{k,1} > a_N\}\,\right]\,|\\
&+ |\,\mathbb{E}\left[\,\xi_{k,1}(Z)\cdot \mathbb{I}\{\rho_{k,1} \leq a_N\} - H(Z)\,\right]\,| + |\,\mathbb{E}\left[\,\xi_{k,1}(Z)\cdot \mathbb{I}\{\rho_{k,1} > a_N\}\,\right]\,|\\
=\ & |\,\mathbb{E}\left[\,(\,\xi_{k,1}(X) - H(X)\,)\cdot \mathbb{I}\{\rho_{k,1} \leq a_N\}\,\right]\,|\\
&+ |\,\mathbb{E}\left[\,(\,\xi_{k,1}(Y) - H(Y)\,)\cdot \mathbb{I}\{\rho_{k,1} \leq a_N\}\,\right]\,|\\
&+ |\,\mathbb{E}\left[\,\xi_{k,1}(Z)\cdot \mathbb{I}\{\rho_{k,1} \leq a_N\} - H(Z)\,\right]\,|\\
&+ (\,|H(X)| + |H(Y)|\,)\,\Pr\left(\,\rho_{k,i} > a_N\,\right). && (3.85)
\end{aligned}$$

First we consider the bias of $\xi_{k,i}(Z)$, which is local d-dimensional Kozachenko-Leonenko entropy estimator [2] . The following lemma gives the convergence rate of truncated KL estimator $\xi_{k,1}(Z)$.

Lemma 6. *Under the Assumption 2.*$(a)-(d)$ *and* (g)*,*

$$\begin{aligned}
&\mathbb{E}\left[\,\xi_{k,1}(Z)\cdot \mathbb{I}\{\rho_{k,1} \leq a_N\} - H(Z)\,\right]\\
\leq\ & O\left(\frac{(\log N)^{(1+\delta)(1+1/d)}}{N^{1/d}}\right), && (3.86)
\end{aligned}$$

Now we consider the bias of $\xi_{k,i}(X)$ and $\xi_{k,i}(Y)$ when $\rho_{k,i} \leq a_N$. The fol-

lowing lemma establishes the convergence rate for marginal entropy estimator $\xi_{k,1}(X)$.

Lemma 7. *Under the Assumption 2.$(c)-(e)$, the bias of marginal entropy estimator $\xi_{k,1}(X)$ is given by:*

$$\begin{aligned} & \mathbb{E}\left[\left(\xi_{k,1}(X) - H(X)\right) \cdot \mathbb{I}\{\rho_{k,1} \leq a_N\}\right] \\ = \; & O\left(\frac{(\log N)^{(1+\delta)(1+1/d)}}{N^{1/d}}\right), \end{aligned} \tag{3.87}$$

for $k \geq d_x/d_y$.

The probability that $\rho_{k,i} > a_N$ is bounded by the following lemma:

Lemma 8. *Under the Assumption 2.(c) and (d), we have:*

$$\begin{aligned} & \Pr(\rho_{k,i} > a_N) \\ \leq \; & C\left(N^{k-1}\exp\{-C(\log N)^{1+\delta}\} + \left(\frac{(\log N)^{1+\delta}}{N}\right)^{1/d}\right). \end{aligned} \tag{3.88}$$

Convergence rate of $\xi_{k,1}(Y)$ is immediate by exchanging X and Y and $k \geq d_x/d_y$. Combining Lemma 6, Lemma 7 and Lemma 8, we obtain the desired statement.

3.6.1 Proof of Lemma 6

We follow closely the proof from [32] of the $\sqrt{N}$-consistency of the one-dimensional entropy estimator introduced in [2]. It was proved in [32] that the KL entropy estimator achieves $\sqrt{N}$-consistency in mean, i.e. $\mathbb{E}[\widehat{H}(Z)] - H(Z) = O(1/\sqrt{N})$,and in variance, i.e. $\mathbb{E}[(\widehat{H}(Z) - \mathbb{E}[\widehat{H}(Z)])^2] = O(1/N)$, under the assumption that the Z is a one-dimensional random variable and the estimator uses only the nearest neighbor distance with $k = 1$. In the process of proving our main result, we prove a generalization of this rate of convergence of the KL entropy estimator for general d-dimensional space and for a general k. Also notice that our proof works for any choice of ℓ_p distance for $1 \leq p \leq \infty$, so we will drop the subscribe p in the proof.

Firstly, we notice that $\xi_{k,i}(Z)$ are identically distributed and $\xi_{k,i} = 0$ if

$\rho_{k,i} > a_N$, so we have:

$$\begin{aligned} \mathbb{E}\left[\,\widehat{H}_{\mathrm{tKL}}(X)\,\right] &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\,\xi_{k,i}(Z)\,\right] = \mathbb{E}\left[\,\xi_{k,1}(Z)\,\right] \\ &= \mathbb{E}\left[\,\xi_{k,1}(Z)\cdot\mathbb{I}\{\rho_{k,i}\le a_N\}\,\right]. \end{aligned} \tag{3.89}$$

We introduce the following notations. Let

$$b_N = e^{-\psi(k)}Nc_d a_N^d = e^{-\psi(k)}c_d(\log N)^{1+\delta}, \tag{3.90}$$

and for every $u > 0$ define

$$r_N(u) = \left(\frac{ue^{\psi(k)}}{c_d N}\right)^{1/d}, \tag{3.91}$$

such that $r_N(e^{\xi_{k,1}(Z)}) = \rho_{k,1}$ for $\rho_{k,1} \le a_N$ and $r_N(b_N) = a_N$. It is easy to check that $\frac{dr_N(u)}{du} = \frac{r_N(u)}{ud}$. These definitions provides a new representation of the expectation in (3.89) using a change of variables $u = r_N^{-1}(\rho_{k,1})$:

$$\begin{aligned} &\mathbb{E}\left[\,\xi_{k,1}(Z)\cdot\mathbb{I}\{\rho_{k,1}\le a_N\}\,\right] = \mathbb{E}\left[\,\log u\cdot\mathbb{I}\{u\le b_N\}\,\right] \\ =\ &\int\left(\int_0^{b_N}\log u\, dF_{N,z}(u)\right)f(z)dx, \end{aligned} \tag{3.92}$$

where we define the following distribution:

$$\begin{aligned} F_{N,z}(u) &= \Pr\left(e^{\xi_{k,1}(Z)} < u \middle| Z_1 = z\right) \\ &= \Pr\left(\rho_{k,1} < r_N(u) \middle| Z_1 = z\right). \end{aligned} \tag{3.93}$$

Similar change of variables holds for the actual entropy as follows.

Lemma 9. *The entropy $H(Z)$ can be rewritten as*

$$H(Z) = \int\left(\int_0^{\infty}\log u dF_z(u)\right)f(z)dx, \tag{3.94}$$

where $F_Z(u)$ is defined as

$$F_z(u) = 1 - \exp\{-ue^{\psi(k)}f(z)\}\sum_{j=0}^{k-1}\frac{(ue^{\psi(k)}f(z))^j}{j!}. \tag{3.95}$$

This allows us to decompose the bias into three terms, each of which can be bounded separately.

$$\begin{aligned} & \left| \mathbb{E}\left[\widehat{H}_{\mathrm{tKL}}(X) \right] - H(Z) \right| = \left| \mathbb{E}\left[\xi_{k,1}(Z) \cdot \mathbb{I}\{\rho_{k,1} \le a_N\} - H(Z) \right] \right| \\ \le & \int \left(I_1(z) + I_2(z) + I_3(z) \right) f(z) dx, \end{aligned} \tag{3.96}$$

where

$$I_1(z) = \left| \int_{b_N}^{\infty} \log u \, dF_z(u) \right|, \tag{3.97}$$

$$I_2(z) = \left| \int_0^1 \log u \, dF_{N,z}(u) - \int_0^1 \log u \, dF_z(u) \right|, \tag{3.98}$$

$$I_3(z) = \left| \int_1^{b_N} \log u \, dF_{N,z}(u) - \int_1^{b_N} \log u \, dF_z(u) \right|. \tag{3.99}$$

We will bound the three terms separately. The main idea is that $I_1(z)$ is small when b_N is sufficiently large, and $I_2(z)$ and $I_3(z)$ are small when $f_z(u)$ and $f_{N,z}(u)$ are close.

$I_1(z)$: We upper bound the tail probability that the k-NN distance is truncated. By plugging in the CDF (3.95) of $F_z(u)$, we get:

$$\begin{aligned} & I_1(z) = \left| \int_{b_N}^{\infty} \log u dF_z(u) \right| = \left| \int_{b_N}^{\infty} \log u \frac{dF_z(u)}{du} du \right| \\ = & \frac{1}{(k-1)!} \left| \int_{b_N}^{\infty} (\log u) \, e^{\psi(k)} f(z) \exp\{-u e^{\psi(k)} f(z)\} (u e^{\psi(k)} f(z))^{k-1} du \right| \\ = & \frac{1}{(k-1)!} \left| \int_{b_N e^{\psi(k)} f(z)}^{\infty} (\log t - \psi(k) - \log f(z)) \, e^{-t} t^{k-1} dt \right|, \end{aligned} \tag{3.100}$$

where the third equality is from (3.150) in later section and the last equality comes from changing of variable $t = u e^{\psi(k)} f(z)$. Now we consider two cases:

1. $b_N e^{\psi(k)} f(z) < 1$. Then (3.100) is upper bounded by:

$$\begin{aligned}
& \frac{1}{(k-1)!} \left| \int_{b_N e^{\psi(k)} f(z)}^{\infty} (\log t - \psi(k) - \log f(z))\, e^{-t} t^{k-1} dt \right| \\
\leq\ & \frac{1}{(k-1)!} \Big(\left| \int_{b_N e^{\psi(k)} f(z)}^{\infty} \log t\, e^{-t} t^{k-1} dt \right| \\
& + |\psi(k) + \log f(z)| \left| \int_{b_N e^{\psi(k)} f(z)}^{\infty} e^{-t} t^{k-1} dt \right| \Big) \\
\leq\ & \frac{1}{(k-1)!} \left(\int_0^{\infty} |\log t|\, e^{-t} t^{k-1} dt + |\psi(k) + \log f(z)| \int_0^{\infty} e^{-t} t^{k-1} dt \right) \\
\leq\ & C_1 (1 + |\psi(k) + \log f(z)|), \qquad (3.101)
\end{aligned}$$

where $C_1 = \max \Big\{ \frac{1}{(k-1)!} \int_0^{\infty} |\log t|\, e^{-t} t^{k-1} dt, \frac{1}{(k-1)!} \int_0^{\infty} e^{-t} t^{k-1} dt \Big\}$.

2. $b_N e^{\psi(k)} f(z) \geq 1$. Then (3.100) is upper bounded by:

$$\begin{aligned}
& \frac{1}{(k-1)!} \left| \int_{b_N e^{\psi(k)} f(z)}^{\infty} (\, \log t - \psi(k) - \log f(z)\,)\ e^{-t} t^{k-1} dt \right| \\
\leq\ & \frac{1}{(k-1)!} \Big(\left| \int_{b_N e^{\psi(k)} f(z)}^{\infty} \log t\, e^{-t} t^{k-1} dt \right| \\
& + |\psi(k) + \log f(z)| \left| \int_{b_N e^{\psi(k)} f(z)}^{\infty} e^{-t} t^{k-1} dt \right| \Big) \\
\leq\ & C_2 (1 + |\psi(k) + \log f(z)|) \int_{b_N e^{\psi(k)} f(z)}^{\infty} e^{-t/2} dt \\
\leq\ & 2C_2 (1 + |\psi(k) + \log f(z)|) \exp\{-b_N e^{\psi(k)} f(z)\}, \qquad (3.102)
\end{aligned}$$

where C_2 is a constant satisfying $\log t \cdot t^{k-1}/(k-1)! < C_2 e^{t/2}$ and $t^{k-1}/(k-1)! < C_2 e^{t/2}$ for all $t > 1$.

Now combining the two cases, $I_1(z)$ is bounded by:

$$\begin{aligned}
& I_1(z) \leq (1 + |\psi(k) + \log f(z)|) \Big(C_1\, \mathbb{I}\{b_N e^{\psi(k)} f(z) < 1\} \\
& + 2C_2 \exp\{-b_N e^{\psi(k)} f(z)\} \Big) \\
\leq\ & C_3 (1 + |\log f(z)|)\, \exp\{-b_N e^{\psi(k)} f(z)\}, \qquad (3.103)
\end{aligned}$$

where we use the fact that $\mathbb{I}\{b_N e^{\psi(k)} f(z) < 1\} \leq \exp\{1 - b_N e^{\psi(k)} f(z)\}$. Here $C_3 = (C_1 e + 2C_2)(1 + |\psi(k)|)$.

$I_2(z)$: $I_2(z)$ can be bounded by:

$$\begin{aligned} I_2(z) &= \left| \int_0^1 \log u \, dF_{N,z}(u) - \int_0^1 \log u \, dF_z(u) \right| \\ &\leq \int_0^1 |\log u| \, |f_{N,z}(u) - f_z(u)| \, du, \end{aligned} \tag{3.104}$$

where $f_{N,z}(u)$ and $f_z(u)$ are the corresponding pdfs of $F_{N,z}(u)$ and $F_z(u)$, respectively. Here we partition the support into two parts. Let

$$S_1 = \{z : \|H_f(z')\|_2 < C_d, \forall z' \in B(z, a_N)\}, \tag{3.105}$$

$$S_2 = \{z : \|H_f(z')\|_2 \geq C_d \text{ for some } z' \in B(z, a_N)\} = S_1^C. \tag{3.106}$$

From Assumption 2.(g), the $(d-1)$-dimensional Hausdorff measure of the set that $\|H_f(z)\| \geq C_d$ is finite, so the Lebegue measure of S_2 is bounded by $2a_N C_e$ for sufficiently large N. For points in S_1 and S_2, In the following lemma we give an upper bound for the difference of $f_{N,z}(u)$ and $f_z(u)$ for x in S_1 and S_2 separately.

Lemma 10. *Under the Assumption 2.(a) and (d), for any $x \in S_1$,*

$$|f_{N,z}(u) - f_z(u)| \leq C_4 \left(N^{-2/d} + N^{-1} \right), \tag{3.107}$$

for $u \leq 1$. For $x \in S_2$, we have

$$|f_{N,z}(u) - f_z(u)| \leq C_4, \tag{3.108}$$

for $u \leq 1$.

Using Lemma 10 and the fact that $\int_0^1 |\log u| du = 1$, $I_2(z)$ is bounded by:

$$\begin{aligned} I_2(z) &\leq C_4(N^{-2/d} + N^{-1}) \int_0^1 |\log u| du \\ &\leq C_4(N^{-2/d} + N^{-1}), \end{aligned} \tag{3.109}$$

for $x \in S_1$ and

$$I_2(z) \leq C_4 \int_0^1 |\log u| du \leq C_4, \tag{3.110}$$

for $x \in S_2$.

$I_3(z)$: $I_3(z)$ can be bounded by:

$$\begin{aligned} I_3(z) &= \left| \int_1^{b_N} \log u \, dF_{N,z}(u) - \int_1^{b_N} \log u \, dF_z(u) \right| \\ &= \left| \int_1^{b_N} \frac{1}{u}(1 - F_{N,z}(u))du - \int_1^{b_N} \frac{1}{u}(1 - F_x(u))du \right| \\ &\leq \int_1^{b_N} \frac{1}{u}|F_{N,z}(u) - F_z(u)|du. \end{aligned} \tag{3.111}$$

In the following lemma we give an upper bound for the difference of $F_{N,z}(u)$ and $F_z(u)$ for x in S_1 and S_2 separately.

Lemma 11. *Under the Assumption 2.(a) and (d),*

$$|F_{N,z}(u) - F_z(u)| \leq C_5 \left(u^{1+2/d} N^{-2/d} + u^2/N \right), \tag{3.112}$$

for $x \in S_1$ and

$$|F_{N,z}(u) - F_z(u)| \leq C_5 \left(u + u^2/N \right), \tag{3.113}$$

for $x \in S_2$.

Using Lemma 11, $I_3(z)$ is upper bounded by:

$$\begin{aligned} I_3(z) &\leq C_5 \int_1^{b_N} \Big((u/N)^{2/d} + u/N \Big) du \\ &\leq C_5 \left(b_N^{1+2/d} N^{-2/d} + b_N^2 N^{-1} \right), \end{aligned} \tag{3.114}$$

for $x \in S_1$ and

$$I_3(z) \leq C_5 \int_1^{b_N} \Big(1 + u/N \Big) du \leq C_5 \left(b_N + b_N^2 N^{-1} \right), \tag{3.115}$$

for $x \in S_2$

Combining the upper bounds of $I_1(z)$, $I_2(z)$ and $I_3(z)$ and defining $C_6 =$

$\max\{C_3, C_4, C_5\}$, the bias is bounded by:

$$
\begin{aligned}
& \mathbb{E}\left[\widehat{H}_{\mathrm{tKL}}(Z)\right] - H(Z) \\
\leq & \int \left(I_1(z) + I_2(z) + I_3(z)\right) f(z)dx \\
\leq & \int I_1(z) f(z)dx + \int_{S_1} \left(I_2(z) + I_3(z) \right) f(z)dx + \int_{S_2} \left(I_2(z) + I_3(z) \right) f(z)dx \\
\leq & \; C_6 \int \left(|1 + \log f(z)| \exp\{-b_N e^{\psi(k)} f(z)\} \right) f(z)dx \\
& + \int_{S_1} \left(N^{-2/d} + N^{-1} + b_N^{1+2/d} N^{-2/d} + b_N^2 N^{-1} \right) f(z)dx \\
& + \int_{S_2} \left(1 + b_N + b_N^2 N^{-1} \right) f(z)dx \\
\leq & \; C_6 \Big(\int f(z) \exp\{-b_N e^{\psi(k)} f(z)\} + \int f(z) |\log f(z)| \exp\{-b_N e^{\psi(k)} f(z)\} \\
& + b_N^{1+2/d} N^{-2/d} + b_N^2 N^{-1} + b_N \Big(\int_{S_2} f(z)dx \Big) \Big). \qquad (3.116)
\end{aligned}
$$

By Assumption 2.(c), the first term is bounded by: $\int f(z) \exp\{-b_N e^{\psi(k)} f(z)\} \leq C_d e^{-b_N C_0}$. The second term is bounded by Hölder inequality as:

$$
\begin{aligned}
& \int f(z) |\log f(z)| \exp\{-b_N e^{\psi(k)} f(z)\} \\
\leq & \left(\int f(z) (\log f(z))^{1+\gamma} dx \right)^{1/(1+\gamma)} \left(\int f(z) e^{-\frac{1+\gamma}{\gamma} b_N e^{\psi(k)} f(z)} dx \right)^{\gamma/(1+\gamma)} \\
\leq & \; C_b^{1/(1+\gamma)} \left(C_d e^{-\frac{1+\gamma}{\gamma} C_0 b_N} \right)^{\gamma/(1+\gamma)}. \qquad (3.117)
\end{aligned}
$$

By choosing $b_N = e^{-\psi(k)} c_d (\log N)^{1+\delta}$ for some $\delta > 0$, we know that $e^{-C_0 b_N}$ decays faster than $N^{-\alpha}$ for any α.

The last term is bounded by $b_N \left(\int_{S_2} f(z)dx \right) \leq b_N C_a m(S_2) \leq 2 b_N a_N C_a C_e$, where $m(S_2)$ is the Lebesgue measure of S_2. Recall that we choose $a_N = ((\log N)^{1+\delta}/N)^{1/d}$, so the proof is complete.

3.6.2 Proof of Lemma 7

Define $r_N = (\log N)^2 N^{-1/d_x}$, we can split the bias of $\xi_{k,i}(X)$ into two parts:

$$\begin{aligned} & |\,\mathbb{E}\,[\,(\,\xi_{k,1}(X) - H(X)\,)\cdot \mathbb{I}\{\rho_{k,1} \le a_N\}\,]\,| \\ \le\ & |\,\mathbb{E}\,[\,(\,\xi_{k,1}(X) - H(X)\,)\cdot \mathbb{I}\{\rho_{k,1} < r_N\}\,]\,| \\ & +|\,\mathbb{E}\,[\,(\,\xi_{k,1}(X) - H(X)\,)\cdot \mathbb{I}\{r_N \le \rho_{k,1} \le a_N\}\,]\,|\,. \end{aligned} \tag{3.118}$$

If $\rho_{k,1} < r_N$, recall that $\xi_{k,1}(X) = -h(n_{x,1}) + \log c_{d_x} + \log N + d_x \log \rho_{k,1}$, where $h(x) = \log(x)$ or $\psi(x+1)$. Notice that $k < n_{x,1} < N$, so $0 \le h(n_{x,1}) \le 2\log N$. Therefore, we can bound the first term of (3.118) by:

$$\begin{aligned} & \mathbb{E}\,[\,(\,\xi_{k,1}(X) - H(X)\,)\cdot \mathbb{I}\{\rho_{k,1} < r_N\}\,] \\ \le\ & \mathbb{E}\,[\,(\,\log N + \log c_{d_x} + d_x \log \rho_{k,1} - H(X)\,)\cdot \mathbb{I}\{\rho_{k,1} < r_N\}\,] \\ \le\ & (\,\log N + \log c_{d_x} - H(X)\,)\,\Pr(\,\rho_{k,1} < r_N\,) \\ & + d_x \int_0^{r_N} \log r\, f_{\rho_{k,1}}(r) dr, \end{aligned} \tag{3.119}$$

where $f_{\rho_{k,1}}(r)$ is the pdf of $\rho_{k,1}$. Similarly, it can be lower bounded by:

$$\begin{aligned} & \mathbb{E}\,[\,(\,\xi_{k,1}(X) - H(X)\,)\cdot \mathbb{I}\{\rho_{k,1} < r_N\}\,] \\ \ge\ & \mathbb{E}\,[\,(\,-\log N + \log c_{d_x} + d_x \log \rho_{k,1} - H(X)\,)\cdot \mathbb{I}\{\rho_{k,1} < r_N\}\,] \\ \ge\ & (\,-\log N + \log c_{d_x} - H(X)\,)\,\Pr(\,\rho_{k,1} < r_N\,) \\ & + d_x \int_0^{r_N} \log r\, f_{\rho_{k,1}}(r) dr. \end{aligned} \tag{3.120}$$

Therefore, we obtain:

$$\begin{aligned} & |\,\mathbb{E}\,[\,(\,\xi_{k,1}(X) - H(X)\,)\cdot \mathbb{I}\{\rho_{k,1} < r_N\}\,]\,| \\ \le\ & (\,\log N + |\,\log c_{d_x} - H(X)\,|\,)\,\Pr(\,\rho_{k,1} < r_N\,) \\ & + d_x \int_0^{r_N} |\log r| f_{\rho_{k,1}}(r) dr. \end{aligned} \tag{3.121}$$

Now we will given an upper bound on the probability $\Pr(\,\rho_{k,1} < r\,)$ for any $r \le r_N$. Given that $Z_1 = z$, Let p_r be the probability inside the ℓ_p ball

centered at $Z_1 = z = (x, y)$ with radius r. For sufficiently large N, we have

$$p_r \;=\; \Pr(\, u \in B_Z(z,r)\,) \le \left(\sup_{t \in B_Z(z,r)} f(t) \right) c_d r^d \le C_a c_d r^d. \quad (3.122)$$

Therefore, $\Pr(\, \rho_{k,1} < r \mid Z_1 = z\,)$ is upper bounded by:

$$\begin{aligned} &\Pr(\, \rho_{k,1} < r \mid Z_1 = z\,) = \sum_{m=k}^{N-1} \tbinom{N-1}{m} p_r^m (1-p_r)^{N-1-m} \\ &\le \sum_{m=k}^{N-1} N^m p_r^m \le \sum_{m=k}^{N-1} (N C_a c_d r^d)^m \le 2(N C_a c_d r^d)^k. \end{aligned} \quad (3.123)$$

Recall that $r \le r_N = (\log N)^2 N^{-1/d_x}$, so for sufficiently large N, we have $N C_a c_d r^d \le 1/2$, which gives us the last inequality. Notice that this probability is independent of z, therefore, we have $\Pr(\, \rho_{k,1} < r\,) \le 2(N C_a c_d r^d)^k$. Plugging in $r_N = (\log N)^2 N^{-1/d_x}$, we obtain:

$$\begin{aligned} &\Pr(\, \rho_{k,1} < r_N\,) \le 2(N C_a c_d (\log N)^{2d} N^{-d/(d_x)})^k \\ =\;& 2 C_a^k c_d^k (\log N)^{2kd} N^{-k d_y / d_x}. \end{aligned} \quad (3.124)$$

Let $F_{\rho_{k,1}}(r)$ be the CDF of $\rho_{k,1}$ and $F_0(r) = 2(N C_a c_d r^d)^k$ be the upper bound for $F_{\rho_{k,1}}(r)$. Then using integration by parts, the integral $\int_0^{r_N} |\log r| f_{\rho_{k,1}}(r) dr$

can be bounded by:

$$\begin{aligned}
&\int_0^{r_N} |\log r| f_{\rho_{k,1}}(r)dr = \int_0^{r_N} (-\log r) dF_{\rho_{k,i}}(r) \\
= \;& -\log(r_N)F_{\rho_{k,i}}(r_N) + \lim_{r\to 0}\big(\log(r)F_{\rho_{k,i}(r)}\big) - \int_0^{r_N}(-\frac{F_{\rho_{k,i}}(r)}{r})dr \\
\leq \;& -\log(r_N)F_0(r_N) + \int_0^{r_N}\frac{F_0(r)}{r}dr \\
= \;& -2\log(r_N)(NC_ac_dr_N^d)^k + \int_0^{r_N}\frac{2(NC_ac_dr^d)^k}{r}dr \\
= \;& -2\log(r_N)(NC_ac_dr_N^d)^k + \frac{2}{kd}(NC_ac_dr_N^d)^k \\
= \;& \frac{2}{kd}(NC_ac_d)^k r_N^{kd}(1-kd\log(r_N)) \\
= \;& \frac{2}{kd}(NC_ac_d)^k(\log N)^{2kd}N^{-\frac{kd}{d_x}}(1-kd(-\frac{1}{d_x}\log N + 2\log\log N)) \\
= \;& \frac{2(C_ac_d)^k}{kd}(\log N)^{2kd}(1+\frac{kd}{d_x}\log N)N^{-\frac{kd_y}{d_x}}. \qquad (3.125)
\end{aligned}$$

If $k \geq d_x/d_y$, then there exists some constant C such that $\Pr(\rho_{k,1} < r_N) \leq C(\log N)^{2kd}/N$ and $\int_0^{r_N}|\log r| f_{\rho_{k,1}}(r)dr \leq C(\log N)^{2kd+1}/N$. Therefore, plug it in (3.121), we have:

$$\begin{aligned}
&|\,\mathbb{E}\,[\,(\,\xi_{k,1}(X) - H(X)\,)\cdot\mathbb{I}\{\rho_{k,1}\leq r_N\}\,]\,| \\
\leq \;& (\,\log N + |\,\log c_{d_x} - H(X)\,|\,)\,C\frac{(\log N)^{2kd}}{N} + d_xC\frac{(\log N)^{2kd+1}}{N} \\
\leq \;& C(1+d_x)\frac{(\log N)^{2kd+1}}{N} + C\,|\,\log c_{d_x} - H(X)\,|\,\frac{(\log N)^{2kd}}{N} \\
\leq \;& N^{-d_y/d}, \qquad (3.126)
\end{aligned}$$

for sufficiently large N.

Now we consider the second term of (3.118). Recall that

$$\xi_{k,1,2}(X) \equiv \log\left(\frac{c_{d_x,2}N\rho_{k,1,2}^{d_x}}{n_{x,1,2}}\right), \qquad (3.127)$$

$$\xi_{k,1,\infty}(X) \equiv \log\left(\frac{c_{d_x,\infty}N\rho_{k,1,\infty}^{d_x}}{\exp\{\psi(n_{x,1,\infty}+1)\}}\right). \qquad (3.128)$$

Given that $r_N \leq \rho_{k,1,2} \leq a_N$, the bias of $\xi_{k,1,2}(X)$ is upper bounded by:

$$\begin{aligned}
& |\,\mathbb{E}\,[\,(\,\xi_{k,1,2}(X) - H(X)\,)\cdot \mathbb{I}\{r_N \leq \rho_{k,1,2} \leq a_N\}\,]\,| \\
= \;& \Big|\mathbb{E}_{Z,\rho_{k,1,2}}\Big[\,\mathbb{E}_{n_{x,1,2}}\Big[\,\Big(\,\xi_{k,1,2}(X) + \int f_X(x)\log f_X(x)dx\,\Big) \\
& \cdot\mathbb{I}\{r_N \leq \rho_{k,1,2} \leq a_N\}\,\Big|\,Z,\rho_{k,1,2}\,\Big]\,\Big]\,\Big| \\
\leq \;& \int\Big(\int_{r_N}^{a_N}\Big|\,\log\big(\,f_X(x)c_{d_x,2}Nr^{d_x}\,\big) \\
& -\mathbb{E}\,[\,\log(n_{x,1,2})|\rho_{k,1,2}=r,Z_1=z\,]\,\Big|f_{\rho_{k,1,2}}(r)dr\Big)f(z)dz, \quad (3.129)
\end{aligned}$$

where we applied the Jensen's inequality. By noticing that $\log(x) < \psi(x+1) < \log(x+1)$ for any integer $x \geq 2$, we have $|\psi(x+1)-y| \leq \max_{\theta\in\{0,1\}} |\log(x+\theta)-y|$. So the bias of $\xi_{k,1,\infty}$ is upper bounded by:

$$\begin{aligned}
& \mathbb{E}\,[\,(\,\xi_{k,1,\infty}(X) - H(X)\,)\cdot \mathbb{I}\{r_N \leq \rho_{k,1,\infty} \leq a_N\}\,] \\
= \;& \Big|\mathbb{E}_{Z,\rho_{k,1,\infty}}\Big[\,\mathbb{E}_{n_{x,1,\infty}}\Big[\,\Big(\,\xi_{k,1,\infty}(X) + \int f_X(x)\log f_X(x)dx\,\Big) \\
& \cdot\mathbb{I}\{r_N \leq \rho_{k,1,\infty} \leq a_N\}\,\Big|\,Z,\rho_{k,1,\infty}\,\Big]\,\Big]\,\Big| \\
\leq \;& \int\Big(\int_{r_N}^{a_N}\Big|\,\mathbb{E}\,[\,\psi(n_{x,1,\infty}+1)|\rho_{k,1,\infty}=r,Z_1=z\,] \\
& -\log\Big(\,f_X(x)c_{d_x,\infty}Nr^{d_x}\,\Big)\,\Big|f_{\rho_{k,1,\infty}}(r)dr\Big)\,f(z)dz. \quad (3.130)
\end{aligned}$$

Combine the arguments for KSG and BI-KSG, we obtain:

$$\begin{aligned}
& \mathbb{E}\,[\,(\,\xi_{k,1}(X) - H(X)\,)\cdot \mathbb{I}\{r_N \leq \rho_{k,1} \leq a_N\}\,] \\
\leq \;& \int\Big(\int_{r_N}^{a_N}\Big|\max_{\theta\in\{0,1\}}\mathbb{E}\,[\,\log(n_{x,1}+\theta)|\rho_{k,1}=r,Z_1=z\,] \\
& -\log\big(\,f_X(x)c_{d_x}Nr^{d_x}\,\big)\,\Big|f_{\rho_{k,1}}(r)dr\Big)f(z)dz. \quad (3.131)
\end{aligned}$$

From now on we drop the subscript 2 or ∞. Now similar to the proof of 6, we divide the support of X into two parts:

$$\begin{aligned}
S_1^{(X)} &= \{x : \|H_{f_x}(x)\| < C_d, \forall x' \in B_X(x,a_N)\}, \quad (3.132)\\
S_2^{(X)} &= \{x : \|H_f(x)\| \geq C_d, \text{ for some } x' \in B_X(x,a_N)\}, \quad (3.133)
\end{aligned}$$

where the Lebesgue measure of $S_2^{(X)}$ is upper bounded by $2C_h a_N$ for suffi-

ciently small a_N. Therefore, we rewrite (3.131) as:

$$\begin{aligned}
& \mathbb{E}\left[\left(\xi_{k,1}(X) - H(X)\right) \cdot \mathbb{I}\{r_N \le \rho_{k,1} \le a_N\}\right] \\
\le\ & \int_{S_1} \Bigg(\int_{r_N}^{a_N} \Big| \max_{\theta \in \{0,1\}} \mathbb{E}\left[\log(n_{x,1} + \theta) | \rho_{k,1} = r, Z_1 = z\right] \\
& - \log\left(f_X(x) c_{d_x} N r^{d_x}\right) \Big| f_{\rho_{k,1}}(r) dr \Bigg) f(z) dz \\
& + \int_{S_2} \Bigg(\int_{r_N}^{a_N} \Big| \max_{\theta \in \{0,1\}} \mathbb{E}\left[\log(n_{x,1} + \theta) | \rho_{k,1} = r, Z_1 = z\right] \\
& - \log\left(f_X(x) c_{d_x} N r^{d_x}\right) \Big| f_{\rho_{k,1}}(r) dr \Bigg) f(z) dz. \qquad (3.134)
\end{aligned}$$

Recall that in Theorem 7, given that $\rho_{k,i} = r$ and $Z_i = z$, $n_{x,i} - k$ is distributed as $\sum_{l=k+1}^{N-1} U_l$, where U_l are i.i.d Bernoulli random variables with mean p satisfying

$$r^{-d_x} \left| p - f_X(x) c_{d_x} r^{d_x} \right| \le C_1 (r^2 + r^{d_y}), \qquad (3.135)$$

for $x \in S_1^{(X)}$. For $x \in S_2^{(X)}$, the Bernoulli property still holds, but the mean p is simply bounded by

$$r^{-d_x} |p - f_X(x) c_{d_x} r^{d_x}| \le r^{-d_x} f_X(x) c_{d_x} r^{d_x} \le C_a c_{d_x}. \qquad (3.136)$$

From now on, we will focus on $x \in S_1^{(X)}$. For $x \in S_2^{(X)}$, the analyses also hold if we replace $C_1(r^2 + r^{d_y})$ by $C_a c_{d_x}$ everywhere. We will skip that for simplicity. For $r > r_N = (\log N)^2 N^{-1/d_x}$, we know that $p \ge f_X(x) c_{d_x} r^{d_x}/2 = f_X(x) c_{d_x} (\log N)^{2d_x}/(2N)$ for sufficiently large N. Therefore, for any $\theta \in \{0, 1\}$, using the Taylor expansion of a logarithm, we obtain:

$$\begin{aligned}
& \mathbb{E}\left[\log(n_{x,1} + \theta) \,|\, \rho_{k,1} = r, Z_1 = z\right] \\
=\ & \log\left(p(N - k - 1) + k + \theta\right) - \frac{1 - p}{2p(N - k - 1)} + O\left(\frac{1}{p^2 (N - k - 1)^2}\right). \qquad (3.137)
\end{aligned}$$

For sufficiently large N, this gives

$$
\begin{aligned}
& \left| \mathbb{E}(\log(n_{x,1}+\theta) \,|\, \rho_{k,1}=r, Z_1=z) - \log\left(f_X(x) c_{d_x} N r^{d_x} \right) \right| \\
\leq\ & \left| \log\left(p(N-k-1)+k+\theta \right) - \log\left(f_X(x) c_{d_x} N r^{d_x} \right) \right| \\
& + \frac{1-p}{2p(N-k-1)} + \frac{C_2}{p^2(N-k-1)^2} \\
\leq\ & \left| \log(pN) - \log\left(f_X(x) c_{d_x} N r^{d_x} \right) \right| + \left| \log \frac{pN}{pN+k(1-p)+\theta-p} \right| \\
& + \frac{1-p}{2p(N-k-1)} + \frac{C_2}{p^2(N-k-1)^2} \\
\leq\ & \left| \log(pN) - \log\left(f_X(x) c_{d_x} N r^{d_x} \right) \right| + \frac{C_3}{pN}. \qquad (3.138)
\end{aligned}
$$

For sufficiently large N we have sufficiently small r such that, from Theorem 7, we get $p > f_X(x) c_{d_x} r^{d_x}/2$. Therefore the first term in (3.138) is bounded by:

$$
\begin{aligned}
& \left| \log(pN) - \log\left(f_X(x) c_{d_x} N r^{d_x} \right) \right| \\
\leq\ & \left| p - f_X(x) c_{d_x} r^{d_x} \right| \left(\frac{1}{2p} + \frac{1}{2 f_X(x) c_{d_x} r^{d_x}} \right) \\
\leq\ & C_1 \left(r^{d_x+2} + r^{d_x+d_y} \right) \frac{3}{2 f_X(x) c_{d_x} r^{d_x}} \leq \frac{3C_1(r^2+r^{d_y})}{2 c_{d_x} f_X(x)}, \qquad (3.139)
\end{aligned}
$$

where we used the fact that $\log x - \log y \leq |x-y|(1/(2x)+1/(2y))$ for any positive x and y and the upper bound on $|p - f_X(x) c_{d_x} r^{d_x}|$ from (3.135). The second term in (3.138) is bounded by $2C_3/(f_X(x) r^{d_x} N)$, which gives, for $C_4 = \max\{3C_1/2c_{d_x}, 2C_3\}$,

$$
\begin{aligned}
& \left| \mathbb{E}(\log(n_{x,1}+\theta) | \rho_{k,1}=r, Z_1=z) - \log\left(f_X(x) c_{d_x} N r^{d_x} \right) \right| \\
\leq\ & \frac{C_4}{f_X(x)} \left(\frac{1}{r^{d_x} N} + r^2 + r^{d_y} \right). \qquad (3.140)
\end{aligned}
$$

To integrate with respect to $\rho_{k,1} = r$, note that $\rho_{k,1}$ is simply the k^{th}-order statistic of $N-1$ i.i.d. random variables $\{ \|Z_2 - z\|, \|Z_3 - z\|, \ldots, \|Z_N - z\| \}$. The corresponding pdf satisfies [121]:

$$
f_{\rho_{k,1}^{(N-1)}}(r) = \frac{N-1}{k-1} f_{\rho_{k-1,1}^{(N-2)}}(r) P(z,r). \qquad (3.141)
$$

For any $\theta \in \{0,1\}$, we have

$$\begin{aligned}
&\int_0^{a_N} \left| \mathbb{E}(\log(n_{x,i}+\theta)|r,z) - \log\left(f_X(x)c_{d_x}Nr^{d_x}\right)\right| f_{\rho_{k,i}^{(N-1)}}(r)dr \\
\leq\ & C_4 \int_0^{a_N} \frac{1}{f_X(x)}\left(\frac{1}{r^{d_x}N} + r^2 + r^{d_y}\right) f_{\rho_{k,i}^{(N-1)}}(r)dr \\
=\ & C_4 \int_0^{a_N} \frac{(N-1)P(z,r)}{(k-1)f_X(x)}\left(\frac{1}{r^{d_x}N} + r^2 + r^{d_y}\right) f_{\rho_{k-1,1}^{(N-2)}}(r)dr \\
\leq\ & C_4 \max_{r\leq a_N} \frac{NP(z,r)}{(k-1)f_X(x)}\left(\frac{1}{r^{d_x}N} + r^2 + r^{d_y}\right). \qquad (3.142)
\end{aligned}$$

By Lemma 12, $|P(z,r) - f(z)c_d r^d| \leq Cr^{d+2}$. Therefore, for sufficiently small a_N, we have $P(z,r) < 2f(z)c_d r^d$ for all $r \leq a_N$. Then we have:

$$\begin{aligned}
&\max_{r\leq a_N} \frac{NP(z,r)}{(k-1)f_X(x)}\left(\frac{1}{r^{d_x}N} + r^2 + r^{d_y}\right) \\
\leq\ & \max_{r\leq a_N} \frac{2f(z)c_d r^d N}{(k-1)f_X(x)}\left(\frac{1}{r^{d_x}N} + r^2 + r^{d_y}\right) \\
=\ & \max_{r\leq a_N} \frac{2c_d f_{Y|X}(y|x)}{k-1}\left(r^{d_y} + Nr^{d+2} + Nr^{d+d_y}\right) \\
\leq\ & C_5\left(a_N^{d_y} + Na_N^{d+2} + Na_N^{d+d_y}\right). \qquad (3.143)
\end{aligned}$$

Since $f_{Y|X}(y|x)$ is upper bounded by C_e, here C_5 is given by $C_5 = 2c_d C_e/(k-1)$. The above upper bound holds for $x \in S_1^{(X)}$, while for $x \in S_2^{(X)}$, we have an upper bound of $C_6(a_N^{d_y} + Na_N^d)$ for some $C_6 > 0$. Now averaging over z, we get:

$$\begin{aligned}
&\mathbb{E}\left[\left(\xi_{k,1}(X) - H(X)\right)\cdot \mathbb{I}\{r_N \leq \rho_{k,1} \leq a_N\}\right] \\
\leq\ & C_4C_5 \int_{S_1} f(z)\left(a_N^{d_y} + Na_N^{d+2} + Na_N^{d+d_y}\right)dz \\
& + C_4C_6 \int_{S_2} f(z)\left(a_N^{d_y} + Na_N^d\right) \\
\leq\ & C_4C_5\left(a_N^{d_y} + Na_N^{d+2} + Na_N^{d+d_y}\right) \\
& + C_aC_4C_6 m(S_2)\left(a_N^{d_y} + Na_N^d\right). \qquad (3.144)
\end{aligned}$$

Here the Lebesgue measure of S_2 is upper bounded by $2C_g a_N$ by Assumption 2.(h). Together with (3.126) and by the choice of a_N in (3.6), the proof is completed.

3.6.3 Proof of Lemma 8

For $Z_1 = z$, the k-NN distance is larger than a_N, i.e. $\rho_{k,1} > a_N$ when at most $k-1$ samples are in $B(z, a_N)$, which gives

$$\Pr\left(\rho_{k,1} > a_N \,\middle|\, Z_1 = z\right) = \sum_{m=0}^{k-1} \binom{N-1}{m} P(z,a_N)^m \left(1 - P(z,a_N)\right)^{N-1-m}. \tag{3.145}$$

Similarly, we divide the support into two parts as follows,

$$\begin{aligned} S_1 &= \{z : \|H_f(z')\| < C_d, \forall z' \in B(z,a_N)\}, \\ S_2 &= \{z : \|H_f(z')\| \geq C_d, \text{ for some } z' \in B(z,a_N)\} = S_1^C. \end{aligned} \tag{3.146}$$

We can see that $\int_{S_2} f(z)dz \leq 2C_a a_N C_g$. For $z \in S_1$, since f is twice continuously differentiable in $B(z,a_N)$ and a_N vanishes as N grows, $f(z)Vol(B(z,a_N))$ approaches $P(z,a_N)$. Precisely, by Lemma 12, for sufficiently large N, we have $P(z,a_N) \geq f(z)c_d a_N^d - C_d a_N^{d+2}$. This provide the following upper bound:

$$\begin{aligned} & \Pr\left(\rho_{k,1} > a_N \,\middle|\, Z_1 = z \in S_1\right) \\ = & \sum_{m=0}^{k-1} \binom{N-1}{m} P(z,a_N)^m \left(1 - P(z,a_N)\right)^{N-1-m} \\ \leq & \sum_{m=0}^{k-1} N^m \left(1 - P(z,a_N)\right)^{N-1-m} \leq kN^{k-1}(1 - P(z,a_N))^{N-k-1} \\ \leq & kN^{k-1} \exp\{-(N-k-1)P(z,a_N)\} \\ \leq & kN^{k-1} \exp\{-(N-k-1)f(z)c_d a_N^d + (N-k-1)C_d a_N^{d+2}\} \\ \leq & kN^{k-1} \exp\{-Cf(z)(\log(N))^{1+\delta}\} \exp\{\log(N)^{(1+\delta)(1+2/d)}/N^{2/d}\} \\ \leq & keN^{k-1} \exp\{-Cf(z)(\log(N))^{1+\delta}\}. \end{aligned} \tag{3.147}$$

The last inequality comes from the fact that $\log(N)^{(1+\delta)(1+2/d)}/N^{2/d} < 1$ for sufficiently large N. For $z \in S_2$, we just use the trivial bound

$$\Pr\left(\rho_{k,1} > a_N \,\middle|\, Z_1 = z \in S_2\right) \leq 1. \tag{3.148}$$

Taking the expectation over Z_1,

$$\begin{aligned}
& \Pr\left(\rho_{k,1} > a_N\right) \\
= & \int_{S_1} f(z) \Pr\left(\rho_{k,i} > a_N \,\middle|\, Z_i = z\right) dz + \int_{S_2} f(z) \Pr\left(\rho_{k,i} > a_N \,\middle|\, Z_i = z\right) dz \\
\leq & keN^{k-1} \int_{S_1} f(z) \exp\{-Cf(z)(\log(N))^{1+\delta}\} dz + \int_{S_2} f(z) dz \\
\leq & keC_c N^{k-1} \exp\{-CC_0(\log(N))^{1+\delta}\} + 2C_a a_N C_g, \qquad (3.149)
\end{aligned}$$

where the last inequality comes from Assumption 2.(c). We complete the proof by plugging in $a_N = ((\log N)^{1+\delta}/N)^{1/d}$.

3.6.4 Proof of Lemma 9

Since $F_z(u)$ is a continuous CDF, the corresponding pdf is given by:

$$\begin{aligned}
& f_z(u) = \frac{dF_z(u)}{du} \\
= & -\exp\{-ue^{\psi(k)} f(z)\} \sum_{j=1}^{k-1} \frac{(ue^{\psi(k)} f(z))^{j-1}}{(j-1)!} e^{\psi(k)} f(z) \\
& + e^{\psi(k)} f(z) \exp\{-ue^{\psi(k)} f(z)\} \sum_{j=0}^{k-1} \frac{(ue^{\psi(k)} f(z))^{j}}{j!} \\
= & \frac{1}{(k-1)!} e^{\psi(k)} f(z) \exp\{-ue^{\psi(k)} f(z)\} (ue^{\psi(k)} f(z))^{k-1}. \qquad (3.150)
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \int_0^\infty \log u dF_z(u) \\
= & \frac{1}{(k-1)!} \int_0^\infty \log u \, e^{\psi(k)} f(z) \exp\{-ue^{\psi(k)} f(z)\} (ue^{\psi(k)} f(z))^{k-1} du \\
= & \frac{1}{(k-1)!} \int_0^\infty (\log t - \psi(k) - \log f(z)) \, e^{-t} t^{k-1} dt \\
= & \psi(k) - \psi(k) - \log f(z) = -\log f(z), \qquad (3.151)
\end{aligned}$$

where the third-to-last equation comes from change of variable $t = ue^{\psi(k)} f(z)$. The last equation comes from $\psi(k) = \frac{1}{(k-1)!} \int_0^\infty (\log t) \, t^{k-1} e^{-t} dt$ and $1 =$

$\frac{1}{(k-1)!}\int_0^\infty t^{k-1}e^{-t}dt$. Therefore,

$$\int\left(\int_0^\infty \log u dF_z(u)\right) f(z)dx = \int\left(-\log f(z)\right) f(z)dx = H(X). \quad (3.152)$$

3.6.5 Proof of Lemma 10

Recall that

$$f_z(u) = \frac{1}{(k-1)!}e^{\psi(k)}f(z)\exp\{-ue^{\psi(k)}f(z)\}(ue^{\psi(k)}f(z))^{k-1}. \quad (3.153)$$

Notice that $r_N(u)$ is the k^{th}-order statistic of $\{\, \|X_1-x\|, \|X_2-x\|, \ldots, \|X_{N-1}-x\| \,\}$. Therefore the density $f_{N,z}(u)$ is given by:

$$\begin{aligned} & f_{N,z}(u) = f_{r_N(u)}\frac{dr_N(u)}{du} \\ = & \frac{(N-1)!}{(k-1)!(N-k-1)!}\left(P(x,r_N(u))\right)^{k-1} \\ & \left(1-P(x,r_N(u))\right)^{N-k-1}\frac{dP(x,r_N(u))}{dr_N(u)}\frac{dr_N(u)}{du} \\ = & \frac{(N-1)!}{(k-1)!(N-k-1)!}\left(P(x,r_N(u))\right)^{k-1} \\ & \left(1-P(x,r_N(u))\right)^{N-k-1}\frac{dP(x,r_N(u))}{du}. \end{aligned} \quad (3.154)$$

Here $P(x,r)(u) = \Pr\{\|X-x\| < r\} = \int_{t\in B(x,r)} f(t)dt$. Since f is twice differentiable and $r_N(u)$ goes to 0 as N goes to infinity, we can use the quantity $f(z)Vol(B(z,r_N(u))$ to estimate $P(z,r_N(u))$. The following lemma bounds the error of this estimation for $x \in S_1$ and $x \in S_2$ separately.

Lemma 12. *Under Assumption 2.(a) amd (d), there exists a constant C such that for sufficiently small r, we have*

$$\left|\, P(x,r) - f(z)c_d r^d \,\right| \;\leq\; Cr^{d+2}, \quad (3.155)$$

and

$$\left|\, \frac{dP(x,r)}{dr} - f(z)dc_d r^{d-1} \,\right| \;\leq\; Cr^{d+1}, \quad (3.156)$$

for $x \in S_1$. For $x \in S_2$, we have

$$\left| P(x,r) - f(z)c_d r^d \right| \;\leq\; Cr^d, \tag{3.157}$$

and

$$\left| \frac{dP(x,r)}{dr} - f(z)dc_d r^{d-1} \right| \;\leq\; Cr^{d-1}. \tag{3.158}$$

Using Lemma 12 and substituting $r = r_N(u) = (ue^{\psi(k)}/(c_d N))^{1/d}$, we have:

$$\begin{aligned} & \left| P(x, r_N(u)) - \frac{ue^{\psi(k)} f(z)}{N} \right| \\ = \;& \left| P(x, r_N(u)) - f(z)c_d (r_N(u))^d \right| \leq C_1 (r_N(u))^{d+2}, \end{aligned} \tag{3.159}$$

for $x \in S_1$. Similarly, $\left| \frac{d}{du} P(x, r_N(u)) - \frac{e^{\psi(k)} f(z)}{N} \right|$ can be bounded by:

$$\begin{aligned} & \left| \frac{d}{du} P(x, r_N(u)) - \frac{e^{\psi(k)} f(z)}{N} \right| \\ = \;& \frac{dr_N(u)}{du} \left| \frac{d}{dr_N(u)} P(x, r_N(u)) - (\frac{dr_N(u)}{du})^{-1} \frac{e^{\psi(k)} f(z)}{N} \right| \\ = \;& \frac{r_N(u)}{u\, d} \left| \frac{d}{dr_N(u)} P(x, r_N(u)) - f(z) d c_d (r_N(u))^{d-1} \right| \\ \leq \;& \frac{C_1 (r_N(u))^{d+2}}{u}, \end{aligned} \tag{3.160}$$

for $x \in S_1$. Analogously we have $\left| P(x, r_N(u)) - \frac{ue^{\psi(k)} f(z)}{N} \right| \leq C_1 (r_N(u))^d$ and $\left| \frac{d}{du} P(x, r_N(u)) - \frac{e^{\psi(k)} f(z)}{N} \right| \leq C_1 (r_N(u))^d / u$ for $x \in S_2$. Now we can write the difference of $f_{N,z}(u)$ and $f_z(u)$ via two terms:

$$|f_{N,z}(u) - f_z(u)| \;\leq\; |f_{N,z}(u) - f^{(1)}_{N,x}(u)| \,+\, |f^{(1)}_{N,x}(u) - f_z(u)|, \tag{3.161}$$

where $f^{(1)}_{N,x}(u)$ defined as:

$$\begin{aligned} f^{(1)}_{N,x}(u) \;=\; & \frac{(N-1)!}{(k-1)!(N-k-1)!} (\frac{ue^{\psi(k)} f(z)}{N})^{k-1} \\ & (1 - \frac{ue^{\psi(k)} f(z)}{N})^{N-k-1} \frac{e^{\psi(k)} f(z)}{N}. \end{aligned} \tag{3.162}$$

Consider the function $g(p) = \frac{(N-1)!}{(k-1)!(N-k-1)!} p^{k-1} (1-p)^{N-k-1}$ for $p \in (0,1)$. By

basic calculus, we can see that $g(p) \le C_2 N$ and $|g'(p)| \le C_3 N^2$ for $p \in (0,1)$. Therefore, the first term in (3.161) can be bounded as:

$$\begin{aligned}
& |f_{N,z}(u) - f^{(1)}_{N,x}(u)| \\
= & \left| g\left(P(x, r_N(u))\right) \frac{dP(x, r_N(u))}{du} - g\left(\frac{ue^{\psi(k)}f(z)}{N}\right) \frac{e^{\psi(k)}f(z)}{N} \right| \\
\le & \; g\left(P(x, r_N(u))\right) \left| \frac{dP(x, r_N(u))}{du} - \frac{e^{\psi(k)}f(z)}{N} \right| \\
& + \left| g\left(P(x, r_N(u))\right) - g\left(\frac{ue^{\psi(k)}f(z)}{N}\right) \right| \frac{e^{\psi(k)}f(z)}{N} \\
\le & \; g\left(P(x, r_N(u))\right) \left| \frac{dP(x, r_N(u))}{du} - \frac{e^{\psi(k)}f(z)}{N} \right| \\
& + \max_{p\in(0,1)} |g'(p)| \left| P(x, r_N(u)) - \frac{ue^{\psi(k)}f(z)}{N} \right| \frac{e^{\psi(k)}f(z)}{N} \\
\le & \; C_1 C_2 N (r_N(u))^{d+2}/u + C_1 C_3 N^2 (r_N(u))^{d+2} \frac{e^{\psi(k)}f(z)}{N} \\
\le & \; C_4 \frac{u^{1+2/d}}{N^{2/d}} (1 + \frac{1}{u}) \le C_4 N^{-2/d}, \qquad (3.163)
\end{aligned}$$

for $u \le 1$ and $x \in S_1$. Here C_4 is the maximum of $C_1 C_2 \left(e^{\psi(k)} C_a\right)^{1+2/d}$ and $C_1 C_3 \left(e^{\psi(k)} C_a\right)^{2+2/d}$, where $C_a = \sup_x f(z)$ by Assumption 2.(a). Similarly, we have $|f_{N,z}(u) - f^{(1)}_{N,z}(u)| \le C_4$ for $u \le 1$ and $x \in S_2$. For the second term, we denote $q = ue^{\psi(k)}f(z)$ for short. Then the second term in (3.161) can be bounded as:

$$\begin{aligned}
& |f^{(1)}_{N,x}(u) - f_z(u)| \\
= & \; \frac{1}{u} \left| \frac{(N-1)!}{(k-1)!(N-k-1)!} \Big(\frac{q}{N}\Big)^k \Big(1 - \frac{q}{N}\Big)^{N-k-1} - \frac{1}{(k-1)!} q^k e^{-q} \right| \\
= & \; \frac{k}{u} \left| \tbinom{N-1}{k} \Big(\frac{q}{N}\Big)^k \Big(1 - \frac{q}{N}\Big)^{N-k-1} - \frac{q^k e^{-q}}{k!} \right|. \qquad (3.164)
\end{aligned}$$

Notice that the difference inside the absolute value is just the difference of $P(X = k)$ under $\text{Bino}(N-1, q/N)$ and $\text{Poisson}(q)$. The difference is bounded by the following lemma.

Lemma 13. *For $q < C\sqrt{N}$, we have:*

$$\left| \tbinom{N-1}{k} \Big(\frac{q}{N}\Big)^k \Big(1 - \frac{q}{N}\Big)^{N-k-1} - \frac{q^k e^{-q}}{k!} \right| \le C_5 q^{k+2} e^{-q} N^{-1}, \qquad (3.165)$$

for some $C_5 > 0$.

Therefore, by Lemma 13, we have:

$$\begin{aligned} |f^{(1)}_{N,x}(u) - f_z(u)| &\leq C_5 \frac{kq^{k+2}e^{-q}}{uN} \\ \leq \; C_5 \frac{k(e^{\psi(k)}f(z))^{k+2}u^{k+1}}{N} &\leq C_6 N^{-1}, \end{aligned} \tag{3.166}$$

for $u \leq 1$, here $C_6 = C_5 k(e^{\psi(k)}C_a)^{k+1}$. Therefore, combining (3.163) and (3.166), we have the desired statement.

3.6.6 Proof of Lemma 11

Recall that

$$F_z(u) = 1 - \exp\{-ue^{\psi(k)}f(z)\} \sum_{j=0}^{k-1} \frac{ue^{\psi(k)}f(z))^j}{j!}. \tag{3.167}$$

The CDF $F_{N,x}(u) = \Pr\left(\, \rho_{k,i} < r_N(u) | X_i = x \,\right)$ is just the probability that at least k samples are inside the ball $B(x, r_N(u))$ and hence

$$\begin{aligned} & F_{N,x}(u) \\ = \; & 1 - \sum_{j=0}^{k-1} \frac{(N-1)!}{j!(N-j-1)!} \left(P(x, r_N(u))\right)^j \left(1 - P(x, r_N(u))\right)^{N-j-1}. \end{aligned} \tag{3.168}$$

So we have:

$$\begin{aligned} & |F_{N,x}(u) - F_z(u)| \\ = \; & \Big| \sum_{j=0}^{k-1} \frac{(N-1)!}{j!(N-j-1)!} \left(P(x, r_N(u))\right)^j \left(1 - P(x, r_N(u))\right)^{N-j-1} \\ & - \exp\{-ue^{\psi(k)}f(z)\} \sum_{j=0}^{k-1} \frac{ue^{\psi(k)}f(z))^j}{j!} \Big| \\ \leq \; & \sum_{j=0}^{k-1} \frac{1}{j!} \Big| \frac{(N-1)!}{(N-j-1)!} \left(P(x, r_N(u))\right)^j \left(1 - P(x, r_N(u))\right)^{N-j-1} \\ & - \exp\{-ue^{\psi(k)}f(z)\}(ue^{\psi(k)}f(z))^j \Big|. \end{aligned} \tag{3.169}$$

Let

$$\begin{aligned} & h_{N,x,j}(u) \\ = & \frac{(N-1)!}{j!(N-j-1)!} \left(P(x, r_N(u))\right)^j \left(1 - P(x, r_N(u))\right)^{N-j-1}, \quad (3.170) \end{aligned}$$

and

$$h_{x,j}(u) = \frac{1}{j!} \exp\{-ue^{\psi(k)} f(z)\} (ue^{\psi(k)} f(z))^j. \quad (3.171)$$

Consider

$$\begin{aligned} & h^{(1)}_{N,x,j}(u) \\ = & \frac{(N-1)!}{j!(N-j-1)!} \left(\frac{ue^{\psi(k)} f(z)}{N}\right)^j \left(1 - \frac{ue^{\psi(k)} f(z)}{N}\right)^{N-j-1}. \quad (3.172) \end{aligned}$$

We will bound $|h_{N,x,j}(u) - h_{x,j}(u)|$ by $|h_{N,x,j}(u) - h^{(1)}_{N,x,j}(u)| + |h^{(1)}_{N,x,j}(u) - h_{x,j}(u)|$. For the first term, consider function $g_j(p) = \frac{(N-1)!}{j!(N-j-1)!} p^j (1-p)^{N-j-1}$. It is easy to see that $|g_j'(p)| \leq C_1 N$ for any $p \in (0,1)$. Therefore, by Lemma 12, we obtain:

$$\begin{aligned} & |h_{N,x,j}(u) - h^{(1)}_{N,x,j}(u)| = \left| g(P(x, r_N(u))) - g(\frac{ue^{\psi(k)} f(z)}{N}) \right| \\ \leq & \max_{p \in (0,1)} |g'(p)| \left| P(x, r_N(u))) - \frac{ue^{\psi(k)} f(z)}{N} \right| \\ \leq & C_1 N (r_N(u))^{d+2} \leq C_2 u^{1+2/d} N^{-2/d}, \quad (3.173) \end{aligned}$$

for $x \in S_1$ and $|h_{N,x,j}(u) - h^{(1)}_{N,x,j}(u)| \leq C_2 u$ for $x \in S_2$, where $C_2 = \max\{C_1 (e^{\psi(k)} C_a)^{1+2/d}, C_1 e^{\psi(k)} C_a\}$. For the second term, let $q = ue^{\psi(k)} f(z)$, and using a similar analysis as (3.166), we obtain:

$$\begin{aligned} & |h^{(1)}_{N,x,j}(u) - h_{x,j}(u)| \\ = & \left| \binom{N-1}{j} (\frac{q}{N})^j (1 - \frac{q}{N})^{N-j-1} - \frac{q^j e^{-q}}{j!} \right| \leq C_3 \frac{q^{j+2} e^{-q}}{N}. \quad (3.174) \end{aligned}$$

Combine (3.173) and (3.174), and we obtain:

$$
\begin{aligned}
& |F_{N,x}(u) - F_x(u)| \leq \sum_{j=0}^{k-1} |h_{N,x,j}(u) - h_{x,j}(u)| \\
\leq & \sum_{j=0}^{k-1} \Big(|h_{N,x,j}(u) - h^{(1)}_{N,x,j}(u)| + |h^{(1)}_{N,x,j}(u) - h_{x,j}(u)| \Big) \\
\leq & kC_2 u^{1+2/d} N^{-2/d} + C_3 \sum_{j=0}^{k-1} \frac{q^{j+2} e^{-q}}{N} \\
\leq & kC_2 u^{1+2/d} (N)^{-2/d} + (k-1)! C_3 q^2 / N \\
\leq & kC_2 u^{1+2/d} (N)^{-2/d} + (k-1)! C_3 (e^{\psi(k)} C_a)^2 u^2 / N, \qquad (3.175)
\end{aligned}
$$

for $x \in S_1$. Here we used the fact that $\sum_{j=1}^{k-1} q^j e^{-q} \leq (k-1)! \sum_{j=1}^{k-1} \frac{q^j e^{-q}}{(k-1)!} \leq (k-1)!$. Analogously, we have

$$
|F_{N,x}(u) - F_x(u)| \leq kC_2 u + (k-1)! C_3 (e^{\psi(k)} C_a)^2 u^2 / N, \qquad (3.176)
$$

for $x \in S_2$. Therefore, we have the desired statement by $C_5 = \max\{kC_2, (k-1)! C_3 (e^{\psi(k)} C_a)^2\}$.

3.6.7 Proof of Lemma 12

We will prove the lemma for $x \in S_1$ and $x \in S_2$ separately. For $x \in S_1$, we have $\|H_f(z)\| \leq C_d$ for every $y \in B(x,r)$ as long as $r \leq a_N$. Hence, there exists a $y = at + (1-a)x$ for some $a \in [0,1]$ such that

$$
\begin{aligned}
& \big| P(x,r) - f(z) c_d r^d \big| = \left| \int_{t \in B(x,r)} \big(f(t) - f(z) \big) \, dt \right| \\
= & \left| \int_{t \in B(x,r)} \big(f(z) + \nabla f(z)^T (t-x) + (t-x)^T H_f(y)(t-x) - f(z) \big) \, dt \right| \\
= & \left| \int_{t \in B(x,r)} \big((t-x)^T H_f(y)(t-x) \big) \, dt \right| \\
\leq & C_d \int_{t \in B(x,r)} \|t-x\|^2 dt \\
\leq & C_d Vol(B(x,r)) \cdot d \cdot r^2 \leq C_1 r^{d+2}, \qquad (3.177)
\end{aligned}
$$

where $Vol(B(x,r))$ is the volume of $B(x,r)$. $\|t-x\|^2 \leq d \cdot r^2$ for all $t \in B(x,r)$ (here $B(x,r)$ can be any p-norm ball with $1 \leq p \leq \infty$). For the second part, let $S(B(x,r))$ be the surface of $B(x,r)$. Consider m^{d-1} be the Lebesgue measure on $\mathbb{R}^{d-1}$, so $m^{d-1}\left(\, S(B(x,r))\,\right) = dc_d r^{d-1}$. Similarly we have:

$$\begin{aligned} & \left| \frac{dP(x,r)}{dr} - f(z) dc_d r^{d-1} \right| = \left| \int_{t \in S(B(x,r))} \left(\, f(t) - f(z) \,\right) dm^{d-1}(t) \right| \\ \leq \;& C_d \int_{t \in S(B(x,r))} \|t-x\|^2 dm^{d-1}(t) \leq C_2 r^{d+1}. \end{aligned} \tag{3.178}$$

For $x \in S_2$, we simply bound the difference by:

$$\left|\, P(x,r) - f(z) c_d r^d \,\right| \leq f(z) c_d r^d \leq C_a c_d r^d, \tag{3.179}$$

and

$$\left| \frac{dP(x,r)}{dr} - f(z) dc_d r^{d-1} \right| \leq f(z) dc_d r^{d-1} \leq C_a dc_d r^{d-1}, \tag{3.180}$$

since $f(z) \leq C_a$ by Assumption 2.(a).

3.6.8 Proof of Lemma 13

We will prove that

$$\left| \log \left(\tbinom{N}{k} \left(\frac{q}{N}\right)^k \left(1 - \frac{q}{N}\right)^{N-k} \right) - \log \left(\frac{q^k e^{-q}}{k!} \right) \right| \;\leq\; Cq^2/N. \tag{3.181}$$

Then for sufficiently small q such that $\exp\{Cq^2/N\} \leq 2Cq^2/N$, we obtain our desired statement by the fact that $|x-y| \leq |\log x - \log y| \cdot \frac{y}{2}$ for small enough $|\log x - \log y|$. Using Stirling's formula: $\log(N!) = N \log N - N +$

$\frac{1}{2}\log(2\pi N) + O(1/N)$, the difference (3.181) is given by:

$$
\begin{aligned}
& \left| \log\left(\binom{N}{k} \left(\frac{q}{N}\right)^k \left(1-\frac{q}{N}\right)^{N-k} \right) - \log\left(\frac{q^k e^{-q}}{k!} \right) \right| \\
= & \Big| \log N! - \log(N-k)! - \log k! + k\log q + (N-k)\log(N-q) \\
& -N\log N - k\log q + q + \log(k!) \Big| \\
= & \left| \log N! - \log(N-k)! + (N-k)\log(N-q) - N\log N + q \right| \\
\le & \Big| N\log N - N + \frac{1}{2}\log(2\pi N) - (N-k)\log(N-k) + (N-k) \\
& -\frac{1}{2}\log(2\pi(N-k)) + (N-k)\log(N-q) - N\log N + q \Big| + C/N \\
= & \left| -k + \frac{1}{2}\log\frac{N}{N-k} + (N-k)\log\frac{N-q}{N-k} + q \right| + C/N \\
= & \left| -k + q + (N-k)\left(\frac{k-q}{N-k} - \frac{(k-q)^2}{2(N-k)^2} + O(\frac{(k-q)^3}{(N-k)^3}) \right) \right| + C/N \\
\le & \frac{(k-q)^2}{2(N-k)} + Cq^3/N^2 + C/N \le Cq^2/N, \qquad (3.182)
\end{aligned}
$$

where we used the assumption that $q < C\sqrt{N}$ for sufficiently small constant $C > 0$.

3.7 Proof of Theorem 5 on the variance of KSG estimator

We use the Efron-Stein inequality to bound the variance of the estimator. Similar as the proof of Theorem 4, we drop the superscript KSG or BI-KSG and subscript 2 or ∞ when the statement holds for both. Reminder that we use $Z = (X, Y)$ for ease of notation through the proof. For simplicity, let $\widehat{I}^{(N)}(Z)$ be the estimate based on original samples $\{Z_1, Z_2, \ldots, Z_N\}$. For the usage of Efron-Stein inequality, we consider another set of i.i.d. samples $\{Z'_1, Z'_2, \ldots, Z'_N\}$ drawn from the same distribution. Let $\widehat{I}^{(N)}(Z^{(j)})$ be the estimate based on $\{Z_1, \ldots, Z_{j-1}, Z'_j, Z_{j+1}, \ldots, Z_N\}$. Then Efron-Stein inequality states that

$$
\mathrm{Var}\left[\widehat{I}^{(N)}(Z) \right] \le \frac{1}{2}\sum_{j=1}^{N} \mathbb{E}\left[\left(\widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)}) \right)^2 \right]. \qquad (3.183)
$$

Now we will give an upper bound for the difference $|\widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)})|$ for given index j. First of all, let $\widehat{I}^{(N)}(Z_{\setminus j}) = (1/N)\sum_{i=1,i\neq j}^{N} \iota_{k,i}(Z_{\setminus j})$ be the estimate based on $\{Z_1, \ldots, Z_{j-1}, Z_{j+1}, \ldots, Z_N\}$, then by triangle inequality, we have:

$$\begin{aligned}
& \sup_{Z_1,\ldots,Z_N,Z_j'} \Big| \widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)}) \Big| \\
\leq & \sup_{Z_1,\ldots,Z_N,Z_j'} \Big(\Big| \widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z_{\setminus j}) \Big| + \Big| \widehat{I}^{(N)}(Z_{\setminus j}) - \widehat{I}^{(N)}(Z^{(j)}) \Big| \Big) \\
\leq & \sup_{Z_1,\ldots,Z_N} \Big| \widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z_{\setminus j}) \Big| \\
& + \sup_{Z_1,\ldots,Z_{j-1},Z_j',Z_{j+1},\ldots,Z_N} \Big| \widehat{I}^{(N)}(Z_{\setminus j}) - \widehat{I}^{(N)}(Z^{(j)}) \Big| \\
= & 2 \sup_{Z_1,\ldots,Z_N} \Big| \widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z_{\setminus j}) \Big|, \qquad (3.184)
\end{aligned}$$

where the last equality is because of $\{Z_1, \ldots, Z_{j-1}, Z_j', Z_{j+1}, \ldots, Z_N\}$ has the same joint distribution as $\{Z_1, \ldots, Z_N\}$. Now recall that $\widehat{I}^{(N)}(Z) = (1/N)\sum_{i=1}^{N} \iota_{k,i}(Z)$. Therefore, we have

$$\begin{aligned}
& \sup_{Z_1,\ldots,Z_N,Z_j'} \Big| \widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)}) \Big| \\
\leq & \frac{2}{N} \sup_{Z_1,\ldots,Z_N} \sum_{i=1}^{N} \Big| \iota_{k,i}(Z) - \iota_{k,i}(Z_{\setminus j}) \Big|, \qquad (3.185)
\end{aligned}$$

by defining $\iota_{k,j}(Z_{\setminus j}) = 0$. Now we need to upper-bound the difference $|\,\iota_{k,i}(Z) - \iota_{k,i}(Z_{\setminus j})\,|$ created by eliminating sample Z_j for different i 's. We consider the following cases of i's as follows:

- **Case I.** $i = j$. In this case $\iota_{k,i}(Z_{\setminus j}) = 0$ and the upper bounds $|\iota_{k,i}(Z)| \leq 2\log N$ always holds, so $|\,\iota_{k,i}(Z) - \iota_{k,i}(Z_{\setminus j})\,| \leq 2\log N$. The number of i's in this case is only 1. So $\sum_{\text{Case I}} |\,\iota_{k,i}(Z) - \iota_{k,i}(Z_{\setminus j})\,| \leq 2\log N$.

- **Case II.** Z_j is in the k-nearest neighbors of Z_i. In this case, we do not know how $n_{x,i}$ and $n_{y,i}$ will change by eliminating Z_j, so we just use the loosest bound $|\,\iota_{k,i}(Z) - \iota_{k,i}(Z_{\setminus j})\,| \leq 4\log N$. However, the number of i's in this case is upper bounded by the following lemma.

  Lemma 14. *Let $Z, Z_1, Z_2, \ldots, Z_N$ be vectors of $\mathbb{R}^d$ and $\mathcal{Z}_i$ be the set*

$\{Z_1, \ldots, Z_{i-1}, Z, Z_{i+1}, \ldots, Z_N\}$. *Then*

$$\sum_{i=1}^{N} \mathbb{I}\{Z \text{ is in the } k\text{-NN of } Z_i \text{ in } \mathcal{Z}_i\} \leq k\gamma_d, \tag{3.186}$$

(distance ties are broken by comparing indices). Here γ_d is the minimum number of cones with angle smaller than $\pi/6$ needed to cover $\mathbb{R}^d$. Moreover, if we allow k to be different for difference i, we have

$$\sum_{i=1}^{N} \frac{1}{k_i} \mathbb{I}\{Z \text{ is in the } k_i\text{-NN of } Z_i \text{ in } \mathcal{Z}_i\} \leq \gamma_d(\log N + 1). \tag{3.187}$$

By the first inequality in Lemma 14, the number of i's in this case is upper bounded by $k\gamma_{d_x+d_y}$. Therefore, $\sum_{\text{Case II}} |\, \iota_{k,i}(Z) - \iota_{k,i}(Z_{\setminus j}) \,| \leq 4k\gamma_{d_x+d_y} \log N$.

- **Case III.** Z_j is not in the k-nearest neighbors of Z_i, but $\|X_j - X_i\| \leq \rho_{k,i}$, i.e., X_j is in the $n_{x,i}$-nearest neighbors of X_i. In this case, $n_{x,i}$ will decrease by 1 and $n_{y,i}$ remains the same. So

$$\begin{aligned} & |\, \iota_{k,i,\infty}(Z) - \iota_{k,i,\infty}(Z_{\setminus j}) \,| \\ \leq\ & |\psi(n_{x,i,\infty} + 1) - \psi(n_{x,i,\infty})| = \frac{1}{n_{x,i,\infty}}, \qquad (3.188) \\ & |\, \iota_{k,i,2}(Z) - \iota_{k,i,2}(Z_{\setminus j}) \,| \\ \leq\ & |\log(n_{x,i,2}) - \log(n_{x,i,2} - 1)| \leq \frac{1}{n_{x,i,2} - 1} \leq \frac{2}{n_{x,i,2}}, \qquad (3.189) \end{aligned}$$

where the last inequality comes from $n_{x,i,2} \geq k \geq 2$. We do not have an upper bound for the number of i's in this case, but from the second inequality in Lemma 14, we have the following upper bound, where $\mathcal{X}_{i,j} = \{X_1, \ldots, X_{i-1}, X_j, X_{i+1}, \ldots, X_N\}$:

$$\begin{aligned} & \sum_{\text{Case III}} |\, \iota_{k,i}(Z) - \iota_{k,i}(Z_{\setminus j}) \,| \\ \leq\ & \sum_{i=1}^{N} \frac{2}{n_{x,i}} \mathbb{I}\{X_j \text{ is in the } n_{x,i}\text{-NN of } X_i \text{ in } \mathcal{X}_{i,j}\} \\ \leq\ & 2\gamma_{d_x}(\log N + 1) \leq 2\gamma_{d_x+d_y}(\log N + 1). \qquad (3.190) \end{aligned}$$

- **Case IV.** Z_j is not in the k-nearest neighbors of Z_i, but $\|Y_j - Y_i\| \leq \rho_{k,i}$, i.e., Y_j is in the $n_{y,i}$-nearest neighbors of Y_i. In this case, $n_{y,i}$ will decrease by 1 and $n_{x,i}$ remains the same. Follow the same analysis in Case III, we have $\sum_{\text{Case IV}} | \iota_{k,i}(Z) - \iota_{k,i}(Z_{\setminus j}) | \leq 2\gamma_{d_x+d_y}(\log N + 1)$ as well.

- **Case V.** Z_j is not in the k-nearest neighbors of Z_i, and $\|X_j - X_i\| > \rho_{k,i}$, $\|Y_j - Y_i\| > \rho_{k,i}$. In this case, neither $n_{x,i}$ nor $n_{y,i}$ will change. So $\sum_{\text{Case V}} | \iota_{k,i}(Z) - \iota_{k,i}(Z_{\setminus j}) | = 0$.

Combining the five cases, we have:

$$\begin{aligned} & \sum_{i=1}^{N} \Big| \iota_{k,i}(Z) - \iota_{k,i}(Z_{\setminus j}) \Big| \\ \leq \; & 2\log N + 4k\gamma_{d_x+d_y} \log N + 4\gamma_{d_x+d_y}(\log N + 1), \end{aligned} \tag{3.191}$$

for $k \geq 1$, $\log N \geq 1$ and all $\{Z_1, \ldots, Z_N\}$. Plug it into (3.185), we obtain,

$$\begin{aligned} & \sup_{Z_1,\ldots,Z_N,Z'_j} \Big| \widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)}) \Big| \\ \leq \; & \frac{2(2\log N + 4k\gamma_{d_x+d_y} \log N + 4\gamma_{d_x+d_y}(\log N + 1))}{N} \\ \leq \; & \frac{28\gamma_{d_x+d_y} k \log N}{N}. \end{aligned} \tag{3.192}$$

Plug it into Efron-Stein inequality (3.183), we obtain:

$$\begin{aligned} & \text{Var}\Big[\widehat{I}^{(N)}(Z)\Big] \leq \frac{1}{2}\sum_{j=1}^{N} \mathbb{E}\Big[\Big(\widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)})\Big)^2\Big] \\ \leq \; & \frac{1}{2}\sum_{j=1}^{N} \sup_{Z_1,\ldots,Z_n,Z'_j} \Big(\widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)})\Big)^2 \\ \leq \; & \frac{1}{2}\sum_{j=1}^{N} (\frac{28\gamma_{d_x+d_y} k \log N}{N})^2 = \frac{392\gamma^2_{d_x+d_y} k^2 \log^2 N}{N}. \end{aligned} \tag{3.193}$$

3.7.1 Proof of Lemma 14

For the first part of the lemma, we refer to Lemma 20.6 in [51]. The second part of the lemma is a consequence of the first part. We reorder the indices

i's by k_i and rewrite the summation as follows,

$$\begin{aligned}
&\sum_{i=1}^{N} \frac{1}{k_i} \mathbb{I}\{Z \text{ is in the } k_i\text{-NN of } Z_i \text{ in } \mathcal{Z}_i\} \\
= &\sum_{k=1}^{N} \frac{1}{k} \sum_{i=1}^{N} \mathbb{I}\{k_i = k\} \mathbb{I}\{Z \text{ is in the } k\text{-NN of } Z_i \text{ in } \mathcal{Z}_i\} \\
= &\sum_{k=1}^{N} \frac{1}{k} \sum_{i=1}^{N} \mathbb{I}\{k_i = k \text{ and } Z \text{ is in the } k\text{-NN of } Z_i \text{ in } \mathcal{Z}_i\}. \quad (3.194)
\end{aligned}$$

Notice that we take the summation over $k = 1$ to N since each k_i can not be more than N. Denote $S_k = \sum_{i=1}^{N} \mathbb{I}\{k_i = k \text{ and } Z \text{ is in the } k\text{-NN of } Z_i \text{ in the set } \{Z_1, \ldots, Z_{i-1}, Z, Z_{i+1}, \ldots, Z_N\}\}$ for simplicity. Then we need to prove that $\sum_{k=1}^{N} (S_k/k) \leq \gamma_d \log N$. By the first part of this lemma, we obtain,

$$\begin{aligned}
&\sum_{\ell=1}^{k} S_\ell = \sum_{\ell=1}^{k} \sum_{i=1}^{N} \mathbb{I}\{k_i = \ell \text{ and } Z \text{ is in the } \ell\text{-NN of } Z_i \text{ in } \mathcal{Z}_i\} \\
= &\sum_{i=1}^{N} \sum_{\ell=1}^{k} \mathbb{I}\{k_i = \ell \text{ and } Z \text{ is in the } \ell\text{-NN of } Z_i \text{ in } \mathcal{Z}_i\} \\
\leq &\sum_{i=1}^{N} \mathbb{I}\{k_i \leq k \text{ and } Z \text{ is in the } k\text{-NN of } Z_i \text{ in } \mathcal{Z}_i\} \\
\leq &\; k\gamma_d. \quad (3.195)
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&\sum_{k=1}^{N} \frac{S_k}{k} = \sum_{k=1}^{N-1} \frac{1}{k(k+1)} \left(\sum_{\ell=1}^{k} S_\ell \right) + \frac{1}{N} \sum_{\ell=1}^{N} S_\ell \\
\leq &\sum_{k=1}^{N-1} \frac{k\gamma_d}{k(k+1)} + \frac{N\gamma_d}{N} = \sum_{k=1}^{N} \frac{\gamma_d}{k} < \gamma_d(\log N + 1), \quad (3.196)
\end{aligned}$$

which completes the proof.

CHAPTER 4

GEOMETRICAL ADAPTIVE ENTROPY ESTIMATION

Unsupervised representation learning is one of the major themes of modern data science; a common theme among the various approaches is to extract maximally "informative" features via *information-theoretic metrics* (entropy, mutual information and their variations) – the primary reason for the popularity of information-theoretic measures is that they are invariant to one-to-one transformations and that they obey natural axioms such as data processing. Such an approach is evident in many applications, as varied as computational biology [29], sociology [11] and information retrieval [122], with the citations representing a mere smattering of recent works. Within mainstream machine learning, a systematic effort at unsupervised clustering and hierarchical information extraction is conducted in recent works of [90, 92]. The basic workhorse in all these methods is the computation of mutual information (pairwise and multivariate) from i.i.d. samples. Indeed, *sample-efficient estimation* of mutual information emerges as the central scientific question of interest in a variety of applications, and is also of fundamental interest to statistics, machine learning and information theory communities.

While these estimation questions have been studied in the past three decades (and summarized in [123]), the renewed importance of estimating information-theoretic measures in a *sample-efficient* manner is persuasively argued in a recent work [79], where the authors note that existing estimators perform poorly in several key scenarios of central interest (especially when the high-dimensional random variables are strongly related to each other). The most common estimators (featured in scientific software packages) are nonparametric and involve k nearest neighbor (NN) distances between the samples. The widely used estimator of mutual information is the one by Kraskov and Stögbauer and Grassberger [4] and christened the KSG estimator (nomenclature based on the authors, cf. [79]) – while this estimator works well in practice (and performs better than other approaches such as

those based on kernel density estimation procedures on a variety of standard distributions [124]), it still suffers in high dimensions. The basic issue is that the KSG estimator (and the underlying differential entropy estimator based on nearest neighbor distances by Kozachenko and Leonenko (KL, not to be confused with Kullback-Leibler) [2]) does not take advantage of the fact that the samples could lie in a smaller dimensional subspace (more generally, manifold) despite the high dimensionality of the data itself. Such lower dimensional structures effectively act as boundaries, causing the estimator to suffer from what is known as boundary biases.

Ameliorating this deficiency is the central theme of recent works [87, 79, 115], each of which aims to improve upon the classical KL (differential) entropy estimator of [2]. A local SVD is used to heuristically improve the density estimate at each sample point in [79], while a local Gaussian density (with empirical mean and covariance weighted by NN distances) is heuristically used for the same purpose in [115]. Both these approaches, while inspired and intuitive, come with no theoretical guarantees (even consistency) and from a practical perspective involve delicate choice of key hyper parameters. An effort toward a systematic study is initiated in [87] which connects the aforementioned heuristic efforts of [79, 115] to the *local log-likelihood* density estimation methods [6, 7] from theoretical statistics.

The local density estimation method is a strong generalization of the traditional kernel density estimation methods, but requires a delicate normalization which necessitates the solution of certain integral equations (cf. Equation (9) of [7]). Indeed, such an elaborate numerical effort is one of the key impediments for the entropy estimator of [87] to be practically valuable. A second key impediment is that theoretical guarantees (such as consistency) can only be provided when the bandwidth is chosen globally (leading to poor sample complexity in practice) and consistency requires the bandwidth h to be chosen such that $nh^d \to \infty$ and $h \to 0$, where n is the sample size and d is the dimension of the random variable of interest.

More generally, it appears that a systematic application of local log-likelihood methods to estimate *functionals* of the unknown density from i.i.d. samples is missing in the theoretical statistics literature (despite local log-likelihood methods for regression and density estimation being standard textbook fare [125, 126]). We resolve each of these deficiencies in this chapter by undertaking a comprehensive study of estimating the entropy and mutual information

from i.i.d. samples using sample dependent bandwidth choices (typically *fixed* k-NN distances). This effort allows us to connect disparate threads of ideas from different arenas: NN methods, local log-likelihood methods, asymptotic order statistics and sample-dependent heuristic, but inspired, methods for mutual information estimation suggested in the work of [4].

Main contributions of Chapter 4:

1. **Density estimation:** Parameterizing the log density by a polynomial of degree p, we derive *simple closed-form* expressions for the local log-likelihood maximization problem for the cases of $p \leq 2$ for arbitrary dimensions, with Gaussian kernel choices. This derivation, posed as an exercise in [126, Exercise 5.2], significantly improves the computational efficiency upon similar endeavors in the recent efforts of [87, 115, 112].

2. **Entropy estimation:** Using resubstitution of the local density estimate, we derive a simple closed-form estimator of the entropy using a sample dependent bandwidth choice (of k-NN distance, where k is a *fixed* small integer independent of the sample size): this estimator outperforms state of the art entropy estimators in a variety of settings. Since the bandwidth is data dependent and vanishes too fast (because k is fixed), the estimator has a bias, which we derive a closed form expression for and show that it is *independent* of the underlying distribution and hence can be easily corrected: this is our main theoretical contribution, and involves new theorems on asymptotic statistics of nearest neighbors generalizing classical work in probability theory [127], which might be of independent mathematical interest.

3. **Generalized view:** We show that seemingly very different approaches to entropy estimation – recent works of [79, 87, 115] and the classical work of fixed k-NN estimator of Kozachenko and Leonenko [2] – can all be cast in the local log-likelihood framework as specific kernel and *sample dependent bandwidth* choices. This allows for a unified view, which we theoretically justify by showing that resubstitution entropy estimation for *any* kernel choice using fixed k-NN distances as bandwidth involves a bias term that is *independent of the underlying distribution* (but depends on the specific choice of kernel and parametric density

family). Thus our work is a strict mathematical generalization of the classical work of [2].

4. **Mutual information estimation:** The inspired work of [4] constructs a mutual information estimator that subtly altered (in a sample dependent way) the three KL entropy estimation terms, leading to superior empirical performance. We show that the underlying idea behind this change can be incorporated in our framework as well, leading to a novel mutual information estimator that combines the two ideas and outperforms state of the art estimators in a variety of settings.

In the rest of this chapter we describe these main results, the sections are organized in roughly the same order as the enumerated list.

4.1 Local likelihood density estimation (LLDE)

Given n i.i.d. samples $X_1, \ldots, X_n$, estimating the unknown density $f_X(\cdot)$ in $\mathbb{R}^d$ is a very basic statistical task. Local likelihood density estimators [7, 6] constitute state of the art and are specified by a suitable nonnegative weight function $K : \mathbb{R}^d \to \mathbb{R}_+$ (also called a kernel), a degree $p \in \mathbb{Z}^+$ of the polynomial approximation, and the bandwidth $h \in \mathbb{R}_+$, and maximizes the local log-likelihood:

$$\mathcal{L}_x(f) \;=\; \sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right) \log f(X_j) - n \int K\left(\frac{u-x}{h}\right) f(u) du, \tag{4.1}$$

where maximization is over an exponential polynomial family, locally approximating $f(u)$ near x:

$$\begin{aligned} & \log_e f_{a,x}(u) \\ = \;& a_0 + \langle a_1, u - x\rangle + \langle u - x, a_2(u-x)\rangle + \cdots + a_p[u-x, \ldots, u-x], \end{aligned} \tag{4.2}$$

parameterized by $a = (a_0, \ldots, a_p) \in \mathbb{R}^{1 \times d \times d^2 \times \cdots \times d^p}$, where $\langle \cdot, \cdot \rangle$ denotes the inner-product and $a_p[u, \ldots, u]$ the p-th order tensor projection. The *local likelihood density estimate* (LLDE) is defined as $\widehat{f}_n(x) = f_{\widehat{a}(x),x}(x) = e^{\widehat{a}_0(x)}$,

where $\widehat{a}(x) \in \arg\max_a \mathcal{L}_x(f_{a,x})$. The maximizer is represented by a series of nonlinear equations, and does not have a closed-form solution in general. We present below a few choices of the degrees and the weight functions that admit closed-form solutions. Concretely, for $p = 0$, it is known that LDDE reduces to the standard Kernel Density Estimator (KDE) [7]:

$$\widehat{f}_n(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \Big/ \int K\left(\frac{u-x}{h}\right) du. \tag{4.3}$$

If we choose the step function $K(u) = \mathbb{I}(\|u\| \le 1)$ ($\|\cdot\|$ denotes Euclidean norm) with a local and data-dependent choice of the bandwidth $h = \rho_{k,x}$ where $\rho_{k,x}$ is the k-NN distance from x, then the above estimator recovers the popular k-NN density estimate as a special case, namely, for $C_d = \pi^{d/2}/\Gamma(d/2+1)$,

$$\widehat{f}_n(x) \;=\; \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\|X_i - x\| \le \rho_{k,x})}{\mathrm{Vol}\{u \in \mathbb{R}^d : \|u - x\| \le \rho_{k,x}\}} \;=\; \frac{k}{n\, C_d\, \rho_{k,x}^d}. \tag{4.4}$$

For higher degree local likelihood, we provide simple closed-form solutions. Somewhat surprisingly, this result has eluded prior works [115, 112] and [87] which specifically attempted the evaluation for $p = 2$. Part of the subtlety in the result is to critically use the fact that the parametric family (e.g., the polynomial family in (4.2)) need not be normalized themselves; the local log-likelihood maximization ensures that the resulting density estimate is correctly normalized so that it integrates to 1.

Proposition 1. *[126, Exercise 5.2] For a degree $p \in \{1, 2\}$, the maximizer of local likelihood (4.1) admits a closed-form solution, when using the Gaussian kernel $K(u) = e^{-\frac{\|u\|^2}{2}}$. In case of $p = 1$,*

$$\widehat{f}_n(x) \;=\; \frac{S_0}{n(2\pi)^{d/2}h^d} \exp\left\{-\frac{1}{2}\frac{1}{S_0^2}\|S_1\|^2\right\}, \tag{4.5}$$

where $S_0 \in \mathbb{R}$ and $S_1 \in \mathbb{R}^d$ are defined for given $x \in \mathbb{R}^d$ and $h \in \mathbb{R}$ as

$$S_0 \equiv \sum_{j=1}^{n} e^{-\frac{\|X_j - x\|^2}{2h^2}}, \quad S_1 \equiv \sum_{j=1}^{n} \frac{1}{h}(X_j - x)\, e^{-\frac{\|X_j - x\|^2}{2h^2}}. \tag{4.6}$$

In case of $p = 2$, for S_0 and S_1 defined as above,

$$\widehat{f}_n(x) \;=\; \frac{S_0}{n(2\pi)^{d/2}h^d|\Sigma|^{1/2}} \exp\Big\{ -\frac{1}{2}\frac{1}{S_0^2} S_1^T \Sigma^{-1} S_1 \Big\}, \tag{4.7}$$

where $|\Sigma|$ is the determinant and $S_2 \in \mathbb{R}^{d\times d}$ and $\Sigma \in \mathbb{R}^{d\times d}$ are defined as

$$S_2 \equiv \sum_{j=1}^{n} \frac{1}{h^2}(X_j - x)(X_j - x)^T e^{-\frac{\|X_j - x\|^2}{2h^2}}, \quad \Sigma \equiv \frac{S_0 S_2 - S_1 S_1^T}{S_0^2}, \tag{4.8}$$

where it follows from Cauchy-Schwarz that Σ is positive semidefinite.

One of the major drawbacks of the KDE and k-NN methods is the increased bias near the boundaries. LLDE provides a principled approach to automatically correct for the boundary bias, which takes effect only for $p \geq 2$ [6, 128]. This explains the performance improvement for $p = 2$ in Figure 4.1 (top panel), and the gap increases with the correlation as boundary effect becomes more prominent. We use the proposed estimators with $p \in \{0, 1, 2\}$ to estimate the mutual information between two jointly Gaussian random variables with correlation r, from $n = 500$ samples, using resubstitution methods explained in the next sections. Each point is averaged over 100 instances.

In the right panel, we generate i.i.d. samples from a two-dimensional Gaussian with correlation 0.9, and find a local approximation $\widehat{f}(u - x^*)$ around x^* denoted by the blue $*$ in the center. Standard k-NN approach fits a uniform distribution over a circle enclosing $k = 20$ nearest neighbors (red circle). The green lines are the contours of the degree-2 polynomial approximation with bandwidth $h = \rho_{20,x}$. The figure illustrates that k-NN method suffers from boundary effect, where it underestimates the probability by over estimating the volume in (4.4). However, degree-2 LDDE is able to correctly capture the local structure of the pdf, correcting for boundary biases.

Despite the advantages of the LLDE, it requires the bandwidth to be data independent and vanishingly small (sublinearly in sample size) for consistency almost everywhere – this is an impediment to practical use since data-independent bandwidth lacks the capability to capture local geometry of data. On the other hand, if we restrict our focus to *functionals* of the density, then both these issues are resolved: this is the focus of Section 4.2 where we show that the bandwidth can be chosen to be based on *fixed* k-NN distances and the resulting universal bias readily corrected.

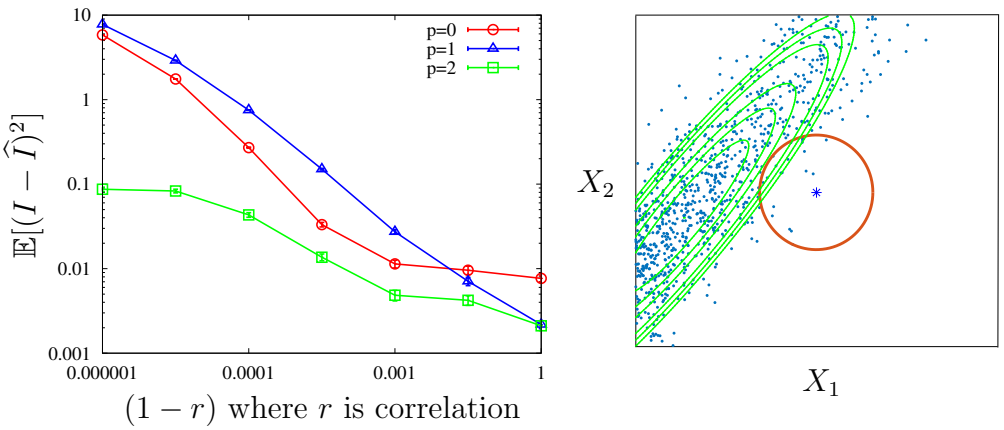

Figure 4.1: The boundary bias becomes less significant and the gap closes as correlation decreases for estimating the mutual information (left). Local approximation around the blue $*$ in the center. The degree-2 local likelihood approximation (contours in green) automatically captures the local structure whereas the standard k-NN approach (uniform distribution in red circle) fails (left).

4.2 Second-order k-LNN entropy estimator

We consider the *resubstitution* entropy estimators which has the form $\widehat{H}(x) = -(1/n)\sum_{i=1}^{n} \log \widehat{f}_n(X_i)$ and propose to use the local likelihood density estimator in (4.7) and a choice of bandwidth that is *local* (varying for each point x) and *adaptive* (based on the data). Concretely, we choose, for each sample point X_i, the bandwidth h_{X_i} to be the the distance to its k-th nearest neighbor ρ_{k,X_i} (we use $\rho_{k,i}$ instead of ρ_{k,X_i} for simplicity of notation for the remainder of the chapter). Precisely, we propose the following k-Local Nearest Neighbor (k-LNN) entropy estimator of degree-2:

$$\widehat{H}^{(n)}_{k\text{LNN}}(X) \;=\; -\frac{1}{n}\sum_{i=1}^{n}\left\{\log\frac{S_{0,i}}{n(2\pi)^{d/2}\rho_{k,i}^{d}|\Sigma_i|^{1/2}} - \frac{1}{2}\frac{1}{S_{0,i}^{2}}S_{1,i}^{T}\Sigma_i^{-1}S_{1,i}\right\} - B_{k,d}, \tag{4.9}$$

where subtracting $B_{k,d}$ defined in Theorem 8 removes the asymptotic bias, and $k \in \mathbb{Z}^+$ is the only hyper parameter determining the bandwidth. In practice k is a small integer fixed to be in the range $4 \sim 8$. We only use the $\lceil \log n \rceil$ nearest subset of samples $\mathcal{S}_i = \{j \in [n] \;:\; j \neq i \text{ and } \|X_i - X_j\| \leq$

$\rho_{\lceil \log n \rceil,i}\}$ in computing the following quantities:

$$S_{0,i} \equiv \sum_{j \in \mathcal{S}_i} e^{-\frac{\|X_j - X_i\|^2}{2\rho_{k,i}^2}}, \quad S_{1,i} \equiv \sum_{j \in \mathcal{S}_i} \frac{1}{\rho_{k,i}} (X_j - X_i) e^{-\frac{\|X_j - X_i\|^2}{2\rho_{k,i}^2}},$$
$$S_{2,i} \equiv \sum_{j \in \mathcal{S}_i} \frac{1}{\rho_{k,i}^2} (X_j - X_i)(X_j - X_i)^T e^{-\frac{\|X_j - X_i\|^2}{2\rho_{k,i}^2}}, \; \Sigma_i \equiv \frac{S_{0,i} S_{2,i} - S_{1,i} S_{1,i}^T}{S_{0,i}^2}. \tag{4.10}$$

The truncation is important for computational efficiency, but the analysis works as long as $m = O(n^{1/(2d)-\varepsilon})$ for any positive ε that can be arbitrarily small. For a larger m, for example of $\Omega(n)$, those neighbors that are further away have a different asymptotic behavior. We show in Theorem 8 that the asymptotic bias is *independent* of the underlying distribution and hence can be *precomputed* and removed, under mild conditions on a twice continuously differentiable pdf $f(x)$ (cf. Lemma 15).

Theorem 8. *For $k \geq 3$ and $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ are i.i.d. samples from a twice continuously differentiable pdf $f(x)$ such that $\mathbb{E}[|\log f(X)|] < \infty$, then*

$$\lim_{n \to \infty} \mathbb{E}[\widehat{H}_{k\mathrm{LNN}}^{(n)}(X)] \;\; = \;\; H(X), \tag{4.11}$$

where $B_{k,d}$ in (4.9) is a constant that only depends on k and d. Further, if $\mathbb{E}[(\log f(X))^2] < \infty$ then the variance of the proposed estimator is bounded by $\mathrm{Var}[\widehat{H}_{k\mathrm{LNN}}^{(n)}(X)] = O((\log n)^2/n)$.

This proves the L_1 and L_2 consistency of the k-LNN estimator. As noted in [113], such an assumption is common in the literature on consistency of k-NN estimators, where it has been implicitly assumed in existing analyses of entropy estimators including [2, 129, 130, 123], without explicitly stating that such assumptions are being made.

Our choice of a local adaptive bandwidth $h_{X_i} = \rho_{k,i}$ is crucial in ensuring that the asymptotic bias $B_{k,d}$ does not depend on the underlying distribution $f(x)$. This relies on a fundamental connection to the theory of asymptotic order statistics made precise in Lemma 15, which also gives the explicit formula for the bias below.

The main idea is that the empirical quantities used in the estimate (4.10) converge in large n limit to similar quantities defined over order statistics.

We make this intuition precise in Section 4.3. We define order statistics over i.i.d. standard exponential random variables $E_1, E_2, \ldots, E_m$ and i.i.d. random variables $\xi_1, \xi_2, \ldots, \xi_m$ drawn uniformly (the Haar measure) over the unit sphere in $\mathbb{R}^d$, for a variable $m \in \mathbb{Z}^+$. We define for $\alpha \in \{0, 1, 2\}$,

$$\tilde{S}_\alpha^{(m)} \equiv \sum_{j=1}^{m} \xi_j^{(\alpha)} \frac{(\sum_{\ell=1}^{j} E_\ell)^\alpha}{(\sum_{\ell=1}^{k} E_\ell)^\alpha} \exp\left\{ -\frac{(\sum_{\ell=1}^{j} E_\ell)^2}{2(\sum_{\ell=1}^{k} E_\ell)^2} \right\}, \tag{4.12}$$

where $\xi_j^{(0)} = 1$, $\xi_j^{(1)} = \xi_j \in \mathbb{R}^d$, and $\xi_j^{(2)} = \xi_j \xi_j^T \in \mathbb{R}^{d \times d}$, and let $\tilde{S}_\alpha = \lim_{m\to\infty} \tilde{S}_\alpha^{(m)}$ and $\tilde{\Sigma} = (1/\tilde{S}_0)^2(\tilde{S}_0\tilde{S}_2 - \tilde{S}_1\tilde{S}_1^T)$. We show that the limiting $\tilde{S}_\alpha$'s are well-defined (in the proof of Theorem 8) and are directly related to the bias terms in the resubstitution estimator of entropy:

$$\begin{aligned} & B_{k,d} \\ = \; & \mathbb{E}[\log(\sum_{\ell=1}^{k} E_\ell) + \frac{d}{2}\log 2\pi - \log(C_d \tilde{S}_0) + \frac{1}{2}\log\left|\tilde{\Sigma}\right| + (\frac{1}{2\tilde{S}_0^2}\tilde{S}_1^T\tilde{\Sigma}^{-1}\tilde{S}_1)\,]. \end{aligned} \tag{4.13}$$

In practice, we propose using a fixed small k such as five. For $k \leq 3$ the estimator has a very large variance, and numerical evaluation of the corresponding bias also converges slowly. For some typical choices of k, we provide approximate evaluations below, where $0.0183(\pm 6)$ indicates empirical mean $\mu = 183 \times 10^{-4}$ with confidence interval 6×10^{-4}. In these numerical evaluations, we truncated the summation at $m = 50,000$. Although we prove that $B_{k,d}$ converges in m, in practice, one can choose m based on the number of samples and $B_{k,d}$ can be evaluated for that m.

Theoretical contribution: Our key technical innovation is a fundamental connection between nearest neighbor statistics and asymptotic order statistics, stated below as Lemma 15: we show that the (normalized) distances $\rho_{\ell,i}$'s jointly converge to the standardized uniform order statistics and the directions $(X_{j_\ell} - X_i)/\|X_{j_\ell} - X_i\|$'s converge to independent uniform distribution (Haar measure) over the unit sphere.

Conditioned on $X_i = x$, the proposed estimator uses nearest neighbor statistics on $Z_{\ell,i} \equiv X_{j_\ell} - x$ where X_{j_ℓ} is the ℓ-th nearest neighbor from x such that $Z_{\ell,i} = ((X_{j_\ell} - X_i)/\|X_{j_\ell} - X_i\|)\rho_{\ell,i}$. Naturally, all the techniques we develop generalize to any estimators that depend on the nearest neighbor

Table 4.1: Numerical evaluation of $B_{k,d}$, via sampling $1,000,000$ instances for each pair (k, d).

		k			
		4	5	6	7
d	1	-0.0183($\pm$6)	-0.0233($\pm$6)	-0.0220($\pm$4)	-0.0200($\pm$4)
	2	-0.1023($\pm$5)	-0.0765($\pm$4)	-0.0628($\pm$4)	-0.0528($\pm$3)

statistics $\{Z_{\ell,i}\}_{i,\ell\in[n]}$ – and the value of such a general result is demonstrated later (in Section 4.3) when we evaluate the bias in similarly inspired entropy estimators [79, 87, 115, 2].

Lemma 15. *Let $E_1, E_2, \ldots, E_m$ be i.i.d. standard exponential random variables and $\xi_1, \xi_2, \ldots, \xi_m$ be i.i.d. random variables drawn uniformly over the unit $(d-1)$-dimensional sphere in d dimensions, independent of the E_i's. Suppose f is twice continuously differentiable and $x \in \mathbb{R}^d$ satisfies that there exists $\varepsilon > 0$ such that $f(a) > 0$, $\|\nabla f(a)\| = O(1)$ and $\|H_f(a)\| = O(1)$ for any $\|a - x\| \leq \varepsilon$. Then for any $m = O(\log n)$, we have the following convergence conditioned on $X_i = x$:*

$$\lim_{n\to\infty} d_{\mathrm{TV}}((c_d n f(x))^{1/d}(\, Z_{1,i}, \ldots, Z_{m,i}\,), (\,\xi_1 E_1^{1/d}, \ldots, \xi_m(\sum_{\ell=1}^{m} E_\ell)^{1/d}\,)) = 0, \tag{4.14}$$

where $d_{\mathrm{TV}}(\cdot,\cdot)$ is the total variation and c_d is the volume of unit Euclidean ball in $\mathbb{R}^d$.

The proof of Theorem 8 consists two parts — upper bound of the bias and the upper bound of the variance. The bias of the estimator can be written as a function of the order statistics $(Z_{1,i}, \ldots, Z_{m,i})$, which converges to standard random variables by Lemma 15. Therefore, the bias converges to a fixed quantity only depends on k and d. The upper bound of the variances comes from Efron-Stein inequality.

Empirical contribution: Numerical experiments suggest that the proposed estimator outperforms state-of-the-art entropy estimators, and the gap increases with correlation. The idea of using k-NN distance as bandwidth for entropy estimation was originally proposed by Kozachenko and Leonenko in [2], and is a special case of the k-LNN method we propose with degree 0 and

a step kernel. We refer to Section 4.3 for a formal comparison. Another popular resubstitution entropy estimator is to use KDE in (4.3) [54], which is a special case of the k-LNN method with degree 0, and the Gaussian kernel is used in simulations. As comparison, we also study a new estimator [31] based on von Mises expansion (as opposed to simple re-substitution) which has an improved convergence rate in the large sample regime. In Figure 4.2 (left), we draw 100 samples i.i.d. from two standard Gaussian random variables with correlation r, and plot resulting mean squared error averaged over 100 instances. The ground truth, in this case is $H(X) = \log(2\pi e) + 0.5\log(1-r^2)$. On the right, we repeat the same simulation for fixed $r = 0.99999$ and varying number of samples and $m = 7\log_e n$.

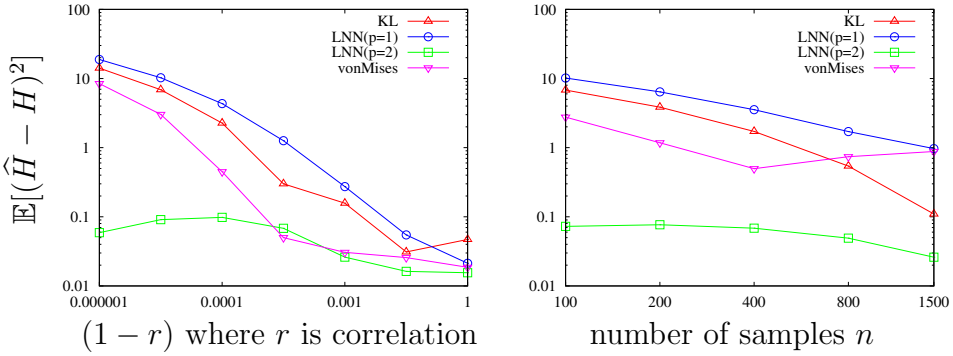

Figure 4.2: Degree-2 k-LNN outperforms other state-of-the-art estimators for entropy estimation.

In Figure 4.3, we repeat the same simulation for 6 standard Gaussian random variables with $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_3, X_4) = \mathrm{Cov}(X_5, X_6) = r$ and $\mathrm{Cov}(X_i, X_j) = 0$ for other pairs (i, j). On the left, we draw 100 i.i.d. samples with various r. We plot resulting mean squared error averaged over 100 instances. The ground truth is $H(X) = 3\log(2\pi e) + 1.5\log(1-r^2)$. On the right, we repeat the same simulation for fixed $r = 0.99999$ and varying number of samples and $m = 7\log_e n$.

In Figure 4.4 (left), we draw 100 samples i.i.d. from a mixture of two joint Gaussian distributions with zero mean and covariance $\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}$, respectively, and plot resulting average estimate over 100 instances. Here we plot an upper bound of the ground truth $H(X) \leq \log(2) +$

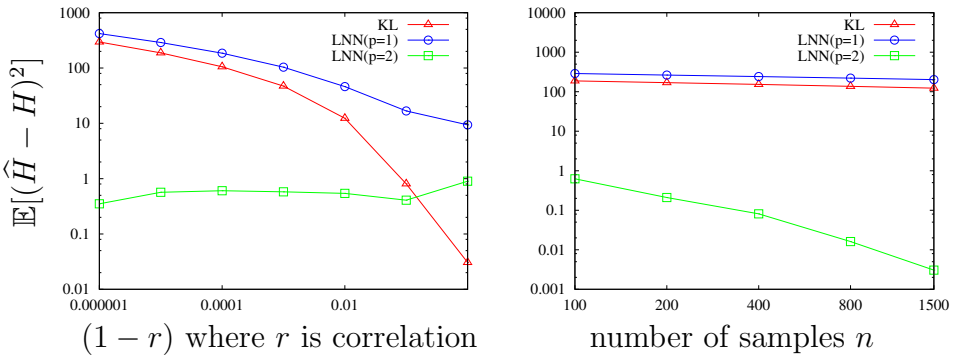

Figure 4.3: Degree-2 k-LNN outperforms other state-of-the-art estimators for high-dimensional entropy estimation.

$\log(2\pi e)+0.5\log(1-r^2)$ for $r \geq 0.9$. On the right, we repeat the same simulation for fixed $r = 0.99999$ and varying number of samples and $m = 7\log_e n$.

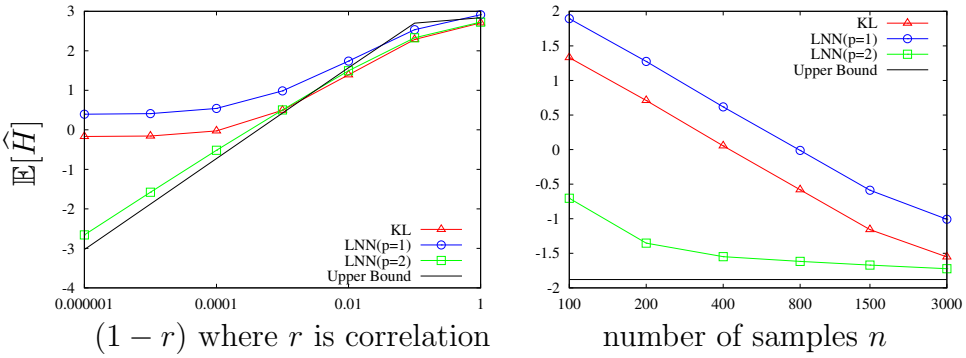

Figure 4.4: Degree-2 k-LNN outperforms other state-of-the-art estimators for non-Gaussian entropy estimation.

4.3 Universality of the k-LNN approach

In this section, we show that Theorem 8 holds universally for a general family of entropy estimators, specified by the choice of $k \in \mathbb{Z}^+$, degree $p \in \mathbb{Z}^+$, and a kernel $K : \mathbb{R}^d \to \mathbb{R}$, thus allowing a unified view of several seemingly disparate entropy estimators [2, 79, 87, 115]. The template of the entropy estimator is the following: given n i.i.d. samples, we first compute the local density estimate by maximizing the local likelihood defined in (4.1) with

bandwidth $\rho_{k,i}$, and then resubstitute it to estimate entropy: $\widehat{H}^{(n)}_{k,p,K}(X) = -(1/n)\sum_{i=1}^{n} \log \widehat{f}_n(X_i)$.

Theorem 9. *For the family of estimators described above, under the hypotheses of Theorem 8, if the solution to the maximization* $\widehat{a}(x) = \arg\max_a \mathcal{L}_x(f_{a,x})$ *exists for all* $x \in \{X_1, \ldots, X_n\}$*, then for any choice of* $k \geq p+1$*,* $p \in \mathbb{Z}^+$*, and* $K : \mathbb{R}^d \to \mathbb{R}$*, the asymptotic bias is independent of the underlying distribution:*

$$\lim_{n\to\infty} \mathbb{E}[\widehat{H}^{(n)}_{k,p,K}(X)] \;\; = \;\; H(X) + \tilde{B}_{k,p,K,d}, \tag{4.15}$$

for some constant $\tilde{B}_{k,d,p,K}$ *that only depends on* k, p, K *and* d*.*

Although in general there is no simple analytical characterization of the asymptotic bias $\tilde{B}_{k,p,K,d}$ it can be readily numerically computed: since $\tilde{B}_{k,p,K,d}$ is independent of the underlying distribution, one can run the estimator over i.i.d. samples from *any* distribution and numerically approximate the bias for *any* choice of the parameters. However, when the maximization $\widehat{a}(x) = \arg\max_a \mathcal{L}_x(f_{a,x})$ admits a closed-form solution, as is the case with proposed k-LNN, then $\widetilde{B}_{k,p,K,d}$ can be characterized explicitly in terms of uniform order statistics.

This family of estimators is general: for instance, the popular KL estimator is a special case with $p = 0$ and a step kernel $K(u) = \mathbb{I}(\|u\| \leq 1)$. Kozachenko and Leonenko [2] showed that the asymptotic bias is independent of the dimension d and can be computed exactly to be $\log n - \psi(n) + \psi(k) - \log k$ and $\psi(k)$ is the digamma function defined as $\psi(x) = \Gamma^{-1}(x)d\Gamma(x)/dx$. The dimension independent nature of this asymptotic bias term (of $O(n^{-1/2})$ for $d = 1$ in [32, Theorem 1] and $O(n^{-1/d})$ for general $d \geq 2$ in [8]) is special to the choice of $p = 0$ and the step kernel; Analogously, the estimator in [79] can be viewed as a special case with $p = 0$ and an ellipsoidal step kernel.

For $p = 0$ and general kernel choice K, this family of estimators become the resubstitution entropy estimator $\widehat{H}(x) = -(1/n)\sum_{i=1}^{n} \log \widehat{f}_n(X_i)$ where $\widehat{f}_n(X_i)$ is simply KDE (or equivalently, zeroth order LLDE) (4.3) combined with an adaptive bandwidth choice. When using KDE, the bandwidth h is typically chosen to satisfy $h \to 0$ and $nh^d \to \infty$ in order to ensure that the mean squared error vanishes as the number of samples n increases. However, the adaptive bandwidth $\rho_{k,i}$ does *not* satisfy $n\rho^d_{k,i} \to \infty$ for fixed small

k, giving the resubstitution entropy estimator an non-zero asymptotic bias. Theorem 9 shows that the asymptotic bias $\tilde{B}_{k,p,K,d}$ is *independent* of the underlying distribution, hence can be precomputed, and removed for any particular choice of kernel K and parameter k. Table 4.2 provides an approximate estimation of $\tilde{B}_{k,p=0,K,d}$ for typical choices of k and two widely used kernels (Gaussian kernel $K(u) \propto \exp\{-\|u\|^2/2\}$ and Epanechnikov kernel $K(u) \propto (1-\|u\|^2)\mathbb{I}\{|u| \leq 1\}$). The main point we would like to highlight is that the asymptotic bias values are all very small and get even smaller as the number of nearest neighbors k increases.

Table 4.2: Numerical evaluation of $\tilde{B}_{k,p=0,K,d}$, via sampling $1,000,000$ instances for each tuple (K,k,d).

	k			
	4	5	6	7
Gau, $d=1$	-0.0543($\pm$4)	-0.0430($\pm$4)	-0.0355($\pm$3)	-0.0302($\pm$3)
Gau, $d=2$	-0.0503($\pm$3)	-0.0393($\pm$2)	-0.0323($\pm$2)	-0.0273($\pm$2)
Epa, $d=1$	0.1976($\pm$6)	0.1480($\pm$5)	0.1183($\pm$5)	0.0987($\pm$4)
Epa, $d=2$	0.2247($\pm$6)	0.1674($\pm$6)	0.1335($\pm$5)	0.1111($\pm$5)

4.4 k-LNN mutual information estimator

Given an entropy estimator $\widehat{H}_{\mathrm{KL}}$, mutual information can be estimated: $\widehat{I}_{\mathrm{3KL}} = \widehat{H}_{\mathrm{KL}}(X) + \widehat{H}_{\mathrm{KL}}(Y) - \widehat{H}_{\mathrm{KL}}(X,Y)$. In [4], Kraskov and Stögbauer and Grassberger introduced $\widehat{I}_{\mathrm{KSG}}(X;Y)$ by coupling the choices of the bandwidths. The joint entropy is estimated in the usual way, but for the marginal entropy, instead of using kNN distances from $\{X_j\}$, the bandwidth $h_{X_i} = \rho_{k,i}(X,Y)$ is chosen, which is the k nearest neighbor distance from (X_i,Y_i) for the joint data $\{(X_j,Y_j)\}$. Consider $\widehat{I}_{\mathrm{3LNN}}(X;Y) = \widehat{H}_{k\mathrm{LNN}}(X) + \widehat{H}_{k\mathrm{LNN}}(Y) - \widehat{H}_{k\mathrm{LNN}}(X,Y)$. Inspired by [4], we introduce the following novel mutual information estimator we denote by $\widehat{I}_{\mathrm{LNN-KSG}}(X;Y)$, where for the joint (X,Y) we use the LNN entropy estimator we proposed in (4.9), and for the marginal entropy we use the bandwidth $h_{X_i} = \rho_{k,i}(X,Y)$ coupled to the joint estimator. Empirically, we observe $\widehat{I}_{\mathrm{KSG}}$ outperforms $\widehat{I}_{\mathrm{3KL}}$ everywhere, validating the use of correlated bandwidths. However, the performance of $\widehat{I}_{\mathrm{LNN-KSG}}$ is similar to $\widehat{I}_{\mathrm{3LNN}}$–sometimes better and sometimes worse.

In Figure 4.5 (left), we estimate mutual information under the same setting as in Figure 4.2 (left). For most regimes of correlation r, both 3LNN and LNN-KSG outperforms other state-of-the-art estimators. The gap increases with correlation r. On the right, we draw i.i.d. samples from two random variables X and Y, where X is uniform over $[0, 1]$ and $Y = X + U$, where U is uniform over $[0, 0.01]$ independent of X. In the large sample limit, all estimators find the correct mutual information. The plot show how sensitive the estimates are, in the small sample regime. Both LNN and LNN-KSG are significantly more robust compared to other approaches. Mutual information estimators have been recently proposed in [79, 87, 115] based on local likelihood maximization. However, they involve heuristic choices of hyper-parameters or solving elaborate optimization and numerical integrations, which are far from being easy to implement.

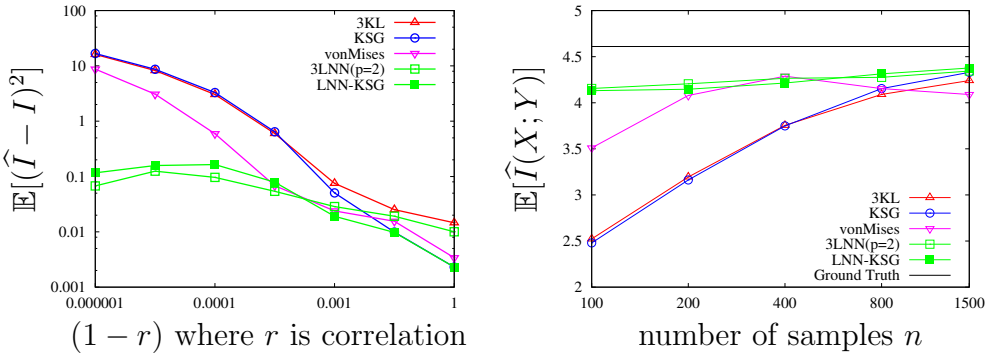

Figure 4.5: Proposed $\widehat{I}_{\text{LNN}-\text{KSG}}$ and $\widehat{I}_{\text{3LNN}}$ outperform other state-of-the-art estimators.

In Figure 4.6, we test the mutual information estimators for $Y = f(X)+U$, where X is uniformly distributed over $[0, 1]$ and U is uniformly distributed over $[0, \theta]$, independent of X, for some noise level θ. Similar simulation were studied in [87]. We draw 2500 i.i.d. sample points for each relationship. The plot show that for small noise level θ, i.e., near-functional related random variables, our proposed estimators $\widehat{I}_{3LNN}$ and $\widehat{I}_{LNN-KSG}$ perform much better than 3KL and KSG estimators. Also our proposed estimators can handle both linear and nonlinear functional relationships.

In Figure 4.7, we test our estimators on linear and nonlinear relationships for both low-dimensional ($D = 2$) and high-dimensional ($D = 5$). Here X_i's are uniformly distributed over $[0, 1]$ and U is uniformly distributed over

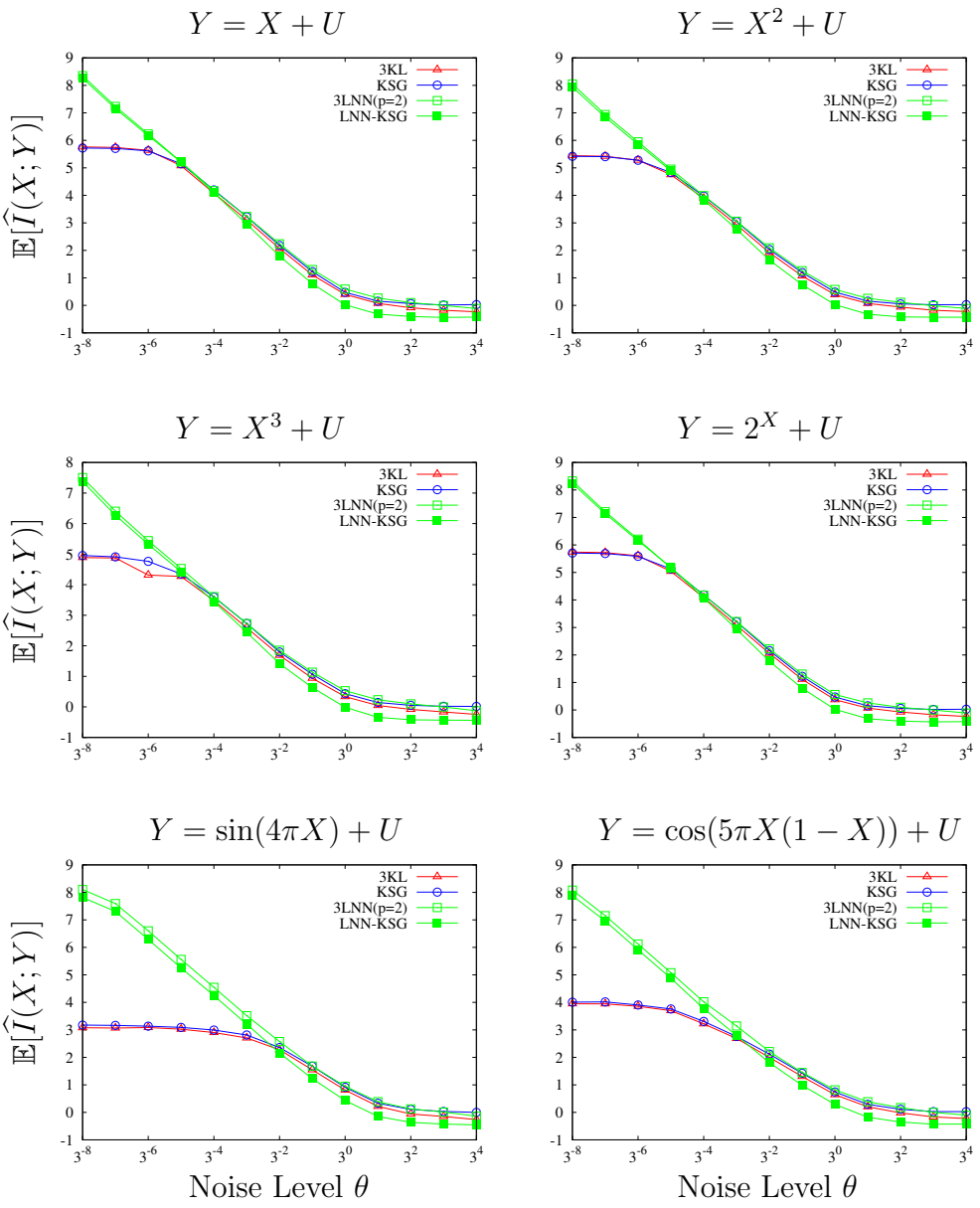

Figure 4.6: Functional relationship test for mutual information estimators. Proposed $\widehat{I}_{\text{LNN}-\text{KSG}}$ and $\widehat{I}_{\text{3LNN}}$ outperform other state-of-the-art estimators.

$[-3^8/2, 3^8/2]$, independently of X_i's. Similar simulation were studied in [79]. We can see that our estimators $\widehat{I}_{3LNN}$ and $\widehat{I}_{LNN-KSG}$ converges much faster than $\widehat{I}_{3KL}$ and $\widehat{I}_{KSG}$.

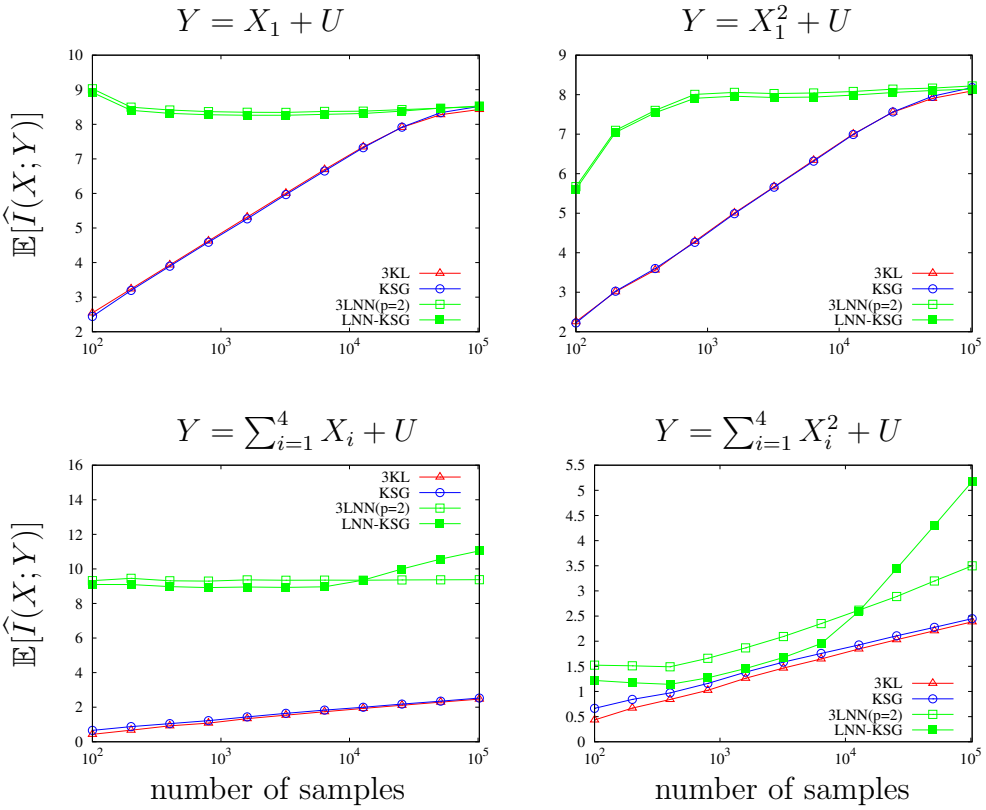

Figure 4.7: Estimated mutual information of low-dimensional and high-dimensional relationships.

4.5 Breaking the bandwidth barrier

While k-NN distance based bandwidth are routine in practical usage [128], the main finding of this work is that they also turn out to be the "correct" mathematical choice for the purpose of asymptotically unbiased estimation of an integral functional such as the entropy: $-\int f(x) \log f(x)$; we briefly discuss the ramifications below. Traditionally, when the goal is to estimate $f(x)$, it is well known that the bandwidth should satisfy $h \to 0$ and $nh^d \to \infty$, for KDEs to be consistent. As a rule of thumb, $h = 1.06\widehat{\sigma}n^{-1/5}$ is suggested when $d = 1$ where $\widehat{\sigma}$ is the sample standard deviation [125, Chapter 6.3]. On

the other hand, when estimating entropy, as well as other integral functionals, it is known that resubstitution estimators of the form $-(1/n)\sum_{i=1}^{n}\log\widehat{f}(X_i)$ achieve variances scaling as $O(1/n)$ independent of the bandwidth [101]. This allows for a bandwidth as small as $O(n^{-1/d})$.

The bottleneck in choosing such a small bandwidth is the bias, scaling as $O(h^2+(nh^d)^{-1}+E_n)$ [101], where the lower-order dependence on n, dubbed E_n, is generally not known. The barrier in choosing a *global* bandwidth of $h = O(n^{-1/d})$ is the strictly positive bias whose value depends on the unknown distribution and cannot be subtracted off. Previous work [100] tried to solve the bias problem for global bandwidth $h = O(n^{-1/d})$ in one-dimensional scenario. However, the proposed local and adaptive choice of the k-NN distance admits an asymptotic bias that is independent of the unknown underlying distribution. Manually subtracting off the non-vanishing bias gives an asymptotically unbiased estimator, with a potentially faster convergence as numerically compared below. Figure 4.8 illustrates how k-NN based bandwidth significantly improves upon, say a rule-of-thumb choice of $O(n^{-1/(d+4)})$ explained above and another choice of $O(n^{-1/(d+2)})$. In the left figure, we use the setting from Figure 4.2 (right) but with correlation $r = 0.999$. On the right, we generate $X \sim \mathcal{N}(0,1)$ and U from uniform $[0, 0.01]$ and let $Y = X+U$ and estimate $I(X;Y)$. Following recent advances in [130, 131], the proposed local estimator has a potential to be extended to, for example, Rényi entropy, but with a multiplicative bias as opposed to additive.

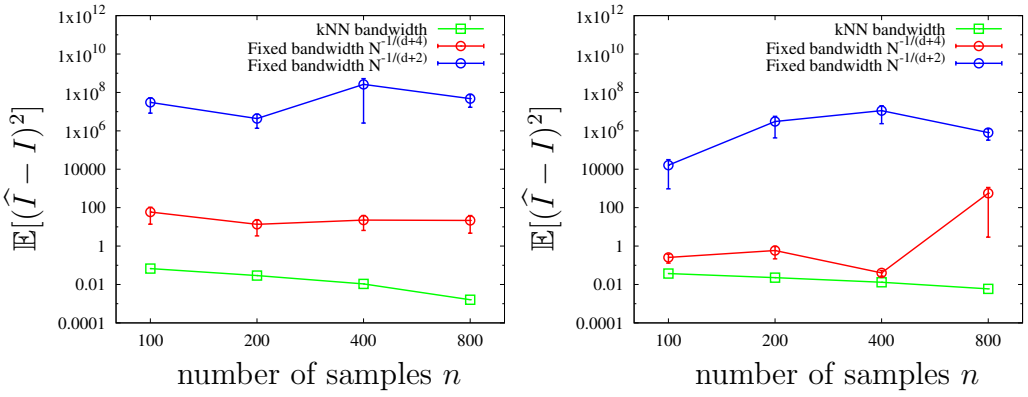

Figure 4.8: Local and adaptive bandwidth significantly improves over rule-of-thumb fixed bandwidth.

4.6 Discussion and review of previous work

The topic of estimation of an integral functional of an unknown density from i.i.d. samples is a classical one in statistics and we tie together a few pertinent topics from the literature in the context of the results of this chapter.

4.6.1 Uniform order statistics and NN distances

The expression for the asymptotic bias in (4.13) which is independent of the underlying distribution forms the main result of this chapter and crucially depends on Lemma 15. Precisely, the lemma implies that the quantities S_i's in (4.10) converge in distribution to $\tilde{S}_i$'s in (4.12). There are two parts to this convergence result: the nearest neighbor distances converge to uniform order statistics and the directions to those nearest neighbors converge independently to Haar measures on the unit sphere. The former has been extensively studied, for example see [127] for a survey of results. The latter is a new result that we state in Lemma 15. Intuitively, assuming smoothness, the probability density f_X in the neighborhood of a sample X_i (as defined by the distance to the k-th nearest neighbor) converges to a uniform distribution over a ball (of radius decreasing at the rate $\rho_{k,i} = \Theta(n^{-1/d})$), as more samples are collected. The nearest neighbor distances and directions converge to those from the uniform distribution over the ball, and Lemma 15 makes this intuition precise for the nearest m neighbors up to $m = O(n^{1/(2d)-\epsilon})$ with any arbitrarily small but positive ε.

Only the convergence analysis of the distances, and not the directions, is required for traditional k-NN based estimators, such as the entropy estimator of [2]. In the seminal paper, [2] introduced *resubstitution* entropy estimators of the form $\widehat{H}(X) = -(1/n)\sum_{i=1}^{n} \log \widehat{f}_n(X_i)$ with $\widehat{f}_n(x) = k/(n\, C_d\, \rho_{k,x}^d)$ (as defined in (4.4)). This k-NN estimator has a non-vanishing asymptotic bias, which was computed as $B_{k,d} = (\psi(k) - \log(k))$ with the digamma function $\psi(\cdot)$ and was suggested to be manually removed. For $k = 1$ this was proved in the original paper of [2], which later was extended in [124, 129] to general k. This mysterious bias term $B_{k,d} = (\psi(k) - \log(k))$ whose original proofs in [2, 124, 129] provided little explanation for, can be alternatively proved with both rigor and intuition by making connections to uniform order statistics. For a special case of $k = 1$, with extra assumptions on the support being

compact, such an elegant proof is provided in [51, Theorem 7.1] which explicitly applies the convergence of the nearest neighbor distance to uniform order statistics. Namely,

$$\mathbb{E}[\widehat{H}(X)] = \mathbb{E}\Big[-\frac{1}{n}\sum_{i=1}^{n} \log\Big(\frac{k}{n\, C_d\, \rho_{k,X_i}^d}\Big)\Big]$$
$$\rightarrow \quad \mathbb{E}\Big[-\log \frac{k\, f(X_i)}{\sum_{j=1}^{k} E_j}\Big] = H(X) + \psi(k) - \log(k), \tag{4.16}$$

where the asymptotic expression follows from $C_d\, n\, f(x)\rho_{k,x}^d \rightarrow \sum_{j=1}^{k} E_j$ as shown, for example, in Lemma 15 and we used $\mathbb{E}[\log \sum_{j=1}^{k} E_j] = \psi(k)$, where $\psi(k) =$ is the digamma function defined as $\psi(x) = \Gamma^{-1}(x)d\Gamma(x)/dx$ and for large x it is approximately $\log(x)$ up to $O(1/x)$, i.e. $\psi(x) = \log x - 1/(2x) + o(1/x)$. Note that this only requires the convergence of the distance and not the direction. Inspired by this modern approach, we extend such a connection in Lemma 15 to prove consistency of our estimator.

4.6.2 Convergence rate of the bias of nearest neighbor based methods

Establishing the convergence rate of the KL estimator is a challenging problem, and is not quite resolved despite work over the past three decades. The $O(1/n)$ convergence rate of the *variance* is established in [132, 130, 51, 35] under various assumptions. Establishing the convergence rate of the *bias* is more challenging. It has been first studied in [53, 133], where root-n consistency is shown in one-dimension with bounded support and assuming $f(x)$ is bounded below. Tsybakov and van der Meulen [32] are the first to prove a root mean squared error convergence rate of $O(1/\sqrt{n})$ for general densities with unbounded support in one-dimension and exponentially decaying tail, such as the Gaussian density. These assumptions are relaxed in [58], where zeroes and fat tails are allowed in $f(x)$. In general d-dimensions, [8, 131] prove bounds on the convergence rate of the bias for finite $k = O(1)$, and [134, 36] for $k = \Omega(\log n)$. Recent papers show that the convergence rate of the bias can reach the minimax lower bound up to a poly-logarithm factor, by either kernel method [49], as well as nearest neighbor method [3]. Establishing the convergence rate for the bias of the proposed local estimator is

an interesting open problem – it is interesting to see if the superior empirical performance of the local estimator is captured in the asymptotics of rate of convergence of the bias.

It is intuitive that kernel density estimators can capture the structure in the distribution if the distribution lies on a lower-dimensional manifold. This is made precise in [135], which also shows improved convergence rates for distributions whose support is on low dimensional manifolds. However, the estimator in [135] critically uses the geodesic distances between the sample points on the manifold. Given that the proposed estimators fit distributions locally, a concrete question of interest is whether such an improvement can be achieved *without* such an explicit knowledge of the geodesic distances, i.e., whether the local estimators automatically adapt to underlying lower-dimensional structures.

4.6.3 Ensemble estimators

Recent works [116, 117, 119, 36] have proposed ensemble estimators, which use known estimators based on kernel density estimators and k-NN methods and construct a new estimate by taking the weighted linear combination of those methods with varying bandwidth or k, respectively. With a proper choice of the weights, which can be computed analytically by solving a simple linear program, a boosting of the convergence rate can be achieved. The key property that allows the design of such ensemble estimators is that the leading terms (in terms of the sample size n) of the bias have a multiplicative constant that only depends on the unknown distribution. An intuitive explanation for this phenomenon is provided in [36] in the context of k-NN methods; it is interesting to explore if such a phenomenon continues in the k-LNN scenario studied in this chapter. Such a study would potentially lead to ensemble-based estimators in the local setting and also naturally allow a careful understanding of the rate of convergence of the bias term.

4.7 Proofs of results in Chapter 4

4.7.1 Proof of Proposition 1

We first prove the derivation of the LLDE with degree $p = 2$ in (4.7). The gradient of the local likelihood evaluated at the maximizer is zero [7], which gives a computational tool for finding the maximizer:

$$\begin{aligned} & \frac{1}{n}\sum_{j=1}^{n} K(\frac{X_j - x}{h}) \\ = & \int K(\frac{u-x}{h}) e^{a_0 + a_1^T(u-x) + (u-x)^T a_2 (u-x)} du, \qquad (4.17) \\ & \frac{1}{n}\sum_{j=1}^{n} \frac{X_j - x}{h} K(\frac{X_j - x}{h}) \\ = & \int \frac{u-x}{h} K(\frac{u-x}{h}) e^{a_0 + a_1^T(u-x) + (u-x)^T a_2 (u-x)} du, \qquad (4.18) \\ & \frac{1}{n}\sum_{j=1}^{n} \frac{(X_j - x)(X_j - x)^T}{h^2} K(\frac{X_j - x}{h}) \\ = & \int \frac{(u-x)(u-x)^T}{h^2} K(\frac{u-x}{h}) e^{a_0 + a_1^T(u-x) + (u-x)^T a_2 (u-x)} du, \qquad (4.19) \end{aligned}$$

where $K(x) = \exp\{-\|x\|^2/2\}$ is the Gaussian kernel. Notice that the left-hand side of the equations are S_0/n, S_1/n and S_2/n, respectively. The RHS can be written in closed forms as:

$$\begin{aligned} \frac{1}{n} S_0 &= (2\pi)^{d/2} |M|^{-1/2} e^{a_0 + \frac{1}{2} a_1^T M^{-1} a_1}, \qquad (4.20) \\ \frac{1}{n} S_1 &= \frac{1}{nh} S_0 M^{-1} a_1, \qquad (4.21) \\ \frac{1}{n} S_2 &= \frac{1}{nh^2} S_0 (M^{-1} + M^{-1} a_1 a_1^T M^{-1}), \qquad (4.22) \end{aligned}$$

where $M = h^{-2} I_{d\times d} - 2a_2$ assuming h sufficiently small such that M is positive definite. We want to derive $\widehat{f}(x) = \exp\{a_0\}$ from the equations. From (4.21) we get $M^{-1}a_1 = S_1(h/S_0)$. Together with (4.22), we get $M^{-1} + M^{-1}a_1 a_1^T M^{-1} = S_2(h^2/S_0)$. Hence, $M^{-1} = (S_2/S_0 - (S_1/S_0)(S_1/S_0)^T)h^2 = h^2\Sigma$. Plug them in (4.20), we obtain the desired expression.

Analogously, for the derivation of the LLDE with degree $p = 1$ in (4.5),

we get

$$\frac{1}{n}S_0 = (2\pi)^{d/2}h^d e^{a_0+\frac{h^2}{2}a_1^T a_1}, \qquad \frac{1}{n}S_1 = \frac{h}{n}S_0 a_1. \tag{4.23}$$

This gives

$$a_1 = \frac{S_1}{hS_0}, \qquad e^{a_0} = \frac{S_0}{n(2\pi)^{d/2}h^d}\,\exp\{\frac{-\|S_1\|^2}{2S_0^2}\}. \tag{4.24}$$

4.7.2 Proof of Lemma 15

Let us introduce some notations and terminologies first. Define $\mathbb{S}^{d-1} \equiv \{x \in \mathbb{R}^d : \|x\| = 1\}$ as the unit $(d-1)$-dimensional sphere and σ^{d-1} as a normalized spherical measure on $\mathbb{S}^{d-1}$. For any $x \in \mathbb{R}^d$ in the Cartesian coordinate system, let $(x_r, x_\theta) \in \mathbb{R}_+ \times \mathbb{S}^{d-1}$ be its representation in the polar coordinate system. Conversely, for any angle $\theta \in \mathbb{S}^{d-1}$ and radius $r \in \mathbb{R}_+$, let $r \cdot \theta \in \mathbb{R}^d$ denote the representation in the Cartesian coordinate system. For any vector of angles $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m) \in (\mathbb{S}^{d-1})^m$ and vector of radiuses $\boldsymbol{r} = (r_1, \ldots, r_m) \in \mathbb{R}_+^m$, define $\boldsymbol{\theta} \cdot \boldsymbol{r} \equiv (\theta_1 \cdot r_1, \ldots, \theta_m \cdot r_m) \in \mathbb{R}^{d\times m}$. For any set $B \subseteq \mathbb{R}^{d\times m}$ and any angles $\boldsymbol{\theta} \in (\mathbb{S}^{d-1})^m$, define $B_{\boldsymbol{\theta}} = \{\boldsymbol{r} \in \mathbb{R}_+^m : \boldsymbol{\theta} \cdot \boldsymbol{r} \in B\}$ be the projection of B onto $\boldsymbol{\theta}$. Let $\{\xi_i\}_{i=1}^m$ be i.i.d. random variables uniformly over $\mathbb{S}^{d-1}$. Then for any joint random variables $(W_1, \ldots, W_m) \in \mathbb{R}_+^m$ which are independent with $\{\xi_i\}_{i=1}^m$, we have

$$\begin{aligned} & \Pr\{(\xi_1 W_1, \ldots, \xi_m W_m) \in B\} \\ = & \int_{\boldsymbol{\theta}\in(\mathbb{S}^{d-1})^m} \Pr\{(W_1, \ldots, W_m) \in B_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta}\}\, d(\sigma^{d-1})^m(\boldsymbol{\theta}) \\ \equiv & \mathbb{E}_{\boldsymbol{\theta}}\left[\Pr\{(W_1, \ldots, W_m) \in B_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta}\}\right], \end{aligned} \tag{4.25}$$

here we write $\int_{\boldsymbol{\theta}\in(\mathbb{S}^{d-1})^m} f(\boldsymbol{\theta})\, d(\sigma^{d-1})^m(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\, f(\boldsymbol{\theta})\,]$ for simplicity of notation. Now let $\boldsymbol{Z} = (Z_{1,i}, \ldots, Z_{m,i})$, $\|\boldsymbol{Z}\| = (\|Z_{1,i}\|, \ldots, \|Z_{m,i}\|)$ and let

$\boldsymbol{E} = (E_1^{1/d}, \ldots, (\sum_{\ell=1}^m E_\ell)^{1/d})$, then for any $B \in \mathbb{R}^{d\times m}$,

$$
\begin{aligned}
& \Big| \Pr\big\{ (c_d n f(x))^{1/d}(Z_{1,i}, \ldots, Z_{m,i}) \in B \big\} \\
& \quad - \Pr\left\{ \left(\xi_1 E_1^{1/d}, \ldots, \xi_m (\sum_{\ell=1}^m E_\ell)^{1/d} \right) \in B \right\} \Big| \\
= & \Big| \Pr\big\{ (c_d n f(x))^{1/d} \boldsymbol{Z} \in B \big\} - \mathbb{E}_{\boldsymbol{\theta}} \left[\Pr\{ \boldsymbol{E} \in B_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \} \right] \Big| \\
\leq & \Big| \Pr\big\{ (c_d n f(x))^{1/d} \boldsymbol{Z} \in B \big\} - \mathbb{E}_{\boldsymbol{\theta}} \left[\Pr\{ (c_d n f(x))^{1/d} \|\boldsymbol{Z}\| \in B_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \} \right] \Big| \qquad (4.26) \\
+ & \mathbb{E}_{\boldsymbol{\theta}} \Big[\Big| \Pr\{ (c_d n f(x))^{1/d} \|\boldsymbol{Z}\| \in B_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \} - \Pr\{ \boldsymbol{E} \in B_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \} \Big| \Big]. \qquad (4.27)
\end{aligned}
$$

We will bound the terms (4.26) and (4.27) separately. For the term (4.26), consider the following event $\mathcal{E} = \{\|Z_{m,i}\| < (\sqrt{n} c_d f(x))^{-1/d}\}$. We will first show that $\mathcal{E}$ happens with high probability and then prove that the difference is small when $\mathcal{E}$ happens.

Firstly, we consider the probability that $\mathcal{E}$ does not happen, which means $\|Z_{m,i}\| \geq (\sqrt{n} c_d f(x))^{-1/d}$. Denote $B(x, r) = \{z : \|z - x\| \leq r\}$ and let $p = \int_{B(x, \|Z_{m,i}\|)} f(t) dt$ be the probability mass of the ball centered at x with radius $\|Z_{m,i}\|$. $\mathcal{E}$ does not happen if and only if there are at most $m-1$ (out of n) samples lying in $B(x, \|Z_{m,i}\|)$. So we need a lower bound for p to get an upper bound for the probability of $\mathcal{E}^C$. Since there exist constants $\epsilon > 0$ such that $\|\nabla f(a)\| = O(1)$ for any $\|a - x\| \leq \epsilon$, for sufficiently large n such that $(\sqrt{n} c_d f(x))^{-1/d} \leq \epsilon$, we have $f(t) \geq f(x) - C_1 \|t - x\| \geq f(x) - C_1 (\sqrt{n} c_d f(x))^{-1/d}$ for any $t \in B(x, (\sqrt{n} c_d f(x))^{-1/d})$, where $C_1 = \sup_{\|a-x\| \leq (\sqrt{n} c_d f(x))^{-1/d}} \|\nabla f(a)\|$. Therefore,

$$
\begin{aligned}
& p = \int_{B(x, \|Z_{m,i}\|)} f(t) dt \geq \int_{B(x, (\sqrt{n} c_d f(x))^{-1/d})} f(t) dt \\
\geq & \int_{B(x, (\sqrt{n} c_d f(x))^{-1/d}))} (f(x) - C_1 (\sqrt{n} c_d f(x))^{-1/d})) dt \\
= & (f(x) - C (\sqrt{n} c_d f(x))^{-1/d})) c_d (\sqrt{n} c_d f(x))^{-1/d})^d \\
= & \frac{1}{\sqrt{n}} - C_1 c_d^{-1/d} f(x)^{-(d+1)/d} n^{-(d+1)/2d}. \qquad (4.28)
\end{aligned}
$$

For sufficiently large n such that $C_1 c_d^{-1/d} f(x)^{-(d+1)/d} n^{-(d+1)/2d} \leq 1/(2\sqrt{n})$,

we have $p \geq 1/(2\sqrt{n})$. Therefore,

$$\begin{aligned} & \Pr\{\|Z_{m,i}\| \geq (\sqrt{n}c_d f(x))^{-1/d}\} \\ = & \sum_{\ell=0}^{m-1} \binom{n}{\ell} p^\ell (1-p)^{n-\ell} \leq \sum_{\ell=0}^{m-1} n^\ell \Big(1 - \frac{1}{2\sqrt{n}}\Big)^{(n-\ell)} \\ \leq & \sum_{\ell=0}^{m-1} n^l e^{-(\sqrt{n}-\ell\sqrt{n})/2} \leq m n^m e^{-(\sqrt{n}-m/\sqrt{n})/2}. \end{aligned} \tag{4.29}$$

Now we consider when $\mathcal{E}$ happens, which means $\|Z_{m,i}\| < (\sqrt{n}c_d f(x))^{-1/d}$. Denote $\overline{B} = \{t : (c_d n f(x))^{1/d} t \in B \text{ and } \|t_m\| < (\sqrt{n}c_d f(x))^{-1/d}\}$ be the truncated scaling of B. Similarly, denote $\overline{B_{\boldsymbol{\theta}}} = \{t : (c_d n f(x))^{1/d} t \in B_{\boldsymbol{\theta}} \text{ and } t_m < (\sqrt{n}c_d f(x))^{-1/d}\}$ be the truncated scaling of $B_{\boldsymbol{\theta}}$. Note that for any $\overline{B}$, the probability that $\boldsymbol{Z} \in \overline{B}$ is the integration of the density of $Z_{1,i}, \ldots, Z_{m,i}$ in B, multiplied by the probability that $Z_{m+1,i}, \ldots, Z_{n,i}$ lying outside the ball $B(x, (\sqrt{n}c_d f(x))^{-1/d})$, therefore,

$$\begin{aligned} & \Pr\{\boldsymbol{Z} \in \overline{B}, \mathcal{E}\} \\ = & \frac{n!}{(n-m)!} \int_{t \in \overline{B}} \left(\prod_{j=1}^{m} f(x + t_j) \right) (\Pr\{\|X - x\| > \|t_m\|\})^{n-m} \, dt \\ = & \frac{n!}{(n-m)!} \int_{t \in \overline{B}} \left(\prod_{j=1}^{m} f(x + t_j) \right) d\mu(t), \end{aligned} \tag{4.30}$$

where the measure $\mu(t)$ satisfies $d\mu(t)/dt = (\Pr\{\|X - x\| > \|t_m\|\})^{n-m}$. Similarly for any $\overline{B_{\boldsymbol{\theta}}}$,

$$\Pr\{\|\boldsymbol{Z}\| \in \overline{B_{\boldsymbol{\theta}}}\} = \frac{n!}{(n-m)!} \int_{t \in \overline{B_{\boldsymbol{\theta}}}} \left(\prod_{j=1}^{m} f(x + t_j \cdot \theta_j) \right) d\mu(t). \tag{4.31}$$

When $\mathcal{E}$ happens, the first term (4.26) can be rewritten by

$$\begin{aligned} & \Big| \Pr\big\{ (c_d n f(x))^{1/d} \boldsymbol{Z} \in B, \mathcal{E} \big\} - \mathbb{E}_{\boldsymbol{\theta}} \big[\Pr\{(c_d n f(x))^{1/d} \|\boldsymbol{Z}\| \in B_{\boldsymbol{\theta}}, \mathcal{E} \,|\, \boldsymbol{\theta}\} \big] \Big| \\ = & \left| 1 - \underbrace{\frac{\mathbb{E}_{\boldsymbol{\theta}} \big[\Pr\{(c_d n f(x))^{1/d} \|\boldsymbol{Z}\| \in B_{\boldsymbol{\theta}}, \mathcal{E} \,|\, \boldsymbol{\theta}\} \big]}{\Pr\big\{ (c_d n f(x))^{1/d} \boldsymbol{Z} \in B, \mathcal{E} \big\}}}_{\mathcal{R}} \right| \\ & \times \Pr\big\{ (c_d n f(x))^{1/d} \boldsymbol{Z} \in B, \mathcal{E} \big\}. \end{aligned} \tag{4.32}$$

Clearly $\Pr\left\{ (c_d n f(x))^{1/d} \boldsymbol{Z} \in B, \mathcal{E} \right\} \leq 1$. And the ratio $\mathcal{R}$ can be bounded by,

$$\begin{aligned}
\mathcal{R} &= \frac{\mathbb{E}_{\boldsymbol{\theta}} \left[\Pr\{ \|\boldsymbol{Z}\| \in \overline{B_{\boldsymbol{\theta}}}, \mathcal{E} \,|\, \boldsymbol{\theta} \} \right]}{\Pr\{ \boldsymbol{Z} \in \overline{B}, \mathcal{E} \}} \\
&= \frac{\frac{n!}{(n-m)!} \mathbb{E}_{\boldsymbol{\theta}} \left[\int_{t \in \overline{B_{\boldsymbol{\theta}}}} \left(\prod_{j=1}^m f(x + t_j \cdot \theta_j) \right) d\mu(t) \right]}{\frac{n!}{(n-m)!} \int_{t \in \overline{B}} \left(\prod_{j=1}^m f(x + t_j) \right) d\mu(t)} \\
&\leq \frac{\sup_{\boldsymbol{\theta} \in (\mathbb{S}^{d-1})^m} \sup_{t \in \overline{B_{\boldsymbol{\theta}}}} \prod_{j=1}^m f(x + t_j \cdot \theta_j)}{\inf_{t \in \overline{B}} \prod_{j=1}^m f(x + t_j)} \\
&\leq \left(\frac{\sup_{\|t\| \leq (\sqrt{n} c_d f(x))^{-1/d}} f(x+t)}{\inf_{\|t\| \leq (\sqrt{n} c_d f(x))^{-1/d}} f(x+t)} \right)^m . \qquad (4.33)
\end{aligned}$$

Since f is continuously differentiable, by mean value theorem, there exists $a, b \in B(x, (\sqrt{n} c_d f(x))^{-1/d})$ such that

$$\begin{aligned}
&\frac{\sup_{\|t\| \leq (\sqrt{n} c_d f(x))^{-1/d}} f(x+t)}{\inf_{\|t\| \leq (\sqrt{n} c_d f(x))^{-1/d}} f(x+t)} = \frac{f(b) + (a-b)^T \nabla f(a)}{f(b)} \\
\leq\ & 1 + \frac{2(\sqrt{n} c_d f(x))^{-1/d} \|\nabla f(a)\|}{f(b)}. \qquad (4.34)
\end{aligned}$$

By the assumption, there exists a ball $B(x, \varepsilon)$ such that $\|\nabla f(a)\| = O(1)$ and $f(a) > 0$ for all $a \in B(x, \varepsilon)$, so for sufficiently large n such that $(\sqrt{n} c_d f(x))^{-1/d} < \varepsilon$, there exists some constant C and $c > 0$ such that $\sup_{a \in B(x, (\sqrt{n} c_d f(x))^{-1/d})} \|\nabla f(a)\| \leq C$ and $\inf_{b \in B(x, (\sqrt{n} c_d f(x))^{-1/d})} f(b) \geq c$. So

$$\begin{aligned}
&\mathcal{R} \leq (1 + \frac{2(\sqrt{n} c_d f(x))^{-1/d} \|\nabla f(a)\|}{f(b)})^m \\
\leq\ & (1 + \frac{2 n^{-1/(2d)} c_d^{-1/d} C}{c^{1+1/d}}) = (1 + C_2 n^{-1/(2d)})^m, \qquad (4.35)
\end{aligned}$$

for some constant C_2. Similarly, (4.33) is lower bounded by $(1 - C_2 n^{-1/(2d)})^m$. So $|1 - \mathcal{R}| \leq \max\{(1 + C_2 n^{-1/(2d)})^m - 1, 1 - (1 - C_2 n^{-1/(2d)})^m\}$.

Combining with (4.29), the first term (4.26) is bounded by:

$$\begin{aligned}
& \Big| \Pr\big\{ (c_d n f(x))^{1/d} \boldsymbol{Z} \in B \big\} - \mathbb{E}_{\boldsymbol{\theta}} \big[\Pr\{(c_d n f(x))^{1/d} \|\boldsymbol{Z}\| \in B_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta}\} \big] \Big| \\
\leq \; & \Pr\big\{ (c_d n f(x))^{1/d} Z \in B, \mathcal{E}^C \big\} + \mathbb{E}_{\boldsymbol{\theta}} \big[\Pr\{(c_d n f(x))^{1/d} \|Z\| \in B_{\boldsymbol{\theta}}, \mathcal{E}^C \,|\, \boldsymbol{\theta}\} \big] \\
+ \; & \Big| \Pr\big\{ (c_d n f(x))^{1/d} Z \in B, \mathcal{E} \big\} - \mathbb{E}_{\boldsymbol{\theta}} \big[\Pr\{(c_d n f(x))^{1/d} \|Z\| \in B_{\boldsymbol{\theta}}, \mathcal{E} \,|\, \boldsymbol{\theta}\} \big] \Big| \\
\leq \; & \Pr\{\mathcal{E}^C\} + \mathbb{E}_{\boldsymbol{\theta}} \big[\Pr\{\mathcal{E}^C\} \big] + \Pr\big\{ (c_d n f(x))^{1/d} \boldsymbol{Z} \in B, \mathcal{E} \big\} |1 - \mathcal{R}| \\
\leq \; & \max\{(1 + C_2 n^{-1/(2d)})^m - 1, 1 - (1 - C_2 n^{-1/(2d)})^m\} \\
& + 2mn^m e^{-(\sqrt{n} - m/\sqrt{n})/2}. \qquad (4.36)
\end{aligned}$$

Now consider the second term (4.27). We will use Corollary 5.5.5 of [127] to show that this term vanishes for $m = O(\log n)$ and as n grows.

Lemma 16 (Corollary 5.5.5, [127])**.** *Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. samples from unknown distribution with pdf $\tilde{f}$. Let $Y_{1:n} \leq Y_{2:n} \leq \cdots \leq Y_{n:n}$ be the order statistics. Assume the density $\tilde{f}$ satisfies $|\log \tilde{f}(y)| \leq L y^{\delta}$ for $0 < y < y_0$ and $\tilde{f}(y) = 0$ for $y < 0$, where L and δ are constants. Then*

$$\begin{aligned}
& d_{\mathrm{TV}}\Big(n \, (Y_{1:n}, \ldots, Y_{m:n}) \, , \big(E_1, \ldots, \sum_{j=1}^{m} E_j \big) \Big) \\
\leq \; & C_0 \big((m/n)^{\delta} m^{1/2} + m/n \big) \, , \qquad (4.37)
\end{aligned}$$

where $C_0 > 0$ is a constant. $E_1, \ldots, E_m$ are i.i.d standard exponential random variables.

Now for fixed x, consider the distribution of $c_d f(x) \|X - x\|^d$ denoted by $\tilde{P}$. Define $Y_1, Y_2, \ldots, Y_n$ drawn i.i.d. from $\tilde{P}$. We can see that $c_d f(x) \|Z\|^d \stackrel{\mathcal{L}}{=} (Y_{1:n}, \ldots, Y_{m:n})$, where $\stackrel{\mathcal{L}}{=}$ denotes equivalence in distribution. We first prove that the pdf $\tilde{f}$ of $\tilde{P}$ satisfies the assumption in Lemma 16. Firstly, it is obvious that $\tilde{f}(t) = 0$ for $t < 0$. For $t > 0$, the pdf $\tilde{f}$ of $\tilde{P}$ is given by:

$$\tilde{f}(t) \;=\; \frac{d}{dt} \Pr\{c_d f(x) \|X - x\|^d \leq t\} = \frac{d}{dt} \int_{y \in B(x, r_t)} f(y) dy, \quad (4.38)$$

where $r_t = (t/(c_d f(x)))^{1/d}$. Here we have

$$\frac{dr_t}{dt} = \frac{t^{1/d - 1} (c_d f(x))^{-1/d}}{d} = \frac{1}{f(x) d c_d r_t^{d-1}}, \qquad (4.39)$$

and

$$\begin{aligned}
& \frac{d}{dr_t}\int_{y\in B(x,r_t)} f(y)dy \\
= & \frac{d}{dr_t}\int_{\theta\in\mathbb{S}^{d-1}}\left(\int_{r=0}^{r_t} f(x+r\cdot\theta)r^{d-1}dr\right)dc_d\, d\sigma^{d-1}(\theta) \\
= & dc_d\int_{\theta\in\mathbb{S}^{d-1}}\left(\frac{d}{dr_t}\int_{r=0}^{r_t} f(x+r\cdot\theta)r^{d-1}dr\right)d\sigma^{d-1}(\theta) \\
= & dc_d r_t^{d-1}\int_{\theta\in\mathbb{S}^{d-1}} f(x+r_t\cdot\theta)d\sigma^{d-1}(\theta). \qquad (4.40)
\end{aligned}$$

Therefore, $\tilde{f}(t) = \int_{\theta\in\mathbb{S}^{d-1}} f(x+r_t\cdot\theta)d\sigma^{d-1}(\theta)/f(x)$. Since f is twice continuously differentiable, by mean value theorem, for any $y \in B(x,r_t)$, there exists $a(y) \in B(x,r_t)$ such that $f(y) - f(x) = (y-x)^T\nabla f(x) + (a(y)-x)^T H_f(a(y))(a(y)-x)$, where $a(y)$ depends on y. Therefore,

$$\begin{aligned}
& |\tilde{f}(t) - 1| \\
\leq & \frac{1}{f(x)}\left|\int_{\theta\in\mathbb{S}^{d-1}} f(x+r_t\cdot\theta)d\sigma^{d-1}(\theta) - f(x)\right| \\
= & \frac{1}{f(x)}\left|\int_{\theta\in\mathbb{S}^{d-1}} (\, f(x+r_t\cdot\theta) - f(x)\,)\, d\sigma^{d-1}(\theta)\right| \\
\leq & \frac{1}{f(x)}\left|\int_{\theta\in\mathbb{S}^{d-1}} (r_t\cdot\theta)^T\nabla f(x)d\sigma^{d-1}(\theta)\right| + \frac{1}{f(x)}\left|\int_{\theta\in\mathbb{S}^{d-1}} (a(x+r_t\cdot\theta)-x)^T\right. \\
& \left. H_f(a(x+r_t\cdot\theta))(a(x+r_t\cdot\theta)-x)d\sigma^{d-1}(\theta)\right| \\
\leq & \frac{\int_{\theta\in\mathbb{S}^{d-1}}\left(\sup_{a\in B(x,r_t)}\|H_f(a)\|\,\|a-x\|^2\right)d\sigma^{d-1}(\theta)}{f(x)} \\
\leq & \frac{r_t^2}{f(x)}\left(\sup_{a\in B(x,r_t)}\|H_f(a)\|\right). \qquad (4.41)
\end{aligned}$$

Since there exists a ball $B(x,\varepsilon)$ such that $\|H_f(a)\| = O(1)$ for all $a \in B(x,\varepsilon)$. Therefore, for sufficiently small t such that $r_t < \varepsilon$, there exists $C_3 > 0$ such that $|\tilde{f}-1| \leq C_3 r_t^2/f(x)$. Recall that $r_t = (t/(c_d f(x)))^{1/d}$, so there exists $L > 0$ such that $|\tilde{f}(t) - 1| \leq Lt^{2/d}$ for sufficiently small t. Hence, $|\log\tilde{f}(t)| \leq L't^{2/d}$ for some $L' > 0$ and sufficiently small t. So $\tilde{f}$ satisfies the

condition in Lemma 16 with $\delta = 2/d$. Therefore, for any $B_{\boldsymbol{\theta}} \subseteq \mathbb{R}_+^m$, we have:

$$\begin{aligned}
& \left| \Pr\{(c_d n f(x))^{1/d} \|\boldsymbol{Z}\| \in B_{\boldsymbol{\theta}}\} - \Pr\{\boldsymbol{E} \in B_{\boldsymbol{\theta}}\} \right| \\
\leq \; & d_{\mathrm{TV}}\Big(c_d n f(x) \left(\|Z_{1,i}\|^d, \ldots, \|Z_{m,i}\|^d \right), \left(E_1, E_1 + E_2, \ldots, \sum_{j=1}^{m} E_j \right) \Big) \\
\leq \; & C_0 \left((\frac{m}{n})^{2/d} m^{1/2} + \frac{m}{n} \right). \qquad (4.42)
\end{aligned}$$

Therefore, by combing (4.36) and (4.42), we have:

$$\begin{aligned}
& \Big| \Pr\left\{ (c_d n f(x))^{1/d} \left(Z_{1,i}, \ldots, Z_{m,i} \right) \in B \right\} \\
& \quad - \Pr\left\{ \left(\xi_1 E_1^{1/d}, \ldots, \xi_m (\sum_{l=1}^{m} E_\ell)^{1/d} \right) \in B \right\} \Big| \\
\leq \; & \max\{(1 + C_2 n^{-\frac{1}{2d}})^m - 1, 1 - (1 - C_2 n^{-\frac{1}{2d}})^m\} \\
& \quad + 2mn^m e^{-\frac{\sqrt{n} - m/\sqrt{n}}{2}} + C_0 \left((\frac{m}{n})^{\frac{2}{d}} m^{\frac{1}{2}} + \frac{m}{n} \right), \qquad (4.43)
\end{aligned}$$

for any set $B \in \mathbb{R}^{d \times m}$. Therefore, the total variation distance between $(c_d n f(x))^{1/d}(Z_{1,i}, Z_{2,i}, \ldots, Z_{m,i})$ and $(\xi_1 E_1^{1/d}, \ldots, \xi_m (\sum_{\ell=1}^{m} E_\ell)^{1/d})$ is bounded by the RHS quantity. By taking $m = O(\log n)$, the RHS converges to 0 as n goes to infinity. Therefore, we have the desired statement.

4.7.3 Proof of Theorem 8

The proof of Theorem 8 is organized as follows:

- First we prove that the estimator is asymptotically unbiased, by rewriting the estimator as $\widehat{H}_k^{(n)} = (1/n) \sum_{i=1}^{n} (h_i - \log f(X_i))$, where h_i is a function of the nearest neighbor statistics $\{Z_{1,i}, Z_{2,i}, \ldots\}$. By Lemma 15, the nearest neighbor statistics converge to some standard random variables jointly, so the h function converges to a certain quantity $B_{k,d}$, whereas $(1/n) \sum_{i=1}^{n} (- \log f(X_i))$ converges to $H(X)$.

- Then we prove that the variance of the estimator is vanishing. We give an upper bound of how much the estimate will change if we change one sample X_i to X_i', and utilize Efron-Stein inequality to give an upper bound of the variance.

Proof of bias

We first compute the asymptotic bias. We define new notations to represent the estimate as $\widehat{H}_k^{(n)} = (1/n)\sum_{i=1}^n H_i$, where

$$H_i \;=\; h\big(\,(c_d n f(X_i))^{1/d} Z_{k,i}, S_{0,i}, S_{1,i}, S_{2,i})\,\big) - \log f(X_i), \tag{4.44}$$

and $h : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^{d\times d} \to \mathbb{R}$ is defined as

$$\begin{aligned} h(t_1,t_2,t_3,t_4) &= d\log\|t_1\| + d\log(2\pi) - \log c_d - \log t_2 \\ &+\frac{1}{2}\log\left(\det\left(\frac{t_4}{t_2} - \frac{t_3 t_3^T}{t_2^2}\right)\right) + \frac{1}{2} t_3^T (t_4 - t_3 t_3^T)^{-1} t_3. \end{aligned} \tag{4.45}$$

Let $H_i \equiv h((c_d n f(X_i))^{1/d} Z_{k,i}, S_{0,i}, S_{1,i}, S_{2,i})) - \log f(X_i)$. Since the terms $H_1, H_2, \ldots, H_n$ are identically distributed, so $\mathbb{E}[\widehat{H}_k^{(n)}]$ converges to

$$\lim_{n\to\infty} \mathbb{E}[\widehat{H}_k^{(n)}] \;=\; \lim_{n\to\infty} \mathbb{E}[H_1] = \lim_{n\to\infty} \mathbb{E}_{X_1}\big[\mathbb{E}[H_1|X_1]\big]. \tag{4.46}$$

The typical approaches of dominated convergence theorem cannot be applied to the above limit, since analyzing $\mathbb{E}[H_1|X_1]$ for finite sample n is challenging. In order to exchange the limit with the (conditional) expectation, we assume the following Ansatz 1 to be true. As noted in [113] this is common in the literature on consistency of k-NN estimators, where the same assumptions have been implicitly made without explicitly stating as such, in existing analyses of entropy estimators including [2, 129, 130, 75]. This assumption can be avoided for Renyi entropy as in the proof of consistency in [113] or for sharper results such as the convergence rate of the bias with respect to the sample size but with more assumptions as in [8, 131, 36].

Ansatz 1. *The following exchange of limit holds if* $\mathbb{E}\left[|\log f(X)|\right] < \infty$

$$\lim_{n\to\infty} \mathbb{E}[H_1] \;=\; \mathbb{E}_{X_1}\left[\lim_{n\to\infty} \mathbb{E}[H_1|X_1]\right]. \tag{4.47}$$

Under this ansatz, perhaps surprisingly, we will show that the expectation inside converges to $-\log f(X_1)$ plus some bias that is independent of the underlying distribution. Precisely, for almost every x and given $X_1 = x$,

$$\begin{aligned} &\mathbb{E}[H_1|X_1 = x] + \log f(x) \\ =\;& \mathbb{E}\left[\,h((c_d n f(x))^{1/d} Z_{k,i}, S_{0,1}, S_{1,i}, S_{2,i})\,\right] \longrightarrow B_{k,d}, \end{aligned} \tag{4.48}$$

as $n \to \infty$ where $B_{k,d}$ is a constant that only depends on k and d, defined in (4.50). This implies that

$$\begin{aligned} & \mathbb{E}_{X_1}\Big[\lim_{n\to\infty}\mathbb{E}[H_1|X_1]\Big] = \mathbb{E}_{X_1}[-\log f(X_1) + B_{k,d}] \\ = \;& H(X) + B_{k,d}. \end{aligned} \tag{4.49}$$

Together with (4.46), this finishes the proof of the desired claim.

We are now left to prove the convergence of (4.48). We first give a formal definition of the bias $B_{k,d}$ by replacing the sample defined quantities by a similar quantities defined from order-statistics, and use Lemma 15 to prove the convergence. Recall that our order-statistics is defined by two sequences of m i.i.d. random variables: i.i.d. standard exponential random variables $E_1, \ldots, E_m$ and i.i.d. random variables $\xi_1, \ldots, \xi_m$ uniformly distributed over $\mathbb{S}^{d-1}$. We define

$$B_{k,d} \;\equiv\; \mathbb{E}\left[h\left(\xi_k \left(\sum_{\ell=1}^{k} E_\ell \right)^{\frac{1}{d}}, \tilde{S}_0^{(\infty)}, \tilde{S}_1^{(\infty)}, \tilde{S}_2^{(\infty)} \right) \right], \tag{4.50}$$

where, as we will show, $\tilde{S}_\alpha^{(\infty)}$ is the limit of empirical quantity $S_{\alpha,i}$ defined from samples for each $\alpha \in \{0,1,2\}$, and we know that $(c_d n f(x))^{1/d} Z_{k,i}$ converges to $\xi_k(\sum_{\ell=1}^k E_\ell)^{1/d}$ for almost every x from Lemma 15. $S^{(\infty)}$ is defined by a convergent random sequence

$$\tilde{S}_\alpha^{(m)} \;\equiv\; \sum_{j=1}^{m} \frac{\xi_j^{(\alpha)} (\sum_{\ell=1}^{j} E_\ell)^{\frac{\alpha}{d}}}{(\sum_{\ell=1}^{k} E_\ell)^{\frac{\alpha}{d}}} \exp\Big\{ -\frac{(\sum_{\ell=1}^{j} E_\ell)^{\frac{2}{d}}}{2(\sum_{\ell=1}^{k} E_\ell)^{\frac{2}{d}}} \Big\}, \tag{4.51}$$

where $\xi_j^{(0)} = 1$, $\xi_j^{(1)} = \xi_j$, $\xi_j^{(2)} = \xi_j \xi_j^T$ and $\tilde{S}_\alpha^{(\infty)} = \lim_{m\to\infty} \tilde{S}_\alpha^{(m)}$. This limit exists, since $\tilde{S}_0^{(m)}$ is non-decreasing in m, and the convergence of $\tilde{S}_1^{(m)}$ and $\tilde{S}_2^{(m)}$ follows from Lemma 17. We introduce simpler notations for the joint random variables: $\tilde{S}^{(m)} = (\xi_k(\sum_{\ell=1}^k E_\ell)^{1/d}, \tilde{S}_0^{(m)}, \tilde{S}_1^{(m)}, \tilde{S}_2^{(m)})$ and $\tilde{S}^{(\infty)} = (\xi_k(\sum_{\ell=1}^k E_\ell)^{1/d}, \tilde{S}_0^{(\infty)}, \tilde{S}_1^{(\infty)}, \tilde{S}_2^{(\infty)})$. Considering the following quantities $S^{(n)} = ((c_d n f(x))^{1/d} Z_{k,i}, S_{0,i}, S_{1,i}, S_{2,i})$ defined from samples, we show that this converges to $\tilde{S}^{(\infty)}$ in distribution. For any set $A \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^{d\times d}$, we need to prove that $|\Pr\{S^{(n)} \in A\} - \Pr\{\tilde{S}^{(\infty)} \in A\}|$ converges to 0 as $n \to \infty$. By

applying triangular inequality,

$$\begin{aligned} & |\Pr\{S^{(n)} \in A\} - \Pr\{\tilde{S}^{(\infty)} \in A\}| \\ \leq \; & |\Pr\{S^{(n)} \in A\} - \Pr\{\tilde{S}^{(m)} \in A\}| + |\Pr\{\tilde{S}^{(m)} \in A\} - \Pr\{\tilde{S}^{(\infty)} \in A\}|, \end{aligned} \tag{4.52}$$

and we show that both terms converge to zero for any $m = \Theta(\log n)$. Given that g is continuous and bounded, this implies that

$$\begin{aligned} & \lim_{n\to\infty} \mathbb{E}[H_1|X_1 = x] = \mathbb{E}[\lim_{n\to\infty} g(S^{(n)}) - \log f(x)|X_1 = x] \\ = \; & -\log f(x) + \mathbb{E}[h(\tilde{S}^{(\infty)})], \end{aligned} \tag{4.53}$$

for almost every x, proving (4.49).

The convergence of the first term follows from Lemma 15. Precisely, consider the function $g_m : \mathbb{R}^{d\times m} \to \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^{d\times d}$ defined as:

$$\begin{aligned} & g_m(t_1, t_2, \ldots, t_m) = \Big(t_k, \sum_{j=1}^{m} \exp\{-\frac{\|t_j\|^2}{2\|t_k\|^2}\}, \\ & \sum_{j=1}^{m} \frac{t_j}{\|t_k\|} \exp\{-\frac{\|t_j\|^2}{2\|t_k\|^2}\}, \sum_{j=1}^{m} \frac{t_j t_j^T}{\|t_k\|^2} \exp\{-\frac{\|t_j\|^2}{2\|t_k\|^2}\} \Big), \end{aligned} \tag{4.54}$$

such that $S^{(n)} = g_m\left((c_d n f(x))^{1/d} (Z_{1,i}, Z_{2,i}, \ldots, Z_{m,i}) \right)$, which follows from the definition of $S^{(n)} = ((c_d n f(x))^{1/d} Z_{k,i}, S_{0,i}, S_{1,i}, S_{2,i})$ in (4.10). Similarly, $\tilde{S}^{(m)} = g_m\left(\xi_1 E_1^{1/d}, \xi_2 (E_1 + E_2)^{1/d}, \ldots \xi_m (\sum_{\ell=1}^{m} E_\ell)^{1/d} \right)$. Since g_m is continuous, so for any set $A \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^{d\times d}$, there exists a set $\tilde{A} \in \mathbb{R}^{d\times m}$ such that $g_m(\tilde{A}) = A$. So for any x such that there exists $\varepsilon > 0$ such that $f(a) > 0$, $\|\nabla f(a)\| = O(1)$ and $\|H_f(a)\| = O(1)$ for any $\|a - x\| < \varepsilon$, we

have:

$$
\begin{aligned}
& |\Pr\{S^{(n)} \in A\} - \Pr\{\tilde{S}^{(m)} \in A\}| \\
= \quad & \Big| \Pr\Big\{ g_m\Big((c_d n f(x))^{\frac{1}{d}} (Z_{1,i}, \ldots, Z_{m,i}) \Big) \in A \Big\} \\
& - \Pr\{ g_m(\xi_1 E_1^{\frac{1}{d}}, \ldots \xi_m (\sum_{l=1}^{m} E_\ell)^{\frac{1}{d}}) \in A \} \Big| \\
= \quad & \Big| \Pr\Big\{ \Big((c_d n f(x))^{\frac{1}{d}} (Z_{1,i}, \ldots, Z_{m,i}) \Big) \in \tilde{A} \Big\} \\
& - \Pr\{ (\xi_1 E_1^{1/d}, \ldots \xi_m (\sum_{\ell=1}^{m} E_\ell)^{1/d}) \in \tilde{A} \} \Big| \\
\leq \quad & d_{\mathrm{TV}}\Big(\Big((c_d n f(x))^{\frac{1}{d}} (Z_{1,i}, \ldots, Z_{m,i}) \Big) \Big(\xi_1 E_1^{1/d}, \ldots \xi_m (\sum_{\ell=1}^{m} E_\ell)^{1/d} \Big) \Big) \\
\overset{n\to\infty}{\longrightarrow} \quad & 0, \qquad (4.55)
\end{aligned}
$$

where the last inequality follows from Lemma 15. By the assumption that f has open support and $\|\nabla f\|$ and $\|H_f\|$ is bounded almost everywhere, this convergence holds for almost every x.

For the second term in (4.52), let $\tilde{T}_\alpha^{(m)} = \tilde{S}_\alpha^{(\infty)} - \tilde{S}_\alpha^{(m)}$ and we claim that $\tilde{T}_\alpha^{(m)}$ converges to 0 in distribution by the following lemma.

Lemma 17. *Assume $m \to \infty$ as $n \to \infty$ and $k \geq 3$, then*

$$\lim_{n\to\infty} \mathbb{E}\| \tilde{T}_\alpha^{(m)} \| = 0, \qquad (4.56)$$

for any $\alpha \in \{0, 1, 2\}$. Hence $(\tilde{T}_0^{(m)}, \tilde{T}_1^{(m)}, \tilde{T}_2^{(m)})$ converges to $(0, 0, 0)$ in distribution.

This implies that $(\tilde{S}_0^{(m)}, \tilde{S}_1^{(m)}, \tilde{S}_2^{(m)})$ converges to $(\tilde{S}_0^{(\infty)}, \tilde{S}_1^{(\infty)}, \tilde{S}_2^{(\infty)})$ in distribution, i.e.,

$$|\Pr\{\tilde{S}^{(m)} \in A\} - \Pr\{\tilde{S}^{(\infty)} \in A\}| \overset{n\to\infty}{\longrightarrow} 0. \qquad (4.57)$$

Combine (4.55) and (4.57) in (4.52), and this implies the desired claim.

Proof of variance

We next prove the upper bound on the variance, following the technique from [51, Section 7.3]. For the usage of Efron-Stein inequality, we need a second set of i.i.d. samples $\{X_1', X_2', \ldots, X_n'\}$. For simplicity, denote $\widehat{H} =$

$\widehat{H}^{(n)}_{kLNN}(X)$ be the kLNN estimate base on original sample $\{X_1, \ldots, X_n\}$ and $\widehat{H}^{(i)}$ be the kLNN estimate based on $\{X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots X_n\}$. Then the Efron-Stein theorem states that

$$\mathrm{Var}\left[\widehat{H}\right] \leq 2\sum_{j=1}^{n} \mathbb{E}\left[\left(\widehat{H} - \widehat{H}^{(j)}\right)^2\right]. \tag{4.58}$$

Recall that $\widehat{H}^{(n)}_k = (1/n)\sum_{i=1}^n H_i$, where

$$H_i \;=\; h\big(\,(c_d n f(X_i))^{1/d} Z_{k,i}, S_{0,i}, S_{1,i}, S_{2,i})\,\big) - \log f(X_i), \tag{4.59}$$

and $h : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^{d\times d} \to \mathbb{R}$ is defined as

$$\begin{aligned} h(t_1, t_2, t_3, t_4) &= d\log\|t_1\| + d\log(2\pi) - \log c_d - \log t_2 \\ &+\frac{1}{2}\log\left(\det\left(\frac{t_4}{t_2} - \frac{t_3 t_3^T}{t_2^2}\right)\right) + \frac{1}{2} t_3^T (t_4 - t_3 t_3^T)^{-1} t_3. \end{aligned} \tag{4.60}$$

Similarly, we can write $\widehat{H}^{(j)} = \frac{1}{n}\sum_{i=1}^n H_i^{(j)}$ for any $j \in \{1, \ldots, n\}$. Therefore, the difference of $\widehat{H}$ and $\widehat{H}^{(j)}$ can be bounded by:

$$\widehat{H} - \widehat{H}^{(j)} = \frac{1}{n}\sum_{i=1}^{n}\left(H_i - H_i^{(j)}\right). \tag{4.61}$$

Notice that H_i only depends on X_i and its m nearest neighbors, so $H_i - H_i^{(j)} = 0$ if none of X_j and X'_j are in m nearest neighbor of X_i. If we denote $Z_{i,j} = \mathbb{I}\{X_j \text{ is in } m \text{ nearest neighbor of } X_i\}$, then $H_i = H_i^{(j)}$ if $Z_{i,j} + Z_{i,j'} = 0$. According to [51, Lemma 20.6], since X has a density, with probability one, $\sum_{i=1}^n Z_{i,j} \leq m\gamma_d$, where γ_d is the minimal number of cones of angle $\pi/6$ that can cover $\mathbb{R}^d$, which only depends on d. Similarly, $\sum_{i=1}^n Z_{i,j'} \leq m\gamma_d$. If we denote $S = \{i : Z_{i,j} + Z_{i,j'} > 0\}$, the cardinality of S satisfy $|S| \leq 2m\gamma_d$. Therefore, we have $\widehat{H} - \widehat{H}^{(j)} = \frac{1}{n}\sum_{i\in S}\left(H_i - H_i^{(j)}\right)$. By Cauchy-Schwarz

inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\widehat{H}-\widehat{H}^{(j)}\right)^2\right]=\mathbb{E}\left[\frac{1}{n^2}\left(\sum_{i\in S}\left(H_i-H_i^{(j)}\right)\right)^2\right]\\
\leq\ &\mathbb{E}\left[\frac{|S|}{n^2}\sum_{i\in S}\left(H_i-H_i^{(j)}\right)^2\right]=\frac{|S|}{n^2}\sum_{i\in S}\mathbb{E}\left[\left(H_i-H_i^{(j)}\right)^2\right]\\
\leq\ &\frac{2|S|}{n^2}\sum_{i\in S}\left(\mathbb{E}\left[H_i^2\right]+\mathbb{E}\left[(H_i^{(j)})^2\right]\right). \qquad (4.62)
\end{aligned}$$

Notice that H_i's and $H_i^{(j)}$'s are identically distributed, so we are left to compute $\mathbb{E}[H_1^2]$. Conditioning on $X_1=x$, similarly to (4.48), we have

$$\begin{aligned}
&\mathbb{E}\left[(H_1+\log f(x))^2|X_1=x\right]\\
=\ &\mathbb{E}\left[h^2((c_dnf(x))^{1/d}Z_{k,i},S_{0,1},S_{1,i},S_{2,i})\right]\longrightarrow B_{k,d}^{(2)}, \qquad (4.63)
\end{aligned}$$

as $n\to\infty$, where

$$B_{k,d}^{(2)}\ \equiv\ \mathbb{E}\left[h^2\left(\xi_k\left(\sum_{\ell=1}^{k}E_\ell\right)^{1/d},\tilde{S}_0^{(\infty)},\tilde{S}_1^{(\infty)},\tilde{S}_2^{(\infty)}\right)\right]. \qquad (4.64)$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[H_1^2|X_1=x\right]=B_{k,d}^{(2)}-2\log f(x)\mathbb{E}\left[H_1|X_1=x\right]-(\log f(x))^2\\
=\ &B_{k,d}^{(2)}-2\log f(x)B_{k,d}+(\log f(x))^2. \qquad (4.65)
\end{aligned}$$

Take expectation over X_1, we obtain:

$$\begin{aligned}
&\mathbb{E}[H_1^2]=\mathbb{E}_{X_1}\left[\lim_{n\to\infty}\mathbb{E}\left[H_1^2|X_1\right]\right]\\
=\ &\mathbb{E}_{X_1}\left[B_{k,d}^{(2)}-2\log f(X_1)B_{k,d}+(\log f(X_1))^2\right]\\
=\ &B_{k,d}^{(2)}+2H(X)B_{k,d}+\int f(x)(\log f(x))^2dx<+\infty, \qquad (4.66)
\end{aligned}$$

where the last inequality comes from the assumption that $\int f(x)(\log f(x))^2dx<$

$+\infty$. Combining with (4.58) and (4.62), we have

$$\begin{aligned}
\mathrm{Var}\left[\widehat{H}\right] &\leq 2\sum_{j=1}^{n}\mathbb{E}\left[\left(\widehat{H}-\widehat{H}^{(j)}\right)^2\right] \\
&\leq \frac{4|S|}{n}\sum_{i\in S}\left(\mathbb{E}\left[H_i^2\right]+\mathbb{E}\left[(H_i^{(j)})^2\right]\right) \\
&\leq \frac{8|S|^2C_2}{n} \leq \frac{32m^2\gamma_d^2C_2}{n}, \qquad (4.67)
\end{aligned}$$

where C_2 is the upper bound for $\mathbb{E}[H_1^2]$. Take $m = O(\log n)$ then the proof is complete.

Proof of Lemma 17

Firstly, since $|\xi_i| = 1$, we can upper bound the expectation of $\mathbb{E}\| \tilde{T}_{\alpha,i}^{(m)} \|$ by

$$\begin{aligned}
&\mathbb{E}\| \tilde{T}_{\alpha,i}^{(m)} \| \\
&= \mathbb{E}\Big\| \sum_{j=m+1}^{\infty} \frac{\xi_j^{(\alpha)}(\sum_{\ell=1}^{j}E_\ell)^{\frac{\alpha}{d}}}{(\sum_{\ell=1}^{k}E_\ell)^{\frac{\alpha}{d}}}\exp\{-\frac{(\sum_{\ell=1}^{j}E_\ell)^{\frac{2}{d}}}{2(\sum_{\ell=1}^{k}E_\ell)^{\frac{2}{d}}}\}\Big\| \\
&\leq \sum_{j=m+1}^{\infty}\mathbb{E}\Big\| \frac{\xi_j^{(\alpha)}(\sum_{\ell=1}^{j}E_\ell)^{\frac{\alpha}{d}}}{(\sum_{\ell=1}^{k}E_\ell)^{\frac{\alpha}{d}}}\exp\{-\frac{(\sum_{\ell=1}^{j}E_\ell)^{\frac{2}{d}}}{2(\sum_{\ell=1}^{k}E_\ell)^{\frac{2}{d}}}\}\Big\| \\
&= \sum_{j=m+1}^{\infty}\mathbb{E}\Big| \frac{(\sum_{\ell=1}^{j}E_\ell)^{\frac{\alpha}{d}}}{(\sum_{\ell=1}^{k}E_\ell)^{\frac{\alpha}{d}}}\exp\{-\frac{(\sum_{\ell=1}^{j}E_\ell)^{\frac{2}{d}}}{2(\sum_{\ell=1}^{k}E_\ell)^{\frac{2}{d}}}\}\Big|. \qquad (4.68)
\end{aligned}$$

Notice that the expression is a function of $(\sum_{\ell=1}^{j}E_\ell/\sum_{\ell=1}^{k}E_\ell)^{1/d} \equiv R_j$ for $j > m$, we will identify the distribution of R_j first. For any fixed $j \geq k$, let $T_k = \sum_{\ell=1}^{k}E_\ell$ and $T_{j-k} = \sum_{\ell=k+1}^{j}E_\ell$, such that $R_j = ((T_k + T_{j-k})/T_k)^{1/d}$. Notice that T_k is the summation of k i.i.d. standard exponential random variables, so $T_k \sim$ *Erlang* $(k, 1)$. Similarly, $T_{j-k} \sim$ *Erlang* $(j-k, 1)$. Also T_k and T_{j-k} are independent. Recall that the pdf of *Erlang* (k, λ) is given by $f_{k,\lambda}(x) = \lambda^k x^{k-1}e^{-\lambda x}/(k-1)!$ for $x \geq 0$. Therefore, the CDF of R_j is

given by:

$$
\begin{aligned}
& F_{R_j}(t) = \Pr\{R_j \le t\} \\
= \;& \Pr\{(\frac{T_k + T_{j-k}}{T_k})^{1/d} \le t\} = \Pr\{\frac{T_{j-k}}{T_k} \le t^d - 1\} \\
= \;& \int_{x\ge 0} \frac{x^{k-1}e^{-x}}{(k-1)!} \left(\int_{y=0}^{(t^d-1)x} \frac{y^{j-k-1}e^{-y}}{(j-k-1)!} dy \right) dx \\
= \;& \int_{x\ge 0} \frac{x^{k-1}e^{-x}}{(k-1)!} \Big(1 - \sum_{\ell=0}^{j-k-1} \frac{1}{\ell!} x^\ell (t^d-1)^\ell e^{-x(t^d-1)} \Big) dx \\
= \;& 1 - \sum_{\ell=0}^{j-k-1} \Big(\int_{x\ge 0} \frac{x^{k-1}e^{-x}}{(k-1)!} \frac{1}{\ell!} x^\ell (t^d-1)^\ell e^{-x(t^d-1)} dx \Big) \\
= \;& 1 - \sum_{\ell=0}^{j-k-1} \left(\frac{(t^d-1)^\ell}{(k-1)!\ell!} \int_{x\ge 0} x^{k-1+\ell} e^{-xt^d} dx \right) \\
= \;& 1 - \sum_{\ell=0}^{j-k-1} \frac{(t^d-1)^\ell}{(k-1)!\ell!} (k-1+\ell)!\, t^{-d(k-1+\ell)} \\
= \;& 1 - \sum_{\ell=0}^{j-k-1} \tbinom{k-1+\ell}{\ell} t^{-d(k-1)} (1-t^{-d})^\ell, \qquad (4.69)
\end{aligned}
$$

for $t \in [1, +\infty)$. Given the CDF of R_j, each term in (4.68) is upper bounded by:

$$
\begin{aligned}
& \mathbb{E}\Big| \frac{(\sum_{\ell=1}^{j} E_\ell)^{\frac{\alpha}{d}}}{(\sum_{\ell=1}^{k} E_\ell)^{\frac{\alpha}{d}}} \exp\{-\frac{(\sum_{\ell=1}^{j} E_\ell)^{\frac{2}{d}}}{2(\sum_{\ell=1}^{k} E_\ell)^{\frac{2}{d}}}\} \Big| = \mathbb{E}_{R_j}\Big| t^\alpha e^{-t^2} \Big| \\
\le \;& \mathbb{E}_{R_j}\Big[t^2 e^{-t^2} \Big] = \int_{t=1}^{\infty} t^2 e^{-t^2} dF_{R_j}(t) \\
= \;& t^2 e^{-t^2} F_{R_j}(t)\Big|_1^\infty - \int_{t=1}^{\infty} F_{R_j}(t) d(t^2 e^{-t^2}) \\
= \;& -\int_{t=1}^{\infty} (2te^{-t^2} - 2t^3 e^{-t^2}) F_{R_j}(t) dt \\
= \;& \int_{t=1}^{\infty} 2t(t^2-1) e^{-t^2} F_{R_j}(t) dt. \qquad (4.70)
\end{aligned}
$$

Therefore, in order to establish an upper bound for (4.68), we need an upper bound for $F_{R_j}(t)$. Here we will consider two cases depending on t. If $t > (j/2k)^{1/d}$, we just use the trivial upper bound $F_{R_j}(t) < 1$. If $1 \le t \le$

$(j/2k)^{1/d}$, since $t^d \geq 1$, we have:

$$\begin{aligned} F_{R_j}(t) &= 1 - \sum_{\ell=0}^{j-k-1} \binom{k-1+\ell}{\ell} t^{-d(k-1)}(1-t^{-d})^\ell \\ &\leq 1 - \sum_{\ell=0}^{j-k-1} \binom{k-1+\ell}{\ell} t^{-dk}(1-t^{-d})^\ell. \end{aligned} \tag{4.71}$$

Notice that $\binom{k-1+\ell}{\ell} t^{-dk}(1-t^{-d})^\ell$ is the pmf of negative binomial distribution $\mathrm{NB}(k, 1-t^{-d})$. Therefore, $F_{R_j}(t) \leq \Pr\{X \geq j-k\}$, where $X \sim \mathrm{NB}(k, 1-t^{-d})$. The mean and variance of X are given by $\mathbb{E}[X] = (1-t^{-d})k/(1-(1-t^{-d})) = (t^d-1)k$ and $\mathrm{Var}(X) = (1-t^{-d})k/(1-(1-t^{-d}))^2 = (t^{2d}-t^d)k$. Therefore, by Chebyshev inequality, the tail probability is bounded by:

$$\begin{aligned} \Pr\{X \geq j-k\} &\leq \frac{\mathrm{Var}(X)}{(j-k-\mathbb{E}[X])^2} \\ &= \frac{(t^{2d}-t^d)k}{(j-k-(t^d-1)k)^2} = \frac{(t^{2d}-t^d)k}{(j-t^dk)^2} \leq \frac{4t^{2d}k}{j^2}, \end{aligned} \tag{4.72}$$

here we use the fact that $t \leq (j/2k)^{1/d}$ so $j - t^dk > j/2$. Therefore, $F_{R_j}(t) \leq 4t^{2d}k/j^2$ for $t > (j/2k)^{1/d}$. Combine the two cases and plug into (4.70), we obtain:

$$\begin{aligned} & \mathbb{E}\left| \frac{(\sum_{\ell=1}^j E_\ell)^{\frac{\alpha}{d}}}{(\sum_{\ell=1}^k E_\ell)^{\frac{\alpha}{d}}} \exp\{-\frac{(\sum_{\ell=1}^j E_\ell)^{\frac{2}{d}}}{2(\sum_{\ell=1}^k E_\ell)^{\frac{2}{d}}}\} \right| \\ = & \int_{t=1}^{\infty} 2t(t^2-1)e^{-t^2} F_{R_j}(t)dt \\ \leq & \int_{t=1}^{(j/2k)^{1/d}} 2t(t^2-1)e^{-t^2}\frac{4t^{2d}k}{j^2}dt + \int_{(j/2k)^{1/d}}^{\infty} 2t(t^2-1)e^{-t^2}dt \\ \leq & \frac{8k}{j^2}\int_{t=1}^{\infty} t^{2d+3}e^{-t^2}dt + 2\int_{(j/2k)^{1/d}}^{\infty} t^3 e^{-t^2}dt \\ \leq & \frac{8kC_d}{j^2} + 2\left(-\frac{1}{2}e^{-t^2}(t^2+1)\Big|_{(j/2k)^{1/d}}^{\infty} \right) \\ = & \frac{8kC_d}{j^2} + e^{-(j/2k)^{2/d}}((\frac{j}{2k})^{2/d}+1), \end{aligned} \tag{4.73}$$

where $C_d = \int_{t=1}^{\infty} t^{2d+3}e^{-t^2}dt$ is a constant only dependent on d. Therefore,

we can see that

$$\mathbb{E}\Big|\frac{(\sum_{\ell=1}^{j}E_\ell)^{\frac{\alpha}{d}}}{(\sum_{\ell=1}^{k}E_\ell)^{\frac{\alpha}{d}}}\exp\{-\frac{(\sum_{\ell=1}^{j}E_\ell)^{\frac{2}{d}}}{2(\sum_{\ell=1}^{k}E_\ell)^{\frac{2}{d}}}\}\Big| = O(1/j^2). \tag{4.74}$$

So

$$\mathbb{E}\Big|\frac{(\sum_{\ell=1}^{j}E_\ell)^{\frac{\alpha}{d}}}{(\sum_{\ell=1}^{k}E_\ell)^{\frac{\alpha}{d}}}\exp\{-\frac{(\sum_{\ell=1}^{j}E_\ell)^{\frac{2}{d}}}{2(\sum_{\ell=1}^{k}E_\ell)^{\frac{2}{d}}}\}\Big| \to 0, \tag{4.75}$$

given $m \to \infty$ as $n \to \infty$.

4.7.4 Proof of Theorem 9

The proposed estimator is a solution to a maximization problem of the local likelihood $\widehat{a} = \arg\max_a \mathcal{L}_{X_i}(f_{a,X_i})$. From [7] we know that the maximizer is a fixed point of a series of non-linear equations of the form

$$\begin{aligned}&\sum_{j\neq i}\frac{(X_j - X_i)^{\otimes\alpha}}{\rho_{k,i}^{\alpha}}K\Big(\frac{X_j - X_i}{\rho_{k,i}}\Big)\\ =\quad & n\,\rho_{k,i}^{d}\,e^{a_0}\int\frac{(u-X_i)^{\otimes\alpha}}{\rho_{k,i}^{\alpha}}K\Big(\frac{u-X_i}{\rho_{k,i}}\Big)e^{\langle u-x,a_1\rangle+\cdots+a_p[(u-x),\cdots,(u-x)]}\frac{1}{\rho_{k,i}^{d}}du,\end{aligned} \tag{4.76}$$

for all $\alpha \in [p]$ where the superscript $\otimes\alpha$ indicates the α-th order tensor product. From the proof of Theorem 8, specifically (4.55) and (4.57), we know that the left-hand side converges to a value that only depends on k, d and K. We denote it by $S_\alpha(k) \in \mathbb{R}^{d^\alpha}$. We make a change of variables $\tilde{a}_0 = a_0 + d\log\rho_{k,i} + \log n$ and $\tilde{a}_\alpha = a_\alpha/\rho_{k,i}^{\alpha}$ for $\alpha \neq 0$. Then, in the limit of growing n, the above equations can be rewritten as

$$S_\alpha(k,d,K) \;=\; e^{\tilde{a}_0}F_\alpha(d,K,\tilde{a}_1,\ldots,\tilde{a}_p), \tag{4.77}$$

for some function F_α. Notice that the dependence on the underlying distribution vanishes in the limit, and the fixed point $\tilde{a}$ only depends on k, p, d,

and K. The desired claim follows from the fact that the estimate is

$$\begin{aligned} \lim_{n\to\infty} \widehat{f}_n(X_i) &= \lim_{n\to\infty} e^{\widehat{a}_0} = \lim_{n\to\infty} \frac{A_{k,d,p,K}}{n\rho_{k,i}^d} \\ &= f(X_i)A_{k,d,p,K}C_d \lim_{n\to\infty} \frac{1}{C_d n\rho_{k,i}^d f(X_i)} = \frac{f(X_i)A_{k,d,p,K}C_d}{\sum_{\ell=1}^k E_\ell}, \end{aligned} \tag{4.78}$$

and plugging in the entropy estimator

$$\widehat{H}(X) \to E_{X_i}[-\log f(X_i)] + B_{k,d,p,K}. \tag{4.79}$$

In the case of the KL estimator, it happens that $S_0 = k$ and $F_0(d) = C_d$ such that $e^{\tilde{a}_0} = k/C_d$, $e^{\widehat{a}_0} = f(X_i)k/(C_d\rho_{k,i}^d f(X_i)n)$ and $B_{k,d,p,K} = -\log k + E[\log(\sum_{\ell=1}^k E_\ell)] = -\log k + \psi(k)$.

CHAPTER 5

ESTIMATING MUTUAL INFORMATION FOR DISCRETE-CONTINUOUS MIXTURES

A fundamental quantity of interest in machine learning is mutual information (MI), which characterizes the shared information between a pair of random variables (X, Y). MI obeys several appealing properties including the data-processing inequality, invariance under one-to-one transformations and the chain rule [71], which led to a wide use in canonical tasks such as classification [21], clustering [22, 23, 24] and feature selection [26, 27]. Mutual information also emerges as the "correct" quantity in several graphical model inference problems (e.g., the Chow-Liu tree [136] and conditional independence testing [137]). MI is also pervasively used in many data science application domains, such as sociology [11], computational biology [29], and computational neuroscience [138].

An important problem in any of these applications is to estimate mutual information effectively from samples. While mutual information has been the *de facto* measure of information in several applications for decades, the estimation of mutual information from samples remains an active research problem. Recently, there has been a resurgence of interest in entropy and mutual information estimators, on both the theoretical as well as practical fronts [116, 120, 34, 139, 85, 140, 79, 87, 5, 8].

The previous estimators focus on either of two cases – the data is either purely discrete or purely continuous. In these special cases, the mutual information can be calculated based on the three (differential) entropies of X, Y and (X, Y). We term estimators based on this principle as $3H$-estimators (since they estimate three entropy terms), and a majority of previous estimators fall under this category [140, 8, 116].

In practical downstream applications, we often have to deal with a *mixture of continuous and discrete* random variables. Random variables can be mixed in several ways. First, one random variable can be discrete whereas the other is continuous. For example, we want to measure the strength of relationship

between children's age and height, here age X is discrete and height Y is continuous. Secondly, a single scalar random variable itself can be a mixture of discrete and continuous components. For example, consider X taking a zero-inflated-Gaussian distribution, which takes value 0 with probability p and is a Gaussian distribution with mean μ with probability $1-p$. This distribution has both a discrete component as well as a component with density. Finally, X and / or Y can be high-dimensional vector, each of whose components may be discrete, continuous or mixed.

In all of the aforementioned *mixed* cases, mutual information is well-defined through the Radon-Nikodym derivative (see Section 5.1) but cannot be expressed as a function of the entropies or differential entropies of the random variables. Crucially, entropy is not well defined when a single scalar random variable comprises of both discrete and continuous components, in which case, $3H$ estimators (the vast majority of prior art) cannot be directly employed. In this chapter, we address this challenge by proposing an estimator that can handle all these cases of mixture distributions. The estimator directly estimates the Radon-Nikodym derivative using the k-nearest neighbor distances from the samples; we prove ℓ_2 consistency of the estimator and demonstrate its excellent practical performance through a variety of experiments on both synthetic and real dataset. Most relevantly, it strongly outperforms natural baselines of discretizing the mixed random variables (by quantization) or making it continuous by adding a small Gaussian noise.

Main contributions of Chapter 5:

- In Section 5.1, we review the general definition of mutual information for the Radon-Nikodym derivative.
- In Section 5.2, we propose our estimator of mutual information for mixed random variables.
- In Section 5.3, we prove that our estimator is ℓ_2 consistent under certain technical assumptions and verify that the assumptions are satisfied for most practical cases.
- Section 5.4 contains the results of our detailed synthetic and real-world experiments testing the efficacy of the proposed estimator.

5.1 Problem formation

In this section, we define mutual information for general distributions as follows (e.g., [141]).

Definition 3. *Let P_{XY} be a probability measure on the space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are both Euclidean spaces. For any measurable set $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$, define $P_X(A) = P_{XY}(A \times \mathcal{Y})$ and $P_Y(B) = P_{XY}(\mathcal{X} \times B)$. Let $P_X P_Y$ be the product measure $P_X \times P_Y$. If P_{XY} is absolutely continuous w.r.t. $P_X P_Y$, then the mutual information $I(X;Y)$ of P_{XY} is defined as*

$$I(X;Y) \equiv \int_{\mathcal{X}\times\mathcal{Y}} \log \frac{dP_{XY}}{dP_X P_Y} dP_{XY}, \tag{5.1}$$

where $\frac{dP_{XY}}{dP_X P_Y}$ is the Radon-Nikodym derivative.

Notice that this general definition includes the following cases of mixtures: (1) X is discrete and Y is continuous (or vice versa); (2) X or Y has many components each, where some components are discrete and some are continuous; (3) X or Y or their joint distribution is a mixture of continuous and discrete distributions.

5.2 Estimators of mutual information

5.2.1 Review of previous works

The estimation problem is quite different depending on whether the underlying distribution is discrete, continuous or mixed. As pointed out earlier, most existing estimators for mutual information are based on the $3H$ principle: they estimate the three entropy terms first. This $3H$ principle can be applied only in the purely discrete or purely continuous case.

Discrete data: For entropy estimation of a discrete variable X, the straightforward approach to plug-in the estimated probabilities $\hat{p}_X(x)$ into the formula for entropy has been shown to be suboptimal [84, 142]. Novel entropy estimators with sub-linear sample complexity have been proposed [81, 77, 140, 85, 143, 144]. MI estimation can then be performed using the $3H$ principle, and such an approach is shown to be worst-case optimal for mutual-information estimation [140].

Continuous data: There are several estimators for differential entropy of continuous random variables, which have been exploited in a $3H$ principle to calculate the mutual information [30]. One family of entropy estimators are based on kernel density estimators [100] followed by re-substitution estimation. An alternate family of entropy estimators is based on k-Nearest Neighbor (k-NN) estimates, beginning with the pioneering work of Kozachenko and Leonenko [2] (the so-called KL estimator). Recent progress involves an inspired mixture of an ensemble of kernel and k-NN estimators [116, 36]. Exponential concentration bounds under certain conditions are in [102].

Mixed random variables: Since the entropies themselves may not be well defined for mixed random variables, there is no direct way to apply the $3H$ principle. However, once the data is quantized, this principle can be applied in the discrete domain. That mutual information in arbitrary measure spaces can indeed be computed as a maximum over quantization is a classical result [145]. However, the choice of quantization is complicated and while some quantization schemes are known to be consistent when there is a joint density [146], the mixed case is complex. Estimator of the average of Radon-Nikodym derivative dP/dQ has been studied in [78, 75]. Very recent work generalizing the ensemble entropy estimator when some components are discrete and others continuous is in [120].

Beyond $3H$ estimation: An inspired work [4] proposed a *direct* method for estimating mutual information (KSG estimator) when the variables have a joint density. The estimator starts with the $3H$ estimator based on differential entropy estimates based on the k-NN estimates, and employs a heuristic to couple the estimates in order to improve the estimator. While the original paper did not contain any theoretical proof, even of consistency, its excellent practical performance has encouraged widespread adoption. Recent work [5] has established the consistency of this estimator along with its convergence rate. Further, recent works [87, 8] involving a combination of kernel density estimators and k-NN methods have been proposed to further improve the KSG estimator. Ross [147] extends the KSG estimator to the case when one variable is discrete and another is scalar continuous.

None of these works consider a case even if one of the components has a mixture of continuous and discrete distribution, let alone for general probability distributions. There are two generic options: (1) one can add small independent noise on each sample to break the multiple samples and apply

a continuous valued MI estimator (like KSG), or (2) quantize and apply discrete MI estimators but the performance for high-dimensional case is poor. These form baselines to compare against in our detailed simulations.

5.2.2 Mixed regime

We first examine the behavior of other estimators in the mixed regime, before proceeding to develop our estimator. Let us consider the case when X is discrete (but real valued) and Y possesses a density. In this case, we will examine the consequence of using the $3H$ principle, with differential entropy estimated by the k-nearest neighbors. To do this, fix a parameter k, that determines the number of neighbors and let $\rho_{i,x}$, $\rho_{i,y}$ and $\rho_{i,xy}$ denote the distance of the k-nearest neighbor of X_i, Y_i and (X_i, Y_i), respectively. Then

$$\widehat{I}^{(N)}_{3\mathrm{H}}(X;Y) = \left(\frac{1}{N} \sum_{i=1}^{N} \log \frac{N c_x \rho_{i,x}^d}{k} + a(k) \right) + \left(\frac{1}{N} \sum_{i=1}^{N} \log \frac{N c_y \rho_{i,y}^d}{k} + a(k) \right) - \left(\frac{1}{N} \sum_{i=1}^{N} \log \frac{N c_{xy} \rho_{i,xy}^d}{k} + a(k) \right), \tag{5.2}$$

where $\psi(\cdot)$ is the digamma function and $a(\cdot) = \log(\cdot) - \psi(\cdot)$. In the case that X is discrete and Y has a density, $I_{3\mathrm{H}}(X;Y) = -\infty + a - b = -\infty$, which is clearly wrong.

The basic idea of the KSG estimator is to ensure that the ρ is the same for both x, y and (x, y) and the difference is instead in the number of nearest neighbors. Let $n_{x,i}$ be the number of samples of X_i's within distance $\rho_{i,xy}$ and $n_{y,i}$ be the number of samples of Y_i's within distance $\rho_{i,xy}$. Then the KSG estimator is given by

$$\widehat{I}^{(N)}_{KSG} \equiv \frac{1}{N} \sum_{i=1}^{N} \left(\psi(k) + \log(N) - \log(n_{x,i} + 1) - \log(n_{y,i} + 1) \right), \tag{5.3}$$

where $\psi(\cdot)$ is the digamma function.

In the case of X being discrete and Y being continuous, it turns out that the KSG estimator does *not* blow up (unlike the $3H$ estimator), since the

distances do not go to zero. However, in the mixed case, the estimator has a non-trivial bias due to discrete points and is no longer consistent.

5.2.3 Proposed estimator

We propose the following estimator for general probability distributions, inspired by the KSG estimator. The intuition is as follows. First notice that MI is the average of the logarithm of Radon-Nikodym derivative, so we compute the Radon-Nikodym derivative for each sample i and take the empirical average. The re-substitution estimator for MI is then given as follows: $\widehat{I}(X;Y) \equiv \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{dP_{XY}}{dP_X P_Y}\right)_{(x_i,y_i)}$. The basic idea behind our estimate of the Radon-Nikodym derivative at each sample point is as follows:

- When the point is discrete (which can be detected by checking if the k-nearest neighbor distance of data i is zero), then we can assert that data i is in a discrete component, and we can use plug-in estimator for Radon-Nikodym derivative.
- If the point is such that there is a joint density (locally), the KSG estimator suggests a natural idea: fix the radius and estimate the Radon-Nikodym derivative by $(\psi(k) + \log(N) - \log(n_{x,i}+1) - \log(n_{y,i}+1))$.
- If k-nearest neighbor distance is not zero, then it may be either purely continuous or mixed. But we show below that the method for purely continuous is also applicable for mixed.

Precisely, let $n_{x,i}$ be the number of samples of X_i's within distance $\rho_{i,xy}$ and $n_{y,i}$ be the number of samples of Y_i's with in $\rho_{i,xy}$. Denote $\tilde{k}_i$ by the number of tuples (X_i, Y_i) within distance $\rho_{i,xy}$. If the k-NN distance is zero, which means that the sample (X_i, Y_i) is a discrete point of the probability measure, we set k to $\tilde{k}_i$, which is the number of samples that have the same value as (X_i, Y_i). Otherwise we just keep $\tilde{k}_i$ as k. Our proposed estimator is described in detail in Algorithm 1.

We note that our estimator recovers previous ideas in several canonical settings. If the underlying distribution is purely discrete, the k-nearest neighbor distance $\rho_{i,xy}$ equals to 0 with high probability, then our estimator recovers the plug-in estimator. If the underlying distribution is purely continuous,

Algorithm 1 Mixed random variable mutual information estimator

Input: $\{X_i, Y_i\}_{i=1}^N$, where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$;
Parameter: $k \in \mathbb{Z}^+$;
for $i = 1$ to N **do**
  $\mathcal{D}_i := \{d_{i,j} := \max\{\|X_j - X_i\|, \|Y_j - Y_i\|\}, j \neq i\}$;
  $\rho_{i,xy} :=$ the k-th smallest element in $\mathcal{D}_i$;
  if $\rho_{i,xy} = 0$ **then**
    $\tilde{k}_i :=$ number of samples such that $d_{i,j} = 0$;
  else
    $\tilde{k}_i := k$;
  end if
  $n_{x,i} :=$ number of samples such that $\|X_j - X_i\| \leq \rho_{i,xy}$;
  $n_{y,i} :=$ number of samples such that $\|Y_j - Y_i\| \leq \rho_{i,xy}$;
  $\xi_i := \psi(\tilde{k}_i) + \log N - \log(n_{x,i} + 1) - \log(n_{y,i} + 1)$;
end for
Output: $\widehat{I}^{(N)}(X;Y) := \frac{1}{N} \sum_{i=1}^N \xi_i$.

then there are no multiple overlapping samples, so $\tilde{k}_i$ equals to k, our estimator recovers the KSG estimator. If X is discrete and Y is single-dimensional continuous and $P_X(x) > 0$ for all x, for sufficiently large dataset, the k-nearest neighbors of sample (x_i, y_i) will be located on the same x_i with high probability. Therefore, our estimator recovers the discrete vs continuous estimator in [147].

5.3 Proof of consistency

We show that under certain technical conditions on the joint probability measure, the proposed estimator is consistent. We begin with the following definitions.

$$P_{XY}(x, y, r) \equiv P_{XY}\left(\{(a, b) \in \mathcal{X} \times \mathcal{Y} : \max\{\|a - x\|, \|b - y\|\} \leq r\}\right), \tag{5.4}$$

$$P_X(x, r) \equiv P_X\left(\{a \in \mathcal{X} : \|a - x\| \leq r\}\right), \tag{5.5}$$

$$P_Y(y, r) \equiv P_Y\left(\{b \in \mathcal{Y} : \|b - y\| \leq r\}\right). \tag{5.6}$$

Theorem 10. *Suppose that the following assumptions hold.*

Assumption 3. *(a) k is chosen to be a function of N such that $k_N \to \infty$*

and $k_N \log N/N \to 0$ *as* $N \to \infty$.

(*b*) *The set of discrete points* $\{(x,y) : P_{XY}(x,y,0) > 0\}$ *is finite.*

(*c*) $\frac{P_{XY}(x,y,r)}{P_X(x,r)P_Y(y,r)}$ *converges to* $f(x,y)$ *as* $r \to 0$ *and* $f(x,y) \le C$ *with probability 1.*

(*d*) $\mathcal{X} \times \mathcal{Y}$ *can be decomposed into countable disjoint sets* $\{E_i\}_{i=1}^{\infty}$ *such that* $f(x,y)$ *is uniformly continuous on each* E_i.

(*e*) $\int_{\mathcal{X}\times\mathcal{Y}} \left| \log f(x,y) \right| dP_{XY} < +\infty$.

Then we have $\lim_{N\to\infty} \mathbb{E}\left[\widehat{I}^{(N)}(X;Y) \right] = I(X;Y)$.

Notice that conditions Assumptions 3.(*b*), (*c*) and (*d*) are satisfied whenever (1) the distribution is (finitely) discrete; (2) the distribution is continuous; (3) some dimensions are (countably) discrete and some dimensions are continuous; (4) a (finite) mixture of the previous cases. Most real-world data can be covered by these cases. A sketch of the proof is below with the full proof in Section 5.5.

We sketch the proof starting with an explicit form of the Radon-Nikodym derivative $dP_{XY}/(dP_X P_Y)$.

Lemma 18. *Under Assumption 3.*(*c*) *and* (*d*) *in Theorem 10,*

$$\frac{dP_{XY}}{dP_X P_Y}(x,y) = f(x,y) = \lim_{r\to 0} \frac{P_{XY}(x,y,r)}{(P_X(x,r)P_Y(y,r)}. \tag{5.7}$$

Notice that $\widehat{I}_N(X;Y) = (1/N)\sum_{i=1}^{N} \xi_i$, where all ξ_i are identically distributed. Therefore, $\mathbb{E}[\widehat{I}^{(N)}(X;Y)] = \mathbb{E}[\xi_1]$. Therefore, the bias can be written as:

$$\begin{aligned} & \left| \mathbb{E}[\widehat{I}^{(N)}(X;Y)] - I(X;Y) \right| = \left| \mathbb{E}_{XY}\left[\mathbb{E}\left[\xi_1 | X, Y \right] \right] - \int \log f(X,Y) P_{XY} \right| \\ \le & \int \left| \mathbb{E}\left[\xi_1 | X, Y \right] - \log f(X,Y) \right| dP_{XY}. \end{aligned} \tag{5.8}$$

Now we upper bound $\left| \mathbb{E}\left[\xi_1 | X, Y \right] - \log f(X,Y) \right|$ for every $(x,y) \in \mathcal{X} \times \mathcal{Y}$ by dividing the domain into three parts as $\mathcal{X} \times \mathcal{Y} = \Omega_1 \bigcup \Omega_2 \bigcup \Omega_3$ where

- $\Omega_1 = \{(x,y) : f(x,y) = 0\}$,

- $\Omega_2 = \{(x,y) : f(x,y) > 0, P_{XY}(x,y,0) > 0\}$,
- $\Omega_3 = \{(x,y) : f(x,y) > 0, P_{XY}(x,y,0) = 0\}$.

We show that $\lim_{N\to\infty} \int_{\Omega_i} \Big| \mathbb{E}\left[\xi_1 | X, Y\right] - \log f(X,Y) \Big| dP_{XY} = 0$ for each $i \in \{1,2,3\}$ separately.

- For $(x,y) \in \Omega_1$, we will show that Ω_1 has zero probability with respect to P_{XY}. Hence, $\int_{\Omega_1} \Big| \mathbb{E}\left[\xi_1 | X, Y\right] - \log f(X,Y) \Big| dP_{XY} = 0$.
- For $(x,y) \in \Omega_2$, $f(x,y)$ equals to $P_{XY}(x,y,0)/P_X(x,0)P_Y(y,0)$, so it can be viewed as a discrete part. We will first show that the k-nearest neighbor distance $\rho_{k,1} = 0$ with high probability. Then we will use the the number of samples on (x,y) as $\tilde{k}_i$, and we will show that the mean of estimate ξ_1 is closed to $\log f(x,y)$.
- For $(x,y) \in \Omega_3$, it can be viewed as a continuous part. We use the similar proof technique as [4] to prove that the mean of estimate ξ_1 is closed to $\log f(x,y)$.

The following theorem bounds the variance of the proposed estimator.

Theorem 11. *Assume in addition that*

(f). $(k_N \log N)^2/N \to 0$ *as* $N \to \infty$.

Then we have

$$\lim_{N\to\infty} \mathrm{Var}\left[\widehat{I}^{(N)}(X;Y) \right] = 0. \tag{5.9}$$

Sketch of proof: We use the Efron-Stein inequality to bound the variance of the estimator. For simplicity, let $\widehat{I}^{(N)}(Z)$ be the estimate based on original samples $\{Z_1, Z_2, \ldots, Z_N\}$, where $Z_i = (X_i, Y_i)$, and $\widehat{I}^{(N)}(Z_{\setminus j})$ is the estimate from $\{Z_1, \ldots, Z_{j-1}, Z_{j+1}, \ldots, Z_N\}$. Then a certain version of Efron-Stein inequality states that:

$$\mathrm{Var}\left[\widehat{I}^{(N)}(Z) \right] \le 2 \sum_{j=1}^{N} \left(\sup_{Z_1,\ldots,Z_N} \left| \widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z_{\setminus j}) \right| \right)^2. \tag{5.10}$$

Now recall that

$$\begin{aligned}\widehat{I}^{(N)}(Z) &= \frac{1}{N}\sum_{i=1}^{N}\xi_i(Z)\\ &= \frac{1}{N}\sum_{i=1}^{N}\Big(\psi(\tilde{k}_i)+\log N-\log(n_{x,i}+1)-\log(n_{y,i}+1)\Big). \quad (5.11)\end{aligned}$$

Therefore, we have

$$\sup_{Z_1,\ldots,Z_N}\Big|\widehat{I}^{(N)}(Z)-\widehat{I}^{(N)}(Z_{\setminus j})\Big| \le \frac{1}{N}\sup_{Z_1,\ldots,Z_N}\sum_{i=1}^{N}\Big|\xi_i(Z)-\xi_i(Z_{\setminus j})\Big|. \quad (5.12)$$

To upper bound the difference $|\,\xi_i(Z)-\xi_i(Z_{\setminus j})\,|$ created by eliminating sample Z_j for different i 's we consider three different cases: (1) $i=j$; (2) $\rho_{k,i}=0$; (3) $\rho_{k,i}>0$, and conclude that $\sum_{i=1}^{N}|\,\xi_i(Z)-\xi_i(Z_{\setminus j})\,| \le O(k\log N)$ for all Z_i's. Plug it into Efron-Stein inequality, we obtain:

$$\begin{aligned}\mathrm{Var}\Big[\widehat{I}^{(N)}(Z)\Big] &\le 2\sum_{j=1}^{N}\Big(\sup_{Z_1,\ldots,Z_N}\Big|\widehat{I}^{(N)}(Z)-\widehat{I}^{(N)}(Z_{\setminus j})\Big|\Big)^2\\ &\le 2\sum_{j=1}^{N}\left(\frac{1}{N}\sup_{Z_1,\ldots,Z_N}\sum_{i=1}^{N}\Big|\xi_i(Z)-\xi_i(Z_{\setminus j})\Big|\right)^2\\ &= O((k\log N)^2/N). \quad (5.13)\end{aligned}$$

By Assumption 3.(f), we have $\lim_{N\to\infty}\mathrm{Var}\Big[\widehat{I}^{(N)}(Z)\Big]=0$.

Combining Theorem 10 and Theorem 11, we have the ℓ_2 consistency of $\widehat{I}^{(N)}(X;Y)$.

5.4 Experiments of Chapter 5

We evaluate the performance of our estimator in a variety of (synthetic and real-world) experiments.

Experiment I. (X,Y) is a mixture of one continuous distribution and one discrete distribution. The continuous distribution is jointly Gaussian with zero mean and covariance $\Sigma=\begin{pmatrix}1 & 0.9\\ 0.9 & 1\end{pmatrix}$, and the discrete distribution is $P(X=1,Y=1)=P(X=-1,Y=-1)=0.45$ and $P(X=1,Y=-1)=$

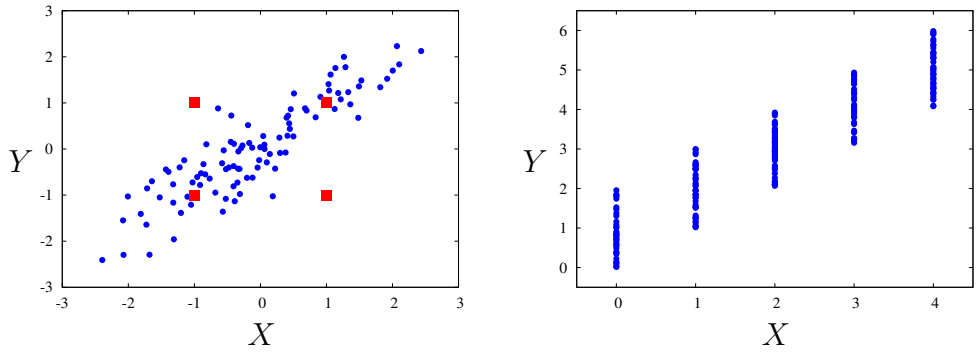

Figure 5.1: Left: An example of samples from a mixture of continuous (blue) and discrete (red) distributions, where red points denote multiple samples. Right: An example of samples from a discrete X and a continuous Y.

$P(X = -1, 1) = 0.05$. These two distributions are mixed with equal probability. The scatter plot of a set of samples from this distribution is shown in the left panel of Figure 5.1, where the red squares denote multiple samples from the discrete distribution. For all synthetic experiments, we compare our proposed estimator with a (fixed) partitioning estimator, an adaptive partitioning estimator [146] implemented by [148], the KSG estimator [4] and noisy KSG estimator (by adding Gaussian noise $N(0, \sigma^2 I)$ on each sample to transform all mixed distributions into continuous one). We plot the mean squared error versus number of samples in Figure 5.2. The mean squared error is averaged over 250 independent trials.

The KSG estimator is entirely misled by the discrete samples as expected. The noisy KSG estimator performs better but the added noise causes the estimate to degrade. In this experiment, the estimate is less sensitive to the noise added and the line is indistinguishable with the line for KSG. The partitioning and adaptive partitioning method quantizes all samples, resulting in an extra quantization error. Note that only the proposed estimator has error decreasing with the sample size.

Experiment II. X is a discrete random variable and Y is a continuous random variable. X is uniformly distributed over integers $\{0, 1, \ldots, m-1\}$ and Y is uniformly distributed over the range $[X, X+2]$ for a given X. The ground truth $I(X;Y) = \log(m) - (m-1)\log(2)/m$. We choose $m = 5$ and a scatter plot of a set of samples is in the right panel of Figure 5.1. Notice that in this case (and the following experiments) our proposed estimator

degenerates to KSG if the hyper parameter k is chosen the same, hence KSG is not plotted. In this experiment our proposed estimator outperforms other methods.

Experiment III. Higher-dimensional mixture. Let (X_1, Y_1) and (Y_2, X_2) have the same joint distribution as in experiment II and independent of each other. We evaluate the mutual information between $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$. Then ground truth $I(X;Y) = 2(\log(m) - (m-1)\log(2)/m)$. We also consider $X = (X_1, X_2, X_3)$ and $Y = (Y_1, Y_2, Y_3)$ where (X_3, Y_3) have the same joint distribution as in experiment II and independent of $(X_1, Y_1), (X_2, Y_2)$. The ground truth $I(X;Y) = 3(\log(m)-(m-1)\log(2)/m)$. The adaptive partitioning algorithm works only for one-dimensional X and Y and is not compared here.

We can see that the performance of partitioning estimator is very bad because the number of partitions grows exponentially with dimension. Proposed algorithm suffers less from the curse of dimensionality. For the right figure, noisy KSG method has smaller error, but we point out that it is unstable with respect to the noise level added: as the noise level is varied from $\sigma = 0.5$ to $\sigma = 0.7$ and the performance varies significantly (far from convergence).

Experiment IV. Zero-inflated Poissonization. Here $X \sim \text{Exp}(1)$ is a standard exponential random variable, and Y is zero-inflated Poissonization of X, i.e., $Y = 0$ with probability p and $Y \sim \text{Poisson}(x)$ given $X = x$ with probability $1 - p$. Here the ground truth is $I(X;Y) = (1-p)(2\log 2 - \gamma - \sum_{k=1}^{\infty} \log k \cdot 2^{-k}) \approx (1-p)0.3012$, where γ is Euler-Mascheroni constant. We repeat the experiment for no zero-inflation ($p = 0$) and for $p = 15\%$. We find that the proposed estimator is comparable to adaptive partitioning for no zero-inflation and outperforms others for 15% zero-inflation.

We conclude that our proposed estimator is consistent for all these four experiments, and the mean squared error is always the best or comparable to the best. Other estimators are either not consistent or have large mean squared error for at least one experiment.

Feature Selection Task. Suppose there are a set of features modeled by independent random variables $(X_1, \ldots, X_p)$ and the data Y depends on a subset of features $\{X_i\}_{i \in S}$, where $\text{card}(S) = q < p$. We observe the features $(X_1, \ldots, X_p)$ and data Y and try to select which features are related to Y. In many biological applications, some of the data is lost due to experimental

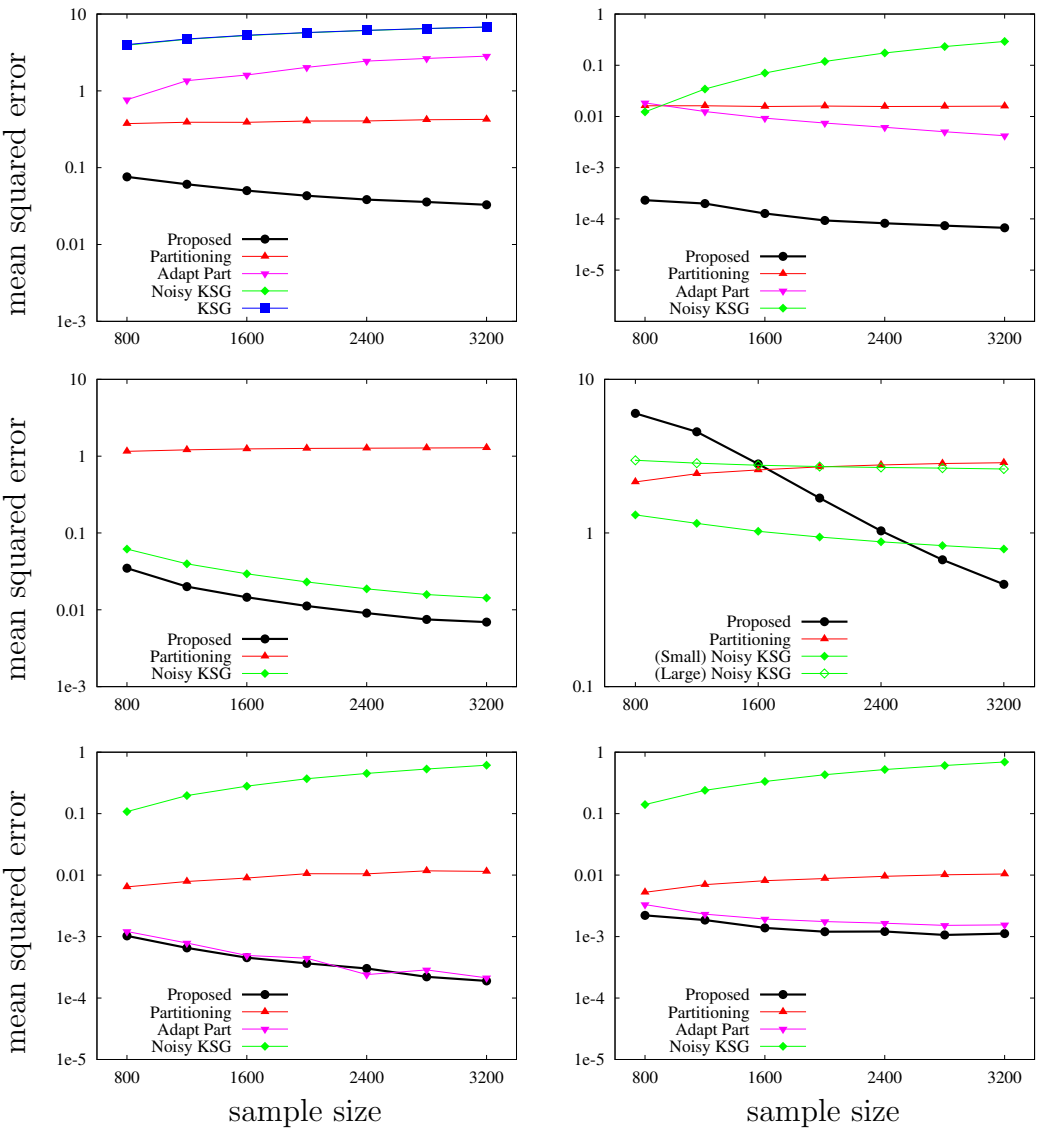

Figure 5.2: Mean squared error vs. sample size for synthetic experiments. Top row (left to right): Experiment I; Experiment II. Middle row (left to right): Experiment III for 4 dimensions and 6 dimensions. Bottom row (left to right): Experiment IV for $p = 0$ and $p = 15\%$.

reasons and set to 0; even the available data is noisy. This setting naturally leads to a mixture of continuous and discrete parts which we model by supposing that the observation is $\tilde{X}_i$ and $\tilde{Y}$, instead of X_i and Y. Here $\tilde{X}_i$ and $\tilde{Y}$ equals to 0 with probability σ and follows Poisson distribution parameterized by X_i or Y (which corresponds to the noisy observation) with probability $1 - \sigma$.

In this experiment, $(X_1, \ldots, X_{20})$ are i.i.d. standard exponential random variables and Y is simply $(X_1, \ldots, X_5)$. $\tilde{X}_i$ equals to 0 with probability 0.15, and $\tilde{X}_i \sim \text{Poisson}(X_i)$ with probability 0.85. $\tilde{Y}_i$ equals to 0 with probability 0.15 and $\tilde{Y}_i \sim \text{Exp}(Y_i)$ with probability 0.85. Upon observing $\tilde{X}_i$'s and $\tilde{Y}$, we evaluate $\text{MI}_i = I(\tilde{X}_i; \tilde{Y})$ using different estimators, and select the features with top-r highest mutual information. Since the underlying number of features is unknown, we iterate over all $r \in \{0, \ldots, p\}$ and observe a receiver operating characteristic (ROC) curve, shown in left of Figure 5.3. Compared to partitioning, noisy KSG and KSG estimators, we conclude that our proposed estimator outperforms other estimators.

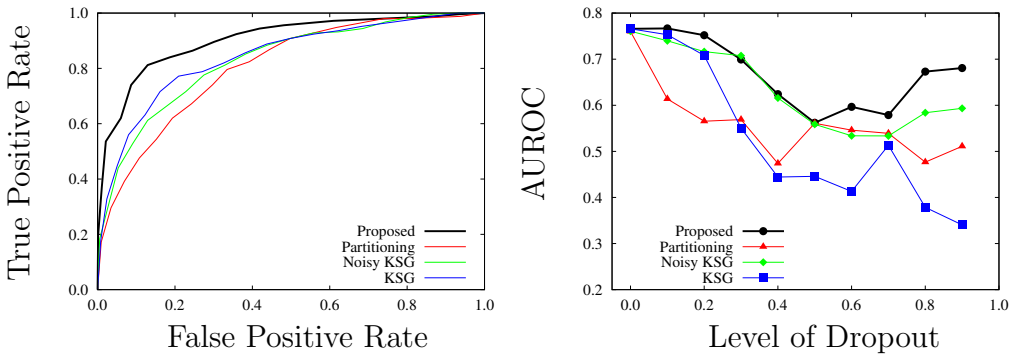

Figure 5.3: Left: ROC curve for the feature selection task. Right: AUROC versus levels of dropout for gene regulatory network inference.

Gene regulatory network inference. Gene expressions form a rich source of data from which to infer gene regulatory networks; it is now possible to sequence gene expression data from single cells using a technology called single-cell RNA-sequencing [149]. However, this technology has a problem called *dropout*, which implies that sometimes, even when the gene is present it is not sequenced [150, 151]. While we tested our algorithm on real single-cell RNA-seq dataset, it is hard to establish the ground truth on these datasets. Instead we resorted to a challenge dataset for reconstructing regulatory net-

works, called the DREAM5 challenge [152]. The simulated (insilico) version of this dataset contains gene expression for 20 genes with 660 data points containing various perturbations. The goal is to reconstruct the true network between the various genes. We used mutual information as the test statistic in order to obtain AUROC for various methods. While the dataset did not have any dropouts, in order to simulate the effect of dropouts in real data, we simulated various levels of dropout and compared the AUROC (area under ROC) of different algorithms in the right of Figure 5.3 where we find the proposed algorithm to outperform the competing ones.

5.5 Proof of Theorem 10 on the bias

To prove the asymptotic unbiasedness of the estimator, we need to write the Radon-Nikodym derivative in an explicit form. The following lemma gives the explicit form of $\frac{dP_{XY}}{dP_X P_Y}$.

Lemma 19. *For almost every* $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\frac{dP_{XY}}{dP_X P_Y} = f(x, y) = \lim_{r \to 0} \frac{P_{XY}(x, y, r)}{P_X(x, r) P_Y(y, r)}. \tag{5.14}$$

Now notice that $\widehat{I}_N(X;Y) = \frac{1}{N}\sum_{i=1}^{N} \xi_i$, where all ξ_i are identically distributed. Therefore, $\mathbb{E}[\widehat{I}_N(X;Y)] = \mathbb{E}[\xi_1]$. Therefore, the bias can be written as:

$$\begin{aligned} & \Big| \mathbb{E}[\widehat{I}_N(X;Y)] - I(X;Y) \Big| = \Big| \mathbb{E}_{XY} \left[\mathbb{E} \left[\xi_1 | X, Y \right] \right] - \int \log f(X,Y) P_{XY} \Big| \\ \leq & \int \Big| \mathbb{E} \left[\xi_1 | X, Y \right] - \log f(X,Y) \Big| dP_{XY}. \end{aligned} \tag{5.15}$$

Now we will give upper bounds for $\Big| \mathbb{E} \left[\, \xi_1 | X, Y \, \right] - \log f(X,Y) \Big|$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We will divide the space into three parts as $\mathcal{X} \times \mathcal{Y} = \Omega_1 \bigcup \Omega_2 \bigcup \Omega_3$ where

- $\Omega_1 = \{(x, y) : f(x, y) = 0\}$,
- $\Omega_2 = \{(x, y) : f(x, y) > 0, P_{XY}(x, y, 0) > 0\}$,
- $\Omega_3 = \{(x, y) : f(x, y) > 0, P_{XY}(x, y, 0) = 0\}$.

We will show that $\lim_{N\to\infty} \int_{\Omega_i} \Big| \mathbb{E}\left[\xi_1 | (X,Y) = (x,y)\right] - \log f(x,y) \Big| \, dP_{XY} = 0$ for each $i \in \{1,2,3\}$ separately.

Case I: $(x,y) \in \Omega_1$. In this case, we will show that Ω_1 has zero probability with respect to P_{XY}.

$$P_{XY}(\Omega_1) = \int_{\Omega_1} dP_{XY} = \int_{\Omega_1} f(X,Y) dP_X P_Y = \int_{\Omega_1} 0 \, dP_X P_Y = 0. \quad (5.16)$$

Therefore, $\int_{\Omega_1} \Big| \mathbb{E}\left[\xi_1 | X, Y\right] - \log f(X,Y) \Big| \, dP_{XY} = 0$.

Case II: $(x,y) \in \Omega_2$. In this case, $f(x,y)$ is just $P_{XY}(x,y,0)/P_X(x,0)P_Y(y,0)$. We will first show that the probability that the k-nearest neighbor distance $\rho_{k,1} > 0$ is small. Then with high probability, we will use the the number of samples on (x,y) as $\tilde{k}_i$, and we will show that the mean of estimate ξ_1 is closed to $\log f(x,y)$.

First, the probability of $\rho_{k,1} > 0$ is upper bounded by:

$$\begin{aligned}
& \Pr\left(\rho_{k,1} > 0 \,|\, (X,Y) = (x,y) \right) \\
= & \sum_{m=0}^{k-1} \tbinom{N-1}{m} P_{XY}(x,y,0)^m (1 - P_{XY}(x,y,0))^{N-1-m} \\
\leq & \sum_{m=0}^{k-1} N^m (1 - P_{XY}(x,y,0))^{N-k} \\
\leq & \; kN^k (1 - P_{XY}(x,y,0))^{N-k} \leq kN^k e^{-(N-k)P_{XY}(x,y,0)}. \qquad (5.17)
\end{aligned}$$

Conditioning on the event that $\rho_{k,1} = 0$, we have $\xi_1 = \psi(\tilde{k}_1) + \log N - \log(n_{x,1}+1) - \log(n_{y,1}+1)$. Then we write $\Big| \mathbb{E}\left[\xi_1 | (X,Y) = (x,y), \rho_{k,1} = 0\right] -$

$\log f(x,y)\Big|$ as

$$
\begin{aligned}
& \Big|\mathbb{E}\left[\xi_1|(X,Y)=(x,y),\rho_{k,1}=0\right]-\log f(x,y)\Big| \\
= & \Big|\mathbb{E}\left[\psi(\tilde{k}_1)+\log\frac{N}{(n_{x,1}+1)(n_{y,1}+1)}|(X,Y)=(x,y),\rho_{k,1}=0\right] \\
& -\log\frac{P_{XY}(x,y,0)}{P_X(x,0)P_Y(y,0)}\Big| \\
\leq & \Big|\mathbb{E}\left[\log(n_{x,1}+1)|(X,Y)=(x,y),\rho_{k,1}=0\right]-\log NP_X(x,0)\Big| \\
& +\Big|\mathbb{E}\left[\log(n_{y,1}+1)|(X,Y)=(x,y),\rho_{k,1}=0\right]-\log NP_Y(y,0)\Big| \\
& +\Big|\mathbb{E}\left[\psi(\tilde{k}_1)|(X,Y)=(x,y),\rho_{k,1}=0\right]-\log NP_{XY}(x,y,0)\Big|. \quad (5.18)
\end{aligned}
$$

Notice that $\tilde{k}_1$ is the number of samples among $\{(X_i,Y_i)\}_{i=2}^N$ such that $(X_i,Y_i)=(x,y)$, where each $(X_i,Y_i)=(x,y)$ with probability $P_{XY}(x,y,0)$. Therefore, the distribution of $\tilde{k}_1$ is $\text{Bino}(N-1,P_{XY}(x,y,0))$. Similarly, $n_{x,1}$ is the number of samples among $\{(X_i,Y_i)\}_{i=2}^N$ such that $X_i=x$, $n_{y,1}$ is the number of samples among $\{(X_i,Y_i)\}_{i=2}^N$ such that $Y_i=y$. Therefore, $n_{x,1}\sim\text{Bino}(N-1,P_X(x,0))$ and $n_{y,1}\sim\text{Bino}(N-1,P_Y(y,0))$. Notice that conditioning on $\rho_{k,i}=0$ is equivalent to conditioning on $\tilde{k}_i\geq k$, or $n_{x,i}\geq k$, $n_{y,i}\geq k$, so we propose the following lemma to deal with (5.18).

Lemma 20. *If X is distributed as Bino(N,p) and $m\geq 0$, then:*

$$
\begin{aligned}
& |\mathbb{E}\left[\log(X+m)|X\geq k\right]-\log(Np)| \\
\leq & \max\left\{\left|\log\left(\frac{1+\frac{m}{Np}}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)}\right)\right|,\frac{1}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)}\frac{3}{2Np}\right\}. \quad (5.19)
\end{aligned}
$$

By Assumption 3.(b), $k/N\to 0$ as $N\to\infty$, then $(Np-k)^2/N=N(p-k/N)^2\to\infty$, So for sufficiently large N, the RHS of Lemma 20 is upper bounded by $\max\{\frac{C_1 m}{Np},\frac{C_2}{Np}\}\leq\frac{C(m+1)}{Np}$, where $C=\max\{C_1,C_2\}$ is some constant not depends on N. Therefore, by applying Lemma 20 with $m=1$, the

first term of (5.18) is bounded by:

$$\begin{aligned} & \left| \mathbb{E}\left[\log(n_{x,1}+1) | (X,Y)=(x,y), \rho_{k,1}=0\right] - \log NP_X(x,0) \right| \\ \leq & \left| \mathbb{E}\left[\log(n_{x,1}+1) | (X,Y)=(x,y), n_{x,i}\geq k\right] - \log(N-1)P_X(x,0) \right| \\ & + \log\frac{N}{N-1} \leq \frac{2C}{(N-1)P_X(x,0)} + \frac{1}{N-1} \\ \leq & \frac{2C+1}{(N-1)P_X(x,0)} \leq \frac{4C+2}{NP_X(x,0)}. \end{aligned} \tag{5.20}$$

Similarly, the second term of (5.18) is bounded by: $(4C+2)/(NP_Y(y,0))$. For the third term, notice that $|\psi(x)-\log(x)| \leq 1/x$ for every integer $x \geq 1$, therefore, $|\psi(\tilde{k}_1)-\log(\tilde{k}_1)| \leq 1/\tilde{k}_1 \leq 1/k$. By applying Lemma 20 with $m=0$, the third term of (5.18) is bounded by: $(2C+2)/(NP_{XY}(x,y,0))+1/k$. By Combining three terms together and noticing that $P_X(x,0) \geq P_{XY}(x,y,0)$ and $P_Y(y,0) \geq P_{XY}(x,y,0)$, we obtain

$$\begin{aligned} & \left| \mathbb{E}\left[\xi_1 | (X,Y)=(x,y), \rho_{k,1}=0\right] - \log f(x,y) \right| \\ \leq & \frac{4C+2}{NP_X(x,0)} + \frac{4C+2}{NP_Y(y,0)} + \frac{2C+2}{NP_{XY}(x,y,0)} + \frac{1}{k} \\ \leq & \frac{10C+6}{NP_{XY}(x,y,0)} + \frac{1}{k}. \end{aligned} \tag{5.21}$$

Combine with the case that $\rho_{i,xy} > 0$, we obtain that:

$$\begin{aligned} & \left| \mathbb{E}\left[\xi_1 | (X,Y)=(x,y)\right] - \log f(x,y) \right| \\ \leq & \left| \mathbb{E}\left[\xi_1 | (X,Y)=(x,y), \rho_{k,1}>0\right] - \log f(x,y) \right| \Pr\left(\rho_{k,1}>0\right) \\ & + \left| \mathbb{E}\left[\xi_1 | (X,Y)=(x,y), \rho_{k,1}=0\right] - \log f(x,y) \right| \Pr\left(\rho_{k,1}=0\right) \\ \leq & (2\log N + |\log f(x,y)|) k N^k e^{-(N-k)P_{XY}(x,y,0)} + \frac{10C+6}{NP_{XY}(x,y,0)} + \frac{1}{k}, \end{aligned} \tag{5.22}$$

where the first term comes from triangle inequality and the fact that $|\xi_1| \leq$

$2\log N$. Integrating over Ω_2, we have:

$$\begin{aligned}
& \int_{\Omega_2} \Big| \mathbb{E}\left[\xi_1 | (X,Y) = (x,y)\right] - \log f(x,y) \Big| \, dP_{XY} \\
\leq & \int_{\Omega_2} (2\log N + |\log f(x,y)|) k N^k e^{-(N-k)P_{XY}(x,y,0)} \, dP_{XY} \\
& + \frac{10C+6}{N} \int_{\Omega_2} \frac{1}{P_{XY}(x,y,0)} dP_{XY} + \frac{1}{k} \\
\leq & (2\log N + \int_{\Omega_2} |\log f(x,y)| dP_{XY}) k N^k e^{-(N-k)\inf_{(x,y)\in\Omega_2} P_{XY}(x,y,0)} \\
& + \frac{10C+6}{N} \mu(\Omega_2) + \frac{1}{k}, \qquad (5.23)
\end{aligned}$$

where μ denotes counting measure. By Assumption 3.(a), k goes to infinity as N goes to infinity, so $1/k$ vanishes as N increases. Assumptions 3.(a) and (b), k/N goes to 0 and Ω_2 has finite counting measure, so the second term also vanishes. Since Ω_2 has finite counting measure, so $\inf_{(x,y)\in\Omega_2} P_{XY}(x,y,0) = \epsilon > 0$. By Assumption 3.$(c)$, $\int_{\Omega_2} |\log f(x,y)| dP_{XY} < +\infty$. Therefore, for sufficiently large N, the first term also vanishes. So,

$$\lim_{N\to\infty} \int_{\Omega_2} \Big| \mathbb{E}\left[\xi_1 | (X,Y) = (x,y)\right] - \log f(x,y) \Big| \, dP_{XY} = 0. \qquad (5.24)$$

Hence the proof of the second case is completed.

Case III: $(x,y) \in \Omega_3$. In this case, $P_{XY}(x,y,r)$ is a monotonic function of r such that $P_{XY}(x,y,0) = 0$ and $\lim_{r\to\infty} P_{XY}(x,y,r) = 1$. Hence, we can view $\log\left(P_{XY}(x,y,r)/P_X(x,r)P_Y(y,r)\right)$ as a function of $P_{XY}(x,y,r)$, and it converges to $\log f(x,y)$ as $P_{XY}(x,y,r) \to 0$, for almost every (x,y). Since $P_{XY}(\Omega_3) \leq 1 < +\infty$ and $\int_{\Omega_3} |\log f(x,y)| dP_{XY} < +\infty$. Then by Egoroff's theorem, for any $\epsilon_N > 0$, there exists a subset $E \subseteq \Omega_3$ with $P_{XY}(E) < \epsilon_N$ and $\int_E |\log f(x,y)| dP_{XY} < \epsilon_N$, such that $\log\left(P_{XY}(x,y,r)/P_X(x,r)P_Y(y,r)\right)$ converges as $P_{XY}(x,y,r) \to 0$, uniformly on $\Omega_3 \setminus E$. For $(x,y) \in E$, notice that $|\xi_1| \leq 2\log N$, so we have:

$$\begin{aligned}
& \int_E \Big| \mathbb{E}\left[\xi_1 | (X,Y) = (x,y)\right] - \log f(x,y) \Big| \, dP_{XY} \\
\leq & \int_E \left(2\log N + |\log f(x,y)|\right) dP_{XY} < (2\log N + 1)\epsilon_N. \qquad (5.25)
\end{aligned}$$

By choosing $\epsilon_N = 1/N$, we will have $\lim_{N\to\infty} \int_E \Big| \mathbb{E}\,[\xi_1 | (X,Y) = (x,y)] - \log f(x,y) \Big| \, dP_{XY} = 0$.

Now for any $(x,y) \in \Omega_3 \setminus E$, since $P_{XY}(x,y,0) = 0$, we know that $\Pr(\, \rho_{k,1} = 0 \,|\, (X,Y) = (x,y)\,) = 0$, so $\tilde{k}_1 = k$ with probability 1. Conditioning on $\rho_{k,1} = r > 0$, the difference $\Big| \mathbb{E}\,[\xi_1 | (X,Y) = (x,y)] - \log f(x,y) \Big|$ can be decomposed into four parts as follows:

$$
\begin{aligned}
& \Big| \mathbb{E}\,[\xi_1 | (X,Y) = (x,y)] - \log f(x,y) \Big| \\
= \; & \Big| \int_{r=0}^{\infty} (\, \mathbb{E}\,[\xi_1 | (X,Y) = (x,y), \rho_{k,1} = r] - \log f(x,y)\,)\, dF_{\rho_{k,1}}(r) \Big| \\
\le \; & \Big| \int_{r=0}^{\infty} \left(\log \frac{P_{XY}(x,y,r)}{P_X(x,r) P_Y(y,r)} - \log f(x,y) \right) dF_{\rho_{k,1}}(r) \Big| \qquad (5.26) \\
+ \; & \Big| \int_{r=0}^{\infty} (\, \psi(k) - \log N - \log P_{XY}(x,y,r)\,)\, dF_{\rho_{k,1}}(r) \Big| \qquad (5.27) \\
+ \; & \Big| \int_{r=0}^{\infty} \Big(\mathbb{E}\,[\log(n_{x,1}+1) | (X,Y) = (x,y), \rho_{k,1} = r] \\
& \quad - \log(N P_X(x,r)) \Big) dF_{\rho_{k,1}}(r) \Big| \qquad (5.28) \\
+ \; & \Big| \int_{r=0}^{\infty} \Big(\mathbb{E}\,[\log(n_{y,1}+1) | (X,Y) = (x,y), \rho_{k,1} = r] \\
& \quad - \log(N P_Y(y,r)) \Big) dF_{\rho_{k,1}}(r) \Big|. \qquad (5.29)
\end{aligned}
$$

Here $F_{\rho_{k,1}}(r)$ is the CDF of the k-nearest neighbor distance $\rho_{k,1}$, given $(X,Y) = (x,y)$. By results of order statistics, its derivative with respect to $P_{XY}(x,y,r)$ is given by:

$$
\begin{aligned}
\frac{dF_{\rho_{k,1}}(r)}{dP_{XY}(x,y,r)} \;\; = \;\; & \frac{(N-1)!}{(k-1)!(N-k-1)!} P_{XY}(x,y,r)^{k-1} \\
& \times (\, 1 - P_{XY}(x,y,r)\,)^{N-k-1}. \qquad (5.30)
\end{aligned}
$$

Now we consider the four terms separately. For (5.26), since the quantity $\log(\, P_{XY}(x,y,r)/P_X(x,r)P_Y(y,r)\,)$ converges as $P_{XY}(x,y,r) \to 0$, uniformly on $\Omega_3 \setminus E$. So for every $(x,y) \in \Omega_3 \setminus E$, there exists an r_N such that $P_{XY}(x,y,r_N) = 4k \log N/N$ and $|\log(\, P_{XY}(x,y,r)/P_X(x,r)P_Y(y,r)\,) - \log f(x,y)| < \delta_N$ for every $r \le r_N$. Here r_N may depend on (x,y), but δ_N does not depend on (x,y) and $\lim_{N\to\infty} \delta_N = 0$. Therefore, (5.26) is upper

bounded by:

$$\begin{aligned}
& \Big| \int_{r=0}^{\infty} \left(\log \frac{P_{XY}(x,y,r)}{P_X(x,r)P_Y(y,r)} - \log f(x,y) \right) dF_{\rho_{k,1}}(r) \Big| \\
\leq & \int_{r=0}^{r_N} \Big| \log \frac{P_{XY}(x,y,r)}{P_X(x,r)P_Y(y,r)} - \log f(x,y) \Big| dF_{\rho_{k,1}}(r) \\
& + \int_{r=r_N}^{\infty} \Big| \log \frac{P_{XY}(x,y,r)}{P_X(x,r)P_Y(y,r)} - \log f(x,y) \Big| dF_{\rho_{k,1}}(r) \\
\leq & \delta_N p + \left(\sup_{r \geq r_N} \Big| \log \frac{P_{XY}(x,y,r)}{P_X(x,r)P_Y(y,r)} - \log f(x,y) \Big| \right) (1-p).
\end{aligned} \tag{5.31}$$

where p is the probability $\Pr(\rho_{k,1} \leq r_N \,|\, (X,Y) = (x,y))$ which is smaller than 1. Secondly, since $P_X(x,y,r) \geq 4k \log N/N > 1/N$ for $r \geq r_N$, so we have $|\log P_{XY}(x,y,r)| \leq \log N$. The same bounds apply for $|\log P_X(x,r)|$ and $|\log P_Y(y,r)|$ as well. By triangle inequality, the supremum is upper bounded by $3 \log N + |\log f(x,y)|$. Finally, the probability $1-p$ is upper bounded by

$$\begin{aligned}
& 1 - p = \Pr(\rho_{k,1} > r_N \,|\, (X,Y) = (x,y)) \\
= & \sum_{m=0}^{k-1} \tbinom{N-1}{m} P_{XY}(x,y,r_N)^m (1 - P_{XY}(x,y,r_N))^{N-1-m} \\
\leq & \sum_{m=0}^{k-1} N^m (1 - P_{XY}(x,y,r_N))^{N-k} = kN^k (1 - \frac{4k \log N}{N})^{N/2} \\
\leq & kN^k e^{-2k \log N} = \frac{k}{N^k},
\end{aligned} \tag{5.32}$$

for sufficiently large N such that $N - k > N/2$. Therefore, (5.26) is upper bounded by

$$\begin{aligned}
& \Big| \int_{r=0}^{\infty} \left(\log \frac{P_{XY}(x,y,r)}{P_X(x,r)P_Y(y,r)} - \log f(x,y) \right) dF_{\rho_{k,1}}(r) \Big| \\
\leq & \delta_N + \frac{k(3 \log N + |\log f(x,y)|)}{N^k}.
\end{aligned} \tag{5.33}$$

For (5.27), we simply plug in $F_{\rho_{k,1}}(r)$ and integrate over $P_{XY}(x,y,r)$ and

obtain

$$\begin{aligned}
&\int_{r=0}^{\infty} (\,\psi(k) - \log N - \log P_{XY}(x,y,r)\,)\, dF_{\rho_{k,1}}(r) \\
= \;& \psi(k) - \log N - \frac{(N-1)!}{(k-1)!(N-k-1)!} \int_{r=0}^{\infty} (\log P_{XY}(x,y,r)) \\
& P_{XY}(x,y,r)^{k-1}\, (\,1 - P_{XY}(x,y,r)\,)^{N-k-1}\, dP_{XY}(x,y,r) \\
= \;& \psi(k) - \log N - \frac{(N-1)!}{(k-1)!(N-k-1)!} \int_{t=0}^{1} (\log t) t^{k-1} (1-t)^{N-k-1} dt \\
= \;& \psi(k) - \log N - (\psi(k) - \psi(N)) = \psi(N) - \log N, \qquad (5.34)
\end{aligned}$$

where we use the fact that $\psi(k) - \psi(N) = \frac{(N-1)!}{(k-1)!(N-k-1)!} \int_{t=0}^{1} (\log t) t^{k-1} (1-t)^{N-k-1} dt$. Notice that $\psi(N) < \log N$ and $\lim_{N\to 0} (\psi(N) - \log N) = 0$.

Now we deal with (5.28) and (5.29). The following lemmas establish the distribution of $n_{x,1}$ and $n_{y,1}$ given $(X,Y) = (x,y)$ and $\rho_{k,1} = r > 0$.

Lemma 21. *Given $(X,Y) = (x,y)$ and $\rho_{k,1} = r > 0$, then $n_{x,1} - k$ is distributed as* $\mathrm{Bino}(N-k-1, \frac{P_X(x,r) - P_{XY}(x,y,r)}{1 - P_{XY}(x,y,r)})$*; $n_{y,1} - k$ is distributed as* $\mathrm{Bino}(N-k-1, \frac{P_Y(y,r) - P_{XY}(x,y,r)}{1 - P_{XY}(x,y,r)})$.

The following lemma is useful to establish the upper bound for (5.28) and (5.29).

Lemma 22. *For integer $m \geq 1$, if X is distributed as* $\mathrm{Bino}(N,p)$*, then $|\mathbb{E}[\log(X+m)] - \log(Np+m)| \leq C/(Np+m)$ for some constant C.*

Now we are ready to upper bound (5.28). First, we rewrite the term (5.28)

as:

$$
\begin{aligned}
& \Big| \int_{r=0}^{\infty} \Big(\mathbb{E}\left[\log(n_{x,1}+1) | (X,Y)=(x,y), \rho_{k,1}=r\right] \\
& - \log N - \log P_X(x,r) \Big) dF_{\rho_{k,1}}(r) \Big| \\
\leq \; & \Big| \int_{r=0}^{\infty} \Big(\mathbb{E}\left[\log(n_{x,1}+1) | (X,Y)=(x,y), \rho_{k,1}=r\right] \\
& - \log \left((N-k-1)\frac{P_X(x,r)-P_{XY}(x,y,r)}{1-P_{XY}(x,y,r)} + k + 1 \right) \Big) dF_{\rho_{k,1}}(r) \Big| \\
& + \Big| \int_{r=0}^{\infty} \left(\log \frac{(N-k-1)\frac{P_X(x,r)-P_{XY}(x,y,r)}{1-P_{XY}(x,y,r)} + k + 1}{N P_X(x,r)} \right) dF_{\rho_{k,1}}(r) \Big| \\
\leq \; & \int_{r=0}^{\infty} \Big| \mathbb{E}\left[\log(n_{x,1}+1) | (X,Y)=(x,y), \rho_{k,1}=r\right] \\
& - \log \left((N-k-1)\frac{P_X(x,r)-P_{XY}(x,y,r)}{1-P_{XY}(x,y,r)} + k + 1 \right) \Big| dF_{\rho_{k,1}}(r) \quad (5.35) \\
& + \Big| \mathbb{E}_r \left[\log \left(\frac{N(P_X(x,r)-P_{XY}(x,y,r)) + (k+1)(1-P_X(x,r))}{N P_X(x,r)(1-P_{XY}(x,y,r))} \right) \right] \Big|, \quad (5.36)
\end{aligned}
$$

where $\mathbb{E}_r$ denotes expectation over $F_{\rho_{i,xy}}$. By Lemma 22, the term (5.35) is upper bounded by

$$
\begin{aligned}
& \int_{r=0}^{\infty} \Big| \mathbb{E}\left[\log(n_{x,1}+1) | (X,Y)=(x,y), \rho_{k,1}=r\right] \\
& - \log \left((N-k-1)\frac{P_X(x,r)-P_{XY}(x,y,r)}{1-P_{XY}(x,y,r)} + k + 1 \right) \Big| dF_{\rho_{k,1}}(r) \\
\leq \; & \int_{r=0}^{\infty} \frac{C}{(N-k-1)\frac{P_X(x,r)-P_{XY}(x,y,r)}{1-P_{XY}(x,y,r)} + k + 1} dF_{\rho_{k,1}}(r) \\
\leq \; & \int_{r=0}^{\infty} \frac{C}{k+1} dF_{\rho_{k,1}}(r) = \frac{C}{k+1}. \quad (5.37)
\end{aligned}
$$

For (5.36), by the fact that $\log(x/y) \leq (x-y)/y$ for all $x, y > 0$ and Cauchy-

Schwarz inequality, we have the following:

$$\begin{aligned}
& \mathbb{E}_r\left[\log\left(\frac{N(P_X(x,r)-P_{XY}(x,y,r))+(k+1)(1-P_X(x,r))}{NP_X(x,r)(1-P_{XY}(x,y,r))}\right)\right] \\
\leq\ & \mathbb{E}_r\left[\frac{N(P_X(x,r)-P_{XY}(x,y,r))+(k+1)(1-P_X(x,r))}{NP_X(x,r)(1-P_{XY}(x,y,r))}-1\right] \\
=\ & \mathbb{E}_r\left[\frac{(k+1-NP_{XY}(x,y,r))(1-P_X(x,r))}{NP_X(x,r)(1-P_{XY}(x,y,r))}\right] \\
\leq\ & \sqrt{\mathbb{E}_r\left[\left(\frac{k+1-NP_{XY}(x,y,r)}{NP_{XY}(x,y,r)}\right)^2\right]} \\
& \times\sqrt{\mathbb{E}_r\left[\left(\frac{P_{XY}(x,y,r)(1-P_X(x,r))}{P_X(x,r)(1-P_{XY}(x,y,r))}\right)^2\right]}. \qquad (5.38)
\end{aligned}$$

Notice that $P_X(x,r) \geq P_{XY}(x,y,r)$ for all r, so the second expectation is always no larger than 1. For the first expectation, we plug in $F_{\rho_{k,1}}(r)$ and integrate over $P_{XY}(x,y,r)$, let $t = P_{XY}(x,y,r)$ and observe,

$$\begin{aligned}
& \mathbb{E}_r\left[\left(\frac{k+1-NP_{XY}(x,y,r)}{NP_{XY}(x,y,r)}\right)^2\right] \\
=\ & \int_{r=0}^{\infty}\left(\frac{k+1-NP_{XY}(x,y,r)}{NP_{XY}(x,y,r)}\right)^2 dF_{\rho_{i,xy}}(r) \\
=\ & \frac{(N-1)!}{(k-1)!(N-k-1)!}\int_{t=0}^{1}\frac{(k+1-Nt)^2}{N^2t^2}t^{k-1}(1-t)^{N-k-1}dt \\
=\ & \frac{(N-1)!}{(k-1)!(N-k-1)!}\frac{(k+1)^2}{N^2}\int_{t=0}^{1}t^{k-3}(1-t)^{N-k-1}dt \\
& -\frac{(N-1)!}{(k-1)!(N-k-1)!}\frac{2(k+1)}{N^2}\int_{t=0}^{1}t^{k-2}(1-t)^{N-k-1}dt \\
& +\frac{(N-1)!}{(k-1)!(N-k-1)!}\int_{t=0}^{1}t^{k-3}(1-t)^{N-k-1}dt \\
=\ & \frac{(N-1)!}{(k-1)!(N-k-1)!}\frac{(k+1)^2}{N^2}\frac{(k-3)!(N-k-1)!}{(N-3)!} \\
& -\frac{(N-1)!}{(k-1)!(N-k-1)!}\frac{2(k+1)}{N^2}\frac{(k-2)!(N-k-1)!}{(N-2)!}+1 \\
=\ & \frac{(N-1)(N-2)(k+1)^2}{N^2(k-1)(k-2)}-\frac{2(N-1)(k+1)}{N(k-1)}+1. \qquad (5.39)
\end{aligned}$$

For sufficiently large N and k, it is upper bounded by $C_1(1/N+1/k)$ for

some constant $C_1 > 0$. Therefore,

$$\begin{aligned} & \mathbb{E}_r \left[\log \left(\frac{N(P_X(x,r) - P_{XY}(x,y,r)) + (k+1)(1 - P_X(x,r))}{N P_X(x,r)(1 - P_{XY}(x,y,r))} \right) \right] \\ \leq \; & \sqrt{C_1(\frac{1}{N} + \frac{1}{k})}. \end{aligned} \tag{5.40}$$

Similarly, by using the fact that $\log(x/y) > (x-y)/x$ and Cauchy-Schwarz inequality again, we conclude that there are some constant $C_2 > 0$ such that

$$\begin{aligned} & \mathbb{E}_r \left[\log \left(\frac{N(P_X(x,r) - P_{XY}(x,y,r)) + (k+1)(1 - P_X(x,r))}{N P_X(x,r)(1 - P_{XY}(x,y,r))} \right) \right] \\ \geq \; & -\sqrt{C_2(\frac{1}{N} + \frac{1}{k})}. \end{aligned} \tag{5.41}$$

Therefore, by combining (5.37), (5.40) and (5.41), we obtain

$$\begin{aligned} & \Big| \int_{r=0}^{\infty} \Big(\mathbb{E}\left[\log(n_{x,1}+1) | (X,Y) = (x,y), \rho_{k,1} = r\right] - \log N \\ & - \log P_X(x,r) \Big) dF_{\rho_{k,1}}(r) \Big| \leq \frac{C}{k+1} + \sqrt{C'(\frac{1}{N} + \frac{1}{k})}, \end{aligned} \tag{5.42}$$

where $C' = \max\{C_1, C_2\}$. Since (5.29) and (5.28) are symmetric, the same upper bound (5.42) also applies to (5.29). Combine (5.33), (5.34) and (5.42), we have

$$\begin{aligned} & \Big| \mathbb{E}\left[\xi_1 | (X,Y) = (x,y)\right] - \log f(x,y) \Big| \leq \delta_N + \frac{k(3\log N + |\log f(x,y)|)}{N^k} \\ & + \log N - \psi(N) + \frac{2C}{k+1} + 2\sqrt{C'(\frac{1}{N} + \frac{1}{k})}, \end{aligned} \tag{5.43}$$

for every $(x, y) \in \Omega_3 \setminus E$. By integration over $\Omega_3 \setminus E$, we have

$$\begin{aligned}
& \int_{\Omega_3 \setminus E} \Big| \mathbb{E}\left[\xi_1 | (X, Y) = (x, y)\right] - \log f(x, y) \Big| \, dP_{XY} \\
\leq \; & \int_{\Omega_3 \setminus E} \Big(\delta_N + \frac{k(3 \log N + |\log f(x, y)|)}{N^k} + \log N - \psi(N) \\
& + \frac{2C}{k+1} + 2\sqrt{C'(\frac{1}{N} + \frac{1}{k})} \Big) dP_{XY} \\
\leq \; & \delta_N + \frac{k(3 \log N + \int_{\mathcal{X} \times \mathcal{Y}} |\log f(x, y)| dP_{XY})}{N^k} + \log N - \psi(N) \\
& + \frac{2C}{k+1} + 2\sqrt{C'(\frac{1}{N} + \frac{1}{k})}. \qquad (5.44)
\end{aligned}$$

By Assumption 3.(c), $\int_{\mathcal{X} \times \mathcal{Y}} |\log f(x, y)| dP_{XY} < +\infty$. By Assumption 3.$(a)$, k increases as $N \to \infty$. Therefore, this quantity vanishes as $N \to \infty$. Combining with the case that $(x, y) \in E$, we have

$$\lim_{N \to \infty} \int_{\Omega_3} \Big| \mathbb{E}\left[\xi_1 | (X, Y) = (x, y)\right] - \log f(x, y) \Big| \, dP_{XY} = 0. \qquad (5.45)$$

5.5.1 Proof of Lemma 19

The proof of this lemma utilizes the Lebesgue-Besicovitch differentiation theorem [70, Theorem 1.32], stated below.

Lemma 23 (Lebesgue-Besicovitch Differentiation Theorem)**.** *Let μ be a Radon measure on $\mathbb{R}^n$. For $f \in L^1_{loc}(\mu)$,*

$$\lim_{r \to 0} \frac{1}{\mu(\bar{B}_r(x))} \int_{\bar{B}_r(x)} f d\mu = f(x), \qquad (5.46)$$

for μ-a.e. x.

For our lemma, let $f = \frac{dP_{XY}}{dP_X P_Y}$ and $\mu = P_X P_Y$. Since μ is a probability measure, it is a Radon measure of Euclidean space. Also, since $\int_{\mathcal{X} \times \mathcal{Y}} |f| d\mu = 1$, so f is globally integrable, hence locally integrable with respect to μ. So the conditions of Lebesgue-Besicovitch differentiation theorem are satisfied

and

$$f(x,y) = \frac{dP_{XY}}{dP_X P_Y}(x,y) = \lim_{r\to 0} \frac{1}{P_X P_Y(\bar{B}_r(x,y))} \int_{\bar{B}_r(x,y)} \frac{dP_{XY}}{dP_X P_Y} dP_X P_Y$$

$$= \lim_{r\to 0} \frac{P_{XY}(\bar{B}_r(x,y))}{P_X P_Y(\bar{B}_r(x,y))} = \lim_{r\to 0} \frac{P_{XY}(x,y,r)}{P_X(x,r)P_Y(y,r)}. \tag{5.47}$$

5.5.2 Proof of Lemma 20

First, we upperbound $\mathbb{E}\left[\log(X)|X \geq k\right] - \log(Np)$. We can see that:

$$\mathbb{E}\left[X+m|X\geq k\right] = \frac{1}{\Pr\left(X\geq k\right)} \sum_{i=k}^{N} (i+m) \left(\begin{array}{c} N \\ i \end{array} \right) p^i (1-p)^{N-i}$$

$$\leq \frac{1}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)} \sum_{i=k}^{N} (i+m) \left(\begin{array}{c} N \\ i \end{array} \right) p^i (1-p)^{N-i}$$

$$\leq \frac{1}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)} \sum_{i=1}^{N} (i+m) \left(\begin{array}{c} N \\ i \end{array} \right) p^i (1-p)^{N-i}$$

$$= \frac{1}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)} \left(\mathbb{E}\left[X\right]+m\right)$$

$$= \frac{Np+m}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)}, \tag{5.48}$$

in which we used the Hoeffding's inequality. Since $\mathbb{E}\left[\log(X+m)|X\geq k\right] \leq \log\left(\mathbb{E}\left[X+m|X\geq k\right]\right)$, thus:

$$\mathbb{E}\left[\log(X)|X\geq k\right] - \log(Np) \leq \log\left(\frac{1+\frac{m}{Np}}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)}\right). \tag{5.49}$$

Second, to give an upper bound over $\log(Np) - \mathbb{E}\left[\log(X+m)|X\geq k\right]$, we first notice that:

$$\log(Np) - \mathbb{E}\left[\log(X+m)|X\geq k\right] \leq \log(Np) - \mathbb{E}\left[\log(X)|X\geq k\right]. \tag{5.50}$$

Then we upperbound $\log(Np) - \mathbb{E}\left[\log(X)|X\geq k\right]$ by applying Taylor's

theorem around $x_0 = Np$, where there exists ζ between x and x_0 such that:

$$\log(x) = \log(Np) + \frac{x - Np}{Np} - \frac{(x - Np)^2}{2\zeta^2}. \tag{5.51}$$

Since $\zeta \geq \min\{x, x_0\} = \min\{x, Np\}$, we have:

$$\begin{aligned} & -\log(x) + \log(Np) + \frac{x - Np}{Np} = \frac{(x - Np)^2}{2\zeta^2} \\ \leq \; & \max\left\{\frac{(x - Np)^2}{2x^2}, \frac{(x - NP)^2}{2(Np)^2}\right\} \\ \leq \; & \frac{(x - Np)^2}{2x^2} + \frac{(x - Np)^2}{2(Np)^2}. \end{aligned} \tag{5.52}$$

Now taking the conditional expectations from both sides, we have:

$$\begin{aligned} & -\mathbb{E}\left[\log(X)|X \geq k\right] + \log(Np) + \frac{\mathbb{E}\left[X|X \geq k\right] - Np}{Np} \\ \leq \; & \mathbb{E}\left[\frac{(X - Np)^2}{2X^2}\middle| X \geq k\right] + \frac{\mathbb{E}\left[(X - Np)^2|X \geq k\right]}{2(Np)^2}. \end{aligned} \tag{5.53}$$

First, we notice that $\mathbb{E}\left[X|X \geq k\right] \geq \mathbb{E}\left[X\right] = Np$. Second,

$$\begin{aligned} & \mathbb{E}\left[(X - Np)^2\middle|X \geq k\right] \leq \frac{1}{1 - \exp\left(-2\frac{(Np-k)^2}{N}\right)}\mathrm{Var}\left[X\right] \\ = \; & \frac{Np(1-p)}{1 - \exp\left(-2\frac{(Np-k)^2}{N}\right)}. \end{aligned} \tag{5.54}$$

Thus we can write:

$$\begin{aligned} & -\mathbb{E}\left[\log(X)|X \geq k\right] + \log(Np) \\ \leq \; & \frac{Np(1-p)}{1 - \exp\left(-2\frac{(Np-k)^2}{N}\right)}\frac{1}{2(Np)^2} + \mathbb{E}\left[\frac{(X - Np)^2}{2X^2}\middle| X \geq k\right]. \end{aligned} \tag{5.55}$$

To deal with the term $\mathbb{E}\left[\frac{(X-Np)^2}{2X^2}\middle| X \geq k\right]$, we have:

$$
\begin{aligned}
& \mathbb{E}\left[\frac{(X-Np)^2}{2X^2}\middle| X \geq k\right] \\
\leq\ & \frac{1}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)} \sum_{i=k}^{N} \frac{(i-Np)^2}{2i^2} \binom{N}{i} p^i (1-p)^{N-i} \\
\leq\ & \frac{1}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)} \sum_{i=k}^{N} \frac{(i-Np)^2}{(i+1)(i+2)} \binom{N}{i} p^i (1-p)^{N-i} \\
=\ & \frac{1}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)} \sum_{i=k}^{N} \frac{(i-Np)^2}{(N+1)(N+2)p^2} \binom{N+2}{i+2} p^{2+i} (1-p)^{N-i} \\
\leq\ & \frac{1}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)} \frac{1}{(N+1)(N+2)p^2} \mathbb{E}_{Y\sim\mathrm{Bino}(N+2,p)}\left[(Y-Np)^2\right] \\
=\ & \frac{1}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)} \frac{(N+2)p(1-p)+4p^2}{(N+1)(N+2)p^2} \\
\leq\ & \frac{1}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)} \frac{(N+2)p}{(N+1)(N+2)p^2} \\
\leq\ & \frac{1}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)} \frac{1}{Np}, \qquad (5.56)
\end{aligned}
$$

in which we used the fact that $2i^2 \geq (i+1)(i+2)$ for $i \geq 4$, and $(N+2)p \geq 4p$ for $N \geq 2$. Plugging it into (5.55), we have the desired result:

$$
-\mathbb{E}\left[\log(X)|X \geq k\right] + \log(Np) \leq \frac{1}{1-\exp\left(-2\frac{(Np-k)^2}{N}\right)} \frac{3}{2Np}. \qquad (5.57)
$$

5.5.3 Proof of Lemma 21

Now we deal with the case that $\rho_{k,1} = r > 0$. Given that $(X_1, Y_1) = (x, y)$ and $\rho_{k,1} = r > 0$, we sort the samples $\{(X_i, Y_i)\}_{i=2}^N$ by their distance to (x, y) defined as $d_i = \max\{\|X_i - x\|, \|Y_i - y\|\}$. To avoid the case that two samples have identical distance, we introduce a set of random variables $\{Z_i\}_{i=2}^N$ i.i.d. samples from $\mathrm{Unif}[0, 1]$ and define a comparison operator $\prec$ as:

$$
i \prec j \iff d_i < d_j \quad \text{or} \quad \{d_i = d_j \quad \text{and} \quad Z_i < Z_j\}. \qquad (5.58)
$$

Since for any $i \neq j$, the probability that $Z_i = Z_j$ is zero, so we can have either $i \prec j$ or $i \succ j$ with probability 1. Now let $\{2, 3, \ldots, N\} = S \cup \{j\} \cup T$ be a partition of the indices with $|S| = k-1$ and $|T| = N-k-1$. Define an event $\mathcal{A}_{S,j,T}$ associated to the partition as:

$$\mathcal{A}_{S,j,T} = \big\{\, s \prec j, \forall s \in S, \text{ and } t \succ j, \forall t \in T \,\big\}. \tag{5.59}$$

Since $(X_j, Y_j) - (x, y)$ are i.i.d. random variables each of the events $\mathcal{A}_{S,j,T}$ has identical probability. The number of all partitions is $\frac{(N-1)!}{(N-k-1)!(k-1)!}$ and thus $\Pr(\mathcal{A}_{S,j,T}) = \frac{(N-k-1)!(k-1)!}{(N-1)!}$. So the cdf of $n_{x,1}$ is given by:

$$\begin{aligned}
& \Pr\big(n_{x,1} \leq k+m \big| \rho_{k,1} = r, (X_1, Y_1) = (x, y)\big) \\
= & \sum_{S,j,T} \Pr(\mathcal{A}_{S,j,T} \,|\, \rho_{k,1} = r, (X_1, Y_1) = (x, y)) \\
& \Pr\big(n_{x,1} \leq k+m \big| \mathcal{A}_{S,j,T}, \rho_{k,1} = r, (X_1, Y_1) = (x, y)\big) \\
= & \frac{(N-k-1)!(k-1)!}{(N-1)!} \sum_{S,j,T} \Pr\Big(n_{x,1} \leq k+m \Big| \mathcal{A}_{S,j,T}, \rho_{k,1} = r, \\
& (X_1, Y_1) = (x, y)\Big).
\end{aligned} \tag{5.60}$$

Now condition on event $\mathcal{A}_{S,j,T}$ and $\rho_{k,1} = r$, namely (X_j, Y_j) is the k-nearest neighbor with distance r, S is the set of samples with distance smaller than (or equal to) r and T is the set of samples with distance greater than (or equal to) r. Recall that $n_{x,1}$ is the number of samples with $\|X_j - x\| \leq r$. For any index $s \in S \cup \{j\}$, $\|X_j - x\| \leq r$ are satisfied. Therefore, $n_{x,1} \leq k+m$ means that there are no more than m samples in T with $\mathcal{X}$-distance smaller than r. Let $U_l = \mathbb{I}\{\|X_l - x\| \leq r \,\big|\, d_l \geq r\}$. Therefore,

$$\begin{aligned}
& \Pr\big(n_{x,1} \leq k+m \big| \mathcal{A}_{S,j,T}, \rho_{k,1} = r, (X_1, Y_1) = (x, y)\big) \\
= & \Pr\Big(\sum_{l \in T} \mathbb{I}\{\|X_l - x\| \leq r\} \leq m \,\Big|\, d_s \leq r, \forall s \in S, d_j = r, d_t \geq r, \forall t \in T\Big) \\
= & \Pr\left(\sum_{l \in T} \mathbb{I}\{\|X_l - x\| \leq r\} \leq m \,\middle|\, d_l \geq r, \forall l \in T\right) \\
= & \Pr\left(\sum_{l \in T} U_l \leq m\right),
\end{aligned} \tag{5.61}$$

where U_l follows Bernoulli distribution with $\Pr\{U_l = 1\} = Pr\{\|X_l - x\| \leq$

$r|d_l \geq r\}$. We can drop the conditioning of (X_s, Y_s)'s for $s \notin T$ since (X_s, Y_s) and (X_t, Y_t) are independent. Therefore, given that $d_l \geq r$ for all $l \in T$, the variables $\mathbb{I}\{\|X_l - x\| \leq r\}$ are i.i.d. and have the same distribution as U_l. We conclude:

$$\begin{aligned}
& \Pr\left(n_{x,1} \leq k+m \middle| \rho_{k,1} = r, (X_1, Y_1) = (x, y) \right) \\
= \; & \frac{(N-k-1)!(k-1)!}{(N-1)!} \sum_{S,j,T} \Pr\Big(n_{x,1} \leq k+m \Big| \mathcal{A}_{S,j,T}, \rho_{i,xy} = r, \\
& (X_1, Y_1) = (x, y) \Big) \\
= \; & \frac{(N-k-1)!(k-1)!}{(N-1)!} \sum_{S,j,T} \Pr\left(\sum_{l \in T} U_l \leq m \right) \\
= \; & \Pr\left(\sum_{l \in T} U_l \leq m \right). \qquad (5.62)
\end{aligned}$$

Thus we have shown that $n_{x,1} - k$ has the same distribution as $\sum_{l \in T} U_l$, which is a binomial random variable with parameter $|T| = N - k - 1$ and $\Pr\{\|X_l - x\| \leq r \,|\, d_l \geq r\} = \frac{P_X(x,r) - P_{XY}(x,y,r)}{1 - P_{XY}(x,y,r)}$. For $n_{y,1}$, we can follow the same proof and conclude that $n_{y,1} - k \sim \text{Bino}(N-k-1, \frac{P_Y(x,r) - P_{XY}(x,y,r)}{1 - P_{XY}(x,y,r)})$.

5.5.4 Proof of Lemma 22

By Jensen's inequality, we know that $\mathbb{E}[\log X] \leq \log \mathbb{E}[X] = \log(Np+m)$. So it suffices to give an upper bound for $\log(Np+m) - \mathbb{E}[\log X]$. We consider two different cases.

Case I: $Np \geq m$. In this case, for any x, by applying Taylor's theorem around $x_0 = Np + m$, there exists ζ between x and x_0 such that

$$\log(x) = \log(Np+m) + \frac{x - Np - m}{Np+m} - \frac{(x - Np - m)^2}{2\zeta^2}. \qquad (5.63)$$

By noticing that $\zeta \geq \min\{x, x_0\} = \min\{x, Np+m\}$, we have

$$\begin{aligned} & -\log(x) + \log(Np+m) + \frac{x-Np-m}{Np+m} = \frac{(x-Np-m)^2}{2\zeta^2} \\ \leq \;& \max\{\frac{(x-Np-m)^2}{2x^2}, \frac{(x-Np-m)^2}{2(Np+m)^2}\} \\ \leq \;& \frac{(x-Np-m)^2}{2x^2} + \frac{(x-Np-m)^2}{2(Np+m)^2}. \end{aligned} \tag{5.64}$$

Now let $X-m$ be a $\text{Bino}(N,p)$ random variable. By taking expectation on both sides, we have:

$$\begin{aligned} & -\mathbb{E}[\log X] + \log(Np+m) + \frac{\mathbb{E}[X]-Np-m}{Np+m} \\ \leq \;& \mathbb{E}\left[\frac{(X-Np-m)^2}{2X^2}\right] + \frac{\mathbb{E}\left[(X-Np-m)^2\right]}{2(Np+m)^2}. \end{aligned} \tag{5.65}$$

Since $\mathbb{E}[X] = Np+m$, $\mathbb{E}\left[(X-Np-m)^2\right] = \text{Var}[X] = Np(1-p)$, and

$$\begin{aligned} & \mathbb{E}\left[\frac{(X-Np-m)^2}{2X^2}\right] = \sum_{j=0}^{N} \frac{(j-Np)^2}{2(j+m)^2} \binom{N}{j} p^j (1-p)^{N-j} \\ \leq \;& \sum_{j=0}^{N} \frac{(j-Np)^2}{(j+2)(j+1)} \binom{N}{j} p^j (1-p)^{N-j} \\ = \;& \sum_{j=0}^{N} \frac{(j-Np)^2}{(N+2)(N+1)p^2} \binom{N+2}{j+2} p^{j+2} (1-p)^{N-j} \\ \leq \;& \frac{1}{(N+2)(N+1)p^2} \mathbb{E}_{Y\sim\text{Bino}(N+2,p)}\left[(Y-Np)^2\right] \\ = \;& \frac{(N+2)p(1-p)+4p^2}{(N+2)(N+1)p} \leq \frac{(N+2)p}{(N+2)(N+1)p} \leq \frac{1}{Np}, \end{aligned} \tag{5.66}$$

for $m \geq 1$ and $N \geq 4$. Plug these in (5.65), we have

$$\begin{aligned} & -\mathbb{E}[\log X] + \log(Np+m) \leq \frac{1}{Np} + \frac{Np(1-p)}{2(Np+m)^2} \\ \leq \;& \frac{2}{Np+m} + \frac{1}{2(Np+m)} = \frac{5}{2(Np+m)}, \end{aligned} \tag{5.67}$$

where $1/(2Np) \leq 1/(Np+m)$ comes from the fact that $Np \geq m$.

Case II: $Np < m$. In this case, for any x, by applying Taylor's theorem around $x_0 = Np + m$, there exists ζ between x and x_0 such that

$$\log(x) = \log(Np+m) + \frac{x - Np - m}{Np+m} - \frac{(x-Np-m)^2}{2\zeta^2}. \tag{5.68}$$

By noticing that $\zeta \geq \min\{x, x_0\} \geq m \geq (Np+m)/2$, we have:

$$-\log(x) + \log(Np+m) + \frac{x-Np-m}{Np+m} \leq \frac{2(x-Np-m)^2}{(Np+m)^2}. \tag{5.69}$$

Similarly, by taking expectation on both sides, we have

$$\begin{aligned} & -\mathbb{E}[\log X] + \log(Np+m) + \frac{\mathbb{E}[X] - Np - m}{Np+m} \\ \leq \; & \frac{\mathbb{E}\left[\, 2(X-Np-m)^2 \,\right]}{(Np+m)^2}. \end{aligned} \tag{5.70}$$

By plugging in $\mathbb{E}[X] = Np + m$ and $\mathbb{E}\left[\, (X-Np-k)^2 \,\right] = \mathrm{Var}[X] = Np(1-p)$, we obtain

$$\begin{aligned} & -\mathbb{E}[\log X] + \log(Np+m) \leq \frac{2Np(1-p)}{(Np+m)^2} \\ \leq \; & \frac{2(Np+m)}{(Np+m)^2} = \frac{2}{Np+m}. \end{aligned} \tag{5.71}$$

Combining the two cases, we obtain the desired statement.

5.6 Proof of Theorem 11 on the variance

We use the Efron-Stein inequality to bound the variance of the estimator. For simplicity, let $\widehat{I}^{(N)}(Z)$ be the estimate based on original samples $\{Z_1, Z_2, \ldots, Z_N\}$, where $Z_i = (X_i, Y_i)$. For the usage of Efron-Stein inequality, we consider another set of i.i.d. samples $\{Z'_1, Z'_2, \ldots, Z'_n\}$ drawn from P_{XY}. Let $\widehat{I}^{(N)}(Z^{(j)})$ be the estimate based on $\{Z_1, \ldots, Z_{j-1}, Z'_j, Z_{j+1}, \ldots, Z_N\}$. Then the Efron-Stein inequality states that

$$\mathrm{Var}\left[\, \widehat{I}^{(N)}(Z) \,\right] \;\leq\; \frac{1}{2}\sum_{j=1}^{N} \mathbb{E}\left[\left(\widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)}) \right)^2 \right]. \tag{5.72}$$

Now we will give an upper bound for the difference $|\widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)})|$ for given index j. First of all, let $\widehat{I}^{(N)}(Z_{\setminus j})$ be the estimate based on the rest of samples $\{Z_1, \ldots, Z_{j-1}, Z_{j+1}, \ldots, Z_N\}$, then by triangle inequality, we have:

$$\begin{aligned}
& \sup_{Z_1,\ldots,Z_N,Z'_j} \Big| \widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)}) \Big| \\
\leq & \sup_{Z_1,\ldots,Z_N,Z'_j} \Big(\Big| \widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z_{\setminus j}) \Big| + \Big| \widehat{I}^{(N)}(Z_{\setminus j}) - \widehat{I}^{(N)}(Z^{(j)}) \Big| \Big) \\
\leq & \sup_{Z_1,\ldots,Z_N} \Big| \widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z_{\setminus j}) \Big| \\
& + \sup_{Z_1,\ldots,Z_{j-1},Z'_j,Z_{j+1},\ldots,Z_N} \Big| \widehat{I}^{(N)}(Z_{\setminus j}) - \widehat{I}^{(N)}(Z^{(j)}) \Big| \\
= & 2 \sup_{Z_1,\ldots,Z_N} \Big| \widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z_{\setminus j}) \Big|, \qquad (5.73)
\end{aligned}$$

where the last equality comes from the fact that $\{Z_1, \ldots, Z_{j-1}, Z'_j, Z_{j+1}, \ldots, Z_N\}$ has the same joint distribution as $\{Z_1, \ldots, Z_N\}$. Now recall that

$$\begin{aligned}
\widehat{I}^{(N)}(Z) &= \frac{1}{N} \sum_{i=1}^{N} \xi_i(Z) \\
&= \frac{1}{N} \sum_{i=1}^{N} \Big(\psi(\tilde{k}_i) + \log N - \log(n_{x,i}+1) - \log(n_{y,i}+1) \Big). \qquad (5.74)
\end{aligned}$$

Therefore, we have

$$\sup_{Z_1,\ldots,Z_N,Z'_j} \Big| \widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)}) \Big| \leq \frac{2}{N} \sup_{Z_1,\ldots,Z_N} \sum_{i=1}^{N} \Big| \xi_i(Z) - \xi_i(Z_{\setminus j}) \Big|. \qquad (5.75)$$

Now we need to upper-bound the difference $|\,\xi_i(Z) - \xi_i(Z_{\setminus j})\,|$ created by eliminating sample Z_j for different i 's. There are three cases of i's as follows,

- **Case I:** $i = j$. Since the upper bounds $|\xi_i(Z)| \leq 2\log N$ and $|\xi_i(Z_{\setminus j})| \leq 2\log(N-1)$ always holds, so $|\,\xi_i(Z) - \xi_i(Z_{\setminus j})\,| \leq 4\log N$. The number of i's in this case is only 1. So $\sum_{\text{Case I}} |\,\xi_i(Z) - \xi_i(Z_{\setminus j})\,| \leq 4\log N$.
- **Case II:** $\rho_{i,xy} = 0$. In this case, recall that $\tilde{k}_i = \Big| \{i' \neq i : Z_i = Z_{i'}\} \Big|$, $n_{x,i} = \Big| \{i' \neq i : X_i = X_{i'}\} \Big|$ and $n_{y,i} = \Big| \{i' \neq i : Y_i = Y_{i'}\} \Big|$. There are four sub-cases in this case.

– **Case II.1:** $Z_i = Z_j$. By eliminating Z_j, $\tilde{k}_i$, $n_{x,i}$, $n_{y,i}$ will all decrease by 1. Therefore,

$$\begin{aligned} & |\,\xi_i(Z) - \xi_i(Z_{\setminus j})\,| \\ = \;& |\,\Big(\psi(\tilde{k}_i) + \log N - \log(n_{x,i}+1) - \log(n_{y,i}+1)\Big) \\ & - \Big(\psi(\tilde{k}_i - 1) + \log(N-1) - \log(n_{x,i}) - \log(n_{y,i})\Big)\,| \\ \leq \;& |\psi(\tilde{k}_i) - \psi(\tilde{k}_i - 1)| + |\log N - \log(N-1)| \\ & + |\log(n_{x,i}+1) - \log(n_{x,i})| + |\log(n_{y,i}+1) - \log(n_{y,i})| \\ \leq \;& \frac{1}{\tilde{k}_i - 1} + \frac{1}{N-1} + \frac{1}{n_{x,i}} + \frac{1}{n_{y,i}} \leq \frac{4}{\tilde{k}_i - 1} = \frac{4}{\tilde{k}_j - 1}. \end{aligned} \tag{5.76}$$

The number of i's in this case is the number if i's such that $Z_i = Z_j$, which is just $\tilde{k}_j$. Therefore, $\sum_{\text{Case II.1}} |\,\xi_i(Z) - \xi_i(Z_{\setminus j})\,| \leq 4\tilde{k}_j/(\tilde{k}_j - 1) \leq 8$, for $\tilde{k}_j \geq k \geq 2$.

– **Case II.2:** $X_i = X_j$ but $Y_i \neq Y_j$. By eliminating Z_j, $\tilde{k}_i$ and $n_{y,i}$ would not change but $n_{x,i}$ will decrease by 1. Therefore,

$$\begin{aligned} & |\,\xi_i(Z) - \xi_i(Z_{\setminus j})\,| \\ \leq \;& |\log N - \log(N-1)| + |\log(n_{x,i}+1) - \log(n_{x,i})| \\ \leq \;& \frac{1}{N-1} + \frac{1}{n_{x,i}} \leq \frac{2}{n_{x,i}} = \frac{2}{n_{x,j}}. \end{aligned} \tag{5.77}$$

The number of i's in this case is the number if i's such that $X_i = X_j$ but $Y_i \neq Y_j$, which is less than $n_{x,j}$. Therefore, $\sum_{\text{Case II.2}} |\,\xi_i(Z) - \xi_i(Z_{\setminus j})\,| \leq 2n_{x,j}/n_{x,j} \leq 2$.

– **Case II.3:** $Y_i = Y_j$ but $X_i \neq X_j$. By eliminating Z_j, $\tilde{k}_i$ and $n_{x,i}$ would not change but $n_{y,i}$ will decrease by 1. Similarly as Case II.2, we have $\sum_{\text{Case II.3}} |\,\xi_i(Z) - \xi_i(Z_{\setminus j})\,| \leq 2$.

– **Case II.4:** $X_i \neq X_j$ and $Y_i \neq Y_j$. In this case, none of $\tilde{k}_i$, $n_{x,i}$, or $n_{y,i}$ will change. So $|\,\xi_i(Z) - \xi_i(Z_{\setminus j})\,| = \log N - \log(N-1) \leq 1/(N-1)$. The number of i's in this case is simply less than $N-1$. Therefore, $\sum_{\text{Case II.4}} |\,\xi_i(Z) - \xi_i(Z_{\setminus j})\,| \leq 1$.

Combining the cases, we conclude that $\sum_{\text{Case II}} |\,\xi_i(Z) - \xi_i(Z_{\setminus j})\,| \leq 13$.

- **Case III:** $\rho_{i,xy} > 0$. In this case, recall that $\tilde{k}_i$ always equals to k, $n_{x,i} = \Big|\,\{i' \neq i : \|X_i - X_{i'}\| \leq \rho_{i,xy}\}\,\Big|$ and $n_{y,i} = \Big|\,\{i' \neq i : \|Y_i - Y_{i'}\| \leq$

$\rho_{i,xy}\}\Big|$. So the analysis will be the same as the analysis of variance of classical KSG estimator in Section 3.7. In this case, we have

$$\begin{aligned} & \sum_{\text{Case III}} |\, \xi_i(Z) - \xi_i(Z_{\setminus j})\,| \\ \leq\ & 2\log N + 4k\gamma_{d_x+d_y}\log N + 4\gamma_{d_x+d_y}(\log N + 1). \end{aligned} \tag{5.78}$$

Combining the three cases, we have:

$$\begin{aligned} & \sum_{i=1}^{N} \Big|\, \xi_i(Z) - \xi_i(Z_{\setminus j})\,\Big| \\ \leq\ & 6\log N + 13 + 4k\gamma_{d_x+d_y}\log N + 4\gamma_{d_x+d_y}(\log N + 1) \\ \leq\ & 31\gamma_{d_x+d_y}k\log N, \end{aligned} \tag{5.79}$$

for $k \geq 1$, $\log N \geq 1$ and all $\{Z_1, \ldots, Z_N\}$. Plug it into (5.75), and we obtain:

$$\sup_{Z_1,\ldots,Z_N,Z'_j} \Big|\, \widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)})\,\Big| \leq \frac{62\gamma_{d_x+d_y}k\log N}{N}. \tag{5.80}$$

Plug it into Efron-Stein inequality (5.72), and we obtain:

$$\begin{aligned} & \mathrm{Var}\Big[\,\widehat{I}^{(N)}(Z)\,\Big] \leq \frac{1}{2}\sum_{j=1}^{N}\mathbb{E}\Big[\,\Big(\,\widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)})\,\Big)^2\,\Big] \\ \leq\ & \frac{1}{2}\sum_{j=1}^{N}\sup_{Z_1,\ldots,Z_n,Z'_j}\Big(\,\widehat{I}^{(N)}(Z) - \widehat{I}^{(N)}(Z^{(j)})\,\Big)^2 \\ \leq\ & \frac{1}{2}\sum_{j=1}^{N}(\frac{62\gamma_{d_x+d_y}k\log N}{N})^2 = \frac{1922\gamma^2_{d_x+d_y}(k\log N)^2}{N}. \end{aligned} \tag{5.81}$$

Since $1922\gamma^2_{d_x+d_y}$ is a constant independent of N, and $(k_N\log N)^2/N \to 0$ as $N \to \infty$ by Assumption 3.(f), we have $\lim_{N\to\infty}\mathrm{Var}\Big[\,\widehat{I}^{(N)}(Z)\,\Big] = 0$.

CHAPTER 6

DISCOVERING POTENTIAL CORRELATIONS VIA INFORMATION BOTTLENECK

Measuring the strength of an association between two random variables is a fundamental topic of broad scientific interest. Pearson's correlation coefficient [153] dates from over a century ago and has been generalized seven decades ago as maximal correlation (mCor) to handle nonlinear dependencies [154, 155, 156]. Novel correlation measures to identify different kinds of associations continue to be proposed in the literature; these include maximal information coefficient (MIC) [11] and distance correlation (dCor) [157]. Despite the differences, a common theme of measurement of the empirical *average* dependence unites the different dependence measures. Alternatively, these are *factual* measures of dependence and their relevance is restricted when we seek a *potential* dependence of one random variable on another. For instance, consider a hypothetical city with very few smokers. A standard measure of correlation on the historical data in this town on smoking and lung cancer will fail to discover the fact that smoking causes cancer, since the average correlation is very small. On the other hand, clearly, there is a potential correlation between smoking and lung cancer; indeed applications of this nature abound in several scenarios in modern data science, including a recent one on genetic pathway discovery [29].

Discovery of a potential correlation naturally leads one to ask for a measure of potential correlation that is statistically well-founded and addresses practical needs. Such is the focus of this work, where our proposed measure of potential correlation is based on a novel interpretation of the *Information Bottleneck* (IB) principle [158]. The IB principle has been used to address one of the fundamental tasks in supervised learning: given samples $\{X_i, Y_i\}_{i=1}^n$, how do we find a *compact* summary of a variable X that is most *informative* in explaining another variable Y. The output of the IB principle is a compact summary of X that is most relevant to Y and has a wide range of applications [159, 160].

We use this IB principle to create a measure of correlation based on the following intuition: if X is (potentially) correlated with Y, then a relatively compact summary of X can still be very informative about Y. In other words, the maximal ratio of how informative a summary can be in explaining Y to how compact a summary is with respect to X is, conceptually speaking, an indicator of potential correlation from X to Y. Quantifying the compactness by $I(U;X)$ and the information by $I(U;Y)$ we consider the *rate of information bottleneck* as a measure of potential correlation:

$$s(X;Y) \equiv \sup_{U-X-Y} \frac{I(U;Y)}{I(U;X)}, \tag{6.1}$$

where $U - X - Y$ forms a Markov chain and the supremum is over all summaries U of X. This intuition is made precise in Section 6.1, where we formally define a natural notion of potential correlation (Axiom 6), and show that the rate of information bottleneck $s(X;Y)$ captures this potential correlation (Theorem 12) while other standard measures of correlation fail (Theorem 13).

This ratio has only recently been identified as the *hypercontractivity* coefficient [10], correcting the former mistaken belief that $s(X;Y) = \mathrm{mCor}^2(X,Y)$, the squared maximal correlation [161]. Hypercontractivity has a distinguished and central role in a large number of technical arenas including quantum physics [162, 163], theoretical computer science [164, 165], mathematics [166, 167] and probability theory [168, 169]. In this chapter, we provide a novel interpretation to the hypercontractivity coefficient as a measure of potential correlation by demonstrating that it satisfies a natural set of axioms such a measure is expected to obey.

For practical use in discovering correlations, the standard correlation coefficients are equipped with corresponding natural sample-based estimators. However, for hypercontractivity coefficient, estimating it from samples is widely acknowledged to be challenging, especially for continuous random variables [170, 171]. There is no existing algorithm to estimate the hypercontractivity coefficient in general, and there is no existing algorithm for solving IB from samples either [170, 171]. We provide a novel estimator of the hypercontractivity coefficient – the first of its kind – by bringing together the recent theoretical discoveries in [10, 172] of an alternate definition of hypercontractivity coefficient as ratio of Kullback-Leibler divergences defined

in (6.16), and recent advances in joint optimization (the maximization step in (6.1)) and estimating information measures from samples using importance sampling [28].

Main contributions of Chapter 6:

- We postulate a set of natural axioms that a measure of potential correlation from X to Y should satisfy (Section 6.1).
- We show that $\sqrt{s(X;Y)}$, our proposed measure of potential correlation, satisfies all the axioms we postulate. In comparison, we prove that existing standard measures of correlation not only fail to satisfy the proposed axioms, but also fail to capture canonical potential correlations captured by $\sqrt{s(X;Y)}$ (Section 6.1). Another natural candidate is mutual information, but it is not clear how to interpret the value of mutual information as it is unnormalized, unlike all other measures of correlation which are between zero and one.
- Computation of the hypercontractivity coefficient from samples is known to be a challenging open problem. We introduce a novel estimator to compute hypercontractivity coefficient from i.i.d. samples in a statistically consistent manner for continuous random variables, using ideas from importance sampling and kernel density estimation (Section 6.2).
- In a series of synthetic experiments, we show empirically that our estimator for the hypercontractivity coefficient is statistically more powerful in discovering a potential correlation than existing correlation estimators; a larger power means a larger successful detection rate for a fixed false alarm rate (Section 6.3.1).
- We show applications of our estimator of hypercontractivity coefficient in two important datasets: In Section 6.3.2, we demonstrate that it discovers hidden potential correlations among various national indicators in WHO datasets, including how aid is potentially correlated with the income growth. In Section 6.3.3, we consider the following gene pathway recovery problem: we are given samples of four gene expressions time series. Assuming we know that gene A causes B, that B causes C, and that C causes D, the problem is to discover that these causations

occur in the sequential order: A to B, and then B to C, and then C to D. We show empirically that the estimator of the hypercontractivity coefficient recovers this order accurately from a vastly smaller number of samples compared to other state-of-the art causal influence estimators.

6.1 Axiomatic approach to measure potential correlations

We propose a set of axioms that we expect a measure of potential correlation to satisfy. We then show that hypercontractivity coefficient, first introduced in [168], satisfies all the proposed axioms, hence propose hypercontractivity coefficient as a measure of potential correlation. We also show that other standard correlation coefficients and mutual information, on the other hand, violates the proposed axioms.

6.1.1 Axioms for potential correlation

We postulate that a *measure of potential correlation* $\rho^* : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ between two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ should satisfy:

1. $\rho^*(X, Y)$ is defined for any pair of non-constant random variables X and Y.
2. $0 \leq \rho^*(X, Y) \leq 1$.
3. $\rho^*(X, Y) = 0$ iff X and Y are statistically independent.
4. For bijective Borel-measurable functions $f, g : \mathbb{R} \to \mathbb{R}$, $\rho^*(X, Y) = \rho^*(f(X), g(Y))$.
5. If $(X, Y) \sim \mathcal{N}(\mu, \Sigma)$, then $\rho^*(X, Y) = |\rho|$, where ρ is the Pearson correlation coefficient.
6. $\rho^*(X, Y) = 1$ if there exists a subset $\mathcal{X}_r \subseteq \mathcal{X}$ such that for a pair of continuous random variables $(X, Y) \in \mathcal{X}_r \times \mathcal{Y}$, $Y = f(X)$ for a Borel-measurable and non-constant continuous function f.

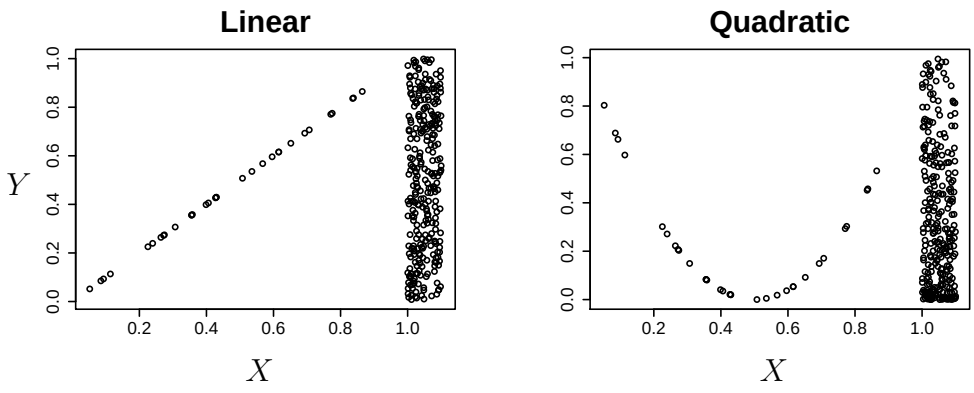

Figure 6.1: A measure of potential correlation should capture the rare correlation in $X \in [0, 1]$ in these examples which satisfy Axiom 6 for a linear and a quadratic function, respectively.

Axioms 1-5 are identical to a subset of the celebrated axioms of Rényi in [156], which ensure that the measure is properly normalized and invariant under bijective transformations, and recovers the Pearson correlation for jointly Gaussian random variables. Rényi's original axioms for a *measure of correlation* in [156] included Axioms 1-5 and also that the measure ρ^* of correlation should satisfy

7. $\rho^*(X, Y) = 1$ if for Borel-measurable functions f or g, $Y = f(X)$ or $X = g(Y)$.

8. $\rho^*(X; Y) = \rho^*(Y; X)$.

The Pearson correlation violates a subset (3, 4, and 7) of Rényi's axioms. Together with recent empirical successes in multimodal deep learning (e.g. [173, 174, 175]), Rényi's axiomatic approach has been a major justification of Hirschfeld-Gebelein-Rényi (HGR) maximal correlation coefficient defined as $\mathrm{mCor}(X, Y) := \sup_{f,g} \mathbb{E}[f(X)g(Y)]$, which satisfies all Rényi's axioms [154]. Here, the supremum is over all measurable functions with $\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$ and $\mathbb{E}[f^2(X)] = \mathbb{E}[g^2(Y)] = 1$. However, maximal correlation is not the only measure satisfying all of Rényi's axioms, as we show in the following.

Proposition 2. *For any function $F : [0, 1]\times[0, 1] \to [0, 1]$ satisfying $F(x, y) = F(y, x)$, $F(x, x) = x$, and $F(x, y) = 0$ only if $xy = 0$, the symmetrized $F(\sqrt{s(X; Y)}, \sqrt{s(Y; X)})$ satisfies all Rényi's axioms.*

This follows from the fact that the hypercontractivity coefficient $\sqrt{s(X;Y)}$ satisfies all but the symmetry in Axiom 7 (Theorem 12), and it follows that a symmetrized version satisfies all axioms, e.g. $(1/2)(\sqrt{s(X;Y)}+\sqrt{s(Y;X)})$ and $(s(X;Y)s(Y;X))^{1/4}$. A formal proof is provided in Section 6.4.1.

From the original Rényi's axioms, for a potential correlation measure, we remove Axiom 8 that ensures symmetry, as directionality is fundamental in measuring the potential correlation from X to Y. We further replace Axiom 7 by Axiom 6, as a variable X has a full potential to be correlated with Y if there exists a domain $\mathcal{X}_r$ such that X and Y are deterministically dependent and non-degenerate (i.e. not a constant function), as illustrated in Figure 6.1 for a linear function and a quadratic function.

6.1.2 The hypercontractivity coefficient satisfies all axioms

We show that the hypercontractivity coefficient defined in (6.1) satisfies all Axioms 1-6. Intuitively, $s(X;Y)$ measures how much potential correlation X has with Y. For example, if X and Y are independent, then $s(X;Y)=0$ as X has no correlation with Y (Axiom 3). By data processing inequality, it follows that it is a measure between zero and one (Axiom 2) and also invariant under bijective transformations (Axiom 4). For jointly Gaussian variables X and Y with the Pearson correlation ρ, we can show that $s(X;Y)=s(Y;X)=\rho^2$. Hence, the squared-root of $s(X;Y)$ satisfies Axiom 5. In fact, $\sqrt{s(X;Y)}$ satisfies all desired axioms for potential correlation, and we make this precise in the following theorem whose proof is provided in Section 6.4.2.

Theorem 12. *Hypercontractivity coefficient $\sqrt{s(X;Y)}$ satisfies Axioms 1-6.*

In particular, the hypercontractivity coefficient satisfies Axiom 6 for potential correlation, unlike other measures of correlation (see Theorem 13 for examples). If there is a potential for X in a possibly rare regime in $\mathcal{X}$ to be fully correlated with Y such that $Y=f(X)$, then the hypercontractivity coefficient is maximum: $s(X;Y)=1$. In the following subsection, we show that existing correlation measures violate the proposed axioms.

6.1.3 Standard correlation coefficients violate the axioms

We analyze existing measures of correlations under the scenario with potential correlation (Axiom 6), where we find that none of the existing correlation measures satisfy Axiom 6. Suppose X and Y are independent (i.e. no correlation) in a subset $\mathcal{X}_d$ of the domain $\mathcal{X}$, and allow X and Y to be arbitrarily correlated in the rest $\mathcal{X}_r$ of the domain, such that $\mathcal{X} = \mathcal{X}_d \cup \mathcal{X}_r$. We further assume that the independent part is dominant and the correlated part is rare; let $\alpha := \Pr(X \in \mathcal{X}_r)$ and we consider the scenario when α is small. A good measure of potential correlation is expected to capture the correlation in $\mathcal{X}_r$ even if it is rare (i.e., α is small). To make this task more challenging, we assume that the conditional distribution of $Y|\{X \in \mathcal{X}_r\}$ is the same as $Y|\{X \notin \mathcal{X}_r\}$. Figure 6.1 illustrates sampled points for two examples from such a scenario. Our main result is the analysis of HGR maximal correlation (mCor) [154], distance correlation (dCor) [157], maximal information coefficients (MIC) [11], which shows that these measures are vanishing with α even if the dependence in the rare regime is very high. Suppose $Y|(X \in \mathcal{X}_r) = f(X)$, then all three correlation coefficients are vanishing as α gets small. This in particular violates Axiom 6. The reason is that standard correlation coefficients measure the *average correlation* whereas the hypercontractivity coefficient measures the *potential correlation*. The experimental comparisons on the power of these measures confirm our analytical predictions in Figure 6.4 in Section 6.3. The formal statement is below and the proof is provided in Section 6.4.3.

Theorem 13. *Consider a pair of continuous random variables* $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. *Suppose* $\mathcal{X}$ *is partitioned as* $\mathcal{X}_r \cup \mathcal{X}_d = \mathcal{X}$ *such that* $P_{Y|X}(S|X \in \mathcal{X}_r) = P_{Y|X}(S|X \in \mathcal{X}_d)$ *for all* $S \subseteq \mathcal{Y}$, $X \perp\!\!\!\perp Y$ *in* $\mathcal{X}_d$, *and* $\alpha = \Pr\{X \in \mathcal{X}_r\}$. *The HGR maximal correlation coefficient is*

$$\mathrm{mCor}(X, Y) \;=\; \sqrt{\alpha}\, \mathrm{mCor}(X_r, Y). \tag{6.2}$$

The distance correlation coefficient is

$$\mathrm{dCor}(X, Y) \;=\; \alpha\, \mathrm{dCor}(X_r, Y). \tag{6.3}$$

The maximal information coefficient is upper bounded by

$$\mathrm{MIC}(X,Y) \;\leq\; \alpha\,\mathrm{MIC}(X_r,Y), \tag{6.4}$$

where X_r is the random variable X conditioned on the rare domain $X \in \mathcal{X}_r$.

Under the rare/dominant scenario considered in Theorem 13, $s(X;Y) \geq \mathrm{mCor}^2(X;Y)$. It is well known that this inequality holds for any X and Y [168]. In particular, [176, Theorem 3] shows that hypercontractivity coefficient is a natural extension of the popular HGR maximal correlation coefficient as follows.

Remark 2. *The squared HGR maximal correlation is a special case of the hypercontractivity optimization in* (6.16) *restricted to searching over a distribution $r(x)$ in a close neighborhood of $p(x)$.*

As $s(X;Y)$ searches over a larger space, it is always larger than or equal to $\mathrm{mCor}^2(X;Y)$. This gives an intuitive justification for using $s(X;Y)$ as a measure of potential influence; we allow search over larger space, but properly normalized by the KL divergence, in a hope to find a potential distribution $r(x)$ that can influence Y significantly. While hypercontractivity coefficient is a natural extension of HGR maximal correlation coefficient, there is an important difference between hypercontractivity coefficient and HGR maximal correlation coefficient (and other correlation measures); hypercontractivity is directional.

Remark 3. *Hypercontractivity coefficient is* asymmetric *in X and Y while HGR maximal correlation, distance correlation, and MIC are* symmetric.

Under the rare/dominant scenario considered in Theorem 13, the hypercontracitivy coefficient $s(X;Y)$ is large because it measures the potential correlation from X to Y. On the other hand, inverse hypercontractivity coefficient $s(Y;X)$, which measures the potential correlation from Y to X, is small as there is no apparent potential correlation from Y to X. This is made precise in the following proposition.

Proposition 3. *Under the hypotheses of Theorem 13, the hypercontractivity coefficient from Y to X is*

$$s(Y;X) = \alpha\; s(Y;X_r). \tag{6.5}$$

Proof is provided in Section 6.4.4.

6.1.4 Mutual information violates the axioms

Beside standard correlation measures, another measure widely used to quantify the strength of dependence is mutual information. We can show that mutual information satisfies Axiom 6 if we replace 1 by ∞. However there are two key problems:

- Mutual information is *unnormalized*, i.e., $I(X;Y) \in [0, \infty)$. Hence, it provides no absolute indication of the strength of the dependence.
- Mathematically, we are looking for a quantity that *tensorizes*, i.e., doesn't change when there are many i.i.d. copies of the same pair of random variables.

  Remark 4. *Hypercontractivity coefficient tensorizes, i.e,*

$$\begin{aligned} &s(X_1, ..., X_n; Y_1, .., Y_n) = s(X_1, Y_1), \\ &\textit{for i.i.d. } (X_i, Y_i), \quad i = 1, \cdots, n. \end{aligned} \tag{6.6}$$

  On the other hand, mutual information is additive, i.e.,

$$\begin{aligned} &I(X_1, \cdots, X_n; Y_1, \cdots, Y_n) = nI(X_1; Y_1), \\ &\text{for i.i.d. } (X_i, Y_i), \quad i = 1, \cdots, n. \end{aligned} \tag{6.7}$$

  Tensorizing quantities capture the strongest relationship among independent copies while additive quantities capture the sum. For instance, mutual information could be large because a small amount of information accumulates over many of the independent components of X and Y (when X and Y are high dimensional) while tensorizing quantities would rule out this scenario, where there is no strong dependence. When the components are not independent, hypercontractivity indeed pools information from different components to find the strongest direction of dependence, which is a desirable property.

One natural way to normalize mutual information is by the log of the cardinality of the input/output alphabets [177]. One can interpret a popular

correlation measure MIC as a similar effort for normalizing mutual information and is one of our baselines.

Given that other correlation measures and mutual information do not satisfy our axioms, a natural question to ask is whether hypercontractivity is a unique solution that satisfies all the proposed axioms. In the following, we show that the hypercontractivity coefficient is not the only one satisfying all the proposed axioms – just as HGR correlation is not the only measure satisfying Rényi's original axioms.

6.1.5 Hypercontractivity ribbon

We show that a family of measures known as *hypercontractivity ribbon*, which includes hypercontractivity coefficient as a special case, satisfy all the axioms. The hypercontractivity ribbon [168, 178] is parametrized by $\alpha > 0$ as

$$r_\alpha(X;Y) = \sup_{r(x,y)\neq p(x,y)} \frac{D(r(y)\|p(y))}{D(r(x)\|p(x)) + \alpha D(r(y|x)\|p(y|x))}. \tag{6.8}$$

An alternative characterization of hypercontractivity ribbon in terms of mutual information is provided in [178, 172];

$$r_\alpha(X;Y) = \sup_{p(u|x,y)} \frac{I(U;Y)}{I(U;X) + \alpha I(U;Y|X)}, \tag{6.9}$$

from which we can see that hypercontractivity coefficient is a special case of hypercontractivity ribbon [10]:

$$s(X;Y) \;\; = \lim_{\alpha\to\infty} r_\alpha(X;Y) = \lim_{\alpha\to\infty} s_\alpha(X;Y). \tag{6.10}$$

Proposition 4. *The (re-parameterized) hypercontractivity ribbon* $s_\alpha(X;Y) := (\alpha\, r_\alpha(X;Y) - 1)/(\alpha - 1)$*, for* $\alpha > 1$*, satisfies Axioms 1-6.*

Proof. By definition, $s_\alpha(X;Y)$ is defined for any pair of non-constant random variables (Axiom 1) and is between 0 and 1 by data processing inequality (Axiom2). We can show that $s_\alpha(X;Y)$ satisfies Axioms 3 and 4, in a similar way to show $s(X;Y)$ satisfies Axioms 3 and 4. Also, $s_\alpha(X;Y) = \rho^2$ for a

jointly Gaussian X,Y with Pearson correlation ρ [172] (Axiom 5). Finally, $s_\alpha(X;Y)$ satisfies Axiom 6 because $r_\alpha(X;Y)$ is non-increasing in α , which implies that $s_\alpha(X;Y) = r_\alpha(X;Y) = 1$ if $s(X;Y) = 1$. □

Although hypercontracitivy ribbon satisfies all axioms, a few properties of the hypercontractivity coefficient makes it more attractive than hypercontractivity ribbon for practical use; hypercontractivity coefficient can be efficiently estimated from samples (see Section 6.2). Hypercontractivity coefficient is a natural extension of the popular HGR maximal correlation coefficient (Remark 2).

6.1.6 Multidimensional X and Y

In this subsection, we discuss potential correlation of multidimensional X and Y. While most of the correlation coefficients, including the hypercontractivity coefficient, are well-defined for multidimensional X and Y, the axioms are specific to univariate X and Y. To bridge this gap, we propose replacing Axiom 5, as this is the only axiom specific to univariate random variables.

Axiom 9. If $(X,Y) \sim \mathcal{N}\left(\mu, \Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}\right)$, then $\rho^*(X,Y) = \|\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\|$, where $\|\cdot\|$ is the spectral norm of a matrix.

This recovers the original Axiom 5 when restricted to univariate X and Y. This naturally generalizes both Rényi's axioms and the proposed potential correlation axioms to multidimensional X and Y.

Proposition 5. *Axiom 9, together with original Rényi's Axioms 1-4, 7, and 8, recovers maximal correlation (mCor) as a measure satisfying all Axioms even in this multi-dimensional case. Axiom 9, together with our proposed Axioms 1-4, and 6, recovers the hypercontractivity coefficient $\sqrt{s(X;Y)}$ as a measure satisfying all axioms.*

The second statement in the proposition follows from the analyses of the hypercontractivity coefficient of Gaussian distributions in [179]. A formal proof is provided in Section 6.4.7.

6.1.7 Noisy, discrete, noisy and discrete potential correlations

In this section, we consider more general scenarios of potential correlation than the one in Axiom 6. We consider (i) noisy potential correlation where $Y = f(X) + Z$ for a Gaussian noise Z for $(X, Y) \in \mathcal{X}_r \times \mathcal{Y}$, (ii) discrete potential correlation, where $\mathcal{X}_r = \{1, \cdots, k\}$, and (iii) noisy discrete potential correlation – a random corruption model. For these three examples, we obtain a lower bound on $s(X;Y)$.

Example 1. *Suppose that for a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, there exists a subset $\mathcal{X}_r \subseteq \mathcal{X}$ for which $\Pr\{X \in \mathcal{X}_r\} = \alpha$ $(\alpha > 0)$, and for $(X, Y) \in \mathcal{X}_r \times \mathcal{Y}$, $(X, Y) \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. Then*

$$s(X;Y) \geq \frac{\log \frac{1}{1-\rho^2} + \log \frac{1}{1+\rho^2}}{\log \frac{1}{1-\rho^2} + \frac{H(\alpha)}{\alpha}}. \tag{6.11}$$

Proof is in Section 6.4.5.

We now consider for discrete (X, Y). We start with the case for which X and Y are perfectly correlated for $(X, Y) \in \mathcal{X}_r \times \mathcal{Y}$.

Example 2. *Suppose that for a pair of discrete random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, there exists a subset $\mathcal{X}_r = \{1, 2, \cdots, k\} \subseteq \mathcal{X}$ for which $\Pr\{X \in \mathcal{X}_r\} = \alpha$ $(\alpha > 0)$, and $X|\{X \in \mathcal{X}_r\} \sim Unif[1:k]$ and $Y = X$ for $X \in \mathcal{X}_r$. Then,*

$$s(X;Y) \geq \frac{\log k}{\log k + \log(1/\alpha)}. \tag{6.12}$$

The inequality holds by considering $r(x) = \mathbb{I}\{X = 1\}$ in (6.16).

We conjecture this lower bound is indeed tight for $\alpha \leq 0.5$ based on numerical simulations. From this lower-bound, we can see the trade-off between k and α. As $k \to \infty$, the lower bounds approaches to 1. As $\alpha \to 1$, the lower bound approaches to 1. As $\alpha \to 0$, the lower bound approaches to 0. In the following, we consider the case where X and Y are not perfectly correlated in $(\mathcal{X}_r \times \mathcal{Y})$ for discrete (X, Y). In particular, we consider a random corruption model for $(\mathcal{X}_r \times \mathcal{Y})$ and obtain a lower bound on $s(X;Y)$.

Example 3. *Suppose that for a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, there exists a subset $\mathcal{X}_r \subseteq \mathcal{X}$ for which $\Pr\{X \in \mathcal{X}_r\} = \alpha$ $(\alpha > 0)$, and for*

$(X, Y) \in \mathcal{X}_r \times \mathcal{Y}$,

$$Y = \begin{cases} X & w.p.\ 1 - \frac{k}{k-1}\epsilon, \\ \mathit{Unif}[1:k] & w.p.\ \frac{1}{k-1}\epsilon. \end{cases} \tag{6.13}$$

Then

$$\begin{aligned} s(X;Y) &\geq \frac{(1-\epsilon)\log k(1-\epsilon) + \epsilon \log k\epsilon/(k-1)}{\log(k/\alpha)} \\ &= \frac{\log k - H_2(\epsilon) - \epsilon \log(k-1)}{\log(k/\alpha)}. \end{aligned} \tag{6.14}$$

On the other hand,

$$\mathrm{mCor}^2(X;Y) = \alpha \left(1 - \frac{k}{k-1}\epsilon\right)^2, \quad 0 \leq \epsilon \leq \frac{k-1}{k}. \tag{6.15}$$

Proof is in Section 6.4.6.

In Figure 6.2, we show plots of lower bounds on $s(X;Y)$ and $\mathrm{mCor}(X;Y)$ in Examples 1-3; from these figures, we can see that $s(X;Y)$ increases as $\rho \to 1$ and $k \to \infty$. In comparison, $\mathrm{mCor}(X;Y)$ remains small.

6.2 Estimator of the hypercontractivity coefficient from samples

In this section, we present an algorithm to compute the hypercontractivity coefficient $s(X;Y)$ from i.i.d. samples $\{X_i, Y_i\}_{i=1}^n$. The computation of the hypercontractivity coefficient from samples is known to be challenging for continuous random variables [170, 171], and to the best of our knowledge, there is no known efficient algorithm to compute the hypercontractivity coefficient from samples. Our estimator is the first efficient algorithm to compute the hypercontractivity coefficient, based on the following equivalent definition of the hypercontractivity coefficient, shown recently in [10]:

$$s(X;Y) \equiv \sup_{r_x \neq p_x} \frac{D(r_y||p_y)}{D(r_x||p_x)}. \tag{6.16}$$

There are two main challenges for computing $s(X;Y)$. The first challenge

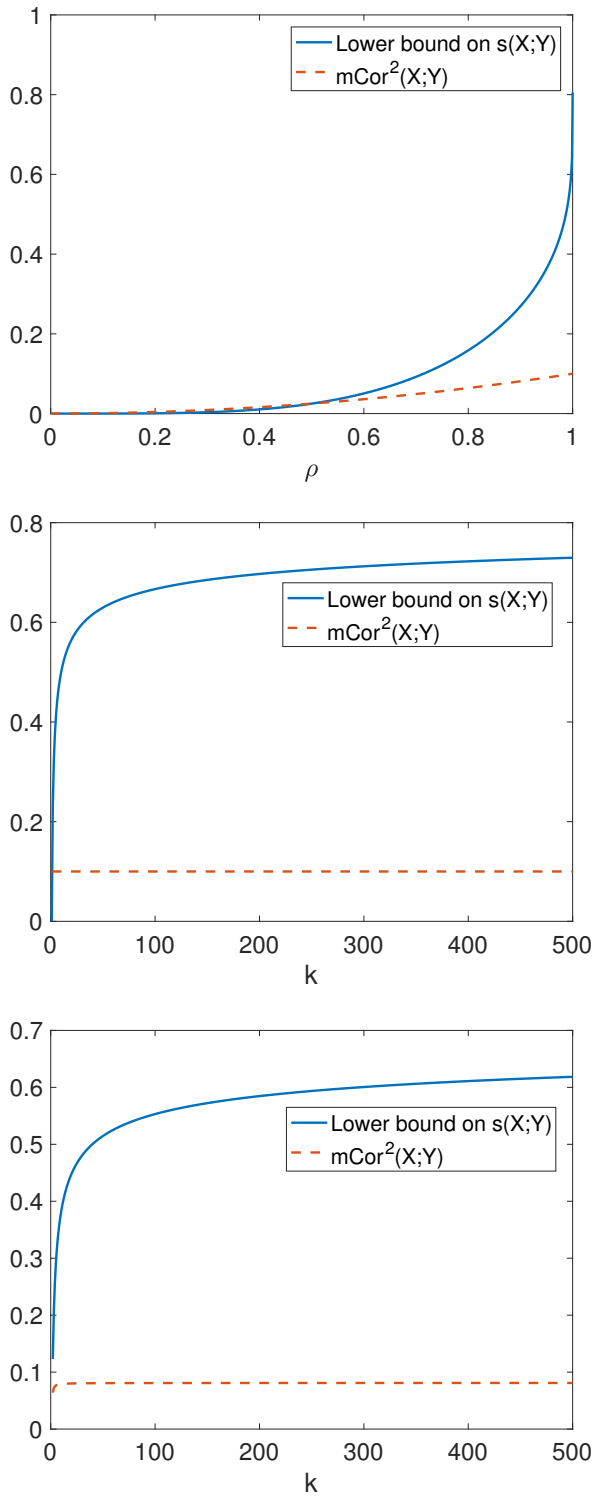

Figure 6.2: Lower bound on $s(X;Y)$ and $\mathrm{mCor}(X;Y)$ for $\alpha = 0.1$ in Example 1 (top), Example 2 (middle) and Example 3 (bottom) for $\epsilon = 0.1$.

is – given a marginal distribution r_x and samples from p_{xy}, how do we estimate the KL divergences $D(r_y||p_y)$ and $D(r_x||p_x)$. The second challenge is the optimization over the infinite dimensional simplex. We need to combine estimation and optimization together in order to compute $s(X;Y)$. Our approach is to combine ideas from traditional kernel density estimates and from importance sampling. Let $w_i = r_x(X_i)/p_x(X_i)$ be the *likelihood ratio* evaluated at sample i. We propose the estimation and optimization be solved jointly as follows.

Estimation: To estimate KL divergence $D(r_x||p_x)$, notice that

$$D(r_x||p_x) \;=\; \mathbb{E}_{X\sim p_x}\left[\frac{r_x(X)}{p_x(X)}\log\frac{r_x(X)}{p_x(X)}\right]. \tag{6.17}$$

Using empirical average to replace the expectation over p_x, we propose

$$\widehat{D}(r_x||p_x) = \frac{1}{n}\sum_{i=1}^{n}\frac{r_x(X_i)}{p_x(X_i)}\log\frac{r_x(X_i)}{p_x(X_i)} = \frac{1}{n}\sum_{i=1}^{n} w_i \log w_i. \tag{6.18}$$

For $D(r_y||p_y)$, we follow the similar idea, but the challenge is in computing $v_j = r_y(Y_j)/p_y(Y_j)$. To do this, notice that $r_{xy} = r_x p_{y|x}$, so

$$r_y(Y_j) \;=\; \mathbb{E}_{X\sim r_x}\left[\,p_{y|x}(Y_j|X)\,\right] = \mathbb{E}_{X\sim p_x}\left[\,p_{y|x}(Y_j|X)\frac{r_x(X)}{p_x(X)}\,\right]. \tag{6.19}$$

Replacing the expectation by empirical average again, we get the following estimator of v_j:

$$\widehat{v}_j \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{p_{y|x}(Y_j|X_i)}{p_y(Y_j)}\frac{r_x(X_i)}{p_x(X_i)} = \frac{1}{n}\sum_{i=1}^{n}\underbrace{\frac{p_{xy}(X_i,Y_j)}{p_x(X_i)p_y(Y_j)}}_{A_{ji}} w_i. \tag{6.20}$$

We can write this expression in matrix form as $\widehat{\mathbf{v}} = \mathbf{A}^T\mathbf{w}$. We use a kernel density estimator from [180] to estimate the matrix $\mathbf{A}$, but our approach is compatible with any density estimator of choice.

Optimization: Given the estimators of the KL divergences, we are able to convert the problem of computing $s(X;Y)$ into an optimization problem over the vector $\mathbf{w}$. Here a constraint of $(1/n)\sum_{i=1}^{n} w_i = 1$ is needed to

satisfy $\mathbb{E}_{p_x}[r_x/p_x] = 1$. To improve numerical stability, we use $\log s(X;Y)$ as the objective function.

Then the optimization problem has the following form:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \log\big((\mathbf{w}^T\mathbf{A}\log(\mathbf{A}^T\mathbf{w})\big) - \log\big(\mathbf{w}^T\log\mathbf{w}\big) \\ \text{subject to} \quad & \frac{1}{n}\sum_{i=1}^{n} w_i = 1 \\ & w_i \geq 0, \forall\, i, \end{aligned} \tag{6.21}$$

where $\mathbf{w}^T\log\mathbf{w} = \sum_{i=1}^{n} w_i \log w_i$ for short. Although this problem is not convex, we apply gradient descent to maximize the objective. In practice, we initialize $w_i = 1 + \mathcal{N}(0,\sigma^2)$ for $\sigma^2 = 0.01$. Hence, the initial r_x is perturbed mildly from p_x. Although we are not guaranteed to achieve the global maximum, we consistently observe in extensive numerical experiments that we have 50%-60% probability of achieving the same maximum value, which we believed to be the global maximum. A theoretical analysis of the landscape of local and global optima and their regions of attraction with respect to gradient descent is an interesting and challenging open question, outside the scope of this chapter.

Consistency of estimation: While a theoretical understanding of the performance of gradient descent on the optimization step (where the number of samples is fixed) above is technically very challenging, we can study the performance of the solution as the number of samples increases. In particular we show below (under suitable simplifying assumptions to get to the essence of the proof) that the optimal solution to the finite sample optimization problem is *consistent*. Suppose that $\mathcal{X}$ is discrete. Further we restrict the optimization over a quantized and bounded set T_Δ, where $\mathbf{w} \in T_\Delta$ is quantized by a gap Δ and satisfies: (1) $C_1 \leq w_i \leq C_2$ for all i; (2) $(1/n)\sum_{i=1}^{n} w_i \log w_i > C_0$. We further assume that we have access of $\mathbf{A} = P_{xy}(X_i,Y_j)/P_x(X_i)P_y(Y_j)$. Define $\widehat{s}_\Delta(X;Y) = \max_{\mathbf{w}\in T_\Delta} \mathbf{w}^T\mathbf{A}\log(\mathbf{A}^T\mathbf{w})/\mathbf{w}^T\log\mathbf{w}$, then with two further simplifying conditions on the joint distribution, we can prove consistency of our estimation procedure:

Theorem 14. *As n goes to infinity, $\widehat{s}_\Delta(X;Y)$ converges to $s(X;Y)$ up to a resolution of quantization in probability, i.e., for any $\varepsilon > 0$, $\Delta > 0$ and*

$s(\Delta) = O(\Delta)$, *we have*

$$\lim_{n\to\infty} \Pr\left(\left| \widehat{s}_{\Delta}(X;Y) - s(X;Y) \right| > \varepsilon + s(\Delta) \right) = 0. \tag{6.22}$$

6.3 Experiments of Chapter 6

We present experimental results on synthetic and real datasets showing that the hypercontractivity coefficient (a) is more powerful in detecting potential correlation compared to existing measures; (b) discovers hidden potential correlations among various national indicators in WHO datasets; and (c) is more robust in discovering pathways of gene interactions from gene expression time series data.

6.3.1 Synthetic data: Power test on potential correlation

As our estimator (and the measure itself) involves a maximization, it is possible that we are sensitive to outliers and may capture spurious noise. Via a series of experiments we show that the hypercontractivity coefficient and our estimator are capturing the true potential correlation. As shown in Figure 6.3, we generate pairs of datasets – one where X and Y are independent and one where there is a potential correlation as per our scenario. We run several experiments with eight types of functional associations, following the examples from [11, 181, 182]. For the correlated datasets, out of n samples $\{(x_i, y_i)\}_{i=1}^{n}$, αn rare but correlated samples are in $\mathcal{X} = [0, 1]$ and $(1-\alpha)n$ dominant but independent samples are in $\mathcal{X} \in [1, 1.1]$. The rare but correlated samples are generated as $x_i \sim \text{Unif}[0, 1], y_i \sim f(x_i) + \mathcal{N}(0, \sigma^2)$ for $i \in [1 : \alpha n]$. The dominant samples are generated as $x_i \sim \text{Unif}[1, 1.1], y_i \sim f(\text{Unif}[0, 1]) + \mathcal{N}(0, \sigma^2)$ for $i \in [\alpha n + 1, n]$.

Table 6.1 shows the hypercontractivity coefficient and the other correlation coefficients for correlated and independent datasets shown in Figure 6.3, along with the chosen value of α and σ^2. Correlation estimates with the largest separation for each row is shown in bold. The hypercontractivity coefficient gives the largest separation between the correlated and the independent dataset for most functional types.

A formal statistical approach to test the robustness as well as accuracy is

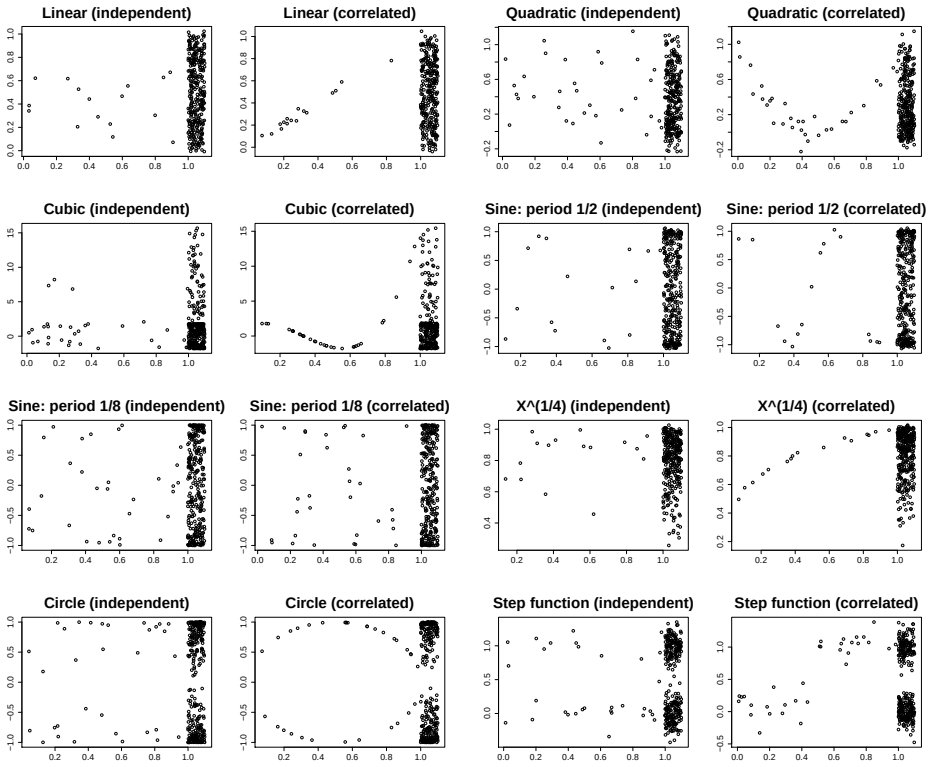

Figure 6.3: Sample data points for eight functions with/without a potential correlation for $n = 320$.

Table 6.1: Comparison of correlation coefficients for independent and correlated samples from Figure 6.3.

				Cor		dCor		mCor	
#	Function	α	σ^2	dep	indep	dep	indep	dep	indep
1	Linear	0.05	0.03	0.03	0.00	0.19	0.11	0.06	0.04
2	Quadratic	0.10	0.10	0.00	0.01	0.09	0.10	**0.07**	**0.02**
3	Cubic	0.10	0.00	0.02	0.00	0.16	0.08	0.09	0.03
4	$\sin(4\pi X)$	0.05	0.03	0.00	0.00	0.10	0.06	0.03	0.01
5	$\sin(16\pi X)$	0.10	0.00	0.00	0.00	0.07	0.08	**0.03**	**0.03**
6	$X^{1/4}$	0.05	0.01	0.01	0.00	0.12	0.07	0.02	0.01
7	Circle	0.10	0.00	0.00	0.00	0.09	0.05	0.01	0.03
8	Step func.	0.10	0.03	0.00	0.00	0.13	0.07	0.04	0.02

				MIC		HC	
#	Function	α	σ^2	dep	indep	dep	indep
1	Linear	0.05	0.03	0.21	0.17	**0.18**	**0.08**
2	Quadratic	0.10	0.10	0.21	0.18	0.08	0.04
3	Cubic	0.10	0.00	**0.26**	**0.17**	0.11	0.04
4	$\sin(4\pi X)$	0.05	0.03	0.20	0.18	**0.10**	**0.04**
5	$\sin(16\pi X)$	0.10	0.00	0.18	0.22	**0.03**	**0.03**
6	$X^{1/4}$	0.05	0.01	0.20	0.20	**0.12**	**0.04**
7	Circle	0.10	0.00	0.16	0.17	**0.06**	**0.01**
8	Step func.	0.10	0.03	0.20	0.17	**0.11**	**0.04**

to run *power tests*: testing for the power of the estimator in binary hypothesis tests. To compute the power of each estimator, we compare the false negative rate at a fixed false positive rate of, say, 5%. We generate 500 independent datasets and 500 correlated datasets. We compute the correlation estimates on 500 independent samples, and take the top 5% as a threshold. We compute the correlation estimates on 500 correlated samples. Power is defined as the fraction of correlated datasets for which the correlation estimate is larger than the threshold.

We show empirically that for linear, quadratic, sine with period 1/2, and the step function, the hypercontractivity coefficient is more powerful as compared to other measures. For a given setting, a larger power means a larger successful detection rate for a fixed false alarm rate. Figure 6.4 shows the power of correlation estimators as a function of the additive noise level, σ^2, for $\alpha = 0.05$ and $n = 320$. The hypercontractivity coefficient is more powerful than other correlation estimators for most functions. The power of all the estimators are very small for sine (period 1/8) and circle functions. This is not surprising given that it is very hard to discern the correlated and independent cases even visually, as shown in Figure 6.3.

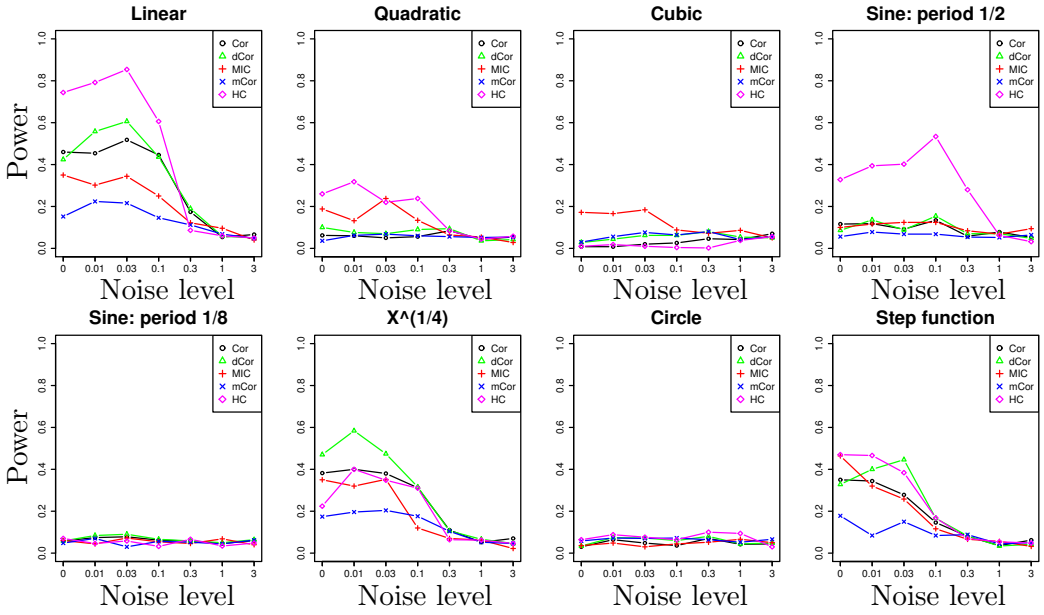

Figure 6.4: Power vs. noise level for $\alpha = 0.05$, $n = 320$.

Figure 6.5 plots the power of correlation estimators as a function of noise level for $\alpha = 0.1$ and $n = 320$. As we can see from these figures, hyper-

contractivity estimator is more powerful than other correlation estimators for most functions. For circle function, the gap between the power of hypercontractivity estimator and the powers of other estimators is significantly large.

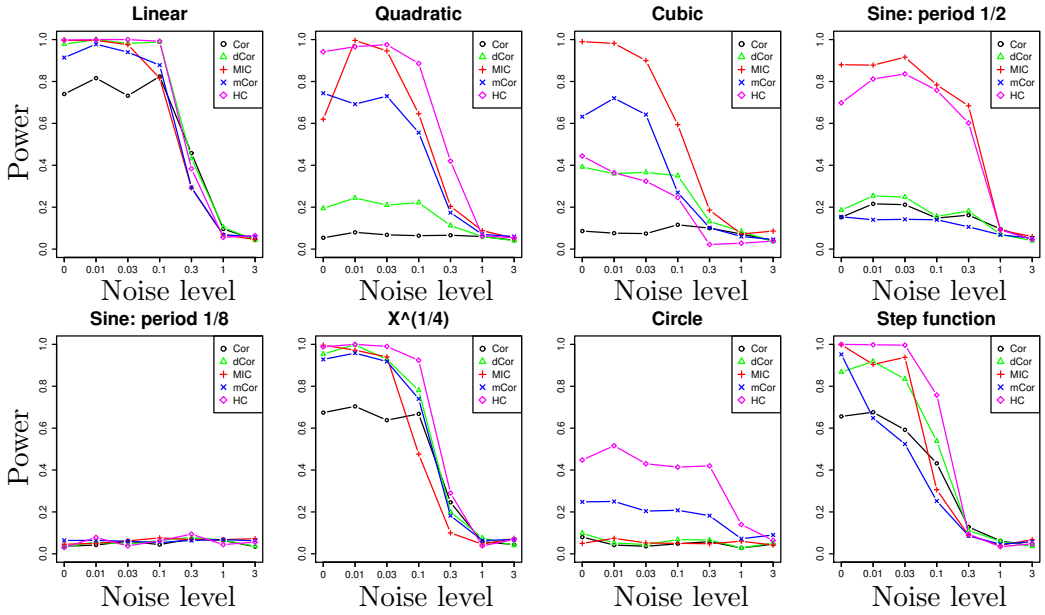

Figure 6.5: Power vs. noise level for $\alpha = 0.1$, $n = 320$.

On the other hand, hypercontractivity estimator is power deficient for the cubic function. This is because in estimating hypercontractivity coefficient, we estimate $p(y_j|x_i)/p(y_j)$ using the kernel density estimator (KDE), which gives a smooth estimate of $p(y_j|x_i)/p(y_j)$, i.e., for x_i and x_j close to each other, estimated $p(y|x_i)$ and $p(y|x_j)$ are close to each other. Hence, for a correlated dataset for a cubic function, shown in Figure 6.6 (bottom right), the estimated $p(y|x)$ does not vary much for x. (Estimated $p(y|x)$ for $x \in [0.8 : 1]$ and $p(y|x)$ for $x \in [1 : 1.1]$ are close to each other). This results in a small hypercontracitivy, which in turn results in a low power in the hypothesis testing. To further analyze this effect, we considered the same dataset but with dominant independent samples appear on the left, as shown in Figure 6.7 (bottom left) and (bottom right), and computed the power of hypercontractivity estimator, shown in Figure 6.7 (top). Hypercontractivity estimator is much more powerful than the one for the original dataset. This is because the estimated $p(y|x)$ for $x \in [0.8, 1]$ is very different from the estimated $p(y|x)$ for $x \in [-0.1, 0]$, which results in a large hypercontractivity

coefficient for the correlated dataset.

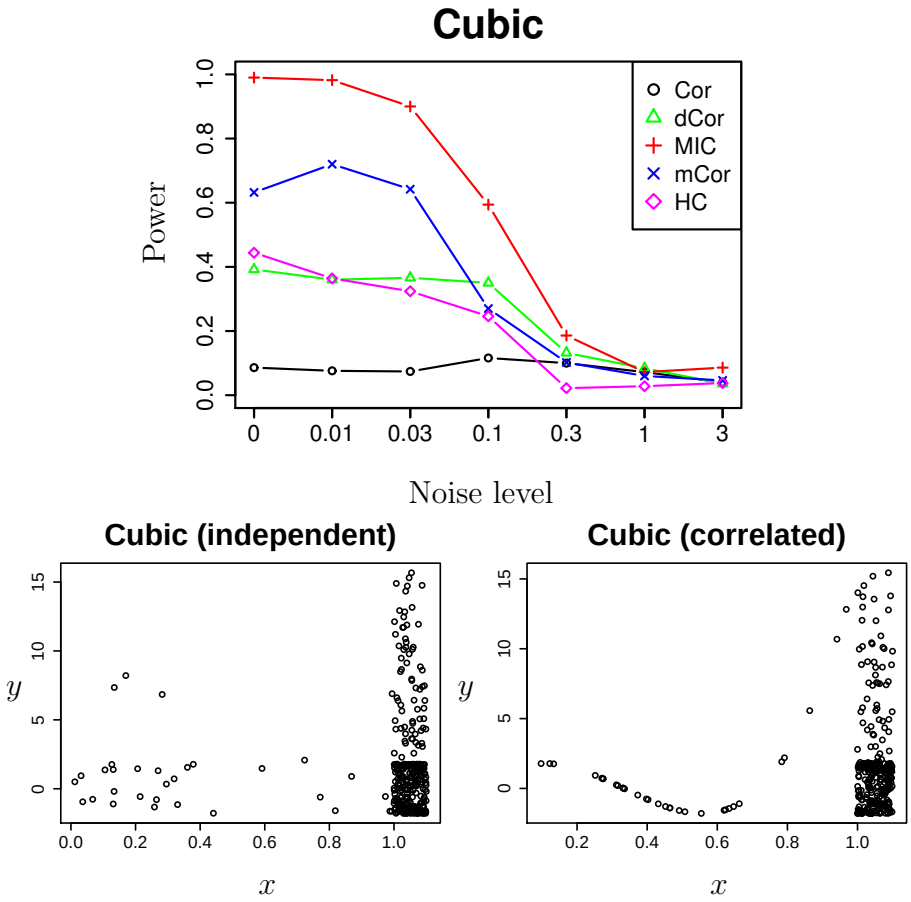

Figure 6.6: Power vs. noise level for $\alpha = 0.1$ and $n = 320$ (top), corresponding examples of an independent dataset (bottom left) and a correlated dataset (bottom right).

To investigate the dependency of power on α more closely, in Figure 6.8, we plot the power vs. α or $n = 320$ and $\sigma^2 = 0.1$. Hypercontractivity estimator is more powerful than other estimators for most α, for all functions except for cubic function. For a sine with period 1/8, due to its high frequency, the powers of all the correlation estimators do not increase as α increases. Figure 6.9 plots the power vs. sample size n for $\alpha = 0.05$ and $\sigma^2 = 0.1$. For sine with period 1/2, hypercontractivity estimator is much more powerful than the other estimators for all sample sizes. We can also see that for sine with period 1/8, powers of all correlation estimators do not increase as sample size increases.

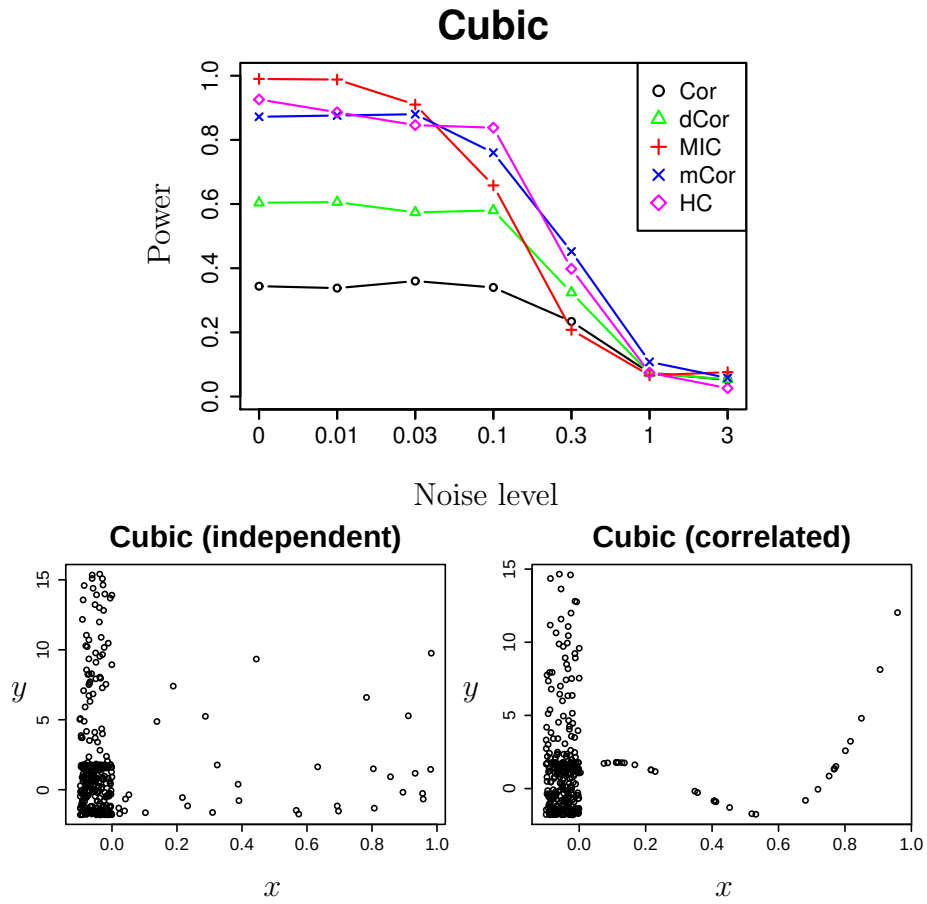

Figure 6.7: Flipped example of power vs. noise level for $\alpha = 0.1$ and $n = 320$ (top), corresponding examples of an independent dataset (bottom left) and a correlated dataset (bottom right).

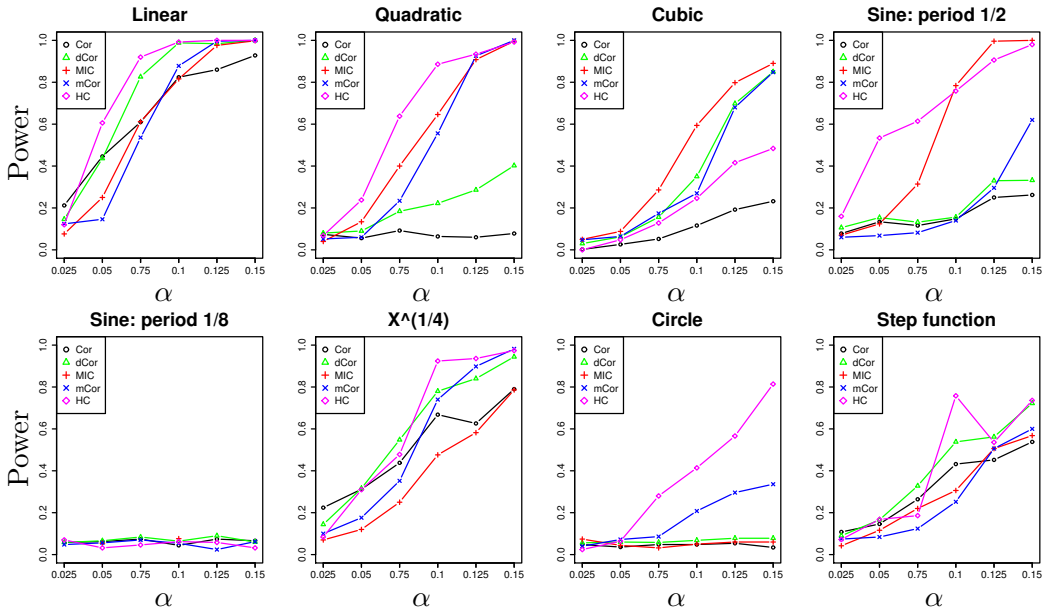

Figure 6.8: Power vs. α for $n = 320$, $\sigma^2 = 0.1$.

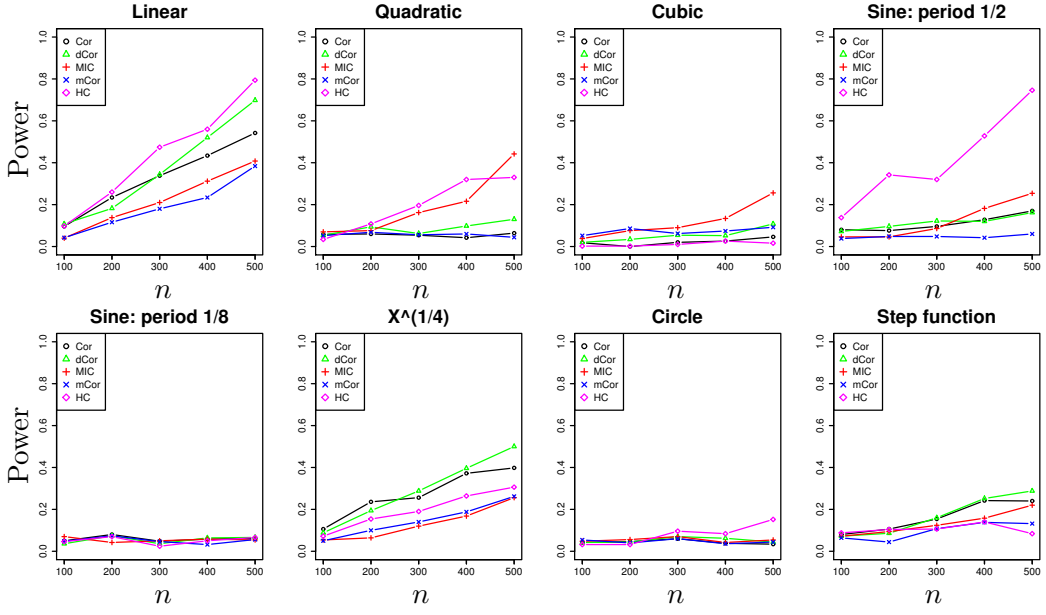

Figure 6.9: Power vs. n (number of samples) for $\alpha = 0.05$, $\sigma^2 = 0.1$.

6.3.2 Real data: Correlation between indicators of WHO datasets

We computed the hypercontractivity coefficient, MIC, and Pearson correlation of 1600 pairs of indicators for 202 countries in the World Health Organization (WHO) dataset [11]. Figure 6.10 illustrates that the hypercontractivity coefficient discovers hidden potential correlation (e.g. in (E) - (H)), whereas other measures fail.Scatter plots of Pearson correlation vs. the hypercontractivity coefficient and MIC vs. the hypercontractivity coefficient for all pairs are presented in (A) and (D). The samples for pairs of indicators corresponding to B, C, E – J are shown in (B), (C), (E) – (J), respectively. In Figure 6.10 (B), it is reasonable to assume that the number of bad teeth per child is uncorrelated with the democracy score. The hypercontractivity coefficient, MIC, and Pearson correlation are all small, as expected. In Figure 6.10 (C), the correlation between CO_2 emissions and energy use is clearly visible, and all three correlation estimates are close to one.

However, only the hypercontractivity coefficient discovers the hidden potential correlation in Figure 6.10 (E) – (H). In Figure 6.10 (E), the data is a mixture of two types of countries – one with small amount of aid received (less than $\$5 \times 10^8$), and the other with large amount of aid received (larger than $\$5 \times 10^8$). Dominantly many countries (104 out of 146) belong to the first type (small aid), and for those countries, the amount of aid received and the income growth are independent. For the remaining countries with larger aid received, although those are rare, there is a clear correlation between the amount of aid received and the income growth.

Similarly in Figure 6.10 (F), there are two types of countries – one with small arms exports (less than $\$2 \times 10^8$) and the other with large arms exports (larger than $\$2 \times 10^8$). Dominantly many countries (71 out of 82) belong to the first type, for which the amount of arms exports and the health expenditure are independent. For the remaining countries that belong to the second type, on the other hand, there is a visible correlation between the arms exports and the health expenditure. This is expected as for those countries that export arms the GDP is positively correlated with both arms exports and health expenditure, whereas for those do not have arms industry, these two will be independent.

In Figure 6.10 (G), for dominant number of countries, the number of male

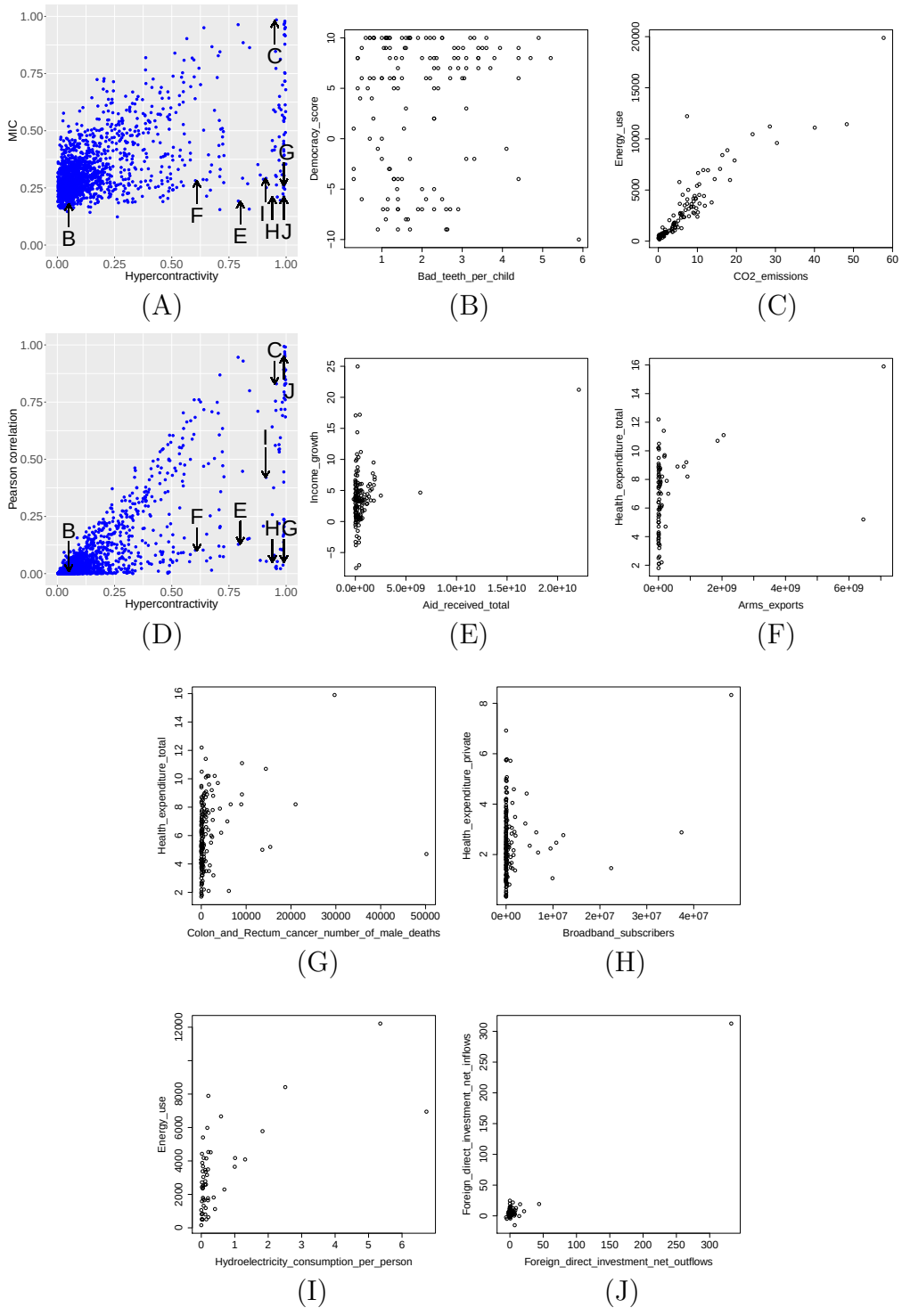

Figure 6.10: (A) and (D): Scatter plot of correlation measures. (B): Correlations are small. (C): Correlations are large. (E),(F),(G),(H): Only the hypercontractivity coefficient discovers potential correlation. (I): Hypercontractivity discovers potential correlation. (J): Hypercontractivity and Pearson correlation are large because of an outlier.

deaths from the colon and rectum cancer is small (145 out of 169 countries have it less than 2000), and it is independent of the amount of health expenditure. On the other hand, for the remaining countries with larger number of male deaths from colon and rectum cancer, the two indicators are positively associated. This is expected as both indicators are positively correlated with the population. Only hypercontractivity discovers this hidden potential correlation. MIC and Pearson correlation are small.

In Figure 6.10 (H), for dominant number of countries, the number of broadband subscribers is very small and is independent of the private health expenditure; 155 out of 180 countries have broadband subscribers less than 10^6. On the other hand, for the remaining countries, the number of broadband subscribers is positively correlated with the private health expenditure. This is as expected because both indicators are positively correlated with the population. Hypercontractivity is large for this dataset, discovering the hidden correlation, whereas all other correlations all small.

In Figure 6.10 (I), most countries do not have large hydroelectricity facilities, and for those countries, energy use and hydroelectricity consumption are independent (41 out of 53 countries have hydroelectricity ≤ 0.25). On the other hand, for the countries which have hydroelectrocity facilities, the amount of total energy use and the amount of hydroelectricity consumption are positively correlated. Hypercontractivity discovers this hidden potential correlation. Unlike in (G) and (H) for which the fraction of correlated samples was only about 14%, in (I), the fraction of correlated samples is about 23%. Hence, Pearson correlation is larger compared to Pearson correlation values for (G) and (H).

In Figure 6.10 (J), there is one country (Luxembourg) with very large amounts of foreign direct investment net inflow and outflow. Due to this outlier, Pearson correlation is close to 1. Hypercontractivity is also close to 1, whereas MIC is small. To analyze the effect of the outlier in correlation measures, in the following, we compute the correlation measures for samples without an outlier.

How hypercontractivity changes as we remove outliers

Figures 6.11–6.16, on the left, are shown samples from Figure 6.10 (E)–(J) respectively. On the middle and on the right are shown all samples but one outlier and all samples but two outliers, respectively. By comparing the hy-

percontractivity coefficients for the three datasets for each pair of indicators, we can analyze the effect of outliers on hypercontractivity. For a comparison, on the top of each figure, we show the estimated hypercontractivity (HC), MIC, Pearson correlation (Cor), distance correlation (dCor), maximal correlation (mCor), and the hypercontractivity for reversed direction (HCR).

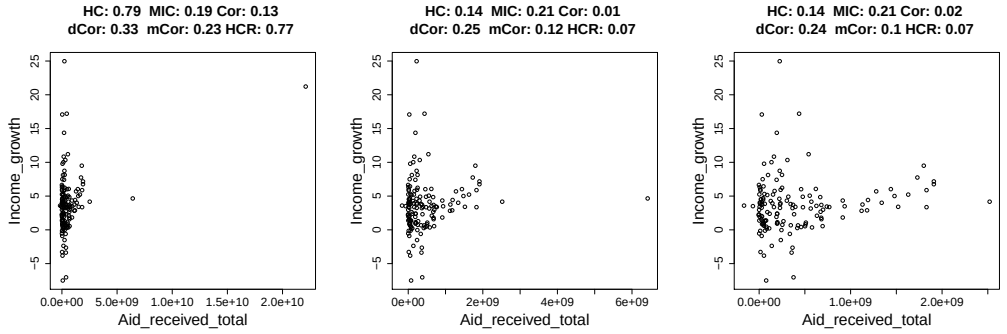

Figure 6.11: Samples for the pair of indicators shown in Figure 6.10-(E) from the entire WHO dataset (left), without one outlier (middle), and without two outliers (right).

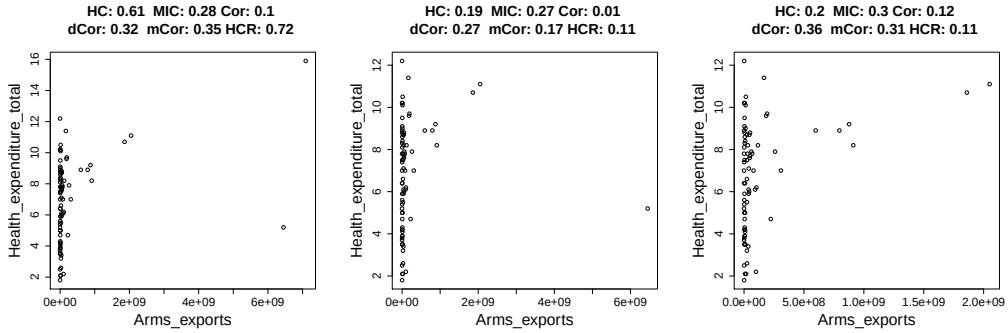

Figure 6.12: Samples for the pair of indicators shown in Figure 6.10-(F) from the entire WHO dataset (left), without one outlier (middle), and without two outliers (right).

In Figure 6.13 (left), the two countries with the largest number of male deaths from the colon and rectum cancer are China and the United States. As China is removed from the dataset, in (middle), hypercontractivity remains unchanged. As we also remove the United States, in (right), hypercontractivity becomes small, 0.17. This value is still larger than the typical coefficient for two independent indicators (≈ 0.05), we can see that hypercontractivity is more sensitive to the outlier than other correlation measures.

In Figure 6.14, the two countries with the largest number of broadband subscribers are the United States and China. When we remove the United

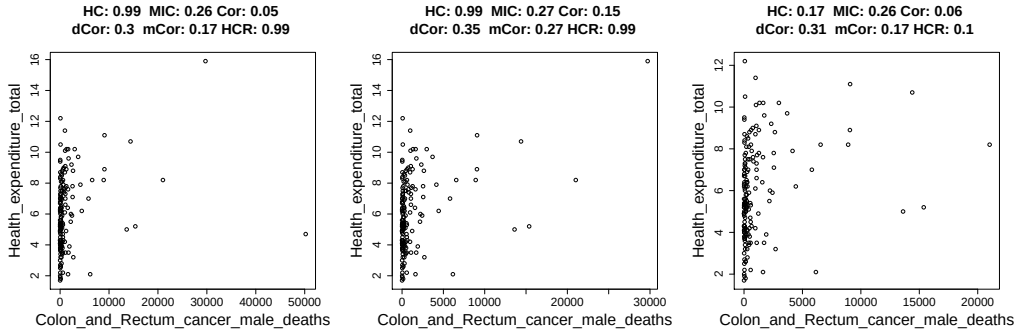

Figure 6.13: Samples for the pair of indicators shown in Figure 6.10-(G) from the entire WHO dataset (left), without one outlier (middle), and without two outliers (right).

States from the samples, hypercontractivity becomes close to zero, which also shows hypercontractivity is sensitive to the outliers.

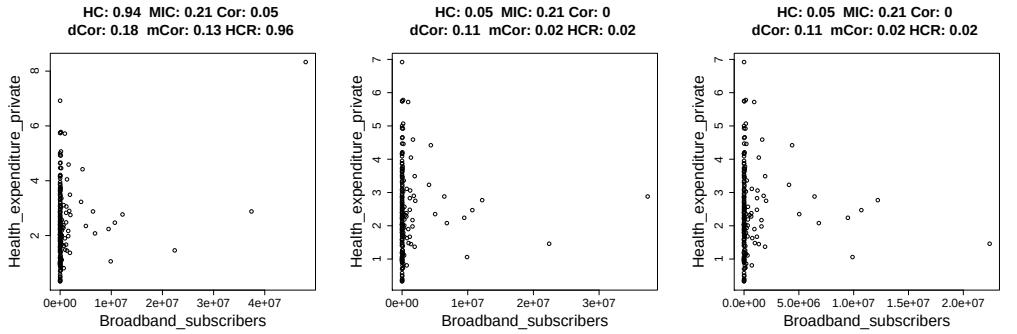

Figure 6.14: Samples for the pair of indicators shown in Figure 6.10-(H) from the entire WHO dataset (left), without one outlier (middle), and without two outliers (right).

In Figure 6.15, hypercontractivity remains large even after we remove outliers. The two countries with the largest amount of hydroelectricity consumption are Norway and Iceland. Even after we remove Norway from the samples, as shown in (middle), hypercontractivity remains large. As we further remove one outlier (Iceland) from the samples, as shown in (right), hypercontractivity becomes 0.49.

In Figure 6.16, (middle), all samples but Luxembourg are shown. We can see that most countries have a very small absolute amount of foreign direct investment net outflows (for 126 out of 157 countries, it is between $[-2, 2]$), and for those countries, the foreign direct investment net outflow is independent of foreign direct investment net inflows. For the remaining countries, there is a positive association between the outflow and the inflow.

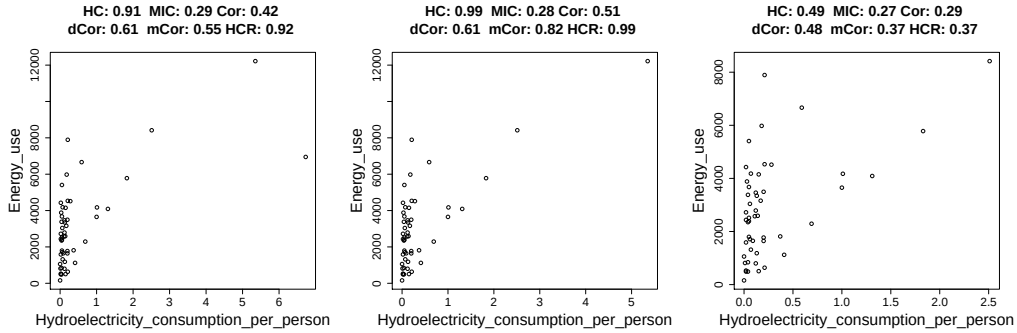

Figure 6.15: Samples for the pair of indicators shown in Figure 6.10-(I) from the entire WHO dataset (left), without one outlier (middle), and without two outliers (right).

Hypercontractivity captures this hidden correlation better than other correlations; hypercontractivity is 0.47, whereas MIC and Pearson correlation are small. If we further remove the rightmost sample, as shown in (right), hypercontractivity becomes small.

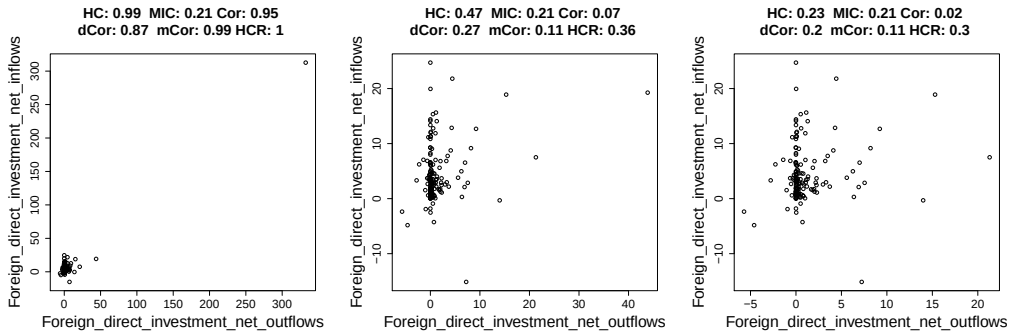

Figure 6.16: Samples for the pair of indicators shown in Figure 6.10-(J) from the entire WHO dataset (left), without one outlier (middle), and without two outliers (right).

Whether we should consider a sample in a rare type as a meaningful sample or as an outlier depends on the application. If we use hypercontractivity to discover a pair of measures for which one variable can be potentially correlated with the other, then we would expect to discover that an aid for a country has potential correlation in the income growth. Other measures will fail. It is possible that hypercontractivity might have a larger false positive rate, and depending on the application, one might prefer to error on the side of having more positive cases to be screened by further experiments, surveys, or human judgments.

Hypercontractivity detecting an outlier

In Figure 6.17 (A) and (B), we show examples of pairs of indicators for which there is one outlier and the remaining samples are independent, but hypercontractivity is large. As shown in Figure 6.17 (A) and (B) (left), hypercontractivity is close to 1, when there is an outlier. As shown in (right), hypercontractivity is close to 0, when the outlier is removed. This implies that one single outlier can make the hypercontractivity large. We can see similar patterns for other correlation measures, such as for Pearson correlation, distance correlation, and maximal correlation for both (A) and (B), and MIC for (B), but are less sensitive than hypercontractivity.

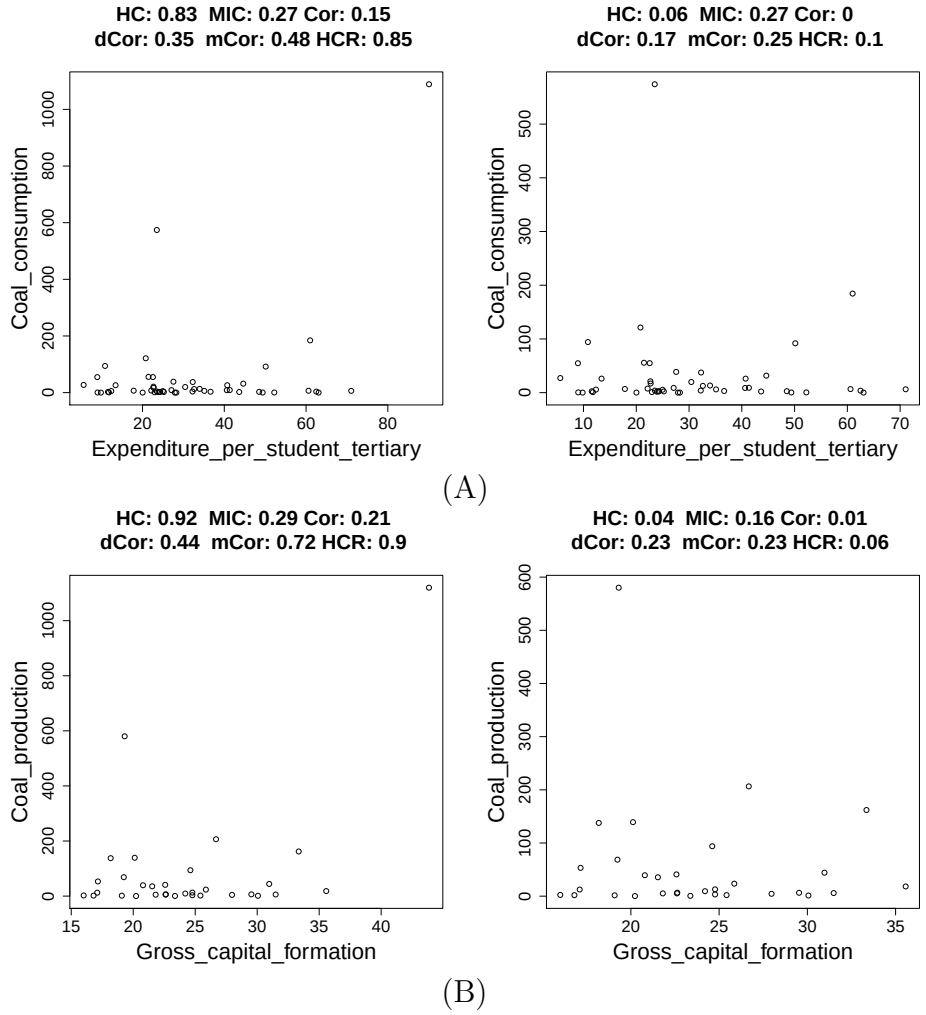

Figure 6.17: Hypercontractivity and other correlation measures become smaller as we remove an outlier.

6.3.3 Gene pathway recovery from single cell data

We replicate the genetic pathway detection experiment from [29], and show that hypercontractivity correctly discovers the genetic pathways from a smaller number of samples. A genetic pathway is a series of genes interacting with each other as a chain. Consider the following setup where four genes whose expression values in a single cell are modeled by random processes X_t, Y_t, Z_t and W_t respectively. These four genes interact with each other following a pathway $X_t \to Y_t \to Z_t \to W_t$; it is biologically known that X_t causes Y_t with a negligible delay, and later at time t', $Y_{t'}$ causes $Z_{t'}$, and so on. Our goal is to recover this known gene pathway from sampled datapoints. For a sequence of time points $\{t_i\}_{i=0}^m$, we observe n_i i.i.d. samples $\{X_{t_i}^{(j)}, Y_{t_i}^{(j)}, Z_{t_i}^{(j)}, W_{t_i}^{(j)}\}_{j=1}^{n_i}$ generated from the random process $P(X_{t_i}, Y_{t_i}, Z_{t_i}, W_{t_i})$. We use the real data obtained by the single-cell mass flow cytometry technique [29].

Given these samples from time series, the goal of [29] is to recover the direction of the interaction along the known pathway using correlation measures as follows, where they proposed a new measure called DREMI. The DREMI correlation measure is evaluated on each pairs on the pathway,$\tau(X_{t_i}, Y_{t_i})$, $\tau(Y_{t_i}, Z_{t_i})$ and $\tau(Z_{t_i}, W_{t_i})$, at each time points t_i. It is declared that a genetic pathway is correctly recovered if the peak of correlation follows the expected trend:

$$\arg\max_{t_i} \tau(X_{t_i}, Y_{t_i}) \leq \arg\max_{t_i} \tau(Y_{t_i}, Z_{t_i}) \leq \arg\max_{t_i} \tau(Z_{t_i}, W_{t_i}). \quad (6.23)$$

In [28], the same experiment has been done with τ evaluated by UMI and CMI estimators. In this chapter, we evaluate τ using our proposed estimator of hypercontractivity.

The Figure 6.18 shows the scatter plots pCD3ζ-pSLP76-pERK-pS6 chain at different time points after TCR activation. The data comes from CD4+ naïve T lymphocytes from B6 mice with CD3, CD28, and CD4 cross-linking. Each row represents a pair of data in the chain, and each column stands for a time point after TCR activation. Estimate of hypercontractivity is shown below the scatter plot for each pair of data and each time point and we highlight the time point where each pair of data is maximally correlated. We can see that the peak of the correlation of pCD3ζ-pSLP76, pSLP76-pERK and pERK-pS6 appears at 0.5 min, 1 min and 2 min respectively, hence the

pathway is correctly identified. In Figure 6.19, the similar plots was shown for T-cells exposed with an antigen. Similarly, hypercontractivity is able to capture the trend.

We subsample the raw data from [29] to evaluate the ability to find the trend from smaller samples. Precisely, given a resampling rate $\gamma \in (0, 1]$, we randomly select a subset of indices $S_i \subseteq [n_i]$ with $\mathrm{card}(S_i) = \lceil \gamma n_i \rceil$, compute pairwise correlations $\tau(X_{t_i}, Y_{t_i})$, $\tau(Y_{t_i}, Z_{t_i})$ and $\tau(Z_{t_i}, W_{t_i})$ from subsamples $\{X_{t_i}^{(j)}, Y_{t_i}^{(j)}, Z_{t_i}^{(j)}, W_{t_i}^{(j)}\}_{j \in S_i}$, and determine whether we can recover the trend successfully, i.e., whether $\arg\max_{t_i} \tau(X_{t_i}, Y_{t_i}) \leq \arg\max_{t_i} \tau(Y_{t_i}, Z_{t_i}) \leq \arg\max_{t_i} \tau(Z_{t_i}, W_{t_i})$. We repeat the experiment several times with independent subsamples and compute the probability of successfully recovering the trend. Figure 6.20 illustrates that when the entire dataset is available, all methods are able to recover the trend correctly. When only fewer samples are available, hypercontractivity improves upon other competing measures in recovering the hidden chronological order of interactions of the pathway. For completeness, we run datasets for both regular T-cells (shown in left figure) and T-cells exposed with an antigen (shown right figure), for which we expect distinct biological trends. Hypercontractivity method can capture the trend for both datasets correctly with fewer samples.

6.4 Proofs of results in Chapter 6

6.4.1 Proof of Proposition 2

Let $S_F(X, Y) = F(\sqrt{s(X;Y)}, \sqrt{s(Y;X)})$ for F satisfying conditions in Proposition 2. We show that $S_F(X, Y)$ satisfies all Rényi's axioms, i.e., Axioms 1-5 and 7-8.

1. $S_F(X, Y)$ is defined for any pair of non-constant random variables X, Y because $s(X;Y) \in [0, 1]$ and $s(Y;X) \in [0, 1]$ are defined for any random variables X, Y by Theorem 12.

2. $S_F(X, Y) \in [0, 1]$ because the output of a function F is in $[0, 1]$ by the condition on F.

3. If X and Y are statistically independent, $s(X;Y) = s(Y;X) = 0$. By the condition on F, it follows that $S_F(X, Y) = 0$. If $S_F(X, Y) = 0$, by

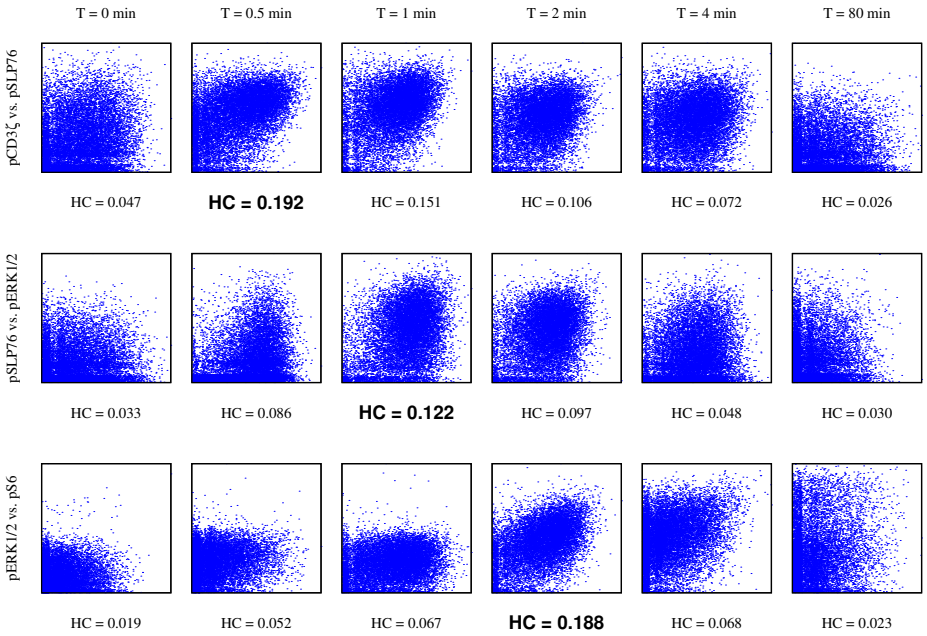

Figure 6.18: Scatter plots of gene pathway data for various pair of data and various time points (regular T-cells).

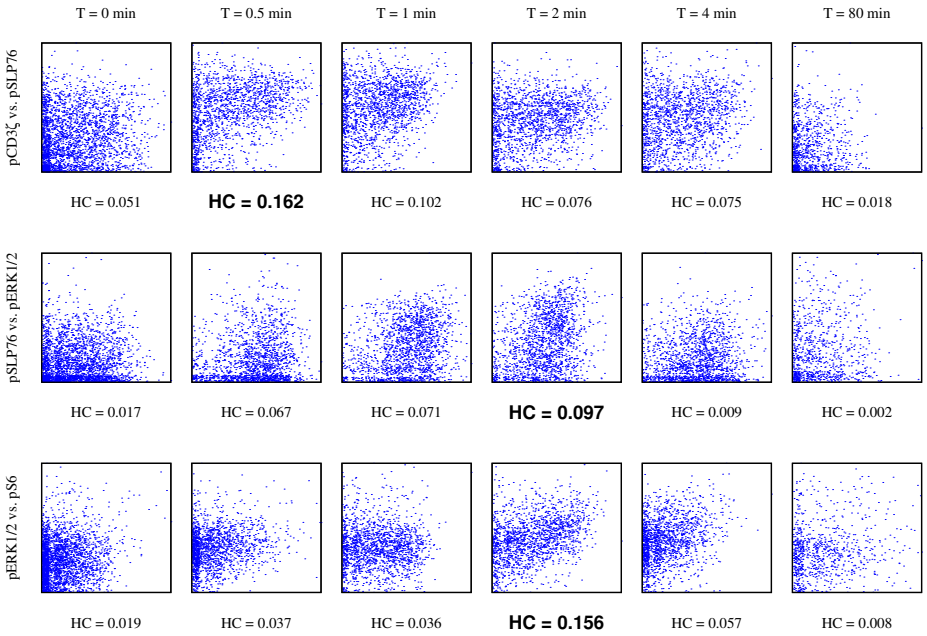

Figure 6.19: Scatter plots of gene pathway data for various pair of data and various time points (T-cells exposed with an antigen).

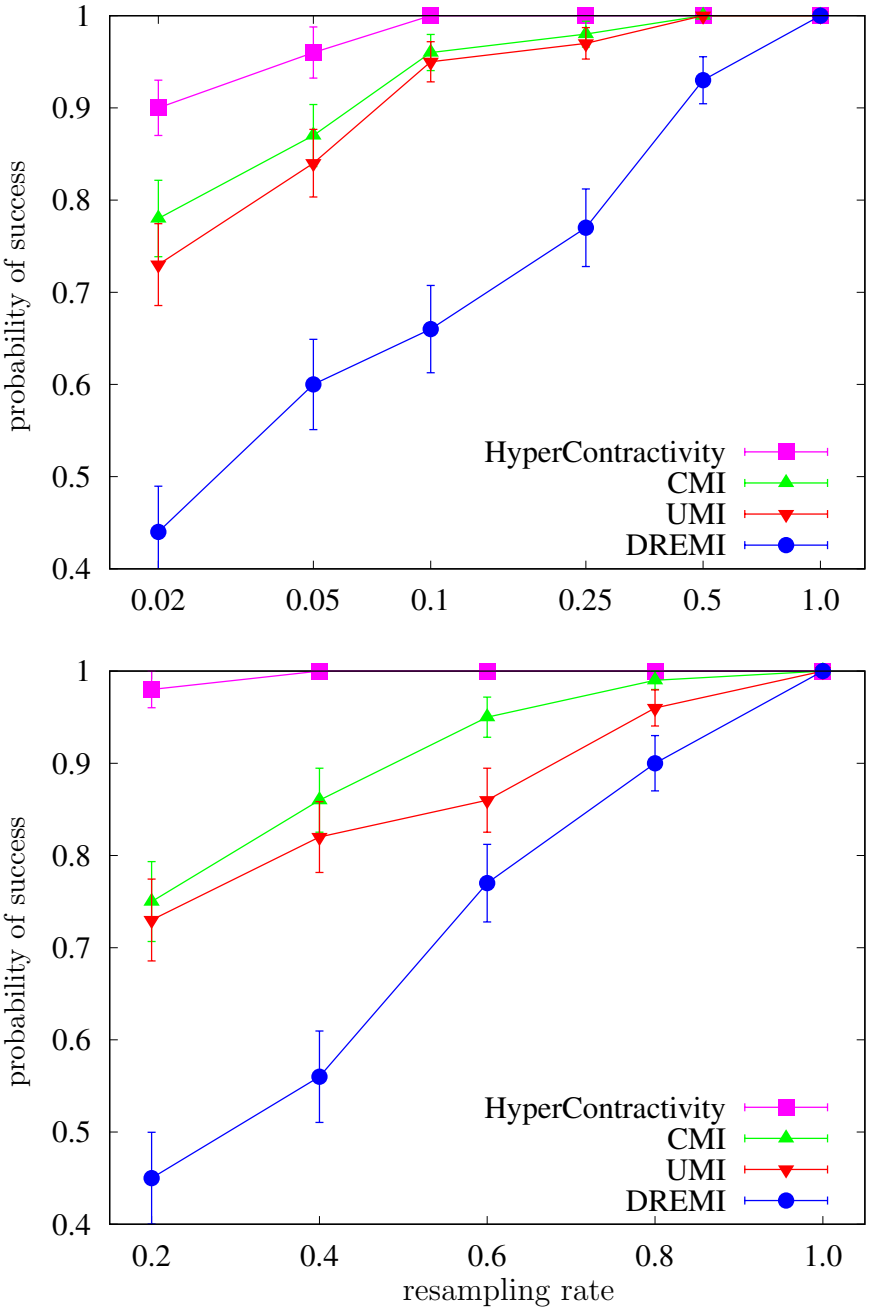

Figure 6.20: Accuracy vs. subsampling rate. Hypercontractivity method has higher probability to recover the trend when data size is smaller compared to other methods. Left: regular T-cells. Right: T-cells exposed with an antigen [29].

the condition on F, $s(X;Y)s(Y;X) = 0$, which implies that X and Y are statistically independent.

4. $S_F(f(X), g(Y)) = S_F(X, Y)$ for any bijective Borel-measurable functions f, g because $\sqrt{s(f(X); g(Y))} = \sqrt{s(X;Y)}$ and $\sqrt{s(g(Y); f(X))} = \sqrt{s(Y;X)}$ by Theorem 12.

5. For $(X, Y) \sim \mathcal{N}(\mu, \Sigma)$ with Pearson correlation ρ, $s(X;Y) = s(Y;X) = \rho^2$. Hence, $S_F(X, Y) = F(|\rho|, |\rho|) = |\rho|$.

7. If $Y = f(X)$ for a non-constant function f, it follows that $I(f(X); f(X)) = I(f(X); X)$ because if $f(X)$ is discrete, $I(f(X); f(X)) = I(f(X); X) = H(f(X))$ and otherwise, $I(f(X); f(X)) = I(f(X); X) = \infty$. Hence

$$\begin{aligned} s(X; f(X)) &= \sup_{U - X - f(X)} I(U; f(X))/I(U; X) \\ = \; & I(f(X); f(X))/I(f(X); X) = 1. \end{aligned} \tag{6.24}$$

   Similarly, $s(f(X); X) = \sup_{U - f(X) - X} I(U; X)/I(U; f(X)) = 1$. Hence, $S_F(X; f(X)) = F(1, 1) = 1$. Likewise, we can show that $S_F(X; Y) = 1$ if $X = g(Y)$.

8. $S_F(X, Y) = S_F(Y, X)$ because $F(x, y) = F(y, x)$.

6.4.2 Proof of Theorem 12

We show that $s(X;Y)$ satisfies Axioms 1-6 in Section 6.1.

1. For any non-constant random variable X, $\exists\, U$ s.t. $I(U; X) > 0$. Hence, $s(X;Y)$ is defined for any pair of non-constant random variables X and Y.

2. Since mutual information is non-negative, $s(X;Y) \geq 0$. By data processing inequality, for any $U - X - Y$, $I(U;X) \leq I(U;Y)$. Hence, $s(X;Y) \leq 1$.

3. If X and Y are independent, for any U, $I(U;Y) \leq I(X;Y) = 0$. Hence, $s(X;Y) = 0$. If X and Y are dependent, $I(X;Y) > 0$, which implies that $s(X;Y) \geq I(X;Y)/H(X) > 0$.

4. For any bijective functions f, g,

$$\begin{aligned} & I(U;g(Y)) = I(U;g(Y),Y) \\ = \;& I(U;Y) + I(U;g(Y)|Y) = I(U;Y). \end{aligned} \tag{6.25}$$

Similarly, $I(U;f(X)) = I(U;X)$. Hence,

$$\begin{aligned} & s(f(X);g(Y)) = \sup_{U:U-f(X)-g(Y),I(U;f(X))>0} \frac{I(U;g(Y))}{I(U;f(X))} \\ = \;& \sup_{U:U-X-f(X)-g(Y)-Y,I(U;X)>0} \frac{I(U;Y)}{I(U;X)} = s(X;Y). \end{aligned} \tag{6.26}$$

5. By Theorem 3.1 in [179], for (X,Y) jointly Gaussian with correlation coefficient ρ,

$$\min_{U:\, U-X-Y} \left(I(U;X) - \beta I(U;Y)\right) = 0, \tag{6.27}$$

for $\beta \leq 1/\rho^2$. Equivalently,

$$\max_{U:\, U-X-Y} \left(I(U;Y) - \rho^2 I(U;X)\right) = 0, \tag{6.28}$$

which implies that $s(X;Y) \leq \rho^2$. To show that $s(X;Y) \geq \rho^2$, let $U_Z = X + Z$ for $Z \sim (0, \sigma_1^2)$. Consider

$$\begin{aligned} & s(X;Y) \geq \lim_{\sigma_1^2 \to \infty} \frac{I(U_Z;Y)}{I(U_Z;X)} \\ = \;& \lim_{\sigma_1^2 \to \infty} \frac{\log\left(\frac{(\sigma_X^2+\sigma_1^2)\sigma_Y^2}{(\sigma_X^2+\sigma_1^2)\sigma_Y^2 - \rho^2\sigma_X^2\sigma_Y^2}\right)}{\log\left(1+\frac{\sigma_X^2}{\sigma_1^2}\right)} \\ = \;& \lim_{\sigma_1^2 \to \infty} \frac{\rho^2\sigma_X^2\sigma_Y^2 / \left((\sigma_X^2+\sigma_1^2)\sigma_Y^2 - \rho^2\sigma_X^2\sigma_Y^2\right)}{\sigma_X^2/\sigma_1^2} = \rho^2. \end{aligned} \tag{6.29}$$

Hence, $s(X;Y) = \rho^2$. An alternative proof is provided in [172].

6. To prove that $s(X;Y)$ satisfies Axiom 6, we first show the following lemma.

Lemma 24. Consider a pair of random variables $(X,Y) \in \mathcal{X} \times \mathcal{Y}$. The

hypercontractivity $s(X;Y)$ is lower bounded by

$$s(X;Y) \geq \frac{I(U;Y|X \in \mathcal{X}_r)}{H(\alpha)/\alpha + I(U;X|X \in \mathcal{X}_r)}, \tag{6.30}$$

for any $\mathcal{X}_r$ such that $\mathcal{X}_r \subseteq \mathcal{X}$ for $\Pr\{X \in \mathcal{X}_r\} =: \alpha > 0$.

Proof. Let

$$U_s = \begin{cases} U \sim p(u|x) & \text{if } X \in \mathcal{X}_r, \\ \emptyset & \text{otherwise.} \end{cases} \tag{6.31}$$

Let $S = \mathbb{I}\{U_s = \emptyset\} = \mathbb{I}\{X \in \mathcal{X}_r\}$. Note that $S - U_s - X - Y$ holds, and that S is a deterministic function of X. Hence,

$$\begin{aligned} & I(U_s;X) = I(U_s,S;X) = I(S;X) + I(U_s;X|S) \\ = \;& H(\alpha) + \alpha I(U;X|X \in \mathcal{X}_r). \end{aligned} \tag{6.32}$$

Consider

$$\begin{aligned} & I(U_s;Y) = I(U_s,S;Y) = I(S;Y) + I(U_s;Y|S) \\ \geq \;& \alpha I(U;Y|X \in \mathcal{X}_r). \end{aligned} \tag{6.33}$$

The proof is completed by combining (6.32) and (6.33). □

Assume that $Y = f(X)$ for $X \in \mathcal{X}_r$. Considering $U = f(X)$ in (6.31) in Lemma 24, we obtain the following lower bound:

$$s(X;Y) \geq \frac{I(f(X);f(X)|X \in \mathcal{X}_r)}{H(\alpha)/\alpha + I(f(X);X|X \in \mathcal{X}_r)}. \tag{6.34}$$

For any continuous random variable X and a non-constant continuous function f, $I(f(X);f(X)|X \in \mathcal{X}_r) = I(f(X);X|X \in \mathcal{X}_r) = \infty$, which implies that $s(X;Y) = 1$.

6.4.3 Proof of Theorem 13

We first prove that $\mathrm{mCor}(X,Y) = \sqrt{\alpha}\,\mathrm{mCor}(X_r,Y)$ in (6.2). Let $S = \mathbb{I}\{X \in \mathcal{X}_r\}$ be the indicator for whether $X \in \mathcal{X}_r$ or not. Let $\mathcal{F}$ be the set of functions

such that $f \in \mathcal{F}$ if $\mathbb{E}[f(X)] = 0$ and $\mathbb{E}[f^2(X)] \leq 1$. Consider

$$\begin{aligned}
& \mathrm{mCor}(X;Y) = \max_{f,g\in\mathcal{F}} \mathbb{E}[f(X)g(Y)] \\
= \;& \max_{f,g\in\mathcal{F}} \mathbb{E}_S[\mathbb{E}[f(X)g(Y)|S]] \\
= \;& \max_{f,g\in\mathcal{F}} (\alpha\mathbb{E}[f(X)g(Y)|X \in \mathcal{X}_r] + \bar{\alpha}\mathbb{E}[f(X)g(Y)|X \in \mathcal{X}_d]) \\
= \;& \max_{f,g\in\mathcal{F}} (\alpha\mathbb{E}[f(X)g(Y)|X \in \mathcal{X}_r] + \bar{\alpha}\mathbb{E}[f(X)|X \in \mathcal{X}_d]\mathbb{E}[g(Y)|X \in \mathcal{X}_d]) \\
\overset{(a)}{=} \;& \alpha \max_{f,g\in\mathcal{F}} \mathbb{E}[f(X)g(Y)|X \in \mathcal{X}_r] \\
\overset{(b)}{=} \;& \sqrt{\alpha}\,\mathrm{mCor}(X_r, Y). \qquad (6.35)
\end{aligned}$$

Step (a) holds since $\mathbb{E}[g(Y)|X \in \mathcal{X}_r] = \mathbb{E}[g(Y)|X \in \mathcal{X}_d]$ from the assumption that marginal distributions are equal, and that $\mathbb{E}[g(Y)] = \alpha\mathbb{E}[g(Y)|X \in \mathcal{X}_r] + \bar{\alpha}\mathbb{E}[g(Y)|X \in \mathcal{X}_d]$. To show step (b), let $c = \mathbb{E}[f(X)|X \in \mathcal{X}_d]$ and note that

$$\begin{aligned}
\alpha\mathbb{E}[f(X)|X \in \mathcal{X}_r] &= -\bar{\alpha}c, \\
\alpha\mathbb{E}[f^2(X)|X \in \mathcal{X}_r] &= \mathbb{E}[f^2(X)] - \bar{\alpha}\mathbb{E}[f^2(X)|X \in \mathcal{X}_d] \leq 1 - \bar{\alpha}c^2, \\
\mathbb{E}[g(Y)|X \in \mathcal{X}_r] &= 0. \qquad (6.36)
\end{aligned}$$

Denote $\mathcal{F}(\mu_1, \mu_2)$ be the set of functions such that $\mathbb{E}[f(X)] = -\mu_1$ and $\mathbb{E}[f^2(X)] \leq \mu_2$. $\mathcal{F} = \mathcal{F}(0, 1)$ for short. Hence,

$$\begin{aligned}
& \max_{f,g\in\mathcal{F}} \mathbb{E}[f(X)g(Y)|X \in \mathcal{X}_r] \\
= \;& \max_{f_r\in\mathcal{F}(-\bar{\alpha}c/\alpha,(1-\bar{\alpha}c^2)/\alpha),g\in\mathcal{F}} \mathbb{E}[f_r(X)g(Y)] \\
= \;& \max_{f_{rc}\in\mathcal{F}(0,(\alpha-\bar{\alpha}c^2)/\alpha^2),g\in\mathcal{F}} \mathbb{E}[(f_{rc}(X)g(Y)] \\
= \;& \max_{f_{rc}\in\mathcal{F}(0,1/\alpha),g\in\mathcal{F}} \mathbb{E}[f_{rc}(X)g(Y)] \\
= \;& \max_{f_{rca},g\in\mathcal{F}} \frac{1}{\sqrt{\alpha}}\mathbb{E}[f_{rca}(X)g(Y)] = \frac{\mathrm{mCor}(X_r, Y)}{\sqrt{\alpha}}, \qquad (6.37)
\end{aligned}$$

where $f_r(X)$, $f_{rc}(X) = f_r(X) + \bar{\alpha}c/\alpha$, and $f_{rca}(X) = \sqrt{\alpha}f_{rc}(X)$ are functions defined only for $X \in \mathcal{X}_r$.

We next show $\mathrm{dCor}(X, Y) = \alpha\,\mathrm{dCor}(X_r, Y)$ in (6.3). Let

$$h_X(s) = \mathbb{E}[e^{isX}],\ h_Y(t) = \mathbb{E}[e^{itY}],\ h_{XY}(s,t) = \mathbb{E}[e^{i(sX+tY)}]. \qquad (6.38)$$

Note that

$$\begin{aligned} & h_{XY}(s,t) = \mathbb{E}[e^{i(sX+tY)}] \\ = \;& \alpha\mathbb{E}[e^{i(sX+tY)}|X \in \mathcal{X}_r] + \bar{\alpha}r\alpha\mathbb{E}[e^{isX}|X \in \mathcal{X}_d]\mathbb{E}[e^{itY}|X \in \mathcal{X}_d] \\ = \;& \alpha\mathbb{E}[e^{i(sX+tY)}|X \in \mathcal{X}_r] + \bar{\alpha}r\alpha\mathbb{E}[e^{isX}|X \in \mathcal{X}_d]\mathbb{E}[e^{itY}], \end{aligned} \tag{6.39}$$

and

$$h_X(s) = \mathbb{E}[e^{isX}] = \alpha\mathbb{E}[e^{isX}|X \in \mathcal{X}_r] + \bar{\alpha}r\alpha\mathbb{E}[e^{isX}|X \in \mathcal{X}_d]. \tag{6.40}$$

By combining (6.39) and (6.40),

$$\begin{aligned} & h_{XY}(s,t) - h_X(s)h_Y(t) \\ = \;& \alpha\mathbb{E}[e^{i(sX+tY)}|X \in \mathcal{X}_r] - \alpha\mathbb{E}[e^{isX}|X \in \mathcal{X}_r]\mathbb{E}[e^{itY}] \\ = \;& \alpha\mathbb{E}[e^{i(sX+tY)}|X \in \mathcal{X}_r] - \alpha\mathbb{E}[e^{isX}|X \in \mathcal{X}_r]\mathbb{E}[e^{itY}|X \in \mathcal{X}_r] \\ = \;& \alpha \, \mathrm{dCor}(X_r, Y). \end{aligned} \tag{6.41}$$

Finally, we show that $\mathrm{MIC}(X,Y) \leq \alpha \, \mathrm{MIC}(X_r, Y)$ in (6.4).

Let $X_Q(X) \in \mathcal{X}_Q(X)$ and $Y_Q(Y) \in \mathcal{Y}_Q(Y)$ denote a quantization of X and Y, respectively. Consider

$$\begin{aligned} & \mathrm{MIC}(X,Y) = \max_{X_Q(X), Y_Q(Y)} \frac{I(X_Q;Y_Q)}{\log\min\{|\mathcal{X}_Q|, |\mathcal{Y}_Q|\}} \\ \leq \;& \max_{X_Q(X), Y_Q(Y)} \frac{I(\mathbb{I}\{X \in \mathcal{X}_r\}, X_Q; Y_Q)}{\log\min\{|\mathcal{X}_Q|, |\mathcal{Y}_Q|\}} \\ \overset{(a)}{=} \;& \alpha \max_{X_Q(X), Y_Q(Y)} \frac{I(X_Q;Y_Q|X \in \mathcal{X}_r)}{\log\min\{|\mathcal{X}_Q|, |\mathcal{Y}_Q|\}} \\ \leq \;& \alpha \max_{X_Q(X_r), Y_Q(Y)} \frac{I(X_Q;Y_Q|X \in \mathcal{X}_r)}{\log\min\{|\mathcal{X}_Q(X_r)|, |\mathcal{Y}_Q|\}} \\ = \;& \alpha \, \mathrm{MIC}(X_r, Y), \end{aligned} \tag{6.42}$$

where step (a) holds because $\mathbb{I}\{X \in \mathcal{X}_r\}$ is independent of Y implies $\mathbb{I}\{X \in \mathcal{X}_r\}$ is independent of Y_Q and X is independent of Y in $X \in \mathcal{X}_d$ implies X_Q is independent of Y_Q in $X \in \mathcal{X}_d$.

6.4.4 Proof of Proposition 3

The inverse hypercontractivity $s(Y;X)$ is defined as

$$s(Y;X) = \sup_{U-Y-X} \frac{I(U;X)}{I(U;Y)}. \tag{6.43}$$

Let $\mathbb{I}_r = \mathbb{I}\{X \in \mathcal{X}_r\}$. Since the marginal distribution of Y given $\{X \in \mathcal{X}_r\}$ and the one given $\{X \notin \mathcal{X}_r\}$ are equivalent, Y and $\mathbb{I}_r$ are independent, i.e., $I(Y;\mathbb{I}_r) = 0$. For any U such that Markov chain $U - Y - X$ holds, the Markov chain $U - Y - X - \mathbb{I}_r$ holds. Hence, $I(U;\mathbb{I}_r) = 0$. Hence, for any $U - Y - X$, consider

$$\begin{aligned} & I(U;X) = I(U;X,\mathbb{I}_r) = I(U;X|\mathbb{I}_r) \\ = \;& (1-\alpha) I(U;X|\mathbb{I}_r = 0) + \alpha I(U;X|\mathbb{I}_r = 1) \\ \overset{(a)}{=} \;& \alpha I(U;X|\mathbb{I}_r = 1). \end{aligned} \tag{6.44}$$

Step (a) holds because Y is independent of X given $\mathbb{I}_r = 0$. Consider

$$\begin{aligned} & I(U;Y) \overset{(a)}{=} I(U;Y,\mathbb{I}_r) = I(U;Y|\mathbb{I}_r) + I(U;\mathbb{I}_r) \\ \overset{(b)}{=} \;& I(U;Y|\mathbb{I}_r) = \alpha I(U;Y|\mathbb{I}_r = 1) + (1-\alpha) I(U;Y|\mathbb{I}_r = 0) \\ \overset{(c)}{=} \;& I(U;Y|\mathbb{I}_r = 1), \end{aligned} \tag{6.45}$$

where step (a) folllows since $U - Y - \mathbb{I}_r$. Step (b) follows from $I(U;\mathbb{I}_r) = 0$. Step (c) holds since $H(U|\mathbb{I}_r = 1) = H(U|\mathbb{I}_r = 0)$ and $U - Y - \mathbb{I}_r$. Therefore, for any $U - Y - X$, it follows that

$$s(Y;X) = \sup_{U-Y-X} \frac{\alpha I(U;X|I_r = 1)}{I(U;Y|I_r = 1)} = \alpha \; s(Y;X_r). \tag{6.46}$$

6.4.5 Noisy potential correlation in Example 1

Let $U = X + Z, \;\; Z \sim \mathcal{N}(0,\sigma_1^2)$. Consider

$$\sup_{U:U-X-Y, I(U;X)>0} \frac{I(U;Y)}{I(U;X)} \geq \sup_{\sigma_1^2 \geq 0} \frac{\log \frac{(1+\sigma_1^2)}{(1+\sigma_1^2)-\rho^2}}{H(\alpha)/\alpha + \log(1+1/\sigma_1^2)}. \tag{6.47}$$

The inequality (6.11) follows by choosing $\sigma_1^2 = (1-\rho^2)/\rho^2$.

6.4.6 Noisy discrete potential correlation in Example 3

The inequality (6.14) follows by choosing $r(x) = \mathbb{I}\{X = 1\}$ in (6.16). To show (6.15), we show that

$$\mathrm{mCor}(X_r, Y) = 1 - \frac{k}{k-1}\epsilon. \tag{6.48}$$

The rest follows because $\mathrm{mCor}(X;Y) = \sqrt{\alpha}\,\mathrm{mCor}(X_r, Y)$ by Proposition 13. To show (6.48), we use the fact that maximal correlation is the second eigenvalue of $Q = P_X^{-1/2} P_{XY} P_Y^{-1/2}$ (see [183] for detailed proof). Hence

$$Q = \left(1 - \frac{k}{k-1}\epsilon\right) I + \frac{\epsilon}{k-1} 1 1^T. \tag{6.49}$$

First singular vector of Q is $P_X^{1/2} = 1/\sqrt{k}$. Second singular vector u_2 is orthogonal to $1/\sqrt{k}$. Therefore (6.48) follows because $\mathrm{mCor}(X_r; Y) = u_2^T Q u_2 = u_2^T(1 - k\epsilon/(k-1))u_2$.

6.4.7 Proof of Proposition 5

We first prove the second part of proposition: the hypercontractivity coefficient $\sqrt{s(X;Y)}$ satisfies Axioms 1-4, 9, and 6. It follows immediately from Theorem 12 that $\sqrt{s(X;Y)}$ satisfies Axioms 1-4 and 6 because in the proof of Theorem 12 – 1-4 and 6, the same argument holds for random vectors X and Y. We can show that that $\sqrt{s(X;Y)}$ satisfies Axiom 9 using results from [179]. In [179], it is shown that that as we increase β starting from zero, $\min\{I(U;X) - \beta I(U;Y)\}$ departs form zero at $\beta = 1/\|\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\|^2$ for jointly Gaussian random vectors X and Y. This result implies that $\sqrt{s(X;Y)} = \|\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\|$.

To show that maximal correlation of two random vectors satisfies Axioms 1-4, 7, and 8, we follow the same arguments for showing that maximal correlation for two random variables satisfies Axioms 1-4, 7, and 8 by [156]. To show that maximal correlation satisfies Axiom 9, note that maximal correlation is upper bounded by hypercontractivity (see Remark 2 in Section 6.1.3). Hence $\mathrm{mCor}(X;Y) \le \|\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\|$ for a jointly Gaussian X, Y. Equality holds because $\mathrm{mCor}(X,Y)$ is lower bounded by its canonical correlation, which is $\|\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1/2}\|$ for jointly Gaussian random vectors (X,Y) [179].

6.4.8 Proof of Theorem 14

We begin with the following assumptions:

(*a*) There exist finite constants $C_1 < C_1' < C_2' < C_2$ such that the ratio of r_x^* and p_x satisfies $r_x^*(x)/p_x(x) \in [C_1', C_2']$ for every $x \in \mathcal{X}$.

(*b*) There exist finite constants $C_0' > C_0 > 0$ such that the KL divergence $D(r_x^*||p_x) > C_0'$.

With a little abuse of notations, we define $s(r_x) = D(r_y||p_y)/D(r_x||p_x)$ and $\widehat{s}(\mathbf{w}) = \mathbf{w}^T\mathbf{A}\log(\mathbf{A}^T\mathbf{w})/\mathbf{w}^T\log\mathbf{w}$. Therefore, $s(X;Y) = \max_{r_x \in R} s(r_x)$ and $\widehat{s}_\Delta(X;Y) = \max_{\mathbf{w}\in T_\Delta} \widehat{s}(\mathbf{w})$. Here R is the probability simplex over all r_x. We want to bound the error $|\widehat{s}_\Delta(X;Y) - s(X;Y)|$. First, consider

$$s_\Delta(X;Y) \equiv \max_{r_x \in T_\Delta(R)} s(r_x), \tag{6.50}$$

where the constraint set $T_\Delta(R)$ is defined as:

$$\begin{aligned} & T_\Delta(R) = \{r_x \in \mathbb{R}^{|\mathcal{X}|} : [(r_x(x)/p_x(x))] \in T_\Delta \\ \text{and} \quad & \sum_{x\in\mathcal{X}} r_x(x) \in [1 - |\mathcal{X}|\Delta, 1 + |\mathcal{X}|\Delta]\}. \end{aligned} \tag{6.51}$$

Now we rewrite the error term as

$$\begin{aligned} & \Big|\, \widehat{s}_\Delta(X;Y) - s(X;Y) \,\Big| \\ \leq \;& |s_\Delta(X;Y) - s(X;Y)| + |\widehat{s}_\Delta(X;Y) - s_\Delta(X;Y)|. \end{aligned} \tag{6.52}$$

The first error comes from quantization. Let r^* be the maximizer of $s(X;Y)$. By assumption, $r^*(x)/p_x(x) \in [C_1, C_2]$, for all x. Since $T_\Delta(R)$ is a quantization of the simplex R, so there exists an $r_0 \in T_\Delta(R)$ such that $|r_0(x) - r^*(x)| < \Delta$ for all $x \in \mathcal{X}$. Now we will bound the difference between $s(r_0)$ and $s(r^*)$ by the following lemma:

Lemma 25. *If $r(x)/p(x) \in [C_1, C_2]$ and $r'(x)/p(x) \in [C_1, C_2]$ for all $x \in \mathcal{X}$, and $D(r_x||p_x) > C_0$ and $D(r_x'||p_x) > C_0$, then*

$$\Big|\, s(r) - s(r') \,\Big| \leq L \max_{x\in\mathcal{X}} |r(x) - r'(x)|, \tag{6.53}$$

for some positive constant L.

Next we have:

$$\begin{aligned} & s(X;Y) = s(r^*) \le s(r_0) + L \max_{x\in\mathcal{X}} |r_0(x) - r^*(x)| \\ \le \quad & \max_{r\in T_\Delta(R)} s(r) + L\Delta = s_\Delta(X;Y) + L\Delta. \end{aligned} \tag{6.54}$$

Similarly, let r^{**} be the maximizer of $s_\Delta(X;Y)$, we can also find a $r_1 \in R$ such that $|r_1(x) - r^{**}(x)| < \Delta$ for all $x \in \mathcal{X}$. Using Lemma 25 again, we will obtain $s_\Delta(X;Y) \le s(X;Y) + L\Delta$. Therefore, the quantization error is bounded by $O(\Delta)$ with probability 1.

Now consider the second term. Upper bound on the second term relies on the convergence of estimation of s. We claim that for given r_x, the estimator is convergent in probability in Lemma 26.

Lemma 26.

$$\lim_{N\to\infty} \Pr\left(\left| \widehat{s}(\mathbf{w}_r) - s(r_x) \right| > \varepsilon \right) = 0. \tag{6.55}$$

Here $\mathbf{w}_r(x) = r_x(x)/p_x(x)$. Since the set $T_\Delta(R)$ is finite, by union bound, we have:

$$\begin{aligned} & \lim_{N\to\infty} \Pr\left(\forall r \in T_\Delta(R), \left| \widehat{s}(\mathbf{w}_r) - s(r_x) \right| \le \varepsilon \right) \\ \ge \quad & 1 - |T_\Delta(R)| \lim_{N\to\infty} \Pr\left(\left| \widehat{s}(\mathbf{w}_r) - s(r_x) \right| \le \varepsilon \right) = 1. \end{aligned} \tag{6.56}$$

Also, by the strong law of large numbers, we have that

$$\lim_{N\to\infty} \Pr\left(\forall x \in \mathcal{X}, |p_x(x) - \frac{n_x}{n}| < \frac{\Delta}{C_2|\mathcal{X}|} \right) = 1, \tag{6.57}$$

where $n_x = \mathrm{card}\{i \in [n] : x_i = x\}$. We claim that if the events inside the probability in (6.56) and (6.57) happen simultaneously, then $|\widehat{s}_\Delta(X;Y) - s_\Delta(X;Y)| < \varepsilon + O(\Delta)$, which implies the desired claim.

Let $\mathbf{w}^* = \arg\max_{\mathbf{w}\in T_\Delta} \widehat{s}(\mathbf{w})$. Define $r_2(x) = \mathbf{w}^*(x)p_x(x)$. Since we have $[r_2(x)/p_x(x)] \in T_\Delta$ for all x and

$$\begin{aligned} & \left| \sum_{x\in\mathcal{X}} r_2(x) - 1 \right| = \left| \sum_{x\in\mathcal{X}} \mathbf{w}_x^*(p_x(x) - \frac{n_x}{n}) \right| + \frac{\Delta|\mathcal{X}|}{2} \\ \le \quad & |\mathcal{X}| \left(\frac{\Delta}{2} + C_2 \max_{x\in\mathcal{X}} \left| p_x(x) - \frac{n_x}{n} \right| \right) \le (|\mathcal{X}|/2 + 1)\Delta. \end{aligned} \tag{6.58}$$

Therefore, $r_2 \in T_\Delta(R)$, so

$$\widehat{s}_\Delta(X;Y) \;=\; \widehat{s}(\mathbf{w}^*) \le s(r_2) + \varepsilon \le s_\Delta(X;Y) + \varepsilon. \tag{6.59}$$

On the other hand, consider $r^{**} = \arg\max_{r_x \in T_\Delta(R)} s(r_x)$ again, and define $\mathbf{w}_0(x) = r^{**}(x)/p_x(x)$. We know that $\mathbf{w}_0 \in T_\Delta^{|\mathcal{X}|}$ but not necessarily $\sum_{i=1}^n \mathbf{w}_0(x_i) = n$. But we claim that the sum is closed to n as follows:

$$\begin{aligned} \Big| \sum_{i=1}^n \mathbf{w}_0(x_i) - n \Big| &= \Big| \sum_{x\in\mathcal{X}} \frac{n_x r^{**}(x)}{p_x(x)} - n| \\ \le\; n \max_{x\in\mathcal{X}} \Big\{ \frac{r^{**}(x)}{p_x(x)} \Big| \frac{n_x}{n} - p_x(x) \Big| \Big\} &\le nC_2 \frac{\Delta}{C_2|\mathcal{X}|} < n\Delta, \end{aligned} \tag{6.60}$$

so we can find a $\mathbf{w}_1 \in T_\Delta(R)$ such that $|\mathbf{w}_1(x) - \mathbf{w}_0(x)| \le \Delta$ for all x. Let $r_4(x) = \mathbf{w}_1(x)p_x(x)$, similar as (6.58), we know that $r_4 \in T_\Delta(R)$. Moreover, $\Big| r_4(x) - r^{**}(x) \Big| \le p_x(x) \Big| \mathbf{w}_1(x) - \mathbf{w}_0(x) \Big| \le \Delta$ for all x. Then we have

$$\begin{aligned} s_\Delta(X;Y) \;&=\; s(r^{**}) \le s(r_4) + L \max_{x\in\mathcal{X}} |r^{**}(x) - r_4(x)| \\ &\le\; \widehat{s}(\mathbf{w}_1) + \varepsilon + L\Delta = \widehat{s}_\Delta(X;Y) + \varepsilon + L\Delta. \end{aligned} \tag{6.61}$$

We conclude that $|\widehat{s}_\Delta(X;Y) - s_\Delta(X;Y)| < \varepsilon + O(\Delta)$.

Proof of Lemma 25

We will show that for any $x \in \mathcal{X}$, we have $|\partial s(r_x)/\partial r_x(x)| \le L/|\mathcal{X}|$ for some L. Therefore,

$$\begin{aligned} |s(r) - s(r')| \;&\le\; \sum_{x\in\mathcal{X}} |\frac{\partial s(r)}{\partial r_x(x)}|\, |r_x(x) - r'_x(x)| \\ &\le\; L \max_{x\in\mathcal{X}} |r_x(x) - r'_x(x)|. \end{aligned} \tag{6.62}$$

The gradient can be written as

$$\begin{aligned} \frac{\partial s(r)}{\partial r_x(x)} &= \frac{\partial}{\partial r_x(x)} \frac{D(r_y||p_y)}{D(r_x||p_x)} \\ &= \frac{\frac{\partial D(r_y||p_y)}{\partial r_x(x)} D(r_x||p_x) - \frac{\partial D(r_x||p_x)}{\partial r_x(x)} D(r_y||p_y)}{D^2(r_x||p_x)}. \end{aligned} \tag{6.63}$$

Since

$$\begin{aligned}
&\frac{\partial D(r_x||p_x)}{\partial r_x(x)} = \log\frac{r_x(x)}{p_x(x)} + 1 \leq \max\{|\log C_1|, |\log C_2|\} + 1, \\
&\frac{\partial D(r_y||p_y)}{\partial r_x(x)} = \int \frac{\partial r_y(y)}{\partial r_x(x)} \frac{\partial D(r_y||p_y)}{\partial r_y(y)} dy \\
= &\int p_{y|x}(y|x)(\log\frac{r_y(y)}{p_y(y)} + 1)dy \\
\leq &\max\{|\log C_1|, |\log C_2|\} + 1. \qquad (6.64)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&|\frac{\partial s(r)}{\partial r_x(x)}| \leq (\max\{|\log C_1|, |\log C_2|\} + 1)\frac{D(p_x||r_x) + D(r_y||p_y)}{D^2(r_x||p_x)} \\
\leq &\frac{2(\max\{|\log C_1|, |\log C_2|\} + 1)}{D(r_x||p_x)} \\
\leq &\frac{2(\max\{|\log C_1|, |\log C_2|\} + 1)}{C_0}. \qquad (6.65)
\end{aligned}$$

Since C_0, C_1, C_2 are constants and $|\mathcal{X}|$ is finite, our proof is complete by letting $L = 2|\mathcal{X}|(\max\{|\log C_1|, |\log C_2|\} + 1)/C_0$.

Proof of Lemma 26

Note that $\widehat{s}(\mathbf{w}_r) = \frac{\mathbf{w}^T\mathbf{A}\log(\mathbf{A}^T\mathbf{w})}{\mathbf{w}^T\log\mathbf{w}}$. Define $\widehat{D}(r_y||p_y) = \mathbf{w}^T\mathbf{A}\log(\mathbf{A}^T\mathbf{w})$ and $\widehat{D}(r_x||p_x) = \mathbf{w}^T\log\mathbf{w}$. We will prove that both $\widehat{D}(r_y||p_y)$ converges to $D(r_y||p_y)$ and $\widehat{D}(r_x||p_x)$ converges to $D(r_x||p_x)$ in probability. Since $D(r_x||p_x) > 0$ and $\widehat{D}(r_x||p_x) > 0$ with probability 1, we obtain that $\widehat{s}(\mathbf{w}_r)$ converges to $D(r_y||p_y)/D(r_x||p_x) = s(r_x)$ in probability.

The convergence $\widehat{D}(r_x||p_x)$ comes from law of large number. Since $\widehat{D}(r_x||p_x) = \frac{1}{n}\sum_{i=1}^n \frac{r_x(X_i)}{p_x(X_i)}\log\frac{r_x(X_i)}{p_x(X_i)}$ and $D(r_x||p_x) = \mathbb{E}_{X\sim p_x}\left[\frac{r_x(X)}{p_x(X)}\log\frac{r_x(X)}{p_x(X)}\right]$, the weak law of large number shows the convergence in probability.

For the convergence of $\widehat{D}(r_y||p_y)$. Consider the vector $\mathbf{v} = \mathbf{A}^T\mathbf{w}$, we have

$$v_j = \frac{1}{n}\sum_{i=1}^n \frac{p_{xy}(X_i, Y_j)}{p_x(X_i)p_y(Y_j)}w_i = \frac{1}{n}\sum_{i=1}^n \frac{p_{y|x}(Y_j|X_i)}{p_y(Y_j)}\frac{r_x(X_i)}{p_x(X_i)}. \qquad (6.66)$$

On the other hand, for fixed $Y_j = y$, we have

$$\frac{r_y(y)}{p_y(y)} = \frac{\mathbb{E}_{X\sim p_x}\left[p_{y|x}(y|X)\frac{r_x(X)}{p_x(X)}\right]}{p_y(y)} = \mathbb{E}_{X\sim p_x}\left[\frac{p_{y|x}(y|X)}{p_y(y)}\frac{r_x(X)}{p_x(X)}\right]. \qquad (6.67)$$

Therefore, by the law of large numbers, we conclude that v_j converges to $\frac{r_y(Y_j)}{p_y(Y_j)}$ in probability. Hence, $\widehat{D}(r_y||p_y) = \frac{1}{n}\sum_{j=1}^{n} v_j \log v_j$ converges to the limit $\frac{1}{n}\sum_{j=1}^{n} \frac{r_y(Y_j)}{p_y(Y_j)} \log \frac{r_y(Y_j)}{p_y(Y_j)}$ in probability. Furthermore, $\frac{1}{n}\sum_{j=1}^{n} \frac{r_y(Y_j)}{p_y(Y_j)} \log \frac{r_y(Y_j)}{p_y(Y_j)}$ converges to $D(r_y||p_y) = \mathbb{E}_{Y\sim p_y}\left[\frac{r_y(Y)}{p_y(Y)} \log \frac{r_y(Y)}{p_y(Y)}\right]$ in probability, by law of large number again. Therefore, we conclude that $\widehat{D}(r_y||p_y)$ converges to $D(r_y||p_y)$ in probability.

CHAPTER 7

LEARNING ONE-HIDDEN-LAYER NEURAL NETWORKS UNDER GENERAL INPUT DISTRIBUTION

Neural networks have made significant impacts over the past decade, thanks to their successful applications across multiple domains including computer vision, natural language processing, and robotics. This success partly owes to the mysterious phenomenon that (stochastic) gradient method applied to highly non-convex loss functions converges to a model parameter that achieves high test accuracy. We are in a dire need of theoretical understanding of such a phenomenon, in order to guide the design of next-generation neural networks and training methods. Significant recent progresses have been made, by asking the question: Can we efficiently learn a neural network model, when there is a ground truth neural network that generated the data?

Suppose the data (x, y) is generated by sampling x from an unknown distribution $f_X(x)$ and y is generated by passing x through an unknown neural network and adding some simple noise. Even if we train neural networks on this "teacher network", it is known to be a hard problem without further assumptions [184]. Significant effort has been on designing new approaches to learn simple neural networks (such as one-hidden-layer neural network) on data from simple distributions (such as Gaussian) [185, 14]. This is followed by analyses on increasingly more complex architectures [184, 186]. However, the analysis techniques critically depend on the Gaussian input assumption, and further the proposed algorithms are tailored specifically to Gaussian inputs. In this chapter, we provide a unified approach to design loss functions that provably learn the true model for a wide range of input distributions with smooth densities.

We consider a scenario where the data is generated from a one-hidden-layer neural network

$$y \;=\; \sum_{i=1}^{k} w_i^* \, g(\langle a_i^*, x\rangle) \;+\; \eta, \tag{7.1}$$

where the true parameters are $w_i^* \in \mathbb{R}$ and $a_i^* \in \mathbb{R}^d$, and η is a zero-mean noise independent of x, with some non-linear activation function $g : \mathbb{R} \to \mathbb{R}$. It has been widely known that first-order methods on the ℓ_2-loss get stuck in bad local minima, even for this simple one-hidden-layer neural networks [187]. If the input x is coming from a Gaussian distribution, [14] proposes a new loss function $G(\cdot)$ with a carefully designed landscape such that Stochastic Gradient Descent (SGD) provably converges to the true parameters. However, the proposed novel loss function is specifically designed for Gaussian inputs, and gets stuck at bad local minima when applied to general non-Gaussian distributions. We showcase this in Section 7.3. Designing the optimization landscape for general input distributions is a practically important and technically challenging problem, as acknowledged in [14] and many existing works in the literature [184, 185, 186].

Our goal is to strictly generalize the approach of [14] and construct a loss function $L(\cdot)$ with a good landscape such that SGD recovers the true parameters with *global* initializations. The main challenge is in estimating the *score function* defined as a functional of the probability density function $f(x)$ of the input data x:

$$\mathcal{S}_m(x) \triangleq \frac{\nabla^{(m)} f_X(x)}{f_X(x)}, \tag{7.2}$$

where $\nabla^{(m)} f_X(x)$ denotes the m-th order derivative for an $m \in \mathbb{Z}$, which plays a crucial role in the landscape design. We need to evaluate this score function at sample points, which is extremely challenging as it involves the *higher order derivatives* of a pdf that we do not know. Standard non-parametric density estimation methods such as the Kernel Density Estimators (KDE) [188] and k-Nearest Neighbor methods (k-NN) all fail to provide an estimator, as they are tailored for density estimation. Existing heuristics do not have even consistency guarantees, which include score matching based methods [189, 190], and de-noising auto-encoder (DAE) based algorithms [191].

In this chapter, we first address this fundamental question of how to estimate the score functions from samples in a principled manner. We introduce a novel approach to adaptively capture the local geometry of the pdf to design a consistent estimator for score functions. To achieve this, we bring ideas from *local likelihood* methods [7, 6] from statistics to the context of

score function estimation and also prove the convergence rate of our estimator (LLSFE), which is of independent mathematical interest. We further introduce a new loss function for training one-hidden-layer neural networks, that builds upon the estimated score functions. We show that this provably has the desired landscape for general input distributions.

Main contributions of Chapter 7:

- **Score function estimation.** In this chapter, we provide the first consistent estimator for score functions (and hence the gradients of $L(\cdot)$), which play crucial roles in several recent model parameter learning problems [189, 190, 191]. Our provably consistent estimation of score functions, LLSFE, from samples, with local geometry adaptations, is of independent mathematical interest.

- **Optimization landscape for general distributions.** For a large class of input distributions, with an appropriate score transformation for the input and appropriate tensor projection, we design a loss function $L(\cdot)$ for one-hidden-layer neural network with good landscape properties. In particular, our result is a strict generalization of [14] which was restricted to Gaussian inputs, in both mathematical and abstract viewpoints.

Related work. Several recent works have provided provable algorithms for training neural networks [192, 193, 194, 195, 196, 197]. An early work on provable learning guarantees on deep generative models for sparse weights was studied in [198]. Brutzkus and Globerson [184] analyzed one-hidden-layer neural network similar to our setting, but restricted to Gaussian input distributions and also assuming hidden variables have disjoint supports. The Gaussian assumption was relaxed in [199], but the analysis technique highly depends on the fact that the non-linear activation function is a ReLU, and the convolutional neural network only uses a single channel. Li and Yuan [186] analyzed the convergence of one-hidden layer neural network with Gaussian input when the true weights are close to identity. The optimization landscape of neural networks for some specific activation functions were studied in [200, 201, 202, 203].

Tensor methods have been used to build provable algorithms for training neural networks [204, 205]. Our work is built upon [14], which uses a fourth-order tensor based objective function and show good landscape properties. Most of the aforementioned works requires specific assumptions on the input distribution (example: Gaussian), while we only require generic smoothness of the underlying (unknown) density.

Notations. We use $\mathcal{T}(x_1, \ldots, x_m)$ to denote the inner product for an m-th order tensor $\mathcal{T}$ and vectors $x_1, \ldots, x_m$. We use $x_1 \otimes x_2 \otimes \cdots \otimes x_m$ to denote outer product of vectors/matrices/tensors. $x^{\otimes j} = x \otimes \cdots \otimes x$ denotes the j-th order tensor power of x and $x^{\otimes 0} = 1$. The spectral norm and Frobenius norm of matrix and high-order tensor are denoted by $\|\mathcal{T}\|_{\mathrm{sp}} = \max_{\|u_i\|_2 \le 1} \mathcal{T}(u_1, u_2, \ldots, u_m)$ and $\|\mathcal{T}\|_{\mathcal{F}} = \sqrt{\sum_{i_1,\ldots,i_m} (\mathcal{T}_{(i_1,\ldots,i_m)})^2}$. $\mathrm{sym}(\mathcal{T})$ denotes the symmtrify operator of a tensor $\mathcal{T}$ defined as $\mathrm{sym}(\mathcal{T})_{(i_1,\ldots,i_m)} = \frac{1}{m!} \sum_{(j_1,\ldots,j_m) \in \pi(i_1,\ldots,i_m)} \mathcal{T}_{(j_1,\ldots,j_m)}$.

7.1 Score function estimation

In this section, we introduce a new approach for estimating score functions defined in (7.2) from i.i.d. samples from a distribution. As the score functions involve higher order derivatives of the pdf, it is critical to capture the rate of *changes* in the pdf. Further, we aim to apply it to data coming from a broad range of distributions. Such sharp estimates for broad class of distributions can only be achieved by combining the strengths of two popular approaches in density estimation: simple parametric density estimators and complex non-parametric density estimators. We bridge this gap by borrowing the techniques from Local Likelihood Density Estimators (LLDE) and bring them to a new light in order to provide the first consistent score function estimators.

7.1.1 Local Likelihood Density Estimator (LLDE)

How do we estimate the normalized derivatives of the density? We address this question in a principled manner utilizing the notion *local likelihood density estimation* (LLDE) from non-parametric methods [7, 6]. LLDE is originally designed for estimating density for distributions with complicated local

geometry, and can be further applied to estimate functionals of density such as information entropy [8]. Inspired by the fact that LLDEs capture the local geometry of the pdf, we build upon the LLDE estimators to design a new estimator of the higher-order derivatives, which is the main bottleneck in score function estimation.

The local likelihood density estimator is specified by a non-negative function $K : \mathbb{R}^d \to \mathbb{R}$ (also called a Kernel function), a degree $p \in \mathbb{Z}^+$ of the polynomial approximation, and a bandwidth $h \in \mathbb{R}^+$. It is the solution of a maximization of the local log-likelihood function:

$$\mathcal{L}_x(f) \;=\; \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big) \log f(X_i) - n \int K\Big(\frac{u-x}{h}\Big) f(u) du. \quad (7.3)$$

For each x, we maximize this function over a parametric family of functions $f(\cdot)$, using the following local polynomial approximation of $\log f(x)$:

$$\begin{aligned} \log f(x) \;&=\; a_0 + a_1^T (u-x) + \frac{1}{2}(u-x)^T A_2 (u-x) \\ &\quad + \cdots + \frac{1}{p!}\mathcal{A}_p(u-x, \ldots, u-x), \end{aligned} \quad (7.4)$$

parameterized by $a = (a_0, a_1, A_2, \ldots, \mathcal{A}_p) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^{d^2} \times \cdots \times \mathbb{R}^{d^p}$. The local likelihood density estimate (LLDE) at point x is defined as $f(x) = e^{\widehat{a}_0}$, where $\widehat{a} = (\widehat{a}_0, \widehat{a}_1, \widehat{A}_2, \ldots, \widehat{\mathcal{A}}_p)$ is the maximizer around a point x: $\widehat{a} \in \arg\max_f \mathcal{L}_x(f)$. The optimization problem can be solved by setting the derivatives $\partial \mathcal{L}_x(p)/\partial \mathcal{A}_j = 0$ for $j \in \{0, \ldots, p\}$. The optimal solution $\widehat{a}$ can be obtained from solving the following equations,

$$\begin{aligned} &\int_{\mathbb{R}^d} \exp\{a_0 + a_1^T(u-x) + \cdots + \frac{1}{p!}\mathcal{A}_p(u-x)^{\otimes p}\}(\frac{u-x}{h})^{\otimes j} K(\frac{u-x}{h}) du \\ =\;& \frac{1}{n}\sum_{i=1}^{n} (\frac{X_i - x}{h})^{\otimes j} K(\frac{X_i - x}{h}). \end{aligned} \quad (7.5)$$

We build upon this idea to first introduce the score function estimator, and focus on the statistical aspect of this estimator. We discuss the computational aspect of finding the solution to this optimization in Section 7.1.4.

7.1.2 From LLDE to local likelihood score function estimator (LLSFE)

We build upon the techniques from LLDE to design our local likelihood score function estimator (LLSFE). Notice that the score function $\mathcal{S}_m(x)$ satisfies the following recursive formula from [15],

$$\mathcal{S}_m(x) \;=\; -\mathcal{S}_{m-1}(x) \otimes \nabla_x \log f(x) - \nabla_x \mathcal{S}_{m-1}(x), \tag{7.6}$$

and $\mathcal{S}_1(x) = -\nabla \log f(x)$. This recursion reveals us that the score function can be represented as a polynomial function of the gradients of log-density $g_1(x) = \nabla \log f(x)$, $G_2(x) = \nabla^{(2)} \log f(x)$ and $\mathcal{G}_m(x) = \nabla^{(m)} \log f(x)$ for $m > 2$. For example, the polynomial for $\mathcal{S}_2(x)$ and $\mathcal{S}_4(x)$ are given below:

$$\mathcal{S}_2(x) \;=\; g_1(x) \otimes g_1(x) + G_2(x), \tag{7.7}$$

$$\begin{aligned}\mathcal{S}_4(x) \;=\;& g_1(x) \otimes g_1(x) \otimes g_1(x) \otimes g_1(x) + 6\,\mathrm{sym}(G_2(x) \otimes g_1(x) \otimes g_1(x)) \\ &+ 3\,(G_2(x) \otimes G_2(x)) + 4\,\mathrm{sym}(\mathcal{G}_3(x) \otimes g_1(x)) + \mathcal{G}_4(x).\end{aligned} \tag{7.8}$$

More generally, the m-th order score function can be represented as:

$$\mathcal{S}_m(x) \;=\; \sum_{\lambda \in \Lambda_m} (-1)^m c_m(\lambda)\, \mathrm{sym}(\bigotimes_{j \in \lambda} \mathcal{G}_j), \tag{7.9}$$

where Λ_m denotes the set of partitions of integer m, and $c_m(\lambda)$ is a positive constant depends on m and the partition. As an example, $\Lambda_4 = \{\{1,1,1,1\},\{2,1,1\},\{2,2\},\{3,1\},\{4\}\}$. Given the polynomial representation of a score function, the LLSFE is given by

$$\widehat{\mathcal{S}_m^{(p)}}(x) \;\triangleq\; \sum_{\lambda \in \Lambda_m} (-1)^m c_m(\lambda)\, \mathrm{sym}(\bigotimes_{j \in \lambda} \widehat{\mathcal{A}}_j^{(p)}), \tag{7.10}$$

where $\mathcal{A}_j^{(p)}$ is the LLDE of $\mathcal{G}_j$ by p-degree polynomial approximation.

7.1.3 Convergence rate of LLSFE

As LLDE captures the local geometry of the pdf, LLSFE inherits this property and is able to consistently estimate the derivatives. This is made precise in the following theorem, where we provide an upper bound of the spectral

norm error of the estimated m-th order score function. First, we formally state our assumptions.

Assumption 4. *(a) The degree of polynomial $p \geq m$.*

(b) The gradients of log-density $\nabla^{(j)} \log f(x)$ at x exist and are bounded by $\|\nabla^{(j)} \log f(x)\|_{\text{sp}} \leq C_j$ for all $j \in [p+1]$.

(c) The non-negative kernel function K satisfies $\int_{\mathbb{R}^d} |x_i|^p K(x)dx < +\infty$ for any $i \in [d]$.

(d) Bandwidth h depends on n s.t. $h \to 0$ and $nh^{d+2m} \to \infty$ as $n \to \infty$.

The following theorem provides an upper bound on the convergence rate of the proposed score function estimator.

Theorem 15. *Under Assumption 4, the spectral norm error of the LLSFE $\widehat{\mathcal{S}_m^{(p)}}(x)$ defined in (7.10) is upper bounded by*

$$\begin{aligned} & \|\widehat{\mathcal{S}_m^{(p)}}(x) - \mathcal{S}_m(x)\|_{\text{sp}} \\ \leq \;& O(d^{m/2} h^{p+1-m}) + O_p(d^{m/2}(nh^{d+2m})^{-1/2}). \end{aligned} \tag{7.11}$$

Proof. (Sketch) Note that the estimator is derived by replacing the truth gradients of log-density $(g_1(x), G_2(x), \ldots, \mathcal{G}_m(x))$ by the estimates of gredients of log-density $(\widehat{a}_0, \widehat{a}_1, \widehat{A}_2, \ldots, \widehat{\mathcal{A}}_p)$. Since we assumes that $\|\mathcal{G}_j\|_{\text{sp}} \leq C_j$, so it suffices to upper bound the spectral norm of the error $\|\widehat{\mathcal{A}}_j^{(p)} - \mathcal{G}_j\|_{\text{sp}}$. The following lemma provides upper bounds for each entry of $\widehat{\mathcal{A}}_j^{(p)} - \mathcal{G}_j$.

Lemma 27. *[7, Theorem 1] Under Assumption 4 we have*

$$\left(\widehat{\mathcal{A}}_j^{(p)}\right)_{(i_1,\ldots,i_j)} - (\mathcal{G}_j)_{(i_1,\ldots,i_j)} \;=\; O(h^{p+1-j}) + O_p((nh^{d+2j})^{-1/2}), \tag{7.12}$$

for any $j \in \{1, \ldots, p\}$ and $i_1, \ldots, i_j \in [d]^j$.

The spectral norm of of the error $\|\widehat{\mathcal{A}}_j^{(p)} - \mathcal{G}_j\|_{\text{sp}}$ is upper bounded by the Frobenius norm. Then applying Lemma 27, we have,

$$\|\widehat{\mathcal{A}}_j^{(p)} - \mathcal{G}_j\|_{\text{sp}} \;\leq\; O(d^{j/2} h^{p+1-j}) + O_p(d^{j/2}(nh^{d+2j})^{-1/2}). \tag{7.13}$$

Substituting this result into the polynomial representations (7.9) and (7.10), we obtain the desired rate. □

Remark 5. *By setting $h = n^{-1/(2p+2+d)}$, we obtain*

$$\|\widehat{\mathcal{S}_m^{(p)}}(x) - \mathcal{S}_m(x)\|_{\text{sp}} \;\leq\; O_p(d^{m/2} n^{-(p+1-m)/(2p+2+d)}). \tag{7.14}$$

Remark 6. *It was shown in [206] that the optimal rate for estimating an entry of $\mathcal{G}_j$ is $O_p(n^{-(p+1-m)/(2p+2+d)})$. We conjecture that LLSFE is also minimax rate-optimal.*

7.1.4 Second degree LLSFE

In Section 7.1.3, we proved the convergence rate of the LLSFE. However, the computational cost of LLSFE can be large since numerical integration is needed to compute the integral in (7.5). To trade off the accuracy and computational cost, we choose Gaussian kernel $K(u) \propto \exp\{\|u\|^2/2\}$ and degree $p = 2$. This makes the integration in the LHS of (7.5) tractable and we obtain closed-form expressions for a_0, a_1 and A_2. Using ideas from [8, Proposition 1], our estimators for a_1 and A_2 are:

$$\widehat{a}_1 \;=\; (\frac{M_2}{M_0} - (\frac{M_1}{M_0})(\frac{M_1}{M_0})^T)^{-1}\frac{M_1}{M_0}, \tag{7.15}$$

$$\widehat{A}_2 \;=\; h^{-2} I_{d\times d} - (\frac{M_2}{M_0} - (\frac{M_1}{M_0})(\frac{M_1}{M_0})^T)^{-1}, \tag{7.16}$$

where $M_j = \sum_{i=1}^n (X_i - x)^{\otimes j} \exp\{-\frac{\|X_i - x\|^2}{2h^2}\}$ for $j \in \{0, 1, 2\}$.

The second degree LLSFE is derived by plugging $\widehat{a}_1$ and $\widehat{A}_2$ into (7.10). The computational complexity of second degree LLDFE is $O(n \cdot d^2)$. In the experiments below, we use this second degree estimator.

7.1.5 Synthetic simulations of LLSFE

In this experiment we validate the performance of LLSFE, for both Gaussian and non-Gaussian distributions. For Gaussian distribution, we choose $x \sim \mathcal{N}(0, I_d)$ and $d = 2$. The ground truth score functions are $\mathcal{S}_2 = xx^T - I_d$ and $\mathcal{S}_4 = x^{\otimes 4} - 6\text{sym}(x \otimes x \otimes I_d) + 3I_d \otimes I_d$. We show the spectral error $\|\widehat{\mathcal{S}}_2 - \mathcal{S}_2\|_{\text{sp}}$ versus number of sample n for estimation of $\mathcal{S}_2$, and the Frobenius error $\|\widehat{\mathcal{S}}_4 - \mathcal{S}_4\|_{\mathcal{F}}$ for estimation of $\mathcal{S}_4$ (since computing spectral norm of high-order tensor is NP-hard [207]). We plot the $\{95\%, 75\%, 50\%, 25\%, 5\%\}$

percentiles of our estimation error over $10,000$ independent trials for the estimation of $\mathcal{S}_2$ and $50,000$ independent trails for the estimation of $\mathcal{S}_4$.

We can see from Figure 7.1 that all the percentiles of the estimation error decrease as n increases. The log-log scale plot is closed to linear, and the average slope is -0.5143 for $\|\widehat{\mathcal{S}}_2 - \mathcal{S}_2\|_{\text{sp}}$ and -0.4984 for $\|\widehat{\mathcal{S}}_4 - \mathcal{S}_4\|_{\mathcal{F}}$. This suggests that LLSFE is consistent and the error decreases at a faster rate than the theoretical upper bound in Remark 5.

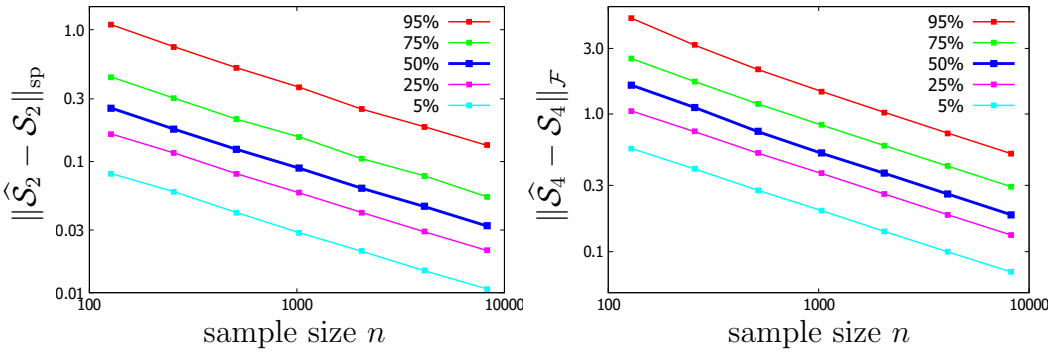

Figure 7.1: Error of score function estimator versus sample size for $x \sim \mathcal{N}(0, I_d)$. Left: $\|\widehat{\mathcal{S}}_2 - \mathcal{S}_2\|_{\text{sp}}$. Right: $\|\widehat{\mathcal{S}}_4 - \mathcal{S}_4\|_{\mathcal{F}}$.

For the non-Gaussian case, we choose $x \sim 0.5\mathcal{N}(\mathbf{1}_d, I_d) + 0.5\mathcal{N}(-\mathbf{1}_d, I_d)$ where $\mathbf{1}_d$ is the all-1 vector and $d = 2$. We also plot the percentiles of the estimation errors in log-log scale in Figure 7.2. We can see that LLSFE gives a consistent estimate for the non-Gaussian case too, and the rate is -0.2587 for $\|\widehat{\mathcal{S}}_2 - \mathcal{S}_2\|_{\text{sp}}$ and -0.1343 for $\|\widehat{\mathcal{S}}_4 - \mathcal{S}_4\|_{\mathcal{F}}$, which are also faster than the upper bound in Remark 5.

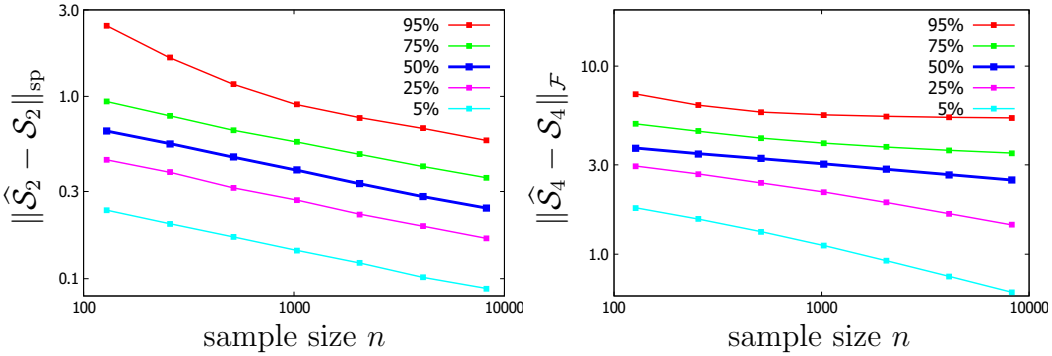

Figure 7.2: Error of score function estimator versus sample size for $x \sim 0.5\mathcal{N}(\mu, I_d) + 0.5\mathcal{N}(-\mu, I_d)$. Left: $\|\widehat{\mathcal{S}}_2 - \mathcal{S}_2\|_{\text{sp}}$. Right: $\|\widehat{\mathcal{S}}_4 - \mathcal{S}_4\|_{\mathcal{F}}$.

7.2 Design of landscape

In this section, we show how the proposed density functional estimators can be applied to design a loss function with desired properties, for regression problems under a neural network model. This gives a novel loss function that does not require the data to be distributed as Gaussian, as typically done in existing literature. Concretely, we consider the problem of training a one-hidden-layer neural network where, for each input $x \in \mathbb{R}^d$, the corresponding output is given by

$$\hat{y}(x) \;=\; \sum_{i=1}^{k} w_i \, g(\langle a_i, x \rangle), \tag{7.17}$$

with weights are $w_i \in \mathbb{R}$ and $a_i \in \mathbb{R}^d$, non-linear activation is $g : \mathbb{R} \to \mathbb{R}$, and the number of hidden neurons is $k \leq d$. Given labeled training data (x, y) coming from some distribution, a standard approach to training such a network is to use the ℓ_2 loss:

$$\ell_2(A) \;=\; \mathbb{E}\left[\|\widehat{y}(x) - y\|^2\right], \tag{7.18}$$

as the training objective, where A denotes the weights of the neural network model. However, traditional optimization techniques on ℓ_2 can easily get stuck in local optima as empirically shown in [187]. This phenomenon can be explained precisely under a canonical scenario where the data is generated from a "teacher neural network":

$$y \;=\; \sum_{i=1}^{k} w_i^* \, g(\langle a_i^*, x \rangle) + \eta, \tag{7.19}$$

where the true parameters are $w_i^* \in \mathbb{R}$ and $a_i^* \in \mathbb{R}^d$, and η is a zero-mean noise independent of x. This assumption that the data also comes from a one-hidden-layer neural network is critical in recent mathematical understanding of neural networks, in showing the gain of a shallow ResNet by [186], various properties of the critical points by [185], and showing that the standard ℓ_2 minimization is prone to get stuck at non-optimal critical points by [14]. A major limitation of this line of research is that they rely critically on the Gaussian assumption on the data x. The analysis techniques

use specific properties of spherical Gaussian random variables such that the theoretical findings do not generalize to any other distributions. Further, the estimators designed as per those analyses fail to give consistent estimates for non-Gaussian data.

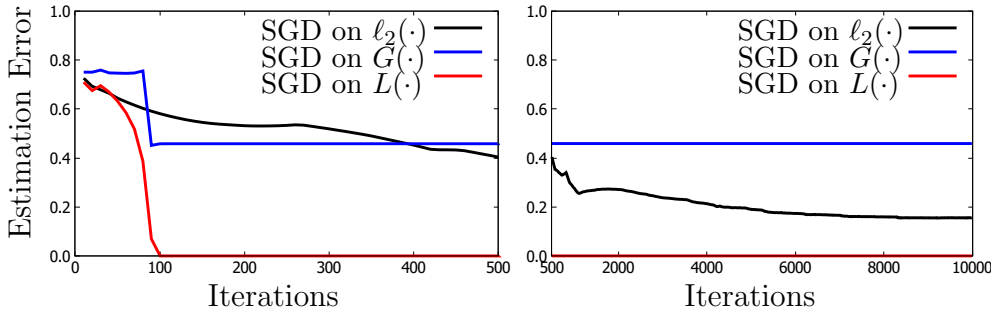

Figure 7.3: SGD to learn a one-layer-ReLU network in (7.19) on the proposed objective function $L(A)$ defined in (7.20) converges to a global minimum with random initialization, whereas on ℓ_2-loss $\ell_2(A)$ and $G(A)$, it gets stuck at bad local minima. Left: First 500 iterations. Right: 500-10,000 iterations.

We showcase this limitation in Figure 7.3, where the data is generated from a Laplacian distribution. The details of this experiment is provided in Section 7.3.1. Minimizing ℓ_2 loss converges slowly and gets stuck at sub-optimal critical points, consistent with previous observations [186]. To overcome this weakness [14] proposed applying Stochastic Gradient Descent (SGD) on a novel loss function $G(A)$ designed from the analysis under the Gaussian assumption. This fails to converge to an optimal critical point for non-Gaussian distributions. To overcome this limitation, we propose a novel loss function $L(A)$ that generalizes to a broad class of distributions.

We focus on the task of recovering the weights a_i^*'s, and denote the set by a matrix $A^\top = [a_1|\dots|a_k] \in \mathbb{R}^{d\times k}$. The scalar weights w_i^*'s can be separately estimated using standard least squares, once A has been recovered. We propose applying SGD on a new loss function $L(A)$, defined as

$$\begin{aligned} L(A) \;=\; & \sum_{i,j\in[k],i\neq j} \mathbb{E}[y\cdot t_1(x,a_i,a_j)] - \mu \sum_{i\in[k]} \mathbb{E}[y\cdot t_2(x,a_i)] \\ & +\lambda \sum_{i\in[k]} (\|a_i\|-1)^2, \end{aligned} \tag{7.20}$$

where $\mu, \lambda > 0$ are regularization coefficients, and

$$\begin{aligned} t_1(x,u,v) &= \mathcal{S}_4(x)(u,u,v,v), \\ t_2(x,u) &= \mathcal{S}_4(x)(u,u,u,u),\ u,v \in \mathbb{R}^d, \end{aligned} \tag{7.21}$$

are the applications of the score functions $\mathcal{S}_m(x) = \nabla^{(m)} f(x)/f(x)$ on the weight vectors a_i's that we are optimizing over,

$$\mathcal{S}_4(x)(u,v,w,z) = \frac{1}{f(x)} \sum_{i_1,i_2,i_3,i_4} \nabla_{x_{i_1}x_{i_2}x_{i_3}x_{i_4}} f(x) u_{i_1} v_{i_2} w_{i_3} z_{i_4}. \tag{7.22}$$

We provide formulas for some simple distributions below.

Example 4 (Gaussian)**.** *If $x \sim \mathcal{N}(0, I_d)$, then $t_1^{(G)}(x,u,v) = (u^\top x)^2(v^\top x)^2 - \|u\|^2(v^\top x)^2 - 4(u^\top x)(v^\top x)(u^\top v) - \|v\|^2(u^\top x)^2 + \|u\|^2\|v\|^2 + 2(u^\top v)^2$ and $t_2^{(G)}(x,u) = (u^\top x)^4 - 6\|u\|^2(u^\top x)^2 + 3\|u\|^4$.*

Example 5 (Mixture of Gaussians)**.** *If $x \sim p\mathcal{N}(\mu_1, I_d) + (1-p)\mathcal{N}(\mu_2, I_d)$, we have that $t_1(x,u,v) = p_1 t_1^{(G)}(x-\mu_1,u,v) + (1-p_1)t_1^{(G)}(x-\mu_2,u,v)$ where the posterior $p_1 \triangleq \frac{p\mathcal{N}(\mu,I_d)}{p\mathcal{N}(\mu_1,I_d)+(1-p)\mathcal{N}(\mu_2,I_d)}$. Similarly for t_2.*

The proposed $L(\cdot)$ is carefully designed to ensure that the loss surface has a desired landscape with no local minima. Here, we give the intuition behind the design principle, and make it precise in the main results of Theorems 16 and 17. This landscape explains the experimental superiority of $L(\cdot)$ in Figures 7.3 and 7.4. Suppose $k = d$ and a_i^*'s are orthogonal vectors. After some calculus, an alternative characterization for L is given by

$$\begin{aligned} L(A) &= \sum_{i\in[d]} w_i^* \mathbb{E}[g^{(4)}(\langle a_i^*, x\rangle)] \sum_{j,k\in[d],j\neq k} \langle a_i^*, a_j\rangle^2 \langle a_i^*, a_k\rangle^2 \\ &\quad -\mu \sum_{i,j\in[d]} w_i^* \mathbb{E}[g^{(4)}(\langle a_i^*, x\rangle)] \langle a_i^*, a_j\rangle^4 + \lambda \sum_{i\in[d]} (\|a_i\| - 1)^2 \\ &= \sum_{i\in[d]} \kappa_i^* \sum_{j,k\in[d],j\neq k} \langle a_i^*, a_j\rangle^2 \langle a_i^*, a_k\rangle^2 - \mu \sum_{i,j\in[d]} \kappa_i^* \langle a_i^*, a_j\rangle^4 \\ &\quad +\lambda \sum_{i\in[d]} (\|a_i\|^2 - 1)^2, \end{aligned} \tag{7.23}$$

for scalar $\kappa_i^* = w_i^* \mathbb{E}[g^{(4)}(\langle a_i^*, x\rangle)]$ that does not depend on the variables we optimize over.

Notice that when the weights are recovered up to a permutation, that is $a_i = \pm a^*_{\pi(i)}$ for some permutation π, the first term in (7.23) equals zero. We can show that these are the only possible local minima in the minimization of the first term under unit-norm constraints, whenever all $\kappa^*_i = 1$. Thus in order to account for this weighted tensor based loss and to avoid spurious local minima, the regularization term $\mu \sum_{i,j\in[d]} \kappa^*_i \langle a^*_i, a_j\rangle^4$ forces these spurious minima to lie close to a permutation of a^*_i up to a sign flip. This is made precise in the characterization of the landscape of $L(\cdot)$ in the proof of Theorem 16. The proof strategy is inspired by the landscape analysis technique of [14], where a similar analysis was done for Gaussian data x.

7.2.1 Theoretical results

We now formally state the assumptions for our theoretical results.

Assumption 5. *(a) The ground-truth parameters w^*_i and a^*_i are such that $w^*_i \mathbb{E}[g^{(4)}(\langle a^*_i, x\rangle)]$ has the same sign for all $i \in [k]$.*

(b) Defining $\kappa^*_i = w^*_i \mathbb{E}[g^{(4)}(\langle a^*_i, x\rangle)]$ and $\kappa^* = \max_i \kappa^*_i / (\min_i \kappa^*_i)$, we choose $\mu < c/\kappa^*$ and $\lambda \geq \kappa^*_{\max}/c$ for $c \leq 0.01$.

(c) $k = d$ and $A \in \mathbb{R}^{d\times d}$ is an orthogonal matrix.

The following theorem characterizes the landscape of $L(\cdot)$.

Theorem 16. *Under Assumption 5, the objective function $L(\cdot)$ satisfies that*

1. *All local minima of L are also global. Furthermore, all approximate local minima are also close to the global minimum. More concretely, for $\varepsilon > 0$, let A satisfy that*

$$\|\nabla L(A)\| \leq \varepsilon \text{ and } \lambda_{\min}\left\{\nabla^2 L(A)\right\} \geq -\tau, \tag{7.24}$$

where $\tau = c\min\{\mu\kappa^*_{\min}/(\kappa^* d), \lambda\}$. Then $A = PDA^* + EA^*$, where P is a permutation matrix, D is a diagonal matrix with $D_{ii} \in \{\pm 1 \pm O(\mu\kappa^*_{\max}/\lambda)\}$, and $|E|_\infty \leq O\{\varepsilon/(\kappa^*_{\min})\}$.

2. *Any saddle point A has a strictly negative curvature, i.e., $\lambda_{\min}(\nabla^2 L(A)) \leq -\tau$.*

Remark 7. *For the case when $a_1, \ldots, a_k$ are linearly independent with $k < d$, similar conclusion hold (see Section 7.4.4).*

7.2.2 Finite sample regime

In the finite sample regime, we replace the population expectation in (7.20) with empirical expectation $\hat{\mathbb{E}}$ and optimize on the corresponding loss $\hat{L}$. The following theorem establishes that $\hat{L}$ also exhibits similar landscape properties as that of L (under some mild technical assumptions outlined in Assumption 6 in Section 7.4.3).

Theorem 17. *Assume that Assumption 5 and Assumption 6 (defined in Section 7.4.3) hold. Then there exists a polynomial* $\text{poly}(d, 1/\varepsilon)$ *such that whenever* $n \geq \text{poly}(d, 1/\varepsilon)$*, with high probability,* $\hat{L}$ *exhibits the same landscape properties as that of* L*, established in Theorem 16.*

A major bottleneck in applying the proposed loss (7.20) directly to real data is that the knowledge of the probability density function of the data x is required. As we saw in the Examples 4 and 5, the loss function t_1 and t_2 depends on the pdf of x. In Section 7.3, we show how we can combine the LLSFE to compute (the gradients of) those functions to introduce a novel consistent estimator with a desirable landscape.

7.3 Experiments of Chapter 7

7.3.1 Landscape of $L(\cdot)$

In this simulation, we show that the landscape of the loss function $L(A)$ is well-behaved, if we know the score function $\mathcal{S}_4(x)$. We choose $x = (x_1, \ldots, x_d)$, where x_i are i.i.d. symmetric exponential distributed random variables, i.e., $f(x_i) = (1/2)\exp\{-|x_i|\}$. The fourth-order score function is given by $\mathcal{S}_4(x) = \text{sgn}(x)^{\otimes 4}$. We compare our loss function $L(A)$ with an ℓ_2-loss, $\ell(\cdot)$, as well as the loss function $G(\cdot)$ proposed in [14], and evaluate the performance through the parameter error (which verifies if $A^{*-1}A$ is close to a

permutation matrix)

$$e(A) = \min\{1 - \min_i \max_j |(A^{*-1}A)_{ij}|, 1 - \min_j \max_i |(A^{*-1}A)_{ij}|\}. \quad (7.25)$$

For the experiment, we choose $A^* = I_d$, $w^* = 1$, $\sigma = \text{ReLU}$, $k = d = 50$ and use full-batch gradient descent with sample size 8192 and learning rate $\eta = 5 \times 10^{-3}$ for ℓ_2 loss and $\eta = 5 \times 10^{-5}$ for $L(A)$ and $G(A)$. Regularization parameter is $\mu = 30$ for both $L(A)$ and $G(A)$. The results are illustrated in Figure 7.3, which shows that (i) $\ell_2(\cdot)$ converges slowly and to a suboptimal critical point indicating the existence of local minima; (ii) $G(\cdot)$ converges to a suboptimal critical point due to the mismatched Gaussian assumption; and (iii) $L(\cdot)$ converges to a global minima.

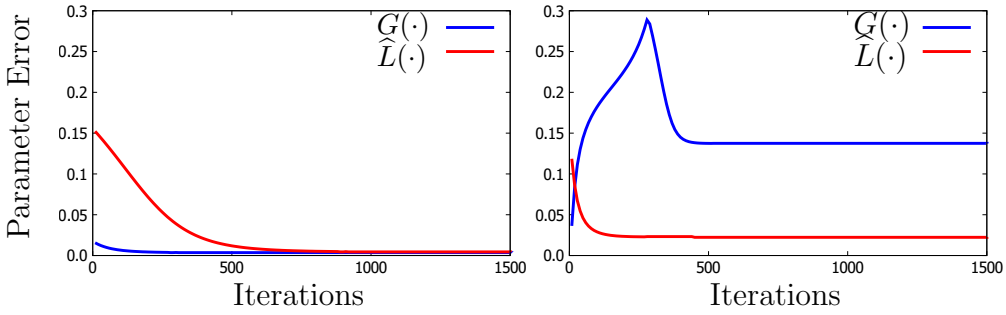

Figure 7.4: Learning curve of objective function $G(A)$ (blue line) and LLSFE based objective function $L(A)$ (7.20) (red line). Left: x is Gaussian. Right: x is Gaussian-mixture.

7.3.2 Combine with LLSFE

Now we use our estimator LLSFE to construct the empirical loss $\widehat{L}(A)$ to train a one-hidden-layer neural network (7.19). The setting of this experiment is same as that of Section 7.3.1 with $k = d = 2$ for simplicity.

In the left panel of Figure. 7.4, we choose Gaussian input $x \sim \mathcal{N}(0, I_d)$ so that the loss $G(A)$ coincides with $L(A)$ if the ground truth $\mathcal{S}_4(x)$ is known. We can see that using estimation error using $L(\cdot)$ operates close to that of the ground-truth $G(\cdot)$ In the right panel of Figure 7.4, we choose $x \sim 0.5\mathcal{N}(\mathbf{1}_d, I_d) + 0.5\mathcal{N}(-\mathbf{1}_d, I_d)$. In this case, $G(A)$ converges to a local minimum, thus incurring higher parameter error, whereas LLSFE-based objective function converges to the global minima very quickly. This confirms

that when the data is not coming from a Gaussian distribution, it is critical to use properly matched estimator, which is provided by the proposed LLSFE approach.

7.4 Proofs of results in Chapter 7

7.4.1 Proof of Theorem 15

Proof. We rewrite the spectral norm error in terms of the polynomial representations (7.9) and (7.10) as

$$\begin{aligned}
&\|\widehat{\mathcal{S}_m}(x) - \mathcal{S}_m(x)\|_{\text{sp}} \leq \sum_{\lambda\in\Lambda_m} c_m(\lambda)\, \|\text{sym}(\bigotimes_{j\in\lambda} \widehat{\mathcal{A}_j^{(p)}}) - \text{sym}(\bigotimes_{j\in\lambda} \mathcal{G}_j)\|_{\text{sp}} \\
\leq\ & \sum_{\lambda\in\Lambda_m} c_m(\lambda)\, \|\bigotimes_{j\in\lambda} \widehat{\mathcal{A}_j^{(p)}} - \bigotimes_{j\in\lambda} \mathcal{G}_j\|_{\text{sp}}, \qquad (7.26)
\end{aligned}$$

where the last inequality comes from the fact that $\|\text{sym}(\mathcal{T})\|_{\text{sp}} \leq \|\mathcal{T}\|_{\text{sp}}$. Then we study each term in (7.26). For simplicity of notation, denote the estimation error $\mathcal{E}_j^{(p)} \triangleq \widehat{\mathcal{A}_j^{(p)}} - \mathcal{G}_j$, then we have

$$\begin{aligned}
&\|\bigotimes_{j\in\lambda} \widehat{\mathcal{A}_j^{(p)}} - \bigotimes_{j\in\lambda} \mathcal{G}_j\|_{\text{sp}} = \|\bigotimes_{j\in\lambda} (\mathcal{E}_j^{(p)} + \mathcal{G}_j) - \bigotimes_{j\in\lambda} \mathcal{G}_j\|_{\text{sp}} \\
=\ & \|\sum_{\nu\subset\lambda} \left(\left(\bigotimes_{j\in\nu} (\mathcal{E}_j^{(p)}) \right) \otimes \left(\bigotimes_{j\in\lambda\setminus\nu} \mathcal{G}_j \right) \right) - \bigotimes_{j\in\lambda} \mathcal{G}_j\|_{\text{sp}} \\
=\ & \|\sum_{\nu\subset\lambda,\nu\neq\emptyset} \left(\left(\bigotimes_{j\in\nu} (\mathcal{E}_j^{(p)}) \right) \otimes \left(\bigotimes_{j\in\lambda\setminus\nu} \mathcal{G}_j \right) \right) \|_{\text{sp}} \\
\leq\ & \sum_{\nu\subset\lambda,\nu\neq\emptyset} \|\left(\bigotimes_{j\in\nu} (\mathcal{E}_j^{(p)}) \right) \otimes \left(\bigotimes_{j\in\lambda\setminus\nu} \mathcal{G}_j \right) \|_{\text{sp}} \\
\leq\ & \sum_{\nu\subset\lambda,\nu\neq\emptyset} \left((\prod_{j\in\nu} \|\mathcal{E}_j^{(p)}\|_{\text{sp}}) \times (\prod_{j\in\lambda\setminus\nu} \|\mathcal{G}_j\|_{\text{sp}}) \right). \qquad (7.27)
\end{aligned}$$

Now we study the spectral norm of $\mathcal{E}_j^{(p)}$, which can be upper bounded by

the Frobenius norm. Then by Lemma 27, we have,

$$\begin{aligned} \|\mathcal{E}_j^{(p)}\|_{\text{sp}} \leq \|\mathcal{E}_j^{(p)}\|_{\mathcal{F}} &= \sqrt{\sum_{i_1,\ldots,i_j} \left(\mathcal{E}_j^{(p)}\right)^2_{(i_1,\ldots,i_j)}} \\ &= O(d^{j/2}h^{p+1-j}) + O_p(d^{j/2}(nh^{d+2j})^{-1/2}). \end{aligned} \tag{7.28}$$

Since for any $j \leq m$, we have $h^{p+1-j} \to 0$ and $nh^{d+2j} \to \infty$ as $n \to \infty$. So for sufficiently large n, we have $\sum_{j\in\lambda} \|\mathcal{E}_j^{(p)}\|_{\text{sp}} \leq 1$ with high probability. Then, plug it into (7.27), we get

$$\begin{aligned} \|\bigotimes_{j\in\lambda} \widehat{\mathcal{A}_j^{(p)}} - \bigotimes_{j\in\lambda} \mathcal{G}_j\|_{\text{sp}} &\leq \sum_{\nu\subset\lambda,\nu\neq\emptyset} \left((\prod_{j\in\nu} \|\mathcal{E}_j^{(p)}\|_{\text{sp}}) \times \prod_{j\in\lambda\setminus\nu} C_j \right) \\ \leq\ & C \sum_{\nu\subset\lambda,\nu\neq\emptyset} \prod_{j\in\nu} \|\mathcal{E}_j^{(p)}\|_{\text{sp}} = C\left(\prod_{j\in\lambda}(1+\|\mathcal{E}_j^{(p)}\|_{\text{sp}}) - 1\right) \\ \leq\ & C\left(\exp\{\sum_{j\in\lambda} \|\mathcal{E}_j^{(p)}\|_{\text{sp}}\} - 1\right) \leq 2C \sum_{j\in\lambda} \|\mathcal{E}_j^{(p)}\|_{\text{sp}} \\ =\ & O(d^{j_{\max}/2}h^{p+1-j_{\max}}) + O_p(d^{j_{\max}/2}(nh^{d+2j_{\max}})^{-1/2}), \end{aligned} \tag{7.29}$$

here constant $C = \max_\nu \prod_{j\in\lambda\setminus\nu} C_j$ and $j_{\max} = \max\{j : j \in \lambda\}$. The last inequality comes from the fact that $e^y - 1 \leq 2y$ for any $y \leq 1$. Since λ is a partition of integer m, we have $j_{\max} \leq m$, and the equation holds if and only if $\lambda = \{m\}$. Therefore the only term in (7.26) that achieves $O(d^{m/2}h^{p+1-m}) + O_p(d^{m/2}(nh^{d+2m})^{-1/2})$ is $\|\widehat{\mathcal{A}}_m^{(p)} - \mathcal{G}_m\|_{\text{sp}}$, with $c_m(\lambda) = 1$. Therefore, we complete the proof. □

7.4.2 Proof of Theorem 16

The key technical lemma behind our results is the Stein's lemma and its generalizations which we present below.

Lemma 28 ([208]). *Let $x \sim \mathcal{N}(0, I_d)$ and $g : \mathbb{R}^d \to \mathbb{R}$ be such that both $\mathbb{E}[\nabla g(x)]$ and $\mathbb{E}[g(x)x]$ exist and are finite. Then*

$$\mathbb{E}[g(x)x] \ = \ \mathbb{E}[\nabla_x g(x)]. \tag{7.30}$$

The following lemma generalizes Stein's lemma to more general distribu-

tions and higher-order derivatives.

Lemma 29 ([209]). *Let $m \geq 1$ and $\mathcal{S}_m(x)$ be defined as in (7.2). Then for any $g : \mathbb{R}^d \to \mathbb{R}$ satisfying some regularity conditions, we have*

$$\mathbb{E}[g(x) \cdot \mathcal{S}_m(x)] \;=\; \mathbb{E}[\nabla_x^{(m)} g(x)]. \tag{7.31}$$

The following theorem gives an alternate characterization of the loss function L and is the key step in the proof of Theorem 16.

Theorem 18. *The loss function $L(\cdot)$ defined in (7.20) satisfies that*

$$\begin{aligned} L(A) = &\sum_{i\in[d]} w_i^* \mathbb{E}[g^{(4)}(\langle a_i^*, x\rangle)] \sum_{j,k\in[d], j\neq k} \langle a_i^*, a_j\rangle^2 \langle a_i^*, a_k\rangle^2 \\ &-\mu \sum_{i,j\in[d]} w_i^* \mathbb{E}[g^{(4)}(\langle a_i^*, x\rangle)] \langle a_i^*, a_j\rangle^4 + \lambda \sum_{i\in[d]} (\|a_i\| - 1)^2 \,. \end{aligned} \tag{7.32}$$

Proof. Since η is zero-mean and independent of x, we have that

$$\mathbb{E}[y \cdot \mathcal{S}_m(x)] \;=\; \sum_{i\in[k]} w_i^* \mathbb{E}[g(\langle a_i^*, x\rangle \cdot \mathcal{S}_m(x))] \,, \tag{7.33}$$

Putting $m = 4$ in Lemma 29, in view of (7.33), we obtain that

$$\mathbb{E}[y \cdot \mathcal{S}_4(x)] \;=\; \sum_{i\in[k]} w_i^* \mathbb{E}[g^{(4)}(\langle a_i^*, x\rangle)] (a_i^*)^{\otimes 4}. \tag{7.34}$$

Thus for any fixed a_j, a_k, we have

$$\begin{aligned} &\mathbb{E}[y \cdot \mathcal{S}_4(x)(a_j, a_j, a_k, a_k)] = \mathbb{E}[y \cdot t_1(x)] \\ = &\sum_{i\in[k]} w_i^* \mathbb{E}[g^{(4)}(\langle a_i^*, x\rangle)] \langle a_i^*, a_j\rangle^2 \langle a_i^*, a_k\rangle^2, \end{aligned} \tag{7.35}$$

$$\begin{aligned} &\mathbb{E}[y \cdot \mathcal{S}_4(x)(a_j, a_j, a_j, a_j)] = \mathbb{E}[y \cdot t_2(x)] \\ = &\sum_{i\in[k]} w_i^* \mathbb{E}[g^{(4)}(\langle a_i^*, x\rangle)] \langle a_i^*, a_j\rangle^4. \end{aligned} \tag{7.36}$$

Now summing over j, k finishes the proof. □

The proof of Theorem 16 follows from Theorem 2.3 of [14] and Theorem 18.

7.4.3 Proof of Theorem 17

We state the assumptions for finite sample landscape analysis below.

Assumption 6. *(a) $\|x\|$ has exponentially decaying tails,*

$$\Pr\left[\|x\|^2 \geq t\right] \;\leq\; K_1 e^{-K_2 t^2}, \quad \forall t \geq 0, \tag{7.37}$$

for some constants $K_1, K_2 > 0$.

(b) Let $l(x, y, A)$ be such that $L(A) = \mathbb{E}[l(x, y, A)] + \lambda \sum_{i\in[k]} (\|a_i\|^2 - 1)^2$. Then there exists a constant $K > 0$ which is at most a polynomial in d and a constant $p \in \mathbb{N}$ such that

$$\begin{aligned} \|\nabla_A l(x, y, A)\| &\leq K\|x\|^p, \\ \|\nabla_A^2 l(x, y, A)\| &\leq K\|x\|^p, \end{aligned} \tag{7.38}$$

for all A such that $\|a_i\| \leq 2$.

In order to establish that the gradient and the Hessian of L are close to their finite sample counterparts, we first consider its truncated version L_T defined as

$$L_T \triangleq \mathbb{E}[l(x, y, A)\mathbb{I}_E], \qquad E \triangleq \{\|x\| \leq R\}, \tag{7.39}$$

where $R = Cd\log(1/\varepsilon)$ for some $\varepsilon < 0$. It follows that L_T is well behaved and exhibits uniform convergence of empirical gradients/Hessians to its population version [14] for A with bounded norm. Then Theorem 17 follows from showing that the gradient and the Hessian of L_T are close to that of L as well in this setting, which we prove in Lemma 30. Next we combine this result with Lemma E.5 of [14] which shows that A with large row norms must also have large gradients and hence cannot be local minima. First we define L_T.

Lemma 30. *Let L_T be defined as in* (7.39) *and 6 hold. Then for a sufficiently large constant C and a sufficiently small $\varepsilon > 0$, we have that*

$$\|\nabla L(A) - \nabla L_T(A)\|_2 \;\leq\; \varepsilon, \tag{7.40}$$

$$\|\nabla^2 L(A) - \nabla^2 L_T(A)\|_2 \;\leq\; \varepsilon, \tag{7.41}$$

for all A with row norm $\|A_i\| \leq 2$.

Proof. We have that

$$\begin{aligned}
& \|\nabla L(A) - \nabla L_T(A)\|_2 = \|\mathbb{E}\left[\nabla l(x,y,A)(1-\mathbb{I}_E)\right]\| \\
\overset{(a)}{\leq} & \ \mathbb{E}\left[\|\nabla l(x,y,A)\|\ \mathbb{I}\{\|x\| \geq R\}\right] \\
= & \ \sum_{i\geq 0} \mathbb{E}\left[\|\nabla l(x,y,A)\|\ \mathbb{I}\{\|x\| \in \left[2^i R, 2^{i+1}R\right]\}\right] \\
\overset{(b)}{\leq} & \ \sum_{i\geq 0} K(2^{i+1}R)^p\ \Pr\left[\|x\| \geq 2^i R\right] \leq \sum_{i\geq 0} K(2^{i+1}R)^p e^{-2^i R} \\
\overset{(c)}{\leq} & \ \sum_{i\geq 0} e^{-2^{i-1}R} = \sum_{i\geq 0} \varepsilon^{Cd2^{i-1}} \\
\overset{(d)}{\leq} & \ \sum_{i\geq 0} \varepsilon/2^{i+1} = \varepsilon,
\end{aligned} \tag{7.42}$$

where (a) follows from the Jensen's inequality, (b) follows from 6, (c) follows from the fact that $K(2x)^p e^{-x} \leq e^{-x/2}$ for x sufficiently large, and (d) follows from choosing C sufficiently large. Similarly for $\|\nabla^2 L(A) - \nabla^2 L_T(A)\|_2$. □

We are now ready to prove Theorem 17.

Proof. Let A be such that norms of all the rows are less than 2. Then we have from Lemma 30 that

$$\|\nabla L(A) - \nabla L_T(A)\|_2 \leq \varepsilon/4, \tag{7.43}$$

$$\|\nabla^2 L(A) - \nabla^2 L_T(A)\|_2 \leq \tau_0/4. \tag{7.44}$$

Notice that the gradient and Hessian of $l(x,y,A)\mathbb{I}_E$ are upper bounded $\tau = \text{poly}(d, 1/\varepsilon)$ for some fixed polynomial poly. Hence using the uniform convergence of the sample gradients/Hessians to their population counterparts [14, Theorem E.3], we have that

$$\|\nabla L_T(A) - \nabla \widehat{L}_T(A)\|_2 \leq \varepsilon/6, \tag{7.45}$$

$$\|\nabla^2 L_T(A) - \nabla^2 \widehat{L}_T(A)\|_2 \leq \tau_0/6, \tag{7.46}$$

whenever $N \geq \text{poly}(d, 1/\varepsilon)$, with high probability. Moreover, from standard

concentration inequalities (such as multivariate Chebyshev) it follows that

$$\|\nabla \widehat{L}(A) - \nabla \widehat{L}_T(A) - (\nabla L(A) - \nabla L_T(A))\|_2 \leq \varepsilon/6, \quad (7.47)$$
$$\|\nabla^2 \widehat{L}(A) - \nabla^2 \widehat{L}_T(A) - (\nabla^2 L(A) - \nabla^2 L_T(A))\|_2 \leq \tau_0/6, \quad (7.48)$$

with high probability, whenever $N \geq \text{poly}(d, 1/\varepsilon)$. Hence, we obtain that

$$\|\nabla L(A) - \nabla \widehat{L}(A)\|_2 \leq \varepsilon/2, \quad (7.49)$$
$$\|\nabla^2 L(A) - \nabla^2 \widehat{L}(A)\|_2 \leq \tau_0/2. \quad (7.50)$$

If A is such that there exists a row A_i with $\|a_i\| \geq 2$, we have from [14, Lemma E.5] that $\langle \nabla \widehat{L}(A), A_i \rangle \geq c\lambda \|a_i\|^4$ for a small constant c and thus A cannot be a local minimum for $\widehat{L}$. Hence all local minima of $\widehat{L}$ must have $\|a_i\| \leq 2$ and thus in view of (7.50) it follows that it also a ε-approximate local minima of L, or more concretely,

$$\|\nabla L(A)\| \leq \varepsilon, \quad \nabla^2 L(A) \succcurlyeq -\tau_0 I_d. \quad (7.51)$$

□

7.4.4 Landscape design for $k < d$

In the setting where $k = d$ and the regressors $a_1^*, \ldots, a_d^*$ are linearly independent, our loss functions $L_4(\cdot)$ can modified in a straightforward manner to arrive at the loss function $F(\cdot)$ defined in Appendix C.2 of [14]. Hence we have the same landscape properties as that of Theorem B.1 of [14]. The proof is exactly similar to that of our Theorem 16.

In a more geneal scenario where $k < d$ and the regressors $a_1^*, \ldots, a_d^*$ are linearly independent, it turns out that our loss function $L_4(\cdot)$ can also be transformed to obtain the loss $\mathcal{F}(\cdot)$ in Appendix C.3 of [14] to arrive at Theorem C.1 of [14] in our setting. The proof is again similar.

CHAPTER 8

RATE DISTORTION FOR MODEL COMPRESSION: FROM THEORY TO PRACTICE

Deep neural networks have been successful, for example, in the application of computer vision [13], machine translation [210] and game playing [211]. With increasing data and computational power, the number of weights in practical neural network model also grows rapidly. For example, in the application of image recognition, the LeNet-5 model [212] only has 400K weights. After two decades, AlexNet [13] has more than 60M weights, VGG-16 net [213] has more than 130M weights and BERT [214] has more than 340M weights. The huge size of neural networks brings many challenges, including large storage, difficulty in training, and large energy consumption. In particular, deploying such extreme models to embedded mobile systems is not feasible.

Several approaches have been proposed to reduce the size of large neural networks while preserving the performance as much as possible. Most of those approaches fall into one of the two broad categories. The first category designs novel network structures with small number of parameters, such as SqueezeNet [215] and MobileNet [216]. The other category directly compresses a given large neural network using pruning, quantization, and matrix factorization, including [217, 218, 18, 17, 219]. There are also advanced methods to train the neural network using Bayesian methods to help pruning or quantization at a later stage, such as [220, 221, 222].

As more and more model compression algorithms are proposed and compression ratio becomes larger and larger, it motivates us to think about the fundamental question — How well can we do for model compression? The goal of model compression is to trade off the *number of bits* used to describe the model parameters, and the *distortion* between the compressed model and original model. We wonder: *At least* how many bits is needed to achieve certain distortion? Despite many successful model compression algorithms, these theoretical questions still remain unclear.

In this chapter, we fill in this gap by bringing tools from rate distor-

tion theory to identify the fundamental limit on how much a model can be compressed. Specifically, we focus on compression of a pretrained model, rather than designing new structures or retraining models. Our approach builds upon rate distortion theory introduced by [19] and further developed by [223]. The approach also connects to modeling neural networks as random variables in [224], which has many practical usages [225].

Our contribution for model compression is twofold: theoretical and practical. We first apply theoretical tools from rate distortion theory to provide a lower bound on the fundamental trade off between *rate* (number of bits to describe the model) and *distortion* between compressed and original models, and prove the tightness of the lower bound for a linear model. This analysis seamlessly incorporate the structure of the neural network architecture into model compression via backpropagation. Motivated by the theory, we design an improved objective for compression algorithms and show that the improved objective gives optimal pruning and quantization algorithm for one-hidden-layer ReLU neural network, and has better performance in real neural networks as well.

Main contributions of Chapter 8:

- In Section 8.1, we briefly review previous work on model compression.
- In Section 8.2, we introduce the background of the rate distortion theory for data compression, and formally state the rate distortion theory for model compression.
- In Section 8.3, we give a lower bound of the rate distortion function, which quantifies the fundamental limit for model compression. We then prove that the lower bound is achievable for linear model.
- In Section 8.4, motivated by the achievable compressor for linear model, we proposed an improved objective for model compression, which takes consideration of the sturcture of the neural network. We then prove that the improved objective gives optimal compressor for one-hidden-layer ReLU neural network.
- In Section 8.5, we demonstrate the empirical performance of the proposed objective on fully connected neural networks on MNIST dataset and convolutional networks on CIFAR dataset.

8.1 Related work on model compression

The study of model compression of neural networks appeared as long as neural network was invented. Here we mainly discuss the literature on directly compressing large models, which are more relevant to our work. They usually contain three types of methods — pruning, quantization and matrix factorization.

Pruning methods set unimportant weights to zero to reduce the number of parameters. Early works of model pruning includes biased weight decay [226], optimal brain damage [217] and optimal brain surgeon [218]. Early methods utilize the Hessian matrix of the loss function to prune the weights, however, Hessian matrix is inefficient to compute for modern large neural networks with millions of parameters. More recently, [18] proposed an iterative pruning and retraining algorithm that works for large neural networks.

Quantization, or weight sharing methods group the weights into clusters and use one value to represent the weights in the same group. This category includes fixed-point quantization by [227], vector quantization by [228], HashedNets by [229], Hessian-weighted quantizaiton by [230] and Diameter-regularized Hessian-weighted quantization by [231].

Matrix factorization assumes the weight matrix in each layer could be factored as a low rank matrix plus a sparse matrix. Hence, storing low rank and sparse matrices is cheaper than storing the whole matrix. This category includes [232] and [219].

There are some recent advanced method beyond pruning, quantization and matrix factorization. [17] assembles pruning, quantization and Huffman coding to achieve better compression rate. Bayesian methods [220, 221, 222] are also used to retrain the model such that the model has more space to be compressed. [233] uses reinforcement learning to design a compression algorithm.

Despite these aforementioned works for model compression, no one has studied the fundamental limit of model compression, as far as we know. More specifically, in this paper, we focus on the study of theory of model compression for pretrained neural network models and then derive practical compression algorithms given the proposed theory.

8.2 Rate distortion theory for model compression

8.2.1 Review of rate distortion theory for data compression

Rate distortion theory, firstly introduced by [19] and further developed by [223], is an important concept in information theory which gives theoretical description of lossy data compression. It addressed the minimum average number of R bits, to transmit a random variable such that the receiver can reconstruct the random variable with distortion D.

Precisely, let $X^n = \{X_1, X_2 \dots X_n\} \in \mathcal{X}^n$ be i.i.d. random variables from distribution P_X. An encoder $f_n : \mathcal{X}^n \to \{1, 2, \dots, 2^{nR}\}$ maps the message X^n into codeword, and a decoder $g_n : \{1, 2, \dots, 2^{nR}\} \to \mathcal{X}^n$ reconstruct the message by an estimate $\hat{X}^n$ from the codeword. See Figure 8.1 for an illustration.

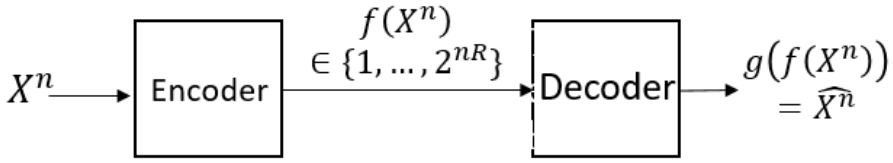

Figure 8.1: An illustration of encoder and decoder.

A distortion function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ quantifies the difference of the original and reconstructed message. Distortion between sequence X^n and $\hat{X}^n$ is defined as the average distortion of X_i's and $\hat{X}_i$'s. Commonly used distortion function includes Hamming distortion function $d(x, \hat{x}) = \mathbb{I}[x \neq \hat{x}]$ for $\mathcal{X} = \{0, 1\}$ and square distortion function $d(x, \hat{x}) = (x - \hat{x})^2$ for $\mathcal{X} = \mathbb{R}$.

Now we are ready to define the rate distortion function for data compression.

Definition 4. *A rate distortion pair (R, D) is* achievable *if there exists a series of (probabilistic) encoder-decoder (f_n, g_n) such that the alphabet of codeword has size 2^{nR} and expected distortion $\lim_{n\to\infty} \mathbb{E}[d(X^n, g_n(f_n(X^n)))] \leq D$.*

Definition 5. *Rate distortion function $R(D)$ equals to the infimum of rate R such that rate distortion pair (R, D) is achievable.*

The main theorem of rate distortion theory ([71, Theorem 10.2.1]) states as follows.

Theorem 19. *Rate distortion theorem for data compression.*

$$R(D) = \min_{P_{\hat{X}|X}:\mathbb{E}[d(X,\hat{X})]\leq D} I(X;\hat{X}). \tag{8.1}$$

The rate distortion quantifies the fundamental limit of data compression, i.e., *at least* how many bits are needed to compress the data, given the quality of the reconstructed data. Here is an example for rate distortion function.

Example 6. *If $X \sim \mathcal{N}(0,\sigma^2)$, the rate distortion function is given by*

$$R(D) = \begin{cases} \frac{1}{2}\log_2(\sigma^2/D) & \text{if } D \leq \sigma^2, \\ 0 & \text{if } D > \sigma^2. \end{cases} \tag{8.2}$$

If the required distortion D is larger than the variance of the Gaussian variable σ^2, we simply transmit $\hat{X} = 0$; otherwise, we will transmit $\hat{X}$ such that $\hat{X} \sim \mathcal{N}(0, \sigma^2 - D)$, $X - \hat{X} \sim \mathcal{N}(0, D)$ where $\hat{X}$ and $X - \hat{X}$ are independent.

8.2.2 Rate distortion theory for model compression

Now we extend the rate distortion theory for data compression to model compression. To apply the rate distortion theory to model compression, we view the weights in the model as a multidimensional random variable $W \in \mathbb{R}^m$ following distribution P_W. The randomness comes from multiple sources including different distributions of training data, randomness of training data and randomness of training algorithm. The compressor can also be random hence we describe the compressor by a conditional probability $P_{\hat{W}|W}$. Now we define the distortion and rate in model compression, analogously to the data compression scenario.

Distortion. Assume we have a neural network f_w that maps input $x \in \mathbb{R}^{d_x}$ to $f_w(x)$. For regression, $f_w(x)$ is defined as the output of the neural network on $\mathbb{R}^{d_y}$. Analogous to the square distortion in data compression, We define the distortion to be the expected ℓ_2 distance between f_w and $f_{\hat{w}}$, i.e.

$$d(w,\hat{w}) \equiv \mathbb{E}_X\left[\|f_w(X) - f_{\hat{w}}(X)\|_2^2\right]. \tag{8.3}$$

For classification, $f_w(x)$ is defined as the output probability distribution over C classes on the simplex Δ^{C-1}. We define the distortion to be the expected distance between f_w and $f_{\hat{w}}$, i.e.

$$d(w, \hat{w}) \equiv \mathbb{E}_X \left[D(f_{\hat{w}}(X) || f_w(X)) \right]. \tag{8.4}$$

Here D could be any statistical distance, including KL divergence, Hellinger distance, total variation distance, etc. Such a definition of distortion captures the difference between the original model and the compressed model, averaged over data X, and measures the quality of a compression algorithm.

Rate. In data compression, the rate is defined as the description length of the bits needed to communicate the compressed data $\hat{X}$. The compressor outputs $\hat{X}$ from a finite *code book* $\mathcal{X}$. The description consists the *code word* which are the indices of $\hat{x}$, and the description of the *code book*.

In rate distortion theory, we ignore the code book length. Since we are transmitting a sequence of data X^n, the code word has to be transmitted for each X_i but the code book is only transmitted once. In asymptotic setting, the description length of code book can be ignored, and the rate is defined as the description length of the code word.

In model compression, we also define the rate as the code word length, by assuming that an underlying distribution P_W of the parameters exists and infinitely many models whose parameters are i.i.d. from P_W will be compressed. In practice, we only compress the parameters once so there is no distribution of the parameters. Nevertheless, the rate distortion theory can also provide important intuitions for one-time compression, explained in Section 8.4.

Now we can define the rate distortion function for model compression. Analogously to Theorem 19, the rate distortion function for model compression is defined as follows.

Definition 6. *Rate distortion function for model compression.*

$$R(D) = \min_{P_{\hat{W}|W}: \mathbb{E}_{W,\hat{W}}\left[d(W,\hat{W})\right] \leq D} I(W; \hat{W}). \tag{8.5}$$

In Section 8.3 we establish a lower bound of the rate distortion function.

8.3 Lower bound and achievability for rate distortion function

8.3.1 Lower bound for linear model

Assume that we are going to compress a linear regression model $f_w(x) = w^T x$. We assume that the mean of data $x \in \mathbb{R}^m$ is zero and the covariance matrix is diagonal, i.e., $\mathbb{E}_X[X_i^2] = \lambda_{x,i} > 0$ and $\mathbb{E}_X[X_i X_j] = 0$ for $i \neq j$. Furthermore, assume that the parameters $W \in \mathbb{R}^m$ are drawn from a Gaussian distribution $\mathcal{N}(0, \Sigma_W)$. The following theorem gives the lower bound of the rate distortion function for the linear regression model.

Theorem 20. *The rate distortion function of the linear regression model $f_w(x) = w^T x$ is lower bounded by*

$$R(D) \geq \underline{R}(D) = \frac{1}{2} \log \det(\Sigma_W) - \sum_{i=1}^{m} \frac{1}{2} \log(D_i), \tag{8.6}$$

where

$$D_i = \begin{cases} \mu/\lambda_{x,i} & \text{if} \mu < \lambda_{x,i} \mathbb{E}_W[W_i^2], \\ \mathbb{E}_W[W_i^2] & \text{if} \mu \geq \lambda_{x,i} \mathbb{E}_W[W_i^2], \end{cases} \tag{8.7}$$

where μ is chosen that $\sum_{i=1}^m \lambda_{x,i} D_i = D$.

This lower bound gives rise to a "weighted water-filling" approach, which differs from the classical "water-filling" for rate distortion of colored Gaussian source in [71, Figure 13.7]. The details and graphical explanation of the "weighted water-filling" can be found in Section 8.6.

8.3.2 Achievability

We show that, the lower bound give in Theorem 20 is achievable. Precisely, we have the following theorem.

Theorem 21. *There exists a class of probabilistic compressors $P^{(D)}_{\hat{W}^*|W}$ such that $\mathbb{E}_{P_W \circ P^{(D)}_{\hat{W}^*|W}} \left[d(W, \hat{W}^*) \right] = D$ and $I(W; \hat{W}^*) = \underline{R}(D)$.*

The optimal compressor is Algorithm 3 in Section 8.6. Intuitively, the optimal compressor does the following:

- Finding the optimal water levels D_i for "weighted water filling", such that the expected distortion $D = \mathbb{E}_{W,\hat{W}}[d(W,\hat{W})] = \mathbb{E}_{W,\hat{W}}[\hat{W}^T\Sigma_X(W-\hat{W})]$ is minimized given certain rate.
- Adding a noise Z_i which is independent of $\hat{W}_i = W_i + Z_i$ and has a variance proportional to the water level. That is possible since W is Gaussian.

We can check that the compressor makes all the inequalities become equality, hence achieve the lower bound. The full proof of the lower bound and achievability can be found in Section 8.6.

8.4 Improved objective for model compression

In traditional rate distortion theory, we assume that there exists a prior distribution P_W on the weights W, and prove the tightness of the lower bound in the asymptotic scenario. However, in practice, we only compress one particular pre-trained model, so there are no prior distribution of W. Nonetheless, we can still learn something important from the achivability of the lower bound, by extracting two "golden rules" from the optimal algorithm for linear regression.

8.4.1 Two golden rules

Recall that for linear regression model, to achieve the smallest rate given certain distortion (or, equivalently, achieve the smallest distortion given certain rate), the optimal compressor need to do the following: (1) find appropriate "water levels" such that the expected distortion $E_{W,\hat{W}}[d(W,\hat{W})] = \mathbb{E}_{W,\hat{W},X}[(W^TX - \hat{W}^TX)^2] = \mathbb{E}_{W,\hat{W}}[(W-\hat{W})^T\Sigma_X(W-\hat{W})]$ is minimized. (2) make sure that $\hat{W}_i$ is orthogonal to $W_i - \hat{W}_i$, i.e., $\mathbb{E}_{W,\hat{W}}[\hat{W}^T\Sigma_X(W-\hat{W})] = 0$. Hence, we extract the following two "golden rules":

1. Orthogonality rule — $\mathbb{E}_{W,\hat{W}}[\hat{W}^T\Sigma_X(W-\hat{W})] = 0$.

2. Minimization rule — $\mathbb{E}_{W,\hat{W}}[(W-\hat{W})^T\Sigma_X(W-\hat{W})]$ should be minimized, given certain rate.

For practical model compression, we adopt these two "golden rules", by making the following amendments. First, we discard the expectation over W and $\hat{W}$ since there is only one model to be compressed. Second, the distortion can be written as $d(w,\hat{w}) = (w-\hat{w})^T\Sigma_X(w-\hat{w})$ only for linear models. For non-linear models, the distortion function is complicated, but can be approximated by a simpler formula. For non-linear regression models, we take first order Taylor expansion of the function $f_{\hat{w}}(x) \approx f_w(x)+(\hat{w}-w)^T\nabla_w f_w(x)$, and have

$$\begin{aligned} & d(w,\hat{w}) = \mathbb{E}_X\left[\,\|f_w(X)-f_{\hat{w}}(X)\|_2^2\,\right] \\ \approx\ & \mathbb{E}_X\left[(w-\hat{w})^T\nabla_w f_w(X)(\nabla_w f_w(X))^T(w-\hat{w})\right] \\ =\ & (w-\hat{w})^T I_w(w-\hat{w}), \end{aligned} \tag{8.8}$$

where the "weight importance matrix" defined as

$$I_w = \mathbb{E}_X\left[\nabla_w f_w(X)(\nabla_w f_w(X))^T\right], \tag{8.9}$$

quantifies the relative importance of each weight to the output. For linear regression models, weight importance matrix I_w equals to Σ_X.

For classification models, we will first approximate the KL divergence. Using the Taylor expansion $x\log(x/a) \approx (x-a)+(x-a)^2/(2a)$ for $x/a \approx 1$, the KL divergence $D_{KL}(P||Q)$ for can be approximated by $D_{KL}(P||Q) \approx \sum_i(P_i-Q_i)+(P_i-Q_i)^2/(2P_i) = \sum_i(P_i-Q_i)^2/(2P_i)$, or in vector form $D_{KL}(P||Q) \approx \frac{1}{2}(P-Q)^T\mathrm{diag}[P^{-1}](P-Q)$. Therefore,

$$\begin{aligned} & d(w,\hat{w}) = \mathbb{E}_X\left[D_{KL}(f_{\hat{w}}(X)||f_w(X))\right] \\ \approx\ & \frac{1}{2}\mathbb{E}_X\left[(f_w(X)-f_{\hat{w}}(X))^T\mathrm{diag}[f_w^{-1}(X)](f_w(X)-f_{\hat{w}}(X))\right] \\ \approx\ & \frac{1}{2}\mathbb{E}_X\left[(w-\hat{w})^T(\nabla_w f_w(X))\mathrm{diag}[f_w^{-1}(X)](\nabla_w f_w(X))^T(w-\hat{w})\right]. \end{aligned} \tag{8.10}$$

So the weight importance matrix is given by

$$I_w = \mathbb{E}_X\left[(\nabla_w f_w(X))\mathrm{diag}[f_w^{-1}(X)](\nabla_w f_w(X))^T\right]. \tag{8.11}$$

This weight importance matrix is also valid for other statistical distances, such as reverse KL divergence, Hellinger distance and Jenson-Shannon distance.

Now we define the two "golden rules" for practical model compression algorithms,

1. Orthogonality rule — $\hat{w}^T I_w(w - \hat{w}) = 0$.
2. Minimization rule — $(w - \hat{w})^T I_w(w - \hat{w})$ is minimized given certain constraints.

In the following subsections we will show the optimality of the "golden rules" for a one-hidden-layer neural network, and apply the "golden rules" to derive new objective function for pruning and quantization.

8.4.2 Optimality for one-hidden-layer ReLU network

We show that if a compressor of a one-hidden-layer ReLU network satisfies the two "golden rules", it will be the optimal compressor, with respect to mean-square-error. Precisely, consider the one-hidden layer ReLU neural network $f_w(x) = ReLU(w^T x)$, where the distribution of input $x \in \mathbb{R}^m$ is $\mathcal{N}(0, \Sigma_X)$. Furthermore, we assume that the covariance matrix $\Sigma_X = \text{diag}[\lambda_{x,1}, \ldots, \lambda_{x,m}]$ is diagonal and $\lambda_{x,i} > 0$ for all i. We have the following theorem.

Theorem 22. *If compressed weight $\hat{w}^*$ satisfies $\hat{w}^* I_w(\hat{w}^* - w) = 0$ and*

$$\hat{w}^* = \arg\min_{\hat{w} \in \hat{\mathcal{W}}} (w - \hat{w})^T I_w (w - \hat{w}), \tag{8.12}$$

where $\hat{\mathcal{W}}$ is some class of compressors, then

$$\hat{w}^* = \arg\min_{\hat{w} \in \hat{\mathcal{W}}} \mathbb{E}_X \left[(f_w(X) - f_{\hat{w}}(X))^2 \right]. \tag{8.13}$$

The proof uses the techniques of Hermite polynomials and Fourier analysis on Gaussian spaces, inspired by [14]. The full proof can be found in Section 8.7. Generalizing this result to other activation functions and deeper neural networks are possible future directions.

Here $\hat{\mathcal{W}}$ denotes a class of compressors, with some constraints. For example, $\hat{\mathcal{W}}$ could be the class of pruning algorithms where no more than 50% weights are pruned, or $\hat{\mathcal{W}}$ could be the class of quantization algorithm where each weight is quantized to 4 bits. Theoretically, it is not guaranteed that the two "golden rules" can be satisfied simultaneously for every $\hat{\mathcal{W}}$, but in the following subsection we show that they can be satisfied simultaneously for two of the most commonly used class of compressors — pruning and quantization. Hence, minimizing the objective $(w-\hat{w})^T I_w (w-\hat{w})$ will be optimal for pruning and quantization.

8.4.3 Improved objective for pruning and quantization

Pruning and quantization are two most basic and useful building blocks of modern model compression algorithms, For example, DeepCompress [17] iteratively prune, retrain and quantize the neural network and achieve state-of-the-art performances on large neural networks.

In pruning algorithms, we choose a subset $S \in [m]$ and set $\hat{w}_i = 0$ for all $i \in S$ and $\hat{w}_i = w_i$ for $i \notin S$. The compression ratio is evaluated by the proportion of unpruned weights $r = (m-|S|)/m$. Since either $\hat{w}_i$ or $w_i - \hat{w}_i$ is zero, so the orthogonality is automatically satisfied, so we have the following corollary.

Corollary 3. *For any fixed r, let*

$$\hat{w}_r^* = \arg \min_{S:\frac{d-|S|}{d}=r} (w-\hat{w})^T I_w (w-\hat{w}). \tag{8.14}$$

Then

$$\hat{w}_r^* = \arg \min_{S:\frac{d-|S|}{d}=r} \mathbb{E}_X \left[(f_w(X) - f_{\hat{w}}(X))^2 \right]. \tag{8.15}$$

In quantization algorithms, we cluster the weights to k centroids $\{c_1, \ldots, c_k\}$. The algorithm optimize the centroids as long as the assignments of each weight $A_i \in [k]$. The final compressed weight is given by $\hat{w}_i = c_{A_i}$. Usually k-means algorithm are utilized to minimize the centroids and assignments

alternatively. The compression ratio of quantization algorithm is given by

$$r = \frac{mb}{m\sum_{j=1}^{k}\frac{m_j}{m}\lceil\log_2\frac{m}{m_j}\rceil + kb}, \tag{8.16}$$

where m is the number of weights and b is the number of bits to represent one weight before quantization (usually 32). By using Huffman coding, the average number of bits for each weight is given by $\sum_{j=1}^{k}(m_j/m)\lceil\log_2(m/m_j)\rceil$, where m_j is the number of weights assigned to the j-th cluster. The definition of compression ratio of pruning and quantization is consistent since both of them equals to the number of bits representing compressed model parameters divided by the number of bits representing original model parameters.

If we can find the optimal quantization algorithm satisfying the minimization rule, then each centroids c_j should be optimal, i.e.

$$0 \;=\; \frac{\partial}{\partial c_j}(w-\hat{w})^T I_w (w-\hat{w}) = -2\left(\sum_{i:A_i=j} e_i^T\right) I_w(w-\hat{w}), \tag{8.17}$$

where e_i is the i-th standard basis. Therefore, we have

$$\begin{aligned}\hat{w}I_w(\hat{w}-w) &= \left(\sum_{j=1}^{k} c_j(\sum_{i:A_i=j} e_i)\right)^T I_w(w-\hat{w}) \\ &= \sum_{j=1}^{k} c_j\left((\sum_{i:A_i=j} e_i^T) I_w(w-\hat{w})\right) = 0.\end{aligned} \tag{8.18}$$

Hence the orthogonality rule is satisfied if the minimization rule is satisfied.

Corollary 4. *For any fixed number of centroids k, let*

$$\hat{w}_k^* = \arg\min_{\{c_1,\ldots,c_k\},A\in[k]^m} (w-\hat{w})^T I_w(w-\hat{w}), \tag{8.19}$$

then

$$\hat{w}_k^* = \arg\min_{\{c_1,\ldots,c_k\},A\in[k]^m} \mathbb{E}_X\left[(f_w(X)-f_{\hat{w}}(X))^2\right]. \tag{8.20}$$

As corollaries of Theorem 22, we proposed to use $(w-\hat{w})^T I_w(w-\hat{w})$ as the objective for pruning and quantization algorithms, which can achieve the minimum MSE for one-hidden-layer ReLU neural network.

8.5 Experiments of Chapter 8

In this section, we show that this objective can also improve pruning and quantization algorithm for larger neural networks on real data. We test the objectives on the following neural network and datasets.[1]

1. Three-layer fully connected neural network on MNIST.

2. Convolutional neural network with five convolutional layers and three fully connected layers on CIFAR 10 and CIFAR 100.

In Section 8.5.1, we use the weight importance matrix for classification in Eq. (8.11), which is derived by approximating the distortion of KL-divergence. This weight importance matrix does not depend on the training labels, so the induced pruning/quantization algorithms is called "unsupervised compression". Furthermore, if the training labels are available, we treat the loss function $\mathcal{L}_w(X, Y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ as the function to be compressed, and derive several pruning/quantization objectives. The induced pruning/quantization methods are called "supervised compression" and are studied in Section 8.5.2.

8.5.1 Unsupervised compression experiments

For classification problems, the weight importance matrix is defined as

$$I_w \quad = \quad \mathbb{E}_X \left[\nabla_w f_w(X) \mathrm{diag}[f_w^{-1}(X)] (\nabla_w f_w(X))^T \right]. \tag{8.21}$$

For computational simplicity, we drop the off-diagonal terms of I_w, and simplify the objective to $\sum_{i=1}^m \mathbb{E}_X[\frac{(\nabla_{w_i} f_w(X))^2}{f_w(X)}](w_i - \hat{w}_i)^2$. To minimize the proposed objective, a pruning algorithm just prune the weights with smaller $\mathbb{E}_X[\frac{(\nabla_{w_i} f_w(X))^2}{f_w(X)}]w_i^2$ greedily. A quantization algorithm uses the weighted k-means algorithm [230] to find the optimal centroids and assignments. We compare the proposed objective with the baseline objective $\sum_{i=1}^m (w_i - \hat{w}_i)^2$, which were used as building blocks in DeepCompress [17]. We compare the objectives in Table 8.1.

For pruning experiment, we choose the same compression rate for every convolutional layer and fully connected layer, and plot the test accuracy and

[1] We load the pretrained models from `https://github.com/aaron-xichen/pytorch-playground`.

Table 8.1: Comparison of unsupervised compression objectives.

Name	Minimizing objective
Baseline	$\sum_{i=1}^{m}(w_i - \hat{w}_i)^2$
Proposed	$\sum_{i=1}^{m} \mathbb{E}_X[\frac{(\nabla_{w_i} f_w(X))^2}{f_w(X)}](w_i - \hat{w}_i)^2$

test cross-entropy loss against compression rate. For quantization experiment, we choose the same number of clusters for every convolutional and fully connected layer. Also we plot the test accuracy and test cross-entropy loss against compression rate. To reduce the variance of estimating the weight importance matrix I_w, we use the *temperature scaling* method introduced by [234] to improve model calibration.

We show that results of pruning experiment in Figure 8.2, and the results of quantization experiment in Figure 8.3. We can see that the proposed objective gives better validation cross-entropy loss than the baseline, for every different compression ratios. The proposed objective also gives better validation accuracy in most scenarios. Occasionally the proposed objective can not improve the accuracy (top left of Figure 8.2). We conjecture that the reason is the ill-calibration of the original model.

8.5.2 Supervised compression experiments

In the previous experiment, we only use the training data to compute the weight importance matrix. But if we can use the training label as well, we can further improve the performance of pruning and quantization algorithms. If the training label is available, we can view the cross-entropy loss function $\mathcal{L}(f_w(x), y) = \mathcal{L}_w(x, y)$ as a function from $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, and define the distortion function as

$$d(w, \hat{w}) \quad = \quad \mathbb{E}_{X,Y}\left[(\mathcal{L}_w(X, Y) - \mathcal{L}_{\hat{w}}(X, Y))^2\right]. \tag{8.22}$$

Taking first-order approximation of the loss function gives the supervised weight importance matrix,

$$I_w \quad = \quad \mathbb{E}\left[\nabla_w \mathcal{L}_w(X, Y)(\nabla_w \mathcal{L}_w(X, Y))^T\right]. \tag{8.23}$$

We write $\mathbb{E}$ instead of $\mathbb{E}_{X,Y}$ for simplicity. Similarly, we drop the off-

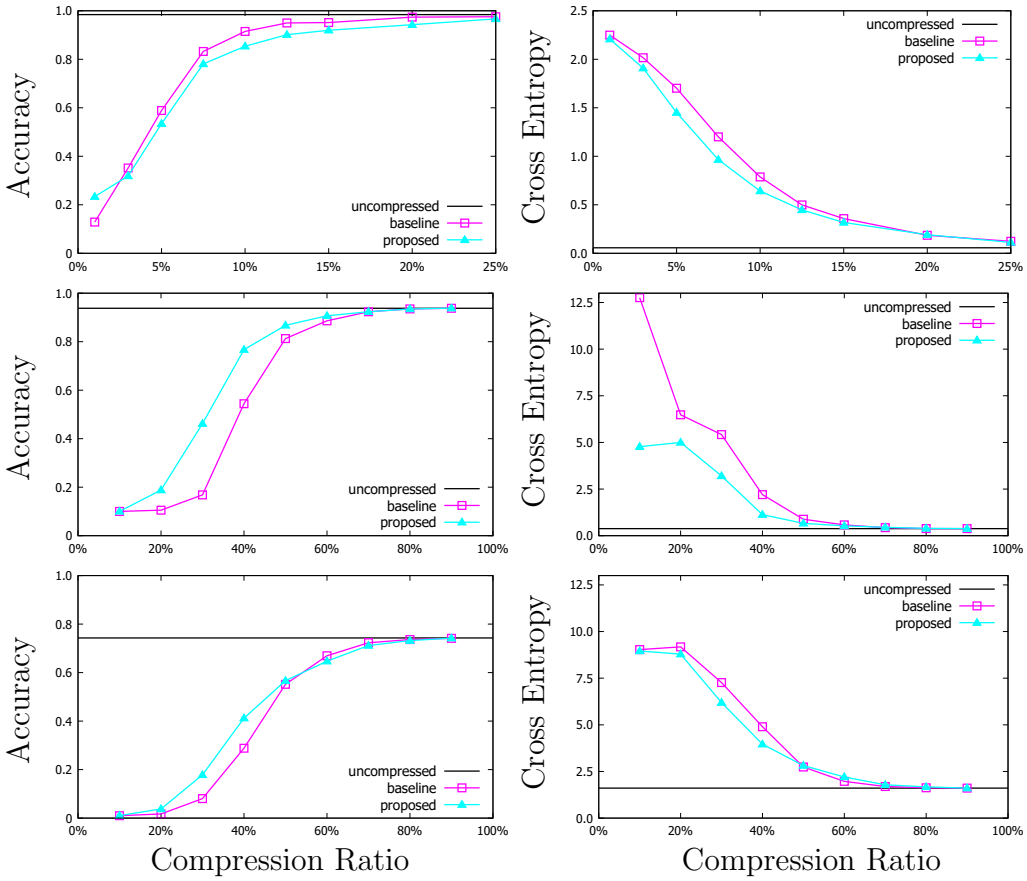

Figure 8.2: Result for unsupervised pruning experiment. Top: fully connected NN on MNIST. Middle: ConvNN on CIFAR 10. Bottom: ConvNN on CIFAR 100. Top: test accuracy, Bottom: test cross entropy.

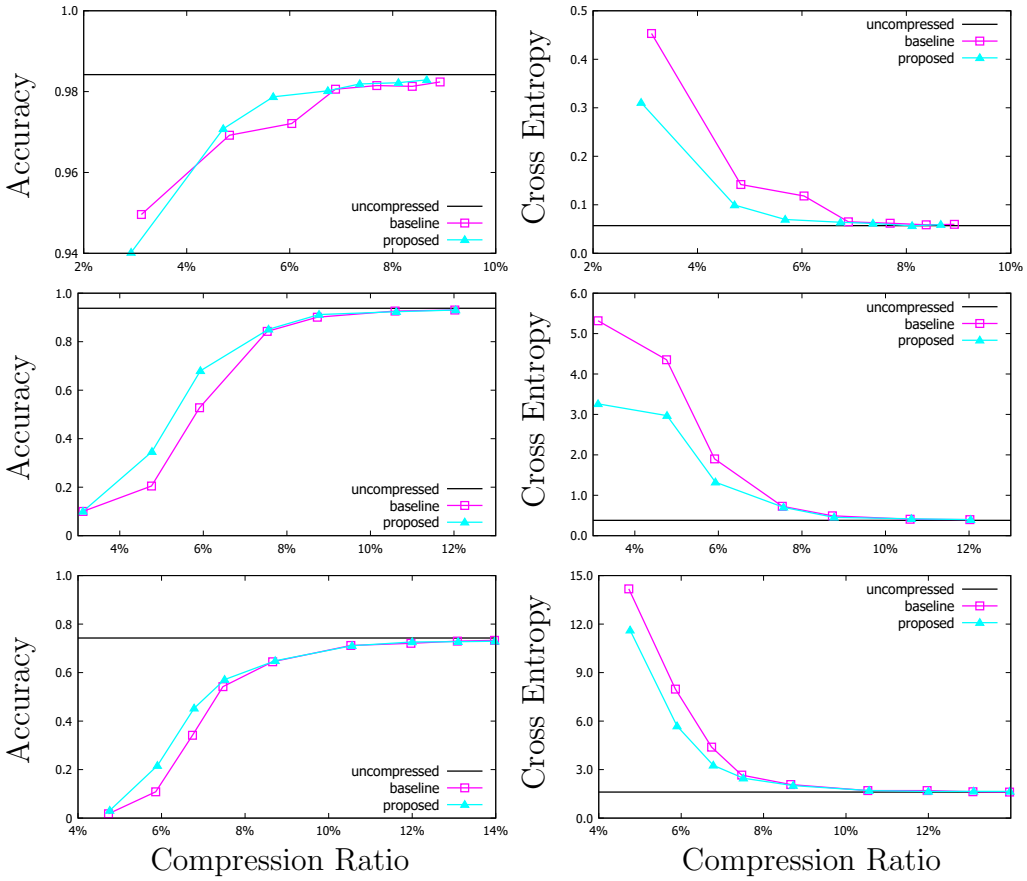

Figure 8.3: Result for unsupervised quantization experiment. Top: fully connected NN on MNIST. Middle: ConvNN on CIFAR 10. Bottom: ConvNN on CIFAR 100. Top: test accuracy, Bottom: test cross entropy.

diagonal terms for ease of computation, and simplify the minimizing objective to $\sum_{i=1}^{m} \mathbb{E}[(\nabla_{w_i} \mathcal{L}_w(X,Y))^2](w_i - \hat{w}_i)^2$, which is called gradient-based objective. Note that for well-trained model, the expected value of gradient $\mathbb{E}[\nabla_w \mathcal{L}_w(X,Y)]$ is closed to zero, but the second moment of the gradient $\mathbb{E}[\nabla_w \mathcal{L}_w(X,Y)(\nabla_w \mathcal{L}_w(X,Y))^T]$ could be large. We compare this objective with the baseline objective $\sum_{i=1}^{m}(w_i - \hat{w}_i)^2$. We also compare with the Hessian-based objective $\sum_{i=1}^{m} \mathbb{E}[\nabla^2_{w_i} \mathcal{L}_w(X,Y)](w_i - \hat{w}_i)^2$, which is used in [217] and [218] for network pruning and [230] for network quantization. To estimate the diagonal entries of the Hessian matrix of the loss function with respect to the model parameters, we implemented curvature propagation [235] treating each layer and activation as a node. The running time is proportional to the running time of the usual gradient backpropagation by a factor independent of the size of the model. Manually optimizing the local Hessian calculation at each node reduces memory usage and allows us to use larger batch size and larger number of samples for more accurate estimates.

Furthermore, if we take second-order approximation of the loss function, and drop the off-diagonal terms of the squared gradient matrix and squared Hessian tensor, we have the following approximation

$$\begin{aligned} & d(w, \hat{w}) \\ = \; & \mathbb{E}\big[(\mathcal{L}_w(X,Y) - \mathcal{L}_{\hat{w}}(X,Y))^2\big] + \frac{1}{2}(w - \hat{w})^T \nabla^2_w \mathcal{L}_w(X,Y)(w - \hat{w}))^2\big] \\ \approx \; & \sum_{i=1}^{m} \mathbb{E}[(\nabla_{w_i} \mathcal{L}_w(X,Y))^2](w_i - \hat{w}_i)^2 + \frac{1}{4} \sum_{i=1}^{m} \mathbb{E}[(\nabla^2_{w_i} \mathcal{L}_w(X,Y))^2](w_i - \hat{w}_i)^4, \end{aligned} \tag{8.24}$$

which is called gradient+Hessian-based objective. We conclude the different supervised objectives in Table 8.2.

Table 8.2: Comparison of supervised compression objectives.

Name	Minimizing objective
Baseline	$\sum_{i=1}^{m}(w_i - \hat{w}_i)^2$
Gradient	$\sum_{i=1}^{m} \mathbb{E}[(\nabla_{w_i} \mathcal{L}_w(X,Y))^2](w_i - \hat{w}_i)^2$
Hessian	$\sum_{i=1}^{m} \mathbb{E}[\nabla^2_{w_i} \mathcal{L}_w(X,Y)](w_i - \hat{w}_i)^2$
Gradient + Hessian	$\sum_{i=1}^{m} \mathbb{E}[(\nabla_{w_i} \mathcal{L}_w(X,Y))^2](w_i - \hat{w}_i)^2$ $+\frac{1}{4}\sum_{i=1}^{m} \mathbb{E}[(\nabla^2_{w_i} \mathcal{L}_w(X,Y))^2](w_i - \hat{w}_i)^4$

Algorithm for Gradient+Hessian objective

For pruning algorithm, we can prune the weights by $\mathbb{E}[(\nabla_{w_i}\mathcal{L}_w(X,Y))^2]w_i^2 + \frac{1}{4}\mathbb{E}[(\nabla^2_{w_i}\mathcal{L}_w(X,Y))^2]w_i^4$ greedily. For quantization algorithm, we present a variation of k-means algorithm which are used to find the optimal quantization for the following objective,

$$\min_{c_1,\ldots,c_k,A\in[k]^m} \sum_{i=1}^{m} \left(I_i(w_i - c_{A_i})^2 + H_i(w_i - c_{A_i})^4\right), \tag{8.25}$$

where I_i is positive weight importance for quadratic term and H_i is positive weight importance for quartic term. Basic idea of the algorithm is — the assignment step finds the optimal assignment given fixed centroids, and the update step finds the optimal centroids given fixed assignments.

Algorithm 2 Quartic weighted k-means

Input: Weights $\{w_1,\ldots,w_m\}$, weight importances $\{I_1,\ldots,I_m\}$, quartic weight importances $\{H_1,\ldots,H_m\}$, number of clusters k, iterations T;
Initialize the centroid of k clusters $\{c_1^{(0)},\ldots,c_k^{(0)}\}$;
for $t = 1$ to T **do**
  Assignment step:
  for $i = 1$ to m **do**
    Assign w_i to the nearest cluster centroid, i.e. $A_i^{(t)} = \arg\min_{j\in[k]}(w_i - c_j^{(t-1)})^2$;
  end for
  Update step:
  for $j = 1$ to k **do**
    Find the only real root x^* of the cubic equation

$$\begin{aligned}&(\sum_{i:A_i^{(t)}=j} 4H_i)x^3 - (\sum_{i:A_i^{(t)}=j} 12H_iw_i)x^2 + (\sum_{i:A_i^{(t)}=j} (12H_iw_i^2 + 2I_i))x \\ &-(\sum_{i:A_i^{(t)}=j} (4H_iw_i^3 + 2I_iw_i)) = 0;\end{aligned}$$

    Update the cluster centroids $c_j^{(t)}$ be the real root x^*;
  end for
end for
Output: Centroids $\{c_1^{(T)},\ldots,c_k^{(T)}\}$ and assignments $A^{(T)} \in [k]^m$.

Here we show that the cubic equation in Algorithm 2 has only one real root. It was know that if the determinant $\Delta_0 = b^2 - 3ac$ of a cubic equation

$ax^3 + bx^2 + cx + d = 0$ is negative, then the cubic equation is strictly increasing or decreasing, hence only have one real root. Now we show that the determinant is negative in this case (we drop the subsripts of the summation for simplicity).

$$\begin{aligned}\Delta_0 &= (\sum_i 12H_i w_i)^2 - 3(\sum_i 4H_i)(\sum_i 12H_i w_i^2 + 2I_i) \\ &= 144\left((\sum_i H_i w_i)^2 - (\sum_i H_i)(\sum_i H_i w_i^2)\right) - 24(\sum_i H_i)(\sum_i I_i).\end{aligned} \tag{8.26}$$

The first term is non-positive because of Cauchy-Schwarz inequality. The second term is negative since H_i's and I_i's are all positive. Hence the determinant is negative.

We show that results of pruning experiment in Figure 8.4, and quantization experiment in Figure 8.5. Generally, the gradient objective and Hessian objective both give better performance than baseline objective, while gradient objective is slightly than Hessian objective at some points. Gradient+Hessian objective gives the best overall performance.

8.5.3 Remarks on the experiments

Analogously to the distortion of regression, we define the distortion function as $d(w, \hat{w}) = \mathbb{E}_{X,Y}\left[(\mathcal{L}_w(X, Y) - \mathcal{L}_{\hat{w}}(X, Y))^2\right]$. However, since the goal of classification is to minimize the loss function, the following definition of distortion function $\tilde{d}(w, \hat{w}) = \mathbb{E}_{X,Y}\left[\mathcal{L}_{\hat{w}}(X, Y) - \mathcal{L}_w(X, Y)\right]$ is also valid and has been adopted in [217] and [230]. The main difference is — $d(w, \hat{w})$ focus on the quality of *compression algorithm*, i.e., how similar is the compressed model compared to uncompressed model, whereas $\tilde{d}(w, \hat{w})$ focus on the quality of *compressed model*, i.e. how good is the compressed model. So $d(w, \hat{w})$ is a better criteria for the compression algorithm. Additionally, by taking second order approximation of $d(w, \hat{w})$, we have gradient+Hessian objective, which shows better empirical performance than Hessian objective, derived by taking second order approximation of $\tilde{d}(w, \hat{w})$.

Here we briefly talk about the hyperparameters used in estimating the gradients $\mathbb{E}[\nabla_{w_i}\mathcal{L}_w(X, Y)]$ and Hessians $\mathbb{E}[\nabla^2_{w_i}\mathcal{L}_w(X, Y)]$.

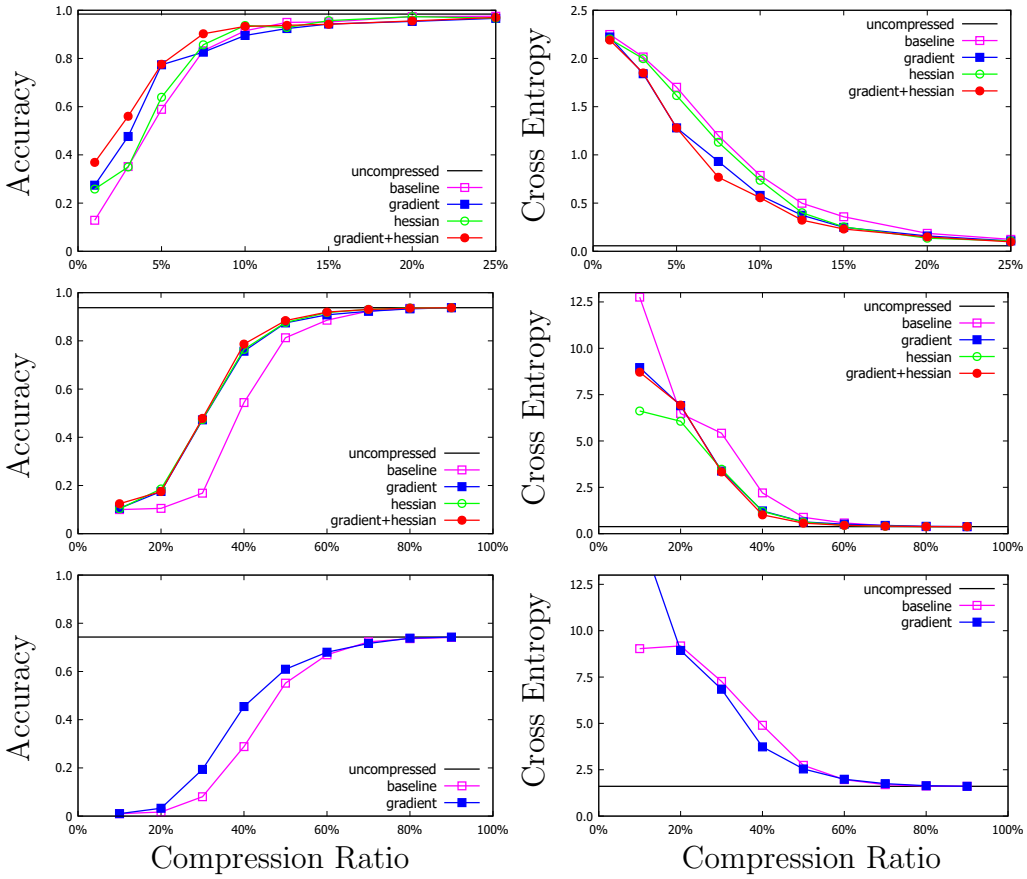

Figure 8.4: Result for supervised pruning experiment. Top: fully connected NN on MNIST. Middle: ConvNN on CIFAR 10. Bottom: ConvNN on CIFAR 100. Top: test accuracy, Bottom: test cross entropy.

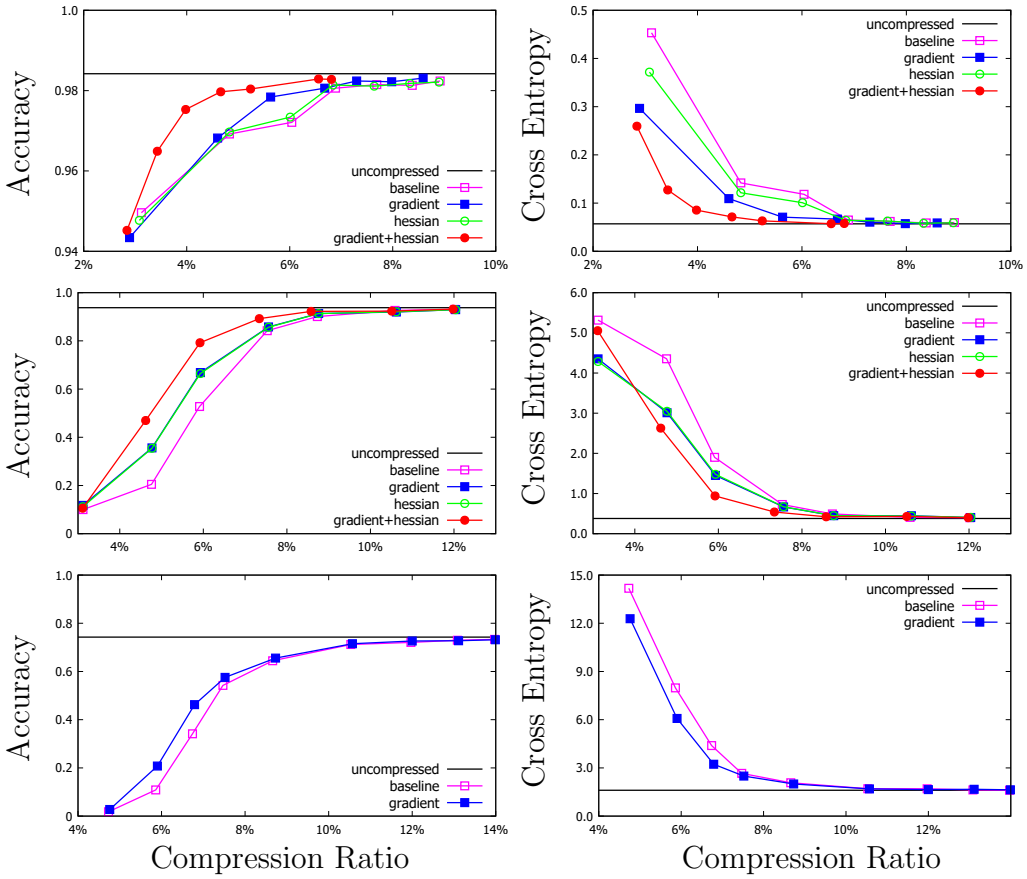

Figure 8.5: Result for supervised quantization experiment. Top: fully connected NN on MNIST. Middle: ConvNN on CIFAR 10. Bottom: ConvNN on CIFAR 100. Top: test accuracy, Bottom: test cross entropy.

Temperature scaling method

The temperature scaling method proposed by [234], aims to improve the confidence calibration of a classification model. Denote $z_w(x) \in \mathbb{R}^C$ is the output of the neural network, and classical softmax gives $f_w^{(c)}(x) = \frac{\exp\{z_w^{(c)}(x)\}}{\sum_{c\in C}\exp\{z_w^{(c)}(x)\}}$. The temperature sclaed softmax gives

$$f_w^{(c)}(x) = \frac{\exp\{z_w^{(c)}(x)/T\}}{\sum_{c\in C}\exp\{z_w^{(c)}(x)/T\}}, \tag{8.27}$$

by choosing different T, the prediction of the model does not change, but the cross entropy loss may change. Hence, we can finetune T to get a better model calibration. In our experiment, we found that in MNIST experiment, the model is poorly calibrated. Hence, the variance of estimating gradient and Hessian is very large. To solve this, we adopt a temperature $T > 1$ such that the loss from correctly predicted data can also be backpropagated.

In Figure 8.6, we show the effect of T for supervised pruning for MNIST. We can see that as T increases from 1, the performance become better at first, then become worse. In our experiment, we choose $T \in \{1.0, 2.0, \ldots, 9.0\}$ which gives best accuracy.

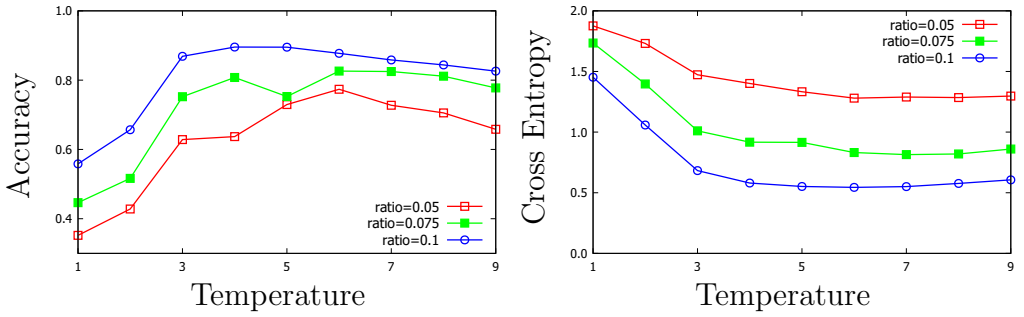

Figure 8.6: Effect of the temperature T. Left: accuracy of supervised pruning for MNIST. Right: cross entropy of supervised pruning for MNIST. Different lines denote different compression ratio $\in \{0.05, 0.075, 0.1\}$.

Regularizer of Hessian

In the experiments, we estimate the Hessians $\mathbb{E}[\nabla^2_{w_i}\mathcal{L}_w(X, Y)]$ using the curvature propagation algorithm [235]. However, due to the sparsity introduced by ReLU, there are many zero entries of the estimated Hessians, which hurts the performance of the algorithm. Hence, we add a constant $\mu > 0$

to the estimated Hessians. In Figure 8.7, we show that effect of μ for supervised pruning for CIFAR10. We can see that as μ increases from 0, the performance increase first then decrease. We use simple binary search to find the best μ.

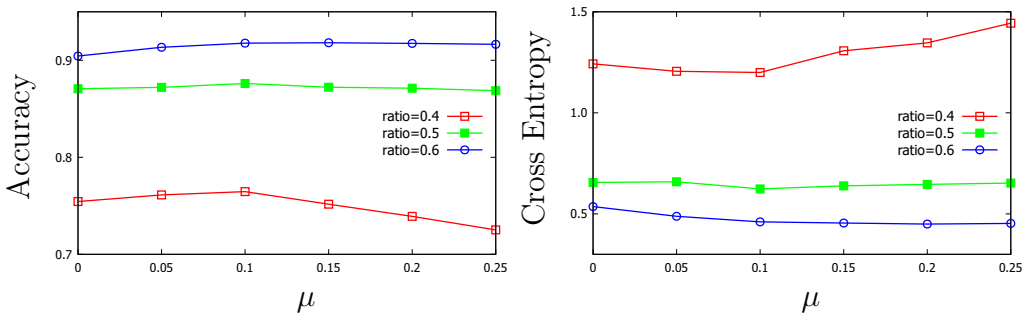

Figure 8.7: Effect of the regularizer μ. Left: accuracy of supervised pruning for CIFAR10. Right: cross entropy of supervised pruning for CIFAR10. Different lines denote different compression ratio $\in \{0.4, 0.5, 0.6\}$.

8.6 Lower bound for rate distortion function

8.6.1 General lower bound

First, we establish establishes a lower bound of the rate distortion function, which works for general models.

Lemma 31. *The rate-distortion function $R(D) \geq \underline{R}(D) = h(W) - C$, where C is the optimal value of the following optimization problem.*

$$\begin{aligned} \max_{P_{\hat{W}|W}} \quad & \sum_{i=1}^{m} \min\left\{h(W_i), \frac{1}{2}\log(2\pi e \mathbb{E}_{W,\hat{W}}[(W_i - \hat{W}_i)^2])\right\} \\ \text{s.t.} \quad & E_{W,\hat{W}}\left[d(W,\hat{W})\right] \leq D. \end{aligned} \tag{8.28}$$

where $h(W) = -\int_{w\in\mathcal{W}} P_W(w)\log P_W(w)dw$ is the differential entropy of W and $h(W_i)$ is the differential entropy of the i-th entry of W.

Proof of Lemma 31 Recall that the rate distortion function for model compression is defined as $R(D) = \min_{P_{\hat{W}|W}:\mathbb{E}_{W,\hat{W}}[d(W,\hat{W})]\leq D} I(W;\hat{W})$. Now we

lower bound the mutual information $I(W, \hat{W})$ by

$$\begin{aligned} & I(W;\hat{W}) = h(W) - h(W \,|\, \hat{W}), \\ = \;& h(W) - \sum_{i=1}^{m} h(W_i \,|\, W_1, \ldots, W_{i-1}, \hat{W}_i, \ldots, \hat{W}_m) \\ \geq \;& h(W) - \sum_{i=1}^{m} h(W_i \,|\, \hat{W}_i). \end{aligned} \tag{8.29}$$

Here the last inequality comes from the fact that conditioning does not increase entropy. Notice that the first term $h(W)$ does not depend on the compressor. For the last term, we upper bound each term $h(W_i \,|\, \hat{W}_i)$ in two ways. On one hand, $h(W_i \,|\, \hat{W}_i)$ is upper bounded by $h(W_i)$ because conditioning does not increase entropy. On the other hand, $h(W_i \,|\, \hat{W}_i) = h(W_i - \hat{W}_i \,|\, \hat{W}_i) \leq h(W_i - \hat{W}_i)$, and by [71, Theorem 8.6.5], differential entropy is maximized by Gaussian distribution, for given second moment. We then have:

$$\begin{aligned} & h(W_i \,|\, \hat{W}_i) \leq \min\left\{ h(W_i), h(W_i - \hat{W}_i) \right\} \\ \leq \;& \min\left\{ h(W_i), \frac{1}{2}\log\left(2\pi e \mathbb{E}_{W,\hat{W}}[(W_i - \hat{W}_i)^2]\right) \right\} \\ = \;& \min\left\{ h(W_i), \frac{1}{2}\log(2\pi e \mathbb{E}_{W,\hat{W}}[(W_i - \hat{W}_i)^2]) \right\}. \end{aligned} \tag{8.30}$$

Therefore, the lower bound of the mutual information is given by,

$$I(W;\hat{W}) \geq h(W) - \sum_{i=1}^{m} \min\left\{ h(W_i), \frac{1}{2}\log(2\pi e \mathbb{E}_{W,\hat{W}}[(W_i - \hat{W}_i)^2]) \right\}. \tag{8.31}$$

8.6.2 Lower bound for linear model

For complex models, the general lower bound in Lemma 31 is difficult to evaluate, due to the large dimension of parameters. It was shown by [3] that the sample complexity to estimate differential entropy is exponential to the dimension. It is even harder to design an algorithm to achieve the lower bound. But for linear model, the lower bound can be simplified. For

$f_w(x) = w^T x$, the distortion function $d(w, \hat{w})$ can be written as

$$\begin{aligned} & d(w, \hat{w}) = \mathbb{E}_X \left[(f_w(X) - f_{\hat{w}}(X))^2 \right] = \mathbb{E}_X \left[(w^T X - \hat{w}^T X)^2 \right] \\ = \; & \mathbb{E}_X \left[(w - \hat{w})^T X X^T (w - \hat{w}) \right] \\ = \; & (w - \hat{w})^T \mathbb{E}_X [X X^T] (w - \hat{w}). \end{aligned} \tag{8.32}$$

Since we assumed that $\mathbb{E}[X] = 0$, $\mathbb{E}[X_i^2] = \lambda_{x,i} > 0$ and $\mathbb{E}[X_i X_j] = 0$, so the constraint in Lemma 31 is given by

$$\begin{aligned} & D \geq \mathbb{E}_{W,\hat{W}} \left[(W - \hat{W})^T \mathbb{E}_X [X X^T] (W - \hat{W}) \right] \\ = \; & \sum_{i=1}^{m} \lambda_{x,i} \underbrace{\mathbb{E}_{W,\hat{W}} \left[(W_i - \hat{W}_i)^2 \right]}_{D_i}. \end{aligned} \tag{8.33}$$

Then the optimization problem in Lemma 31 can be written as follows,

$$\begin{aligned} \max_{p(\hat{w}|w)} \quad & \sum_{i=1}^{m} \min\{h(W_i), \frac{1}{2} \log(2\pi e D_i)\} \\ \text{s.t.} \quad & \sum_{i=1}^{m} \lambda_{x,i} D_i \leq D. \end{aligned} \tag{8.34}$$

Here W_i is a Gaussian random variable, so $h(W_i) = \frac{1}{2} \log(2\pi e \mathbb{E}[W_i^2])$. The Lagrangian function of the problem is given by

$$\begin{aligned} & \mathcal{L}(D_1, \ldots, D_m, \mu) \\ = \; & \sum_{i=1}^{m} \Big(\min\{\frac{1}{2} \log \mathbb{E}[W_i^2], \frac{1}{2} \log D_i\} + \frac{1}{2} \log(2\pi e) - \mu \lambda_{x,i} D_i \Big). \end{aligned} \tag{8.35}$$

By setting the derivative w.r.t. D_i to 0, we have

$$0 = \frac{\partial \mathcal{L}}{\partial D_i} = \frac{1}{2 D_i} - \mu \lambda_{x,i}, \tag{8.36}$$

for all D_i such that $D_i < \mathbb{E}[W_i^2]$. So the optimal D_i should satisfy that $D_i \lambda_{x,i}$ is constant, for all D_i such that $D_i < \mathbb{E}[W_i^2]$. Also the optimal D_i is at most $\mathbb{E}[W_i^2]$. Also, since $h(W) = \frac{m}{2} \log(2\pi e) + \frac{1}{2} \log \det(\Sigma_W)$ the lower bound is

given by

$$R(D) \geq \frac{1}{2}\log\det(\Sigma_W) - \sum_{i=1}^{m}\frac{1}{2}\log(D_i), \tag{8.37}$$

where

$$D_i = \begin{cases} \mu/\lambda_{x,i} & \text{if} \mu < \lambda_{x,i}\mathbb{E}_W[W_i^2], \\ \mathbb{E}_W[W_i^2] & \text{if} \mu \geq \lambda_{x,i}\mathbb{E}_W[W_i^2], \end{cases} \tag{8.38}$$

where μ is chosen that $\sum_{i=1}^{m}\lambda_{x,i}D_i = D$.

This lower bound gives rise to a "weighted water-filling", which differs from the classical "water-filling" for rate-distortion of colored Gaussian source in [71, Figure 13.7], since the water level's D_i are proportional to $1/\lambda_{x,i}$, which is related to the input of the model rather than the parameters to be compressed. To illustrate the "weighted water-filling" process, we choose a simple example where $\Sigma_W = \Sigma_X = \text{diag}[3, 2, 1]$. In Figure 8.8, the widths of each rectangle are proportional to $\lambda_{x,i}$, and the heights are proportional to $\Sigma_W = [3, 2, 1]$. The water level in each rectangle is D_i and the volume of water is μ. As D starts to increase from 0, each rectangle is filled with same volume of water (μ is the same), but the water level D_i's increase with speed $1/\lambda_{x,i}$ respectively (Figure 8.8.(a)). This gives segment (a) of the rate distortion curve in Figure 8.8.(d). If D is large enough such that the third rectangle is full, then D_3 is fixed to be $\mathbb{E}[W_3^2] = 1$, whereas D_1 and D_2 continuously increase (Figure 8.8.(b)). This gives segment (b) in Figure 8.8.(d). Keep increasing D until the second rectangle is also full, then D_2 is fixed to be $\mathbb{E}[W_2^2] = 2$ and D_1 continuous increasing (Figure 8.8 (c)). This gives segment (c) in Figure 8.8.(d). The entire rate-distortion function is shown in Figure 8.8(d), where the first red dot corresponds to the moment that the third rectangle is exactly full, and the second red dot corresponds to moment that the second rectangle is exactly full.

8.6.3 Achievability

We prove that this lower bound is achievable. To achieve the lower bound, we construct the compression algorithm in Algorithm 3,

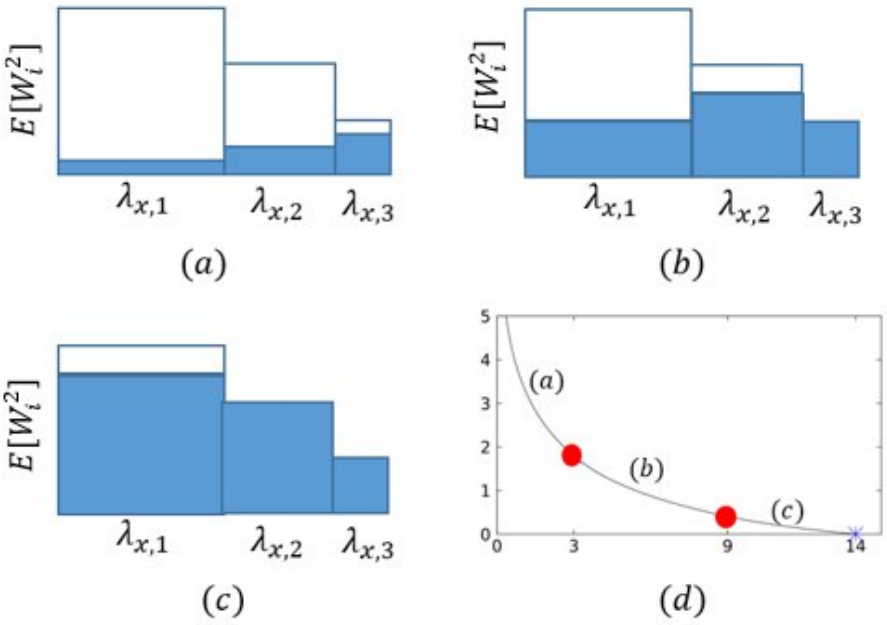

Figure 8.8: Illustration of "weighted water-filling" process.

Algorithm 3 Optimal compression algorithm for linear regression

Input: Distortion D, covariance matrix of parameters Σ_W, covariance matrix of data $\Sigma_X = \text{diag}[\lambda_{x,1}, \ldots, \lambda_{x,m}]$;
Choose D_i's such that

$$D_i = \begin{cases} \mu/\lambda_{x,i} & \text{if} \mu < \lambda_{x,i}\mathbb{E}_W[W_i^2], \\ \mathbb{E}_W[W_i^2] & \text{if} \mu \geq \lambda_{x,i}\mathbb{E}_W[W_i^2], \end{cases}$$

where $\sum_{i=1}^m \lambda_{x,i} D_i = D$;
for $i = 1$ to m **do**
  if $D_i = \mu/\lambda_{x,i}$ **then**
    Choose $\hat{W}_i = 0$;
  else
    Choose a conditional distribution $P_{\hat{W}_i|W_i}$ such that $W_i = \hat{W} + Z_i$ where $Z_i \sim \mathcal{N}(0, D_i)$, $\hat{W}_i \sim \mathcal{N}(0, \mathbb{E}_W[W_i^2] - D_i)$ and $\hat{W}_i$ is independent of Z_i;
  end if
end for
Combine the conditional probability distributions by $P_{\hat{W}|W} = \prod_{i=1}^m P_{\hat{W}_i|W_i}$.

Intuitively, the optimal compressor does the following: (1) Find the optimal water levels D_i for "weighted water filling". (2) For the entries where the corresponding rectangles are full, simply discard the entries. (3) For the entries where the corresponding rectangles are not full, add a noise which is independent of $\hat{W}_i$ and has a variance proportional to the water level. That is possible since W is Gaussian. (4) Combine the conditional probabilities.

To see that this compressor is optimal, we will check that the compressor makes all the inequalities become equality. Here is all the inequalities used in the proof.

- $h(W_i \,|\, W_1, \ldots, W_{i-1}, \hat{W}_i, \ldots, \hat{W}_m) \leq h(W_i|\hat{W}_i)$ for all $i = 1...m$. It becomes equality by $P_{\hat{W}|W} = \prod_{i=1}^m P_{\hat{W}_i|W}$.
- Either
  - $h(W_i|\hat{W}_i) \leq h(W_i)$. It becomes equality for those $\hat{W}_i = 0$.
  - $h(W_i - \hat{W}_i|\hat{W}_i) \leq h(W_i - \hat{W}_i) \leq \frac{1}{2}\log(2\pi e \mathbb{E}_{W,\hat{W}}[(W_i - \hat{W})^2])$. It becomes equality for those $\hat{W}_i$'s such that $W_i - \hat{W}_i$ is independent of $\hat{W}_i$ and $W_i - \hat{W}_i$ is Gaussian.
- The "water levels" D_i. It becomes equality by choosing the D_i's according to Lagrangian conditions.

Therefore, Algorithm 3 gives a compressor $P^{(D)}_{\hat{W}|W}$ such that $\mathbb{E}_{P_W \circ P^{(D)}_{\hat{W}|W}}[d(W, \hat{W})] = D$ and $I(W;\hat{W}) = \underline{R}(D)$, hence the lower bound is tight.

8.7 Proof of results in Chapter 8

In this section, we provide the proof of Theorem 22. For simplicity let $\sigma(t) = t\mathbb{I}\{t \geq 0\}$ denotes the ReLU activation function. First we deal with the objective of the compression algorithm,

$$\begin{aligned}
& (w - \hat{w})^T I_w (w - \hat{w}) \\
= \; & (w - \hat{w})^T \mathbb{E}_X \left[\nabla_w f_w(x) \nabla_w f_w(x)^T\right] (w - \hat{w}) \\
= \; & (w - \hat{w})^T \mathbb{E}_X \left[\nabla_w \sigma(w^T x) \nabla_w \sigma(w^T x)^T\right] (w - \hat{w}) \\
= \; & (w - \hat{w})^T \mathbb{E}_X \left[x^T (\sigma'(w^T x))^2 x\right] (w - \hat{w}) \\
= \; & \mathbb{E}_X \left[\mathbb{I}\{w^T x \geq 0\}((w - \hat{w})^T x)^2\right]. \qquad (8.39)
\end{aligned}$$

Notice that x is jointly Gaussian random variable with zero mean and non-degenerate variance, so the distribution of x is equivalent to the distribution of $-x$. Therefore,

$$\begin{aligned}
&\mathbb{E}_X[\mathbb{I}\{w^T x \geq 0\}((w-\hat{w})^T x)^2] = \int_{x:w^T x\geq 0}((w-\hat{w}^T)x)^2 dx \\
= \;& \frac{1}{2}\left(\int_{x:w^T x\geq 0}((w-\hat{w}^T)x)^2 dx + \int_{x:w^T x\leq 0}((w-\hat{w}^T)x)^2 dx\right) \\
= \;& \frac{1}{2}\int_{x\in\mathbb{R}}^{d}((w-\hat{w}^T)x)^2 dx = \frac{1}{2}(w-\hat{w})^T\Sigma_X(w-\hat{w}). \qquad (8.40)
\end{aligned}$$

So minimizing the gradient-squared based loss is equivalent to minimizing $(w-\hat{w})^T\Sigma_X(w-\hat{w})$. Similarly, the condition $\hat{w}I_w(w-\hat{w})=0$ is equivalent to $\hat{w}\Sigma_X(w-\hat{w})=0$. Now we deal with the MSE loss function $\mathbb{E}[(f_w(x)-f_{\hat{w}}(x))^2]$. We utilize the Hermite polynomials and Fourier analysis on Gaussian space. We use the following key lemma.

Lemma 32. *([14, Claim 4.3]) Let f, g be two functions from $\mathbb{R}$ to $\mathbb{R}$ such that $f^2, g^2 \in L^2(\mathbb{R}, e^{-x^2/2})$. The for any unit vectors u, v, we have that*

$$\mathbb{E}_{x\in\mathcal{N}(0,I_{d\times d})}[f(u^T x)g(v^T x)] = \sum_{p=0}^{\infty}\hat{f}_p\hat{g}_p(u^T v)^p, \qquad (8.41)$$

where $\hat{f}_p = \mathbb{E}_{x\in\mathcal{N}(0,1)}[f(x)h_p(x)]$ is the p-th order coefficient of f, where h_p is the p-th order probabilists' Hermite polynomial.

Since $X\sim\mathcal{N}(0,\Sigma_X)$, we can write $x=\Sigma_X^{1/2}z$, where $z\sim\mathcal{N}(0,I_d)$. So for any compressed weight $\hat{w}$, we have

$$\begin{aligned}
&\mathbb{E}_X\left[(f_w(x)-f_{\hat{w}}(x))^2\right] = \mathbb{E}_X\left[(\sigma(w^T x)-\sigma(\hat{w}^T x))^2\right] \\
= \;& \mathbb{E}_{z\in\mathcal{N}(0,I_d)}[(\sigma(w^T\Sigma_X^{1/2}z)-\sigma(\hat{w}^T\Sigma_X^{1/2}z))^2] \\
= \;& \mathbb{E}_{z\in\mathcal{N}(0,I_d)}[\sigma(w^T\Sigma_X^{1/2}z)^2] - 2\mathbb{E}_{z\in\mathcal{N}(0,I_d)}[\sigma(w^T\Sigma_X^{1/2}z)\sigma(\hat{w}^T\Sigma_X^{1/2}z)] \\
& +\mathbb{E}_{z\in\mathcal{N}(0,I_d)}[\sigma(\hat{w}^T\Sigma_X^{1/2}z)^2] \\
= \;& \sum_{p=0}^{\infty}\hat{\sigma}_p^2(w^T\Sigma_X w)^p - 2\sum_{p=0}^{\infty}\hat{\sigma}_p^2(w^T\Sigma_X\hat{w})^p + \sum_{p=0}^{\infty}\hat{\sigma_p}^2(\hat{w}^T\Sigma_X\hat{w})^p \\
= \;& \sum_{p=0}^{\infty}\hat{\sigma}_p^2\left(\underbrace{(w^T\Sigma_X w)^p - 2(w^T\Sigma_X\hat{w})^p + (\hat{w}^T\Sigma_X\hat{w})^p}_{D_p(w,\hat{w})}\right). \qquad (8.42)
\end{aligned}$$

Now we can see that $D_0(w, \hat{w}) = 0$. $D_1(w, \hat{w}) = w^T \Sigma_X w - 2w^T \Sigma_X \hat{w} + \hat{w}^T \Sigma_X w = (w - \hat{w})^T \Sigma_X (w - \hat{w})$, is just the objective. The following lemma gives the minimizer of $D_p(w, \hat{w})$ for higher order p.

Lemma 33. *If $\hat{w}^*$ satisfies $\hat{w}^* \Sigma_X (\hat{w} - w) = 0$ and $\hat{w}^* = \arg\min_{\hat{s} \in \mathcal{W}} D_1(w, \hat{w})$ for some constrained set $\mathcal{W}$. Then for any $p \geq 2$ and even, we have $\hat{w}^* = \arg\min_{\hat{w} \in \mathcal{W}} D_p(w, \hat{w})$.*

For ReLU function, the coefficients are $\hat{\sigma}_0 = \frac{1}{\sqrt{2\pi}}$, $\hat{\sigma}_1 = \frac{1}{2}$. For $p \geq 2$ and even, $\hat{\sigma}_p = \frac{((p-3)!!)^2}{\sqrt{2\pi p!}}$. For $p \geq 3$ and odd, $\hat{\sigma}_p = 0$. Since the coefficients $\hat{\sigma}_p$ is zero for $p \geq 3$ and odd, so if a compressed weight $\hat{w}$ satisfied $\hat{w} \Sigma_X (\hat{w} - w) = 0$ and minimizes $D_1(\hat{w}, w) = (\hat{w} - w)^T \Sigma_X (\hat{w} - w)$, then it is the minimizer for all $D_p(w, \hat{w})$ for even p, therefore a minimizer of the MSE loss.

8.7.1 Proof of Lemma 33

For simplicity of notation, define $A = w^T \Sigma_X w$, $B = \hat{w}^T \Sigma_X (\hat{w} - w)$ and $C = D_1(w, \hat{w}) = (\hat{w} - w)^T \Sigma_X (\hat{w} - w)$. For all compressors, we have $C \leq A$. Therefore, $w^T \Sigma_X \hat{w} = A + B - C$ and $\hat{w}^T \Sigma_X \hat{w} = A + 2B - C$. So

$$D_p(w, \hat{w}) \quad = \quad A^p - 2(A + B - C)^p + (A + 2B - C)^p. \tag{8.43}$$

First notice that

$$\frac{\partial D_p(w, \hat{w})}{\partial B} = 2p((A + 2B - C)^{p-1} - (A + B - C)^{p-1}). \tag{8.44}$$

For even $p \geq 2$, x^{p-1} is monotonically increasing, so $(A + 2B - C)^{p-1} > (A + B - C)^{p-1}$ if $B > 0$ and vice versa. Therefore, for fixed A and C, $D_p(w, \hat{w})$ is monotonically increasing for positive B and decreasing for negative B. Therefore, $D_p(w, \hat{w})$ is minimized when $B = 0$, and the minimal value is $D_p(w, \hat{w}) = A^p - 2(A - C)^p + (A - C)^p = A^p - (A - C)^p$, which is monotonically increasing with respect to C. So if $\hat{w}^*$ satisfies $B = 0$ and is a minimzer of $C = D_1(w, \hat{w})$, it is also a minimizer for $D_p(w, \hat{w})$ for all $p \geq 2$ and even.

CHAPTER 9

CONCLUSION

In this dissertation, we investigated various aspects of application of information theory at the sample level, by studying the theoretical properties of k-nearest neighbor based information-theoretic quantity estimators, proposing new algorithms and measures for discovering complex relationships among data, and improving deep learning algorithms via information theory.

We have shown that k-nearest neighbor estimator of differential entropy has near optimal convergence rate, and KSG estimator of mutual information has theoretical guarantee. We proposed new estimators of differential entropy when the data has smaller intrinsic dimension, and new estimators of mutual information when the data is a mixture of discrete and continuous regimes. We proposed using hypercontractivity to discover underlying relationships among data and provided corresponding estimator of hypercontractivity. Finally, we improved SGD training algorithm and model compression algorithms of deep learning based on our understanding of information theory.

Our work suggested that information theory does not only provide a mathematical understanding of information, but also has a wide usage in the era of big data. Information theory can be helpful in the area of sociology, computational biology and neural networks by studying information theory at a sample level. By building the bridge between information theory and big data, we can further apply the spirit of information theory in many emerging research areas in the future.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] L. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, 1987.

[3] J. Jiao, W. Gao, and Y. Han, "The nearest neighbor information estimator is adaptively near minimax rate-optimal," in *Advances in Neural Information Processing Systems*, 2018, pp. 3156–3167.

[4] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, 2004.

[5] W. Gao, S. Oh, and P. Viswanath, "Demystifying fixed k-nearest neighbor information estimators," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5629–5661, 2018.

[6] N. L. Hjort and M. C. Jones, "Locally parametric nonparametric density estimation," *The Annals of Statistics*, pp. 1619–1647, 1996.

[7] C. R. Loader, "Local likelihood density estimation," *The Annals of Statistics*, vol. 24, no. 4, pp. 1602–1618, 1996.

[8] W. Gao, S. Oh, and P. Viswanath, "Breaking the bandwidth barrier: Geometrical adaptive entropy estimation," in *Advances in Neural Information Processing Systems*, 2016, pp. 2460–2468.

[9] W. Gao, S. Kannan, S. Oh, and P. Viswanath, "Estimating mutual information for discrete-continuous mixtures," in *Advances in Neural Information Processing Systems*, 2017, pp. 5988–5999.

[10] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," *arXiv preprint arXiv:1304.6133*, 2013.

[11] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[12] H. Kim, W. Gao, S. Kannan, S. Oh, and P. Viswanath, "Discovering potential correlations via hypercontractivity," in *Advances in Neural Information Processing Systems*, 2017, pp. 4577–4587.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[14] R. Ge, J. D. Lee, and T. Ma, "Learning one-hidden-layer neural networks with landscape design," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=BkwHObbRZ

[15] M. Janzamin, H. Sedghi, and A. Anandkumar, "Score function features for discriminative learning: Matrix and tensor framework," *arXiv preprint arXiv:1412.2863*, 2014.

[16] W. Gao, A. Makkuva, S. Oh, and P. Viswanath, "Learning one-hidden-layer neural networks under general input distributions," in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 1950–1959.

[17] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[18] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Advances in Neural Information Processing Systems*, 2015, pp. 1135–1143.

[19] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec*, vol. 4, no. 142-163, p. 1, 1959.

[20] W. Gao, Y.-H. Liu, C. Wang, and S. Oh, "Rate distortion for model compression: From theory to practice," in *International Conference on Machine Learning*, 2019, pp. 2102–2111.

[21] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226–1238, 2005.

[22] A. C. Müller, S. Nowozin, and C. H. Lampert, "Information theoretic clustering using minimum spanning trees," in *Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium*. Springer, 2012, pp. 205–215.

[23] G. Ver Steeg and A. Galstyan, "Maximally informative hierarchical representations of high-dimensional data," in *Artificial Intelligence and Statistics*, 2015, pp. 1004–1012.

[24] C. Chan, A. Al-Bashabsheh, J. B. Ebrahimi, T. Kaced, and T. Liu, "Multivariate mutual information inspired by secret-key agreement," *Proceedings of the IEEE*, vol. 103, no. 10, pp. 1883–1913, 2015.

[25] P. Li and O. Milenkovic, "Inhomogeneous hypergraph clustering with applications," in *Advances in Neural Information Processing Systems*, 2017, pp. 2308–2318.

[26] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[27] F. Fleuret, "Fast binary feature selection with conditional mutual information," *The Journal of Machine Learning Research*, vol. 5, pp. 1531–1555, 2004.

[28] W. Gao, S. Kannan, S. Oh, and P. Viswanath, "Conditional dependence via Shannon capacity: Axioms, estimators and applications," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. JMLR. org, 2016, pp. 2780–2789.

[29] S. Krishnaswamy, M. H. Spitzer, M. Mingueneau, S. C. Bendall, O. Litvin, E. Stone, D. Pe'er, and G. P. Nolan, "Conditional density-based analysis of T cell signaling in single-cell data," *Science*, vol. 346, no. 6213, p. 1250689, 2014.

[30] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. Van der Meulen, "Nonparametric entropy estimation: An overview," *International Journal of Mathematical and Statistical Sciences*, vol. 6, no. 1, pp. 17–39, 1997.

[31] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman et al., "Nonparametric von Mises estimators for entropies, divergences and mutual informations," in *Advances in Neural Information Processing Systems*, 2015, pp. 397–405.

[32] A. B. Tsybakov and E. Van der Meulen, "Root-n consistent estimators of entropy for densities with unbounded support," *Scandinavian Journal of Statistics*, pp. 75–83, 1996.

[33] K. Sricharan, R. Raich, and A. O. Hero, "Estimation of nonlinear functionals of densities with confidence," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4135–4159, 2012.

[34] S. Singh and B. Póczos, "Finite-sample analysis of fixed-k nearest neighbor density functional estimators," in *Advances in Neural Information Processing Systems*, 2016, pp. 1217–1225.

[35] S. Delattre and N. Fournier, "On the Kozachenko–Leonenko entropy estimator," *Journal of Statistical Planning and Inference*, vol. 185, pp. 69–93, 2017.

[36] T. B. Berrett, R. J. Samworth, M. Yuan et al., "Efficient multivariate entropy estimation via k-nearest neighbour distances," *The Annals of Statistics*, vol. 47, no. 1, pp. 288–318, 2019.

[37] P. Hall and J. S. Marron, "Estimation of integrated squared density derivatives," *Statistics & Probability Letters*, vol. 6, no. 2, pp. 109–115, 1987.

[38] P. J. Bickel and Y. Ritov, "Estimating integrated squared density derivatives: sharp best order of convergence estimates," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 381–393, 1988.

[39] D. L. Donoho and M. Nussbaum, "Minimax quadratic estimation of a quadratic functional," *Journal of Complexity*, vol. 6, no. 3, pp. 290–323, 1990.

[40] J. Fan, "On the estimation of quadratic functionals," *The Annals of Statistics*, pp. 1273–1294, 1991.

[41] L. Birgé and P. Massart, "Estimation of integral functionals of a density," *The Annals of Statistics*, pp. 11–29, 1995.

[42] G. Kerkyacharian and D. Picard, "Estimating nonquadratic functionals of a density using Haar wavelets," *The Annals of Statistics*, vol. 24, no. 2, pp. 485–507, 1996.

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

[44] A. Nemirovski, "Topics in non-parametric," *Ecole dEté de Probabilités de Saint-Flour*, vol. 28, p. 85, 2000.

[45] T. T. Cai, M. G. Low et al., "A note on nonparametric estimation of linear functionals," *The Annals of Statistics*, vol. 31, no. 4, pp. 1140–1153, 2003.

[46] T. T. Cai and M. G. Low, "Nonquadratic estimators of a quadratic functional," *The Annals of Statistics*, pp. 2930–2956, 2005.

[47] E. Tchetgen, L. Li, J. Robins, and A. van der Vaart, "Minimax estimation of the integral of a power of a density," *Statistics & Probability Letters*, vol. 78, no. 18, pp. 3307–3311, 2008.

[48] O. Lepski, A. Nemirovski, and V. Spokoiny, "On estimation of the L_r norm of a regression function," *Probability Theory and Related Fields*, vol. 113, no. 2, pp. 221–253, 1999.

[49] Y. Han, J. Jiao, T. Weissman, and Y. Wu, "Optimal rates of entropy estimation over Lipschitz balls," *arXiv preprint arXiv:1711.02141*, 2017.

[50] Y. Han, J. Jiao, R. Mukherjee, and T. Weissman, "On estimation of L_r-norms in Gaussian white noise models," *arXiv preprint arXiv:1710.03863*, 2017.

[51] G. Biau and L. Devroye, *Lectures on the Nearest Neighbor Method*. Springer, 2015.

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

[53] P. Hall, "Limit theorems for sums of general functions of m-spacings," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 96, no. 3. Cambridge University Press, 1984, pp. 517–532.

[54] H. Joe, "Estimation of entropy and other functionals of a multivariate density," *Annals of the Institute of Statistical Mathematics*, vol. 41, no. 4, pp. 683–697, 1989.

[55] B. Van Es, "Estimating functionals related to a density by a class of statistics based on spacings," *Scandinavian Journal of Statistics*, pp. 61–72, 1992.

[56] P. Hall and S. C. Morton, "On the estimation of entropy," *Annals of the Institute of Statistical Mathematics*, vol. 45, no. 1, pp. 69–88, 1993.

[57] B. Y. Levit, "Asymptotically efficient estimation of nonlinear functionals," *Problemy Peredachi Informatsii*, vol. 14, no. 3, pp. 65–72, 1978.

[58] F. El Haje Hussein and Y. Golubev, "On entropy estimation by m-spacing method," *Journal of Mathematical Sciences*, vol. 163, no. 3, pp. 290–309, 2009.

[59] J. Robins, L. Li, E. Tchetgen, and A. van der Vaart, "Higher order influence functions and minimax estimation of nonlinear functionals," in *Probability and Statistics: Essays in Honor of David A. Freedman*. Institute of Mathematical Statistics, 2008, pp. 335–421.

[60] J. Robins, L. Li, R. Mukherjee, E. T. Tchetgen, and A. van der Vaart, "Higher order estimating equations for high-dimensional models," *Annals of Statistics*, vol. 45, no. 5, p. 1951, 2017.

[61] R. Mukherjee, W. K. Newey, J. Robins et al., "Semiparametric efficient empirical higher order influence function estimators," Centre for Microdata Methods and Practice, Institute for Fiscal Studies, Tech. Rep., 2017.

[62] A. Tsybakov, *Introduction to Nonparametric Estimation*. Springer-Verlag, 2008.

[63] O. Lepskii, "On a problem of adaptive estimation in Gaussian white noise," *Theory of Probability & Its Applications*, vol. 35, no. 3, pp. 454–466, 1991.

[64] E. Giné and R. Nickl, "A simple adaptive estimator of the integrated square of a density," *Bernoulli*, pp. 47–61, 2008.

[65] R. Mukherjee, E. Tchetgen Tchetgen, and J. Robins, "Lepski's method and adaptive estimation of nonlinear integral functionals of density," *arXiv preprint arXiv:1508.00249*, 2015.

[66] R. Mukherjee, E. T. Tchetgen, and J. Robins, "On adaptive estimation of nonparametric functionals," *arXiv preprint arXiv:1608.01364*, 2016.

[67] A. Krishnamurthy, K. Kandasamy, B. Poczos, and L. Wasserman, "Nonparametric estimation of Rényi divergence and friends," in *International Conference on Machine Learning*, 2014, pp. 919–927.

[68] R. J. Karunamuni and T. Alberts, "On boundary correction in kernel density estimation," *Statistical Methodology*, vol. 2, no. 3, pp. 191–212, 2005.

[69] B. Efron and C. Stein, "The jackknife estimate of variance," *The Annals of Statistics*, pp. 586–596, 1981.

[70] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*. Chapman and Hall/CRC, 2015.

[71] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[72] J. Jiao, T. A. Courtade, K. Venkat, and T. Weissman, "Justification of logarithmic loss via the benefit of side information," *IEEE Transactions on Information Theory*, vol. 61, no. 10, pp. 5357–5365, 2015.

[73] W. M. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Medical Image Analysis*, vol. 1, no. 1, pp. 35–51, 1996.

[74] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 417–424.

[75] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2392–2405, 2009.

[76] L. Paninski, "Estimating entropy on m bins given fewer than m samples," *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 2200–2203, 2004.

[77] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3702–3720, 2016.

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

[79] S. Gao, G. Ver Steeg, and A. Galstyan, "Efficient estimation of mutual information for strongly dependent variables," in *Artificial Intelligence and Statistics*, 2015, pp. 277–286.

[80] J. B. Kinney and G. S. Atwal, "Equitability, mutual information, and the maximal information coefficient," *Proceedings of the National Academy of Sciences*, vol. 111, no. 9, pp. 3354–3359, 2014.

[81] G. Valiant and P. Valiant, "Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs," in *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*. ACM, 2011, pp. 685–694.

[82] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing that distributions are close," in *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. IEEE, 2000, pp. 259–269.

[83] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi, "Estimating Rényi entropy of discrete distributions," *IEEE Transactions on Information Theory*, vol. 63, no. 1, pp. 38–56, 2016.

[84] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[85] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2835–2885, 2015.

[86] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson III, V. Protopopescu, and G. Ostrouchov, "Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data," *Physical Review E*, vol. 76, no. 2, p. 026209, 2007.

[87] S. Gao, G. V. Steeg, and A. Galstyan, "Estimating mutual information by local Gaussian approximation," in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2015, pp. 278–287.

[88] C. A. Rogers, *Hausdorff Measures*. Cambridge University Press, 1998.

[89] J. Zhu, J.-J. Bellanger, H. Shu, C. Yang, and R. L. B. Jeannès, "Bias reduction in the estimation of mutual information," *Physical Review E*, vol. 90, no. 5, p. 052714, 2014.

[90] G. Ver Steeg and A. Galstyan, "Discovering structure in high-dimensional data through correlation explanation," in *Advances in Neural Information Processing Systems*, 2014, pp. 577–585.

[91] C. Chan and T. Liu, "Clustering by multivariate mutual information under Chow-Liu tree approximation," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2015, pp. 993–999.

[92] G. V. Steeg and A. Galstyan, "The information sieve," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. JMLR. org, 2016, pp. 164–172.

[93] I. Csiszár and P. Narayan, "Secrecy capacities for multiple terminals," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3047–3061, 2004.

[94] P. Narayan and H. Tyagi, "Multiterminal secrecy by public discussion," *Foundations and Trends® in Communications and Information Theory*, vol. 13, no. 2-3, pp. 129–275, 2016.

[95] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf, "Quantifying causal influences," *The Annals of Statistics*, vol. 41, no. 5, pp. 2324–2358, 2013.

[96] A. V. Makkuva and Y. Wu, "On additive-combinatorial affine inequalities for Shannon entropy and differential entropy," in *2016 IEEE International Symposium on Information Theory*. IEEE, 2016, pp. 1053–1057.

[97] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Routledge, 2018.

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

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

[100] L. Paninski and M. Yajima, "Undersmoothed kernel entropy estimators," *IEEE Transactions on Information Theory*, vol. 54, no. 9, pp. 4384–4388, 2008.

[101] H. Liu, L. Wasserman, and J. D. Lafferty, "Exponential concentration for mutual information estimation with application to forests," in *Advances in Neural Information Processing Systems*, 2012, pp. 2537–2545.

[102] S. Singh and B. Póczos, "Exponential concentration of a density functional estimator," in *Advances in Neural Information Processing Systems*, 2014, pp. 3032–3040.

[103] S. Singh and B. Póczos, "Generalized exponential concentration inequality for Rényi divergence estimation," in *International Conference on Machine Learning*, 2014, pp. 333–341.

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

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

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

[107] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *The Annals of Mathematical Statistics*, vol. 36, no. 3, pp. 1049–1051, 1965.

[108] E. Fix and J. L. Hodges Jr, "Discriminatory analysis-nonparametric discrimination: Consistency properties," California Univ Berkeley, Tech. Rep., 1951.

[109] G. Biau, F. Chazal, D. Cohen-Steiner, L. Devroye, and C. Rodriguez, "A weighted k-nearest neighbor density estimate for geometric inference," *Electronic Journal of Statistics*, vol. 5, pp. 204–237, 2011.

[110] S. Kpotufe and U. von Luxburg, "Pruning nearest neighbor cluster trees," in *28th International Conference on Machine Learning (ICML 2011)*. International Machine Learning Society, 2011, pp. 225–232.

[111] S. Dasgupta and S. Kpotufe, "Optimal rates for k-NN density and mode estimation," in *Advances in Neural Information Processing Systems*, 2014, pp. 2555–2563.

[112] Y. Bengio, P. Vincent et al., "Locally weighted full covariance gaussian density estimation," CIRANO, Tech. Rep., 2004.

[113] D. Pál, B. Póczos, and C. Szepesvári, "Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs," in *Advances in Neural Information Processing Systems*, 2010, pp. 1849–1857.

[114] F. Pérez-Cruz, "Estimation of information theoretic measures for continuous random variables," in *Advances in Neural Information Processing Systems*, 2009, pp. 1257–1264.

[115] D. Lombardi and S. Pant, "Nonparametric k-nearest-neighbor entropy estimator," *Physical Review E*, vol. 93, no. 1, p. 013310, 2016.

[116] K. Sricharan, D. Wei, and A. O. Hero, "Ensemble estimators for multivariate entropy estimation," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4374–4388, 2013.

[117] K. R. Moon and A. O. Hero, "Ensemble estimation of multivariate f-divergence," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 356–360.

[118] K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero, "Improving convergence of divergence functional ensemble estimators," in *2016 IEEE International Symposium on Information Theory*. IEEE, 2016, pp. 1133–1137.

[119] K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero III, "Nonparametric ensemble estimation of distributional functionals," *arXiv preprint arXiv:1601.06884*, 2016.

[120] K. R. Moon, K. Sricharan, and A. O. Hero, "Ensemble estimation of mutual information," in *2017 IEEE International Symposium on Information Theory*. IEEE, 2017, pp. 3030–3034.

[121] H. A. David and H. N. Nagaraja, "Order statistics," *Encyclopedia of Statistical Sciences*, 2004.

[122] C. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," *Natural Language Engineering*, vol. 16, no. 1, pp. 100–103, 2010.

[123] Q. Wang, S. R. Kulkarni, and S. Verdú, "Universal estimation of information measures for analog sources," *Foundations and Trends® in Communications and Information Theory*, vol. 5, no. 3, pp. 265–353, 2009.

[124] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, "Nearest neighbor estimates of entropy," *American Journal of Mathematical and Management Sciences*, vol. 23, no. 3-4, pp. 301–321, 2003.

[125] L. Wasserman, *All of Nonparametric Statistics*. Springer Science & Business Media, 2006.

[126] C. Loader, *Local Regression and Likelihood*. Springer Science & Business Media, 2006.

[127] R.-D. Reiss, *Approximate Distributions of Order Statistics: With Applications to Nonparametric Statistics*. Springer Science & Business Media, 2012.

[128] S. J. Sheather, "Density estimation," *Statistical Science*, pp. 588–597, 2004.

[129] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. Novi Inverardi, "A new class of random vector entropy estimators and its applications in testing statistical hypotheses," *Journal of Nonparametric Statistics*, vol. 17, no. 3, pp. 277–297, 2005.

[130] N. Leonenko, L. Pronzato, V. Savani et al., "A class of Rényi information estimators for multidimensional densities," *The Annals of Statistics*, vol. 36, no. 5, pp. 2153–2182, 2008.

[131] S. Singh and B. Póczos, "Analysis of k-nearest neighbor distances with application to entropy estimation," *arXiv preprint arXiv:1603.08578*, 2016.

[132] P. J. Bickel, L. Breiman et al., "Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test," *The Annals of Probability*, vol. 11, no. 1, pp. 185–214, 1983.

[133] P. Hall, "On powerful distributional tests based on sample spacings," *Journal of Multivariate Analysis*, vol. 19, no. 2, pp. 201–224, 1986.

[134] R. Mnatsakanov, N. Misra, S. Li, and E. Harner, "K_n-nearest neighbor estimators of entropy," *Mathematical Methods of Statistics*, vol. 17, no. 3, pp. 261–277, 2008.

[135] A. Ozakin and A. G. Gray, "Submanifold density estimation," in *Advances in Neural Information Processing Systems*, 2009, pp. 1375–1382.

[136] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[137] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[138] F. Rieke, D. Warland, R. d. R. Van Steveninck, W. S. Bialek et al., *Spikes: Exploring the Neural Code*. MIT press Cambridge, 1999, vol. 7, no. 1.

[139] S. Singh and B. Póczos, "Nonparanormal information estimation," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3210–3219.

[140] Y. Han, J. Jiao, and T. Weissman, "Adaptive estimation of Shannon entropy," in *2015 IEEE International Symposium on Information Theory*. IEEE, 2015, pp. 1372–1376.

[141] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and Bayesian networks," in *Convexity and Concentration*. Springer, 2017, pp. 211–249.

[142] J. Acharya, H. Das, A. Orlitsky, and A. T. Suresh, "A unified maximum likelihood approach for estimating symmetric properties of discrete distributions," in *International Conference on Machine Learning*, 2017, pp. 11–21.

[143] Y. Han, J. Jiao, and T. Weissman, "Minimax rate-optimal estimation of divergences between discrete distributions," *arXiv preprint arXiv:1605.09124*, 2016.

[144] J. Jiao, K. Venkat, and T. Weissman, "Non-asymptotic theory for the plug-in rule in functional estimation," *arXiv preprint arXiv:1406.6959*, 2014.

[145] A. Perez, "Information theory with abstract alphabets," *Theory of Probability and its Applications*, vol. 4, no. 1, 1959.

[146] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.

[147] B. C. Ross, "Mutual information between discrete and continuous data sets," *PloS One*, vol. 9, no. 2, p. e87357, 2014.

[148] Z. Szabó, "Information theoretical estimators toolbox," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 283–287, 2014.

[149] A. R. Wu, N. F. Neff, T. Kalisky, P. Dalerba, B. Treutlein, M. E. Rothenberg, F. M. Mburu, G. L. Mantalas, S. Sim, M. F. Clarke et al., "Quantitative assessment of single-cell RNA-sequencing methods," *Nature Methods*, vol. 11, no. 1, p. 41, 2014.

[150] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nature Methods*, vol. 11, no. 7, p. 740, 2014.

[151] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic et al., "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data," *Genome Biology*, vol. 16, no. 1, p. 278, 2015.

[152] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, G. Stolovitzky et al., "Wisdom of crowds for robust gene network inference," *Nature Methods*, vol. 9, no. 8, pp. 796–804, 2012.

[153] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.

[154] H. O. Hirschfeld, "A connection between correlation and contingency," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 31, no. 4. Cambridge University Press, 1935, pp. 520–524.

[155] H. Gebelein, "Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung," *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 21, no. 6, pp. 364–379, 1941.

[156] A. Rényi, "On measures of dependence," *Acta Mathematica Hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[157] G. J. Székely, M. L. Rizzo, N. K. Bakirov et al., "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[158] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[159] I. S. Dhillon, S. Mallela, and R. Kumar, "A divisive information-theoretic feature clustering algorithm for text classification," *Journal of Machine Learning Research (JMLR)*, vol. 3, 2003.

[160] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, "Distributional word clusters vs. words for text categorization," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1183–1208, 2003.

[161] E. Erkip and T. M. Cover, "The efficiency of investment information," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1026–1040, 1998.

[162] E. B. Davies, L. Gross, and B. Simon, "Hypercontractivity: A bibliographic review," *Ideas and Methods in Quantum and Statistical Physics (Oslo, 1988)*, pp. 370–389, 1992.

[163] E. Nelson, "Construction of quantum fields from Markoff fields," *Journal of Functional Analysis*, vol. 12, no. 1, pp. 97–112, 1973.

[164] J. Kahn, G. Kalai, and N. Linial, "The influence of variables on Boolean functions," in *Proceedings of the 29th Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, 1988, pp. 68–80.

[165] R. O'Donnell, *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[166] W. Beckner, "Inequalities in Fourier analysis." Ph.D. dissertation, Princeton., 1975.

[167] L. Gross et al., "Hypercontractivity and logarithmic Sobolev inequalities for the Clifford-Dirichlet form," *Duke Mathematical Journal*, vol. 42, no. 3, pp. 383–396, 1975.

[168] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the Markov operator," *The Annals of Probability*, pp. 925–939, 1976.

[169] E. Mossel, K. Oleszkiewicz, and A. Sen, "On reverse hypercontractivity," *Geometric and Functional Analysis*, vol. 23, no. 3, pp. 1062–1097, 2013.

[170] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[171] A. Achille and S. Soatto, "Information dropout: Learning optimal representations through noisy computation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2897–2905, 2018.

[172] C. Nair, "An extremal inequality related to hypercontractivity of Gaussian random variables," in *Information Theory and Applications Workshop*, 2014.

[173] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 689–696.

[174] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Advances in Neural Information Processing Systems*, 2012, pp. 2222–2230.

[175] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.

[176] A. Makur and L. Zheng, "Linear bounds between contraction coefficients for f-divergences," *arXiv preprint arXiv:1510.01844*, 2015.

[177] C. Bell, "Mutual information and maximal correlation as measures of dependence," *The Annals of Mathematical Statistics*, pp. 587–595, 1962.

[178] C. Nair, *Equivalent Formulations of Hypercontractivity Using Information Measures*. ETH-Zürich, 2014.

[179] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for Gaussian variables," *Journal of Machine Learning Research*, vol. 6, no. Jan, pp. 165–188, 2005.

[180] T. Michaeli, W. Wang, and K. Livescu, "Nonparametric canonical correlation analysis," in *International Conference on Machine Learning*, 2016, pp. 1967–1976.

[181] N. Simon and R. Tibshirani, "Comment on "Detecting Novel Associations in Large Data Sets" by Reshef et al, Science Dec 16, 2011," *arXiv preprint arXiv:1401.7645*, 2014.

[182] M. Gorfine, R. Heller, and Y. Heller, "Comment on detecting novel associations in large data sets," *Unpublished (available at http://emotion. technion. ac. il/˜ gorfinm/filesscience6. pdf on 11 Nov. 2012)*, 2012.

[183] G. Kumar, "Binary Rényi correlation: A simpler proof of witsenhausens result and a tighter upper bound," *Manuscript, available at http://www. stanford. edu/˜ gowthamr/research/binary_ renyi_correlation. pdf*, vol. 2, no. 3, p. 13, 2010.

[184] A. Brutzkus and A. Globerson, "Globally optimal gradient descent for a convnet with Gaussian inputs," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 605–614.

[185] Y. Tian, "An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3404–3413.

[186] Y. Li and Y. Yuan, "Convergence analysis of two-layer neural networks with relu activation," in *Advances in Neural Information Processing Systems*, 2017, pp. 597–607.

[187] R. Livni, S. Shalev-Shwartz, and O. Shamir, "On the computational efficiency of training neural networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 855–863.

[188] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 32–40, 1975.

[189] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, no. Apr, pp. 695–709, 2005.

[190] K. Swersky, D. Buchman, N. D. Freitas, B. M. Marlin et al., "On autoencoders and score matching for energy based models," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 1201–1208.

[191] M. Janzamin, H. Sedghi, U. Niranjan, and A. Anandkumar, "Feast at play: Feature extraction using score function tensors," in *Feature Extraction: Modern Questions and Challenges*, 2015, pp. 130–144.

[192] S. Liang, R. Sun, Y. Li, and R. Srikant, "Understanding the loss surface of neural networks for binary classification," in *International Conference on Machine Learning*, 2018, pp. 2840–2849.

[193] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Artificial Intelligence and Statistics*, 2015, pp. 192–204.

[194] D. Soudry and E. Hoffer, "Exponentially vanishing sub-optimal local minima in multilayer neural networks," 2018. [Online]. Available: https://openreview.net/forum?id=Hkfmn5n6W

[195] S. Goel and A. Klivans, "Learning neural networks with two nonlinear layers in polynomial time," *arXiv preprint arXiv:1709.06010*, 2017.

[196] C. D. Freeman and J. Bruna, "Topology and geometry of half-rectified network optimization," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Bk0FWVcgx

[197] Q. Nguyen and M. Hein, "The loss surface and expressivity of deep convolutional neural networks," 2018. [Online]. Available: https://openreview.net/forum?id=BJjquybCW

[198] S. Arora, A. Bhaskara, R. Ge, and T. Ma, "Provable bounds for learning some deep representations," in *International Conference on Machine Learning*, 2014, pp. 584–592.

[199] S. S. Du, J. D. Lee, and Y. Tian, "When is a convolutional filter easy to learn?" in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=SkA-IE06W

[200] A. Andoni, R. Panigrahy, G. Valiant, and L. Zhang, "Learning polynomials with neural networks," in *International Conference on Machine Learning*, 2014, pp. 1908–1916.

[201] R. Panigrahy, A. Rahimi, S. Sachdeva, and Q. Zhang, "Convergence results for neural networks via electrodynamics," in *LIPIcs-Leibniz International Proceedings in Informatics*, vol. 94. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[202] S. S. Du and J. D. Lee, "On the power of over-parametrization in neural networks with quadratic activation," in *International Conference on Machine Learning*, 2018, pp. 1328–1337.

[203] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 742–769, 2018.

[204] M. Janzamin, H. Sedghi, and A. Anandkumar, "Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods," *arXiv preprint arXiv:1506.08473*, 2015.

[205] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon, "Recovery guarantees for one-hidden-layer neural networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 4140–4149.

[206] C. J. Stone, "Optimal rates of convergence for nonparametric estimators," *The Annals of Statistics*, pp. 1348–1360, 1980.

[207] C. J. Hillar and L.-H. Lim, "Most tensor problems are NP-hard," *Journal of the ACM (JACM)*, vol. 60, no. 6, p. 45, 2013.

[208] C. Stein et al., "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

[209] H. Sedghi, M. Janzamin, and A. Anandkumar, "Provable tensor methods for learning mixtures of generalized linear models," in *Artificial Intelligence and Statistics*, 2016, pp. 1223–1231.

[210] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[211] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton et al., "Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, p. 354, 2017.

[212] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[213] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[214] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[215] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and $<$ 0.5 MB model size," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=S1xh5sYgx

[216] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[217] Y. LeCun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in Neural Information Processing Systems*, 1990, pp. 598–605.

[218] B. Hassibi and D. G. Stork, "Second order derivatives for network pruning: Optimal brain surgeon," in *Advances in Neural Information Processing Systems*, 1993, pp. 164–171.

[219] Y. Cheng, F. X. Yu, R. S. Feris, S. Kumar, A. Choudhary, and S.-F. Chang, "An exploration of parameter redundancy in deep networks with circulant projections," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2857–2865.

[220] K. Ullrich, E. Meeds, and M. Welling, "Soft weight-sharing for neural network compression," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=HJGwcKclx

[221] C. Louizos, K. Ullrich, and M. Welling, "Bayesian compression for deep learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 3288–3298.

[222] M. Federici, K. Ullrich, and M. Welling, "Improved Bayesian compression," *arXiv preprint arXiv:1711.06494*, 2017.

[223] T. Berger, "Rate distortion theory for sources with abstract alphabets and memory," *Information and Control*, vol. 13, no. 3, pp. 254–273, 1968.

[224] S. Mandt, M. D. Hoffman, and D. M. Blei, "Stochastic gradient descent as approximate Bayesian inference," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4873–4907, 2017.

[225] W. Cao, X. Wang, Z. Ming, and J. Gao, "A review on neural networks with random weights," *Neurocomputing*, vol. 275, pp. 278–287, 2018.

[226] S. J. Hanson and L. Y. Pratt, "Comparing biases for minimal network construction with back-propagation," in *Advances in Neural Information Processing Systems*, 1989, pp. 177–185.

[227] V. Vanhoucke, A. Senior, and M. Z. Mao, "Improving the speed of neural networks on CPUs," in *Proc. Deep Learning and Unsupervised Feature Learning NIPS Workshop*, vol. 1. Citeseer, 2011, p. 4.

[228] Y. Gong, L. Liu, M. Yang, and L. Bourdev, "Compressing deep convolutional networks using vector quantization," *arXiv preprint arXiv:1412.6115*, 2014.

[229] W. Chen, J. Wilson, S. Tyree, K. Weinberger, and Y. Chen, "Compressing neural networks with the hashing trick," in *International Conference on Machine Learning*, 2015, pp. 2285–2294.

[230] Y. Choi, M. El-Khamy, and J. Lee, "Towards the limit of network quantization," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=rJ8uNptgl

[231] Y. Bu, W. Gao, S. Zou, and V. V. Veeravalli, "Information-theoretic understanding of population risk improvement with model compression," *arXiv preprint arXiv:1901.09421*, 2019.

[232] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus, "Exploiting linear structure within convolutional networks for efficient evaluation," in *Advances in Neural Information Processing Systems*, 2014, pp. 1269–1277.

[233] Y. He, J. Lin, Z. Liu, H. Wang, L.-J. Li, and S. Han, "AMC: AutoML for model compression and acceleration on mobile devices," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 784–800.

[234] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1321–1330.

[235] J. Martens, I. Sutskever, and K. Swersky, "Estimating the Hessian by back-propagating curvature," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*. Omnipress, 2012, pp. 963–970.