Concentration Inequalities for Density Functionals

by

Shashank Singh

Submitted to the Department of Mathematical Sciences
in partial fulfillment of the requirements for the degree of

Master of Science in Mathematical Sciences

at

CARNEGIE MELLON UNIVERSITY

May 2014

Author . . . .
Department of Mathematical Sciences
May 7, 2014

Certified by . . . .
Barnabás Póczos
Assistant Professor, Machine Learning Department
Thesis Supervisor

This page intentionally left blank.

Concentration Inequalities for Density Functionals

by

Shashank Singh

Submitted to the Department of Mathematical Sciences
on May 7, 2014, in partial fulfillment of the
requirements for the degree of
Master of Science in Mathematical Sciences

Abstract

Estimating various kinds of entropy, mutual information, divergence, and other kinds of integral functionals of probability densities in a statistically consistent manner is of great importance in many machine learning tasks. Although this is a fundamental problem in nonparametric statistics, to the best of our knowledge there have been no finite sample exponential concentration bounds derived for estimators of many of these functionals. The main contribution of our work is to provide such bounds for estimators of a broad class of integral functionals on a smooth Hölder class of densities on the d-dimensional unit cube.

Thesis Supervisor: Barnabás Póczos
Title: Assistant Professor, Machine Learning Department

Contents

List of Figures

Table 1: Notation used throughout this thesis, sorted alphabetically. * indicates typical usage of this notation, with some exceptional uses in specific sections.

Notation	Semantic Meaning
B	bound on $\lvert g_p''(\alpha)\rvert$ (Appendix B only)
B_p	bias of $\hat{p}$ at a point
C^γ	general Hölder space of exponent γ
$C^\gamma_{L,r}$	bounded Hölder space
d	dimension of unit cube $[0,1]^d$
D_α	Rényi-α divergence
$\hat{D}_\alpha$	plugin Rényi-α divergence estimate
$D^{\vec{i}}$	mixed partial derivative indexed by $\vec{i}$
f	integrand of density functional*
F	density functional*
g_p	function such that $\lim_{\alpha\to 1} H_\alpha(p) = g_p'(\alpha)\big\rvert_{\alpha=1}$ (Appendix B only)
H	Shannon entropy
H_α	Rényi-α entropy
$\vec{i}$	multi-index in $\mathbb{N}^d$
I_α	Rényi-α mutual information
$\hat{I}_\alpha$	plugin Rényi-α mutual information estimate
k	number of probability densities $p_1, \cdots, p_k$
K	smoothing kernel
ℓ	$l = \lfloor\beta\rfloor$ greatest integer *strictly* less than β
L	multiplicative Hölder constant
n	sample size
p	a probability density function
$\hat{p}$	clipped mirrored kernel density estimate of p
$\widetilde{p}$	mirrored kernel density estimate of p
q	reference probability density function for Renyi-α divergence
X, Y, Z	random variables taking values in $[0,1]^d$
$\mathcal{X}$	domain of density, typically $[0,1]^d$
α	Rényi-α parameter
β	density smoothness parameter
γ	Hölder condition exponent
κ_1	positive density lower bound
κ_2	density upper bound
$\Sigma(\beta, L, r, d)$	bounded Hölder space on $[0,1]^d$ with vanishing boundary derivatives
ξ	intermediate point in domain (from Mean Value or Taylor's Theorem)
$\asymp$	asymptotic order (i.e., big-Θ)

Chapter 1

Introduction

In this thesis we study the convergence of a certain estimator of density functionals. In general, let X be an absolutely continuous (with respect to Lebesgue measure) d-dimensional random vector with Radon-Nikodym derivative (henceforth, “density”) $p : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$. Given a function $f : \mathbb{R} \to \mathbb{R}$ and a real function space $\mathcal{S}$ over $\mathcal{X}$ in which p lies, we are interested in estimating the integral functional

$$F(p) := \int_{\mathcal{X}} f(p(x))\, dx \tag{1.1}$$

(again, the integral is with respect to Lebesgue measure, which is used henceforth always as the reference measure). For simplicity, we refer to functionals such as (1.1) as “density functionals”.

In our framework, we assume that the underlying density p is not known explicitly. Only a finite, independent and identically distributed (i.i.d.) sample is given from p. In addition, f, $\mathcal{X}$, and $\mathcal{S}$ (typically a Hölder, Sobolev, or similar space) are known.

Our main contribution is to derive error bounds, including an exponential concentration bound, for a particular consistent, nonparametric density functional estimator. We also apply our estimator to derive error bounds for estimating certain functionals arising from information theory.

Organization

In the remainder of this chapter, we discuss some applications motivating our study of density functionals, and also survey related work. In Chapter 2, we present our main theoretical results, first discussing the special motivating case of Rényi-α divergence, and then proving a general result and applying it to Rényi Conditional Mutual Information. Finally, in Chapter 3, we conclude by summarizing our work and discussing potential directions for future work. In Appendix A, we present the results of numerical experiments conducted to verify the performance of our estimators. Finally, Appendix B presents an analytical observation that may allow extension of results on Rényi-α information-theoretic quantities to Shannon quantities.

1.1 Motivations

1.1.1 Divergences

There are several important problems in machine learning and statistics that require the estimation of the distance or divergence between distributions. In the past few decades many different kinds of divergences have been defined to measure the discrepancy between distributions, including the Kullback–Leibler (KL) [15], Rényi-α [32, 33], Tsallis-α [39], Bregman [6], Jensen–Shannon [22], L_p, and Csiszár's-f divergences [8], maximum mean discrepancy [5], and many others. Under certain conditions, divergences can estimate entropy and mutual information. Entropy estimators are important in goodness-of-fit testing [9], parameter estimation in semi-parametric models [41], studying fractal random walks [3], and texture classification [11, 12]. Mutual information estimators have been used in feature selection [28], clustering [2], optimal experimental design [21], fMRI data processing [7], prediction of protein structures [1], and boosting and facial expression recognition [34]. Both entropy estimators and mutual information estimators have been used for independent component and subspace analysis [18, 37], as well as for image registration [16, 11, 12]. For further applications, see [20].

A particular divergence estimation application of interest is Distribution-Based Machine Learning. Many applications call for representation and analysis of 'distributional' data sets where each data point is a collection of samples from a high dimensional distribution (as opposed to valuations of a typically vector valued random variable). In this setting, each data point can be modeled by a collection of distributions, one for each measured attribute. Using divergence estimators one can develop machine learning algorithms (such as regression, classification, and clustering algorithms) that can operate on distributions [31, 26].

1.1.2 Renyi-α Information-Theoretic Functionals

A primary motivation for studying density functional estimators is the estimation of Rényi-α information theoretic quantities. These include Rényi-α entropy, Rényi-α divergence, Rényi-α mutual information, and Rényi-α conditional mutual information. Rényi-α divergence contains the Kullback–Leibler divergence as the $\alpha \to 1$ limit case and can also be related to the Tsallis-α, Jensen-Shannon, and Hellinger divergences. Many information theoretic quantities (including entropy, conditional entropy, and mutual information) can be computed as special cases of Rényi-α divergence.

Although many of the above mentioned divergences were defined decades ago, many questions about the properties of their estimators remain open. In particular, rates of convergence are largely unknown, and no finite sample exponential concentration bounds have been derived for divergence estimators. Hence, one of the primary novel applications of our work is the derivation of an exponential concentration bound for a particular consistent, nonparametric, Rényi-α divergence estimator.

1.2 Related Work on Information-Theoretic Functionals

Probably the closest work to ours is that of [23], who derived an exponential-concentration bound for estimators of one- and two-dimensional Shannon entropy and mutual information, over a class of densities obeying a specific Hölder condition.

To the best of our knowledge, only a few consistent nonparametric estimators exist for Rényi-α divergences: [30] proposed a k-nearest neighbour based estimator and proved the weak consistency of the estimator but did not study the convergence rate of the estimator. [40] provided an estimator for the $\alpha \to 1$ limit case only, i.e., for the KL-divergence. They did not study the convergence rate either, and there is also an apparent error in this work; they applied the reverse Fatou lemma under conditions when it does not hold. This error originates in the work [14] and can also be found in other works. Recently, [29] has proposed another consistency proof for this estimator, but it also contains some errors: the strong law of large numbers is applied under conditions when it does not hold and almost sure convergence of an entire sequence is used in a case when only convergence in probability is assumed. [11, 12] also investigated the Rényi divergence estimation problem but assumed that one of the two density functions is known. [10] developed algorithms for estimating the Shannon entropy and the KL divergence for certain parametric families.

Recently, [25] developed methods for estimating f-divergences using their variational characterization properties. They estimate the likelihood ratio of the two underlying densities and plug that into the divergence formulas. This approach involves solving a convex minimization problem over an infinite-dimensional function space. For certain function classes defined by reproducing kernel Hilbert spaces (RKHS), however, they were able to reduce the computational load from solving infinite-dimensional problems to solving n-dimensional problems, where n denotes the sample size. When n is large, solving these convex problems can still be very demanding. They studied the convergence rate of the estimator, but did not derive exponential concentration bounds for the estimator.

[35, 17, 4] studied the estimation of non-linear functionals of density. They, however, did not study the Rényi divergence estimation and did not derive exponential concentration bounds either. Using ensemble estimators, [36] derived fast rates for entropy estimation but did not investigate the divergence estimation problem. [20] and [9] considered Shannon and Rényi-α entropy estimation from a single sample.[1] Recently, [27] proposed a method for consistent Rényi information estimation, but this estimator also uses one sample only and cannot be used for estimating Rényi divergences. Further information and useful reviews of several different divergences can be found, e.g., in [39].

[1] The original presentations of these works contained some errors; [19] provide corrections for some of these theorems.

Chapter 2

Theoretical Results

In this chapter, we present our main theoretical results concerning error bounds for a density functional estimator, along with proofs of these results. This work began with an attempt to derive error bounds for an estimator of Rényi-α divergence. We then generalized the methods used in the Rényi-α divergence case to derive bounds for a more general class of density estimates, and also applied these general results to the case of Rényi-α Conditional Mutual Information, which has important applications.

Although the results for the Rényi-α divergence case are special cases of the results for general density functionals, in order to motivate the general case, and to portray the work as it developed, after introducing some notation in Section 2.1, we begin by presenting the Rényi-α divergence case in Section 2.2. We then proceed to present the general case and the application to the Rényi-α Conditional Mutual Information in Sections 2.3 and 2.4, respectively.

2.1 Notation

Multi-indices: We use the notation of multi-indices common in multivariable calculus. As a reminder of this notation by example, for analytic functions $f : \mathbb{R}^d \to \mathbb{R}, \forall x, y \in \mathbb{R}$,

$$f(y) = \sum_{\vec{i} \in \mathbb{N}^d} \frac{D^{\vec{i}} f(x)}{\vec{i}!} (y - x)^{\vec{i}},$$

where $\mathbb{N}^d$ is the set of d-tuples of natural numbers,

$$\vec{i}! := \prod_{k=1}^{d} i_k!, \qquad (y-x)^{\vec{i}} := \prod_{k=1}^{d} (y_k - x_k)^{i_k}$$

and

$$D^{\vec{i}} f := \frac{\partial^{|\vec{i}|} f}{\partial^{i_1} x_1 \cdots \partial^{i_d} x_d}, \qquad \text{for} \qquad |\vec{i}| := \sum_{k=1}^{d} i_k.$$

We also use the Multinomial Theorem, which states that, $\forall k \in \mathbb{N}, x \in \mathbb{R}^d$,

$$\left(\sum_{j=1}^{d} x_j \right)^k = \sum_{|\vec{i}|=k} \frac{k!}{\vec{i}!} x^{\vec{i}}. \tag{2.1}$$

Measures: It should also be assumed that all reference measures (for integration, Radon-Nikodym differentiation, etc.) are Lebesgure measure, unless otherwise specified.

Bounded Hölder Space: For a fixed bounded domain $D \subseteq \mathbb{R}^d$ and $\beta \in (0,1]$, it is common in analysis to work over the linear space of (uniformly) β-Hölder [1] continuous functions:

$$C^\beta(D) := \left\{ f : D \to \mathbb{R} \text{ s.t. } \sup_{x \neq y \in D} \frac{|f(x) - f(y)|}{\|x - y\|^\beta} < \infty \right\}. \tag{2.2}$$

(since finite-dimensional norms are equivalent, the choice of norm on $\mathbb{R}^d$ is irrelevant). It is useful but perhaps less common to consider the following generalization: for $\beta > 0$ and $\ell := \lfloor \beta \rfloor$ is the greatest integer *strictly* less than β,

$$C^\beta(D) := \left\{ f : D \to \mathbb{R} \text{ s.t. }, \sup_{\substack{x \neq y \in D \\ |\vec{i}| = \ell}} \frac{|D^{\vec{i}} f(x) - D^{\vec{i}} f(y)|}{\|x - y\|^{(\beta - \ell)}} < \infty \right\}.^2 \tag{2.3}$$

[1] α is more commonly used than β for the Hölder exponent. Here, we reserve α for the Rényi-α parameter, although these quantities are related (for example, for $\alpha \in [0,1)$, Rényi-α entropy is in some sense *precisely* α-Hölder continuous).

[2] Despite the potential notational confusion with the use of C^k as the space of k-times continuously differentiable functions in the case that β is an integer, we use the notation C^β, common in nonparametric statistics, rather than the notation $C^{\ell,\beta}$, common in analysis, for the appropriate Hölder spaces. The statistical notation emphasizes the role of β in our context as a continuous parameter (as opposed to emphasizing the number of available derivatives, ℓ), and is more natural for expressing, for example, our bias convergence rates ($O(n^\beta)$, as opposed to $O(n^{\ell+\beta})$.

i.e., functions whose derivatives of order ℓ all exist and are $(\beta - \ell)$-Hölder continuous. [3] By Taylor's Theorem, this condition is equivalent to requiring the order ℓ Taylor approximation of f centered at x to have error of order $O(\|y - x\|^\beta)$ at y. This property makes this generalized Hölder class very convenient for proving error bounds, and, furthermore, this class is sufficiently large for many interesting applications. Since we are interested proving in numerical finite-sample bounds rather than simply proving convergence, it is necessary to assume some numerical bounds on the class of functions. Thus, we work with density functions in a bounded subset of a generalized Hölder space. [4] In particular, for a fixed $L \geq 0$ and $r \geq 1$, we work within the class

$$C^\beta_{L,r}(D) := \left\{ f : D \to \mathbb{R} \text{ s.t. }, \sup_{\substack{x \neq y \in D \\ |\vec{i}| = \ell}} \frac{|D^{\vec{i}} f(x) - D^{\vec{i}} f(y)|}{\|x - y\|_r^{(\beta - \ell)}} \leq L \right\}. \qquad (2.4)$$

Note that, because L is arbitrary and, for and $r \geq 1$, any $f \in C^\beta$ is in $C^\beta_{L,r}$ for sufficiently large L, the restriction to $C^\beta_{L,r}$ is superficial, and is essentially to fix a value of L. Due to the fixed choice of L, $C^\beta_{L,r}(D)$ is not a linear space. However, $C^\beta_{L,r}(D)$ is convex (this will be important for using the Mean Value Theorem).

Vanishing Boundary Derivatives: One of the primary complications in our work is that we work with random variables taking values in a bounded domain (see Section 2.5 for further discussion of the reasons for and consequencess of this). Due to sparsity of data near the boundary, we must perform some sort of boundary correction to reduce the bias of our estimator near the boundary. We choose to do this by reflecting our data set across each boundary and assuming the density function is approximately constant near the boundary. In order to do this is a reasonably simple manner, we specifically work over the unit cube $[0, 1]^d$. We cast our assumption of a density function that is roughly constant near the boundary in terms of derivatives vanishing near the boundary. Hence, we define the

[3] simply considering $\beta > 1$ without differentiating is unhelpful, as it is easy to show that, if D is nice (e.g., open and connected), $C^\beta(D)$ as in equation (2.2) contains only constant functions when $\beta > 1$.

[4] Since, eventually we are interested in estimating an integral, we may allow the the density to violate the Hölder and boundedness conditions on null sets; that is, as is typical in probability theory, we identify densities which differ on sets of Lebesgue measure 0.

bounded Hölder class with vanishing boundary derivatives:

$$\Sigma(\beta, L, r, d) := \left\{ f \in C^{\beta}_{L,r}([0,1]^d) : \max_{1 \leq |\vec{i}| \leq \ell} |D^{\vec{i}} f(x)| \to 0 \text{ as } \operatorname{dist}(x, \partial[0,1]^d) \to 0 \right\}, \tag{2.5}$$

where

$$\partial[0,1]^d = \{x \in [0,1]^d : x_j \in \{0,1\} \text{ for some } j \in [d]\} \tag{2.6}$$

is the boundary of $[0,1]^d$.

Higher-Order Kernels: Our first step in estimating Rényi-α is to estimate each density function using kernel density estimation. This can be viewed as smoothing the data set, represented as a uniformly weighted sum of point (Dirac delta) distributions, by convolving it with a smooth kernel function $K^d : \mathbb{R}^d \to \mathbb{R}$ (similar to mollification in analysis) [5] . For simplicity, we assume K is supported in $[-1,1]$, and, in order for the result to be a probability density function, the kernel ought to have unit mass. Another useful property is having ℓ orders of symmetry (this will allow us to drop ℓ terms from a certain Taylor approximation, a key step in bounding the bias of our estimator). In particular, we assume

$$\int_{-1}^{1} K(u)\, du = 1, \quad \text{and} \quad \int_{-1}^{1} u^j K(u)\, du = 0, \quad \forall j \in \{1, \dots, \ell\}. \tag{2.7}$$

The existence of such kernels is not immediately apparent. However, they can be constructed in terms of Legendre polynomials (see section 1.2.2 of [38] for such a construction). If $\ell \geq 2$, then such a kernel will necessarily be negative on a set of positive Lebesgue measure. It is possible that the kernel density estimator arising from using such a kernel will take negative values. However, taking only the positive part of the estimator will not increase the error, since the density is non-negative.

[5] To clarify our notation, K^d will denote the d-dimensional product kernel based on $K : \mathbb{R} \to \mathbb{R}$, defined by $K^d(u) = \prod_{j=1}^{d} K(u_j), \forall u \in \mathbb{R}^d$.

2.2 The Rényi-α Divergence Estimation Case

Here, we discuss the case where the density functional is the Rényi-α divergence, as introduced in Section 1.1. After formally introducing the Rényi-α divergence estimation problem, we present our assumptions, our mirrored kernel density estimator, and our main results, followed by some preliminary lemmas and proofs of our main results. The presentation of the estimator, as well as some of the preliminary lemmas and then proofs of our main results in this section are particularly important, as the estimator is nearly the same as will be used in the general case, and some of the lemmas and proof techniques given here will simply be referenced in the discussion of the general case.

2.2.1 Problem Statement

For a given $d \geq 1$, consider random d-dimensional real vectors X and Y in the unit cube $\mathcal{X} := [0,1]^d$, distributed according to densities $p, q : \mathcal{X} \to \mathbb{R}$, respectively. For a given $\alpha \in (0,1) \cup (1,\infty)$, we are interested in using a random sample of n i.i.d. points from p and n i.i.d. points from q to estimate the Rényi-α divergence

$$D_\alpha(p\|q) = \frac{1}{\alpha - 1} \log \left(\int_{\mathcal{X}} p^\alpha(x) q^{1-\alpha}(x)\, dx \right).$$

2.2.2 Assumptions

Density Assumptions: We assume that p and q are in a bounded Hölder class with vanishing boundary derivatives, $\Sigma(\beta, L, r, d)$ (defined in 2.5), [6] and also assume p and q are bounded above and away from 0; i.e., $\exists \kappa_2, \kappa_1 > 0$ with $\kappa_1 \leq \inf_{x\in\mathcal{X}} p(x), \inf_{x\in\mathcal{X}} q(x)$ and $\kappa_2 \geq \sup_{x\in\mathcal{X}} p(x), \sup_{x\in\mathcal{X}} q(x)$. [7]

Kernel Assumptions: We assume the kernel $K : \mathbb{R} \to \mathbb{R}$ has bounded support $[-1, 1]$ and is of order ℓ, as defined in (2.7).

[6] We could take p and q to be in different Hölder classes $\Sigma(\beta_p, L_p, r_p, d)$ and $\Sigma(\beta_q, L_q, r_q, d)$, but the bounds we show depend, asymptotically, only on the weaker of the conditions on p and q (i.e., $\min\{\beta_p, \beta_q\}, \max\{L_p, L_q\}$, etc.).

[7] The need for a lower bound κ_1 (because, essentially, of the explosion of the logarithm at 0) is one of the reasons for working on the domain $\mathcal{X} = [0,1]^d$, a set of finite Lebesgue measure.

There are many interesting points of discussion regarding our assumptions. To avoid redundancy, we withold these until Section 2.5, after discussing the general density functional estimator.

2.2.3 Estimator

Let $[d] := \{1, 2, \ldots, d\}$, and let

$$\mathcal{S} := \{(S_1, S_2, S_3) \,:\, S_1 \cup S_2 \cup S_3 = [d], S_i \cap S_j = \emptyset \text{ for } i \neq j\}$$

denote the set of partitions of $[d]$ into 3 distinguishable parts. For a small $h > 0$ (to be specified later), for each $S \in \mathcal{S}$, define the region

$$\begin{aligned} C_S = \{x \in \mathcal{X} : {} & \forall i \in S_1, 0 \leq x_i \leq h, \\ & \forall j \in S_2, h < x_j < 1 - h, \\ & \forall k \in S_3, 1 - h \leq x_k \leq 1\} \end{aligned}$$

and the regional kernel $K_S : [-1, 2]^d \times \mathcal{X} \to \mathbb{R}$ by

$$K_S(x, y) := \prod_{j \in S_1} K\left(\frac{x_j + y_j}{h}\right) \cdot \prod_{j \in S_2} K\left(\frac{x_j - y_j}{h}\right) \cdot \prod_{j \in S_3} K\left(\frac{x_j - 2 + y_j}{h}\right).$$

Note that $\{C_S : S \in \mathcal{S}\}$ partitions $\mathcal{X}$ (as illustrated in Figure 2-1), up to intersections of measure zero, and that K_S is supported only on $[-1, 2]^d \times C_S$. The term $K\left(\frac{x_j+y_j}{h}\right)$ corresponds to reflecting y across the hyperplane $x_j = 0$, whereas the term $K\left(\frac{x_j-2+y_j}{h}\right)$ reflects y across $x_j = 1$, so that $K_S(x, y)$ is the product kernel (in x), with uniform bandwidth h, centered around a reflected copy of y.

We now define the "mirror image" kernel density estimator

$$\widetilde{p}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \sum_{S \in \mathcal{S}} K_S(x, x^i),$$

where x^i denotes the i^{th} sample. Since the derivatives of p and q vanish near $\partial\mathcal{X}$, p and q

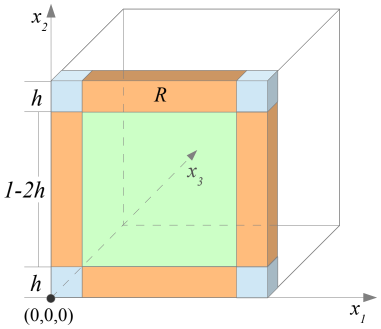

Figure 2-1: Illustration of regions $C_{(S_1,S_2,S_3)}$ with $3 \in S_1$. The region labeled R corresponds to $S_1 = \{3\}, S_2 = \{1\}, S_3 = \{2\}$.

are approximately constant near $\partial\mathcal{X}$, and so the mirror image estimator attempts to reduce boundary bias by mirroring data across $\partial\mathcal{X}$ before kernel-smoothing. We then clip the estimator at our lower and upper bounds κ_1 and κ_2:

$$\widehat{p}_h(x) = \min(\kappa_2, \max(\kappa_1, \widetilde{p}_h(x))).$$

Finally, we plug our clipped density estimate into the following plug-in estimator for Rényi α-divergence:

$$D_\alpha(p\|q) = \frac{1}{\alpha-1}\log\left(\int_{\mathcal{X}} p^\alpha(x)q^{1-\alpha}(x)\,dx\right) = \frac{1}{\alpha-1}\log\left(\int_{\mathcal{X}} f(p(x),q(x))\,dx\right) \tag{2.8}$$

for $f : [\kappa_1, \kappa_2]^2 \to \mathbb{R}$ defined by $f(x_1, x_2) := x_1^\alpha x_2^{1-\alpha}$. Our α-divergence estimate is then $D_\alpha(\widehat{p}_h\|\widehat{q}_h)$.

2.2.4 Main Result

Rather than the usual decomposition of mean squared error into variance and squared bias, we decompose the error $|D_\alpha(\widehat{p}_h\|\widehat{q}_h) - D_\alpha(p\|q)|$ of our estimatator into a bias term and a variance-like term via the triangle inequality:

$$|D_\alpha(\widehat{p}_h\|\widehat{q}_h) - D_\alpha(p\|q)| \leq \underbrace{|D_\alpha(\widehat{p}_h\|\widehat{q}_h) - \mathbb{E}D_\alpha(\widehat{p}_h\|\widehat{q}_h)|}_{\text{variance-like term}} + \underbrace{|\mathbb{E}D_\alpha(\widehat{p}_h\|\widehat{q}_h) - D_\alpha(p\|q)|}_{\text{bias term}}.$$

We will prove the “variance” bound

$$\mathbb{P}\left(|D_\alpha(\widehat{p}_h, \widehat{q}_h) - \mathbb{E}D_\alpha(\widehat{p}_h, \widehat{q}_h)| > \varepsilon\right) \leq 2\exp\left(-\frac{k_1\varepsilon^2 n}{\|K\|_1^{2d}}\right),$$

and the bias bound

$$|\mathbb{E}D_\alpha(\widehat{p}_h\|\widehat{q}_h) - D_\alpha(p\|q)| \leq k_2\left(h^\beta + h^{2\beta} + \frac{1}{nh^d}\right),$$

where k_1, k_2 are constant in the sample size n and bandwidth h (see (2.18) and (2.20) for exact values of these constants). While the variance bound does not depend on h, differentiation shows that the bias bound is minimized by $h \asymp n^{-\frac{1}{d+\beta}}$, giving the convergence rate

$$|\mathbb{E}D_\alpha(\widehat{p}_h\|\widehat{q}_h) - D_\alpha(p\|q)| \in O\left(n^{-\frac{\beta}{d+\beta}}\right).$$

Note that we can use this exponential concentration bound to bound the variance of $D(\widehat{p}_h\|\widehat{q}_h)$. If $F : [0,\infty) \to \mathbb{R}$ is the cumulative distribution of the squared deviation of $D_\alpha(\widehat{p}_h\|\widehat{q}_h)$ from its mean, then

$$1 - F(\varepsilon) = \mathbb{P}\left((D_\alpha(\widehat{p}_h, \widehat{q}_h) - \mathbb{E}D_\alpha(\widehat{p}_h, \widehat{q}_h))^2 > \varepsilon\right) \leq 2\exp\left(-\frac{k_1 n}{\|K\|_1^{2d}}\right).$$

Thus,

$$\begin{aligned}\mathbb{V}[D_\alpha(\widehat{p}_h\|\widehat{q}_h)] = \mathbb{E}\left[(D_\alpha(\widehat{p}_h, \widehat{q}_h) - \mathbb{E}D_\alpha(\widehat{p}_h, \widehat{q}_h))^2\right] &= \int_0^\infty (1 - F(\varepsilon))\, d\varepsilon \\ &\leq \int_0^\infty 2\exp\left(-\frac{k_1 n\varepsilon}{\|K\|_1^{2d}}\right) d\varepsilon = 2\frac{\|K\|_1^{2d}}{k_1}n^{-1}.\end{aligned}$$

We then have a mean squared-error of

$$\mathbb{E}\left[(D(\widehat{p}_h\|\widehat{q}_h) - D(p\|q))^2\right] \in O\left(n^{-1} + n^{-\frac{2\beta}{d+\beta}}\right).$$

which is in $O(n^{-1})$ if $\beta \geq d$ and in $O\left(n^{-\frac{2\beta}{d+\beta}}\right)$ otherwise. This asymptotic rate is consistent with previous bounds in density functional estimation [4, 35].

2.2.5 Preliminaries

Here we establish a few minor points which will smooth the proofs of the main results.

Bound on Derivatives of f: Let f be as in (2.8). Since f is twice continuously differentiable on the compact domain $[\kappa_1, \kappa_2]^2$, there is a constant $C_f \in \mathbb{R}$, depending only on κ_1, κ_2, and α, such that, $\forall \xi \in (\kappa_1, \kappa_2)^2$,

$$\left|\frac{\partial f}{\partial x_1}(\xi)\right|, \left|\frac{\partial f}{\partial x_2}(\xi)\right|, \left|\frac{\partial^2 f}{\partial x_1^2}(\xi)\right|, \left|\frac{\partial^2 f}{\partial x_2^2}(\xi)\right|, \left|\frac{\partial^2 f}{\partial x_1 x_2}(\xi)\right| \leq C_f. \tag{2.9}$$

C_f can be computed explicitly by differentiating f and observing that the derivatives of f are monotone in each argument. We will use this bound later in conjunction with the Mean Value and Taylor's theorems.

Logarithm Bound: If $g, \hat{g} : \mathcal{X} \to \mathbb{R}$ with $0 < c \leq g, \hat{g}$ for some $c \in \mathbb{R}$ depending only on κ_1 and α, then, by the Mean Value Theorem, there exists $C_{\log}$ depending only on κ_1 and α such that

$$\left|\log\left(\int_{\mathcal{X}} \hat{g}(x)\,dx\right) - \log\left(\int_{\mathcal{X}} g(x)\,dx\right)\right| \leq C_{\log} \int_{\mathcal{X}} |\hat{g}(x) - g(x)|\,dx. \tag{2.10}$$

We will use this bound to eliminate logarithms from our calculations.

Bounds on Derivatives of p: Combining the assumption that the derivatives of p vanish on $\partial\mathcal{X}$ and the Hölder condition on p, we bound the derivatives of p *near* $\partial\mathcal{X}$. In particular, we show that, if $\vec{i} \in \mathbb{N}^d$ has $1 \leq |\vec{i}| \leq \ell$, then, $\forall x \in \mathcal{B} := \{x \in \mathcal{X} : \operatorname{dist}(x, \partial\mathcal{X}) \leq h\}$

$$|D^{\vec{i}} p(x)| \leq \frac{L h^{\beta - |\vec{i}|}}{(\ell - |\vec{i}|)!}. \tag{2.11}$$

Proof: We proceed by induction on $|\vec{i}|$, as $|\vec{i}|$ decreases from ℓ to 0. The case $|\vec{i}| = \ell$ is precisely the Hölder assumption (2.4). Now suppose that we have the desired bound for derivatives of order $|\vec{i}|+1$. Let $x \in \partial\mathcal{X}$, $u = (0, \ldots, 0, \pm 1, 0, \ldots, 0) \in \mathbb{R}^d$, where $u_j = \pm 1$.

If $y + hu \in \mathcal{X}$ (any $x \in \mathcal{B}$ is clearly of this form, for some $j \in [d]$), then

$$|D^{\vec{i}}p(y+u)| \leq \int_0^h \left| \frac{\partial}{\partial x_j} D^{\vec{i}} p(y+tu) \right| dt \leq \int_0^h \frac{Lt^{\beta - (|\vec{i}|+1)}}{(\ell - |\vec{i}| - 1)!} \, dt$$
$$= \frac{Lh^{\beta - |\vec{i}|}}{(\beta - |\vec{i}|)(\ell - |\vec{i}| - 1)!} \leq \frac{Lh^{\beta - |\vec{i}|}}{(\ell - |\vec{i}|)!}.$$

The desired result follows by induction on $|\vec{i}|$. ■

Integral of Mirrored Kernel: A key property of the mirrored kernel is that the mass of the kernel over $\mathcal{X}$ is preserved, even near the boundary of $\mathcal{X}$, as the kernels about the reflected data points account exactly for the mass of the kernel about the original data point that is not in $\mathcal{X}$. In particular, $\forall y \in \mathcal{X}$,

$$\sum_{S \in \mathcal{S}} \int_{\mathcal{X}} |K_S(x,y)| \, dx = h^d \|K\|_1^d. \tag{2.12}$$

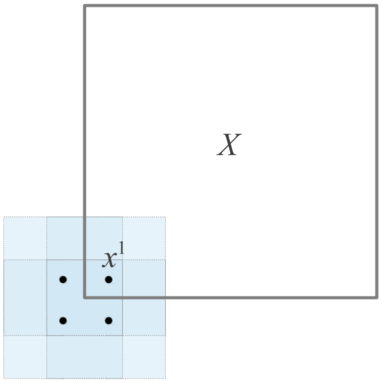

Figure 2-2: A data point $x^1 \in C_{(\{1,2\},\emptyset,\emptyset)} \subset [0,1]^2$, along with its three reflected copies. The sum of the integrals over $\mathcal{X}$ of (the absolute values of) the four kernels (with shaded support) is $\|K\|_1^2$.

Proof: For each $S \in \mathcal{S}$, the change of variables

$$u_j = -x_j, \text{ for } j \in S_1 \quad u_j = x_j, \text{ for } j \in S_2 \quad \text{and} \quad u_j = 2 - x_j, \text{ for } j \in S_3$$

returns the reflected data point created by K_S back onto its original data point. Applying this change of variables gives

$$\sum_{S \in \mathcal{S}} \int_{\mathcal{X}} |K_S(x,y)| \, dx = \int_{[-1,2]^d} \left| K^d \left(\frac{u-y}{h} \right) \right| \, du,$$

where $K^d(x) := \prod_{i=1}^{d} K(x_i)$ denotes the product kernel. Rescaling, translating, and applying Fubini's Theorem,

$$\sum_{S \in \mathcal{S}} \int_{\mathcal{X}} |K_S(x,y)| \, dx = h^d \int_{[-1,1]^d} |K^d(x)| \, dx = h^d \left(\int_{-1}^{1} |K(u)| \, du \right)^d = h^d \|K\|_1^d. \quad \blacksquare$$

2.2.6 Bias Bound

For an arbitrary $p \in \Sigma(\beta, L, r, d)$ let $B_p(x) := \mathbb{E}\widetilde{p}_h(x) - p(x)$ denote the bias of (the unclipped estimator) $\widetilde{p}_h$ at $x \in \mathcal{X}$. The following lemma bounds the integrated squared bias of $\widetilde{p}_h$. For x in the interior of $\mathcal{X}$ (x with distance greater than h from the boundary of $\mathcal{X}$), a standard result bounds this quantity. Near the boundary of $\mathcal{X}$, the proof is more complicated, because the support of the kernel is not fully contained in $\mathcal{X}$, and hence we cannot simply use the symmetry of the kernel to drop the first ℓ terms of the Taylor approximation. Instead, we use the fact that the derivatives of p vanish near the boundary of X and the mirroring of our kernel density estimator, together with the Hölder condition, to bound the estimator's bias near the boundary of $\mathcal{X}$.

Bias Lemma: There exists a constant $C > 0$ such that

$$\int_{\mathcal{X}} B_p^2(x) \, dx \leq C h^{2\beta}. \tag{2.13}$$

Proof: We consider separately the "h-interior" $\mathcal{I}_h := (h, 1-h)^d$ and the "h-boundary" $\mathcal{B}_h = \mathcal{X} \backslash \mathcal{I}_h$. By a standard result [8] for kernel density estimates of Hölder continuous functions (see, for example, Proposition 1.2 of [38]),

$$\int_{\mathcal{I}} B_p^2(x) \, dx \leq C_2 h^{2\beta}, \quad \text{where} \quad C_2 := \frac{L}{\ell!} \|K\|_1^d.$$

[8] The assumption that the kernel K is of order ℓ is used in the proof of this standard result.

We now show that $\int_{\mathcal{B}} B_p^2(x)\,dx \leq C_3^2 h^{2\beta}$ (C_3 will be specified in (2.16)).

Let $S = (S_1, S_2, S_3) \in \mathcal{S}\backslash\{(\emptyset, [d], \emptyset)\}$ (as $C_{(\emptyset,[d],\emptyset)} = \mathcal{I}$). We bound $|B_p(x)|$ on C_S. To simplify notation, by geometric symmetry, we assume $S_3 = \emptyset$. Let $u \in [-1,1]^d$, and define $y_S \in \mathcal{X}$ by $(y_S)_i := hu_i - x_i, \forall i \in S_1$ and $(y_S)_i := x_i - hu_i, \forall i \in S_2$ (we use this choice in a change of variables in (2.17)). By the Hölder condition (2.4) and choice of y_S,

$$\left| p(y_S) - \sum_{|\vec{i}| \leq \ell} \frac{D^{\vec{i}} p(x)}{\vec{i}!} (y_S - x)^{\vec{i}} \right| \leq L \|y_S - x\|_r^\beta = L \left(\sum_{j \in S_1} |2x_j + hu_j|^r + \sum_{j \in S_2} |hu_j|^r \right)^{\beta/r}$$

Since each $|u_j| \leq 1$ and, for each $i \in S_1$, $0 \leq x_j \leq h$,

$$\begin{aligned} \left| p(y_S) - \sum_{|\alpha| \leq \ell} \frac{D^{\vec{i}} p(x)}{\vec{i}!} (y_S - x)^{\vec{i}} \right| &= L \left(\sum_{j \in S_1} (3h)^r + \sum_{j \in S_2} h^r \right)^{\beta/r} \\ &\leq L \left(d\,(3h)^r \right)^{\beta/r} = L \left(3d^{1/r} h \right)^\beta . \end{aligned}$$

Rewriting this using the triangle inequality

$$|p(y_S) - p(x)| \leq L \left(3d^{1/r} h \right)^\beta + \left| \sum_{1 \leq |\alpha| \leq \ell} \frac{D^\alpha p(x)}{\alpha!} (y_S - x)^\alpha \right| . \tag{2.14}$$

Observing $(y - x)^{\vec{i}} \leq (3h)^{|\vec{i}|}$ and applying the bound in (2.11) on p's derivatives near $\partial\mathcal{X}$,

$$\begin{aligned} \left| \sum_{1 \leq |\vec{i}| \leq \ell} \frac{D^{\vec{i}} p(x)}{\vec{i}!} (y_S - x)^{\vec{i}} \right| &\leq \sum_{|\vec{i}| \leq \ell} \left| \frac{L h^{\beta - |\vec{i}|}}{(\ell - |\vec{i}|)! \vec{i}!} (3h)^{|\vec{i}|} \right| \\ &= L h^\beta \sum_{k=0}^{\ell} \sum_{|\vec{i}| = k} \frac{3^{|\vec{i}|}}{(\ell - k)! \vec{i}!} \leq L h^\beta \sum_{k=0}^{\ell} \frac{1}{k!(\ell - k)!} \sum_{|\vec{i}| = k} \frac{k! 3^{|\vec{i}|}}{\vec{i}!} . \end{aligned}$$

Then, applying the multinomial theorem (2.1) followed by the binomial theorem gives

$$\begin{aligned} \left| \sum_{1 \leq |\vec{i}| \leq \ell} \frac{D^{\vec{i}} p(x)}{\vec{i}!} (y_S - x)^{\vec{i}} \right| &\leq L h^\beta \sum_{k=0}^{\ell} \frac{(3d)^k}{k!(\ell - k)!} \\ &= L h^\beta \frac{1}{\ell!} \sum_{k=0}^{\ell} \frac{\ell!}{(\ell - k)! k!} (3d)^k = L h^\beta \frac{(3d + 1)^\ell}{\ell!} . \end{aligned}$$

Combining this bound with (2.14) gives

$$|p(y_S) - p(x)| \leq C_3 h^\beta, \tag{2.15}$$

$$\text{where} \quad C_3 := L\left(\left(3d^{1/r}\right)^\beta + \frac{(3d+1)^\ell}{\ell!}\right). \tag{2.16}$$

For $x \in C_S$, we have $\widetilde{p}_h(x) = \frac{1}{nh^d}\sum_{i=1}^n K_S(x, x^i)$, and thus, by a change of variables, recalling that $K^d(x)$ denotes the product kernel,

$$\mathbb{E}\widetilde{p}_h(x) = \frac{1}{h^d}\int_{\mathcal{X}} K_S(x,u)p(u)\,du = \int_{[-1,1]^d} K^d(v)p(y_S)\,dv, \tag{2.17}$$

Since $\displaystyle\int_{[-1,1]^d} K^d(v)\,dv = 1$, by the bound in (2.15),

$$\begin{aligned}
|B_p(x)| = |\mathbb{E}\widetilde{p}_h(x) - p(x)| &= \left|\int_{[-1,1]^d} K^d(v)p(y_S)\,dv - \int_{[-1,1]^d} K^d(v)p(x)\,dv\right| \\
&\leq \int_{[-1,1]^d} K^d(v)|p(y_S) - p(x)|\,dv \\
&\leq \int_{[-1,1]^d} K^d(v)C_3 h^\beta\,dv = C_3 h^\beta.
\end{aligned}$$

Then, $\int_{\mathcal{B}} B_p^2(x) \quad dx \;\leq\; C_3^2 h^{2\beta}$ ($\mathcal{B}$ has measure less than 1), proving the Bias Lemma. ■

We now return to bounding the bias of $D(\widehat{p}_h, \widehat{q}_h)$, by reducing part of the bias of our divergence estimator to the bias of the kernel density estimate, which we have just bounded.

By Taylor's Theorem, $\forall x \in \mathcal{X}$, for some $\xi : \mathcal{X} \to \mathbb{R}^2$ on the line segment between $(\widehat{p}_h(x), \widehat{q}_h(x))$ and $(p(x), q(x))$,

$$\begin{aligned}
&|\mathbb{E}f(\widehat{p}_h(x), \widehat{q}_h(x)) - f(p(x), q(x))| \\
&= \left|\mathbb{E}\frac{\partial f}{\partial x_1}(p(x),q(x))(\widehat{p}_h(x) - p(x)) + \frac{\partial f}{\partial x_2}(p(x),q(x))(\widehat{q}_h(x) - q(x))\right. \\
&\qquad + \frac{1}{2}\left[\frac{\partial^2 f}{\partial x_1^2}(\xi)(\widehat{p}_h(x) - p(x))^2 + \frac{\partial^2 f}{\partial x_2^2}(\xi)(\widehat{q}_h(x) - q(x))^2\right] \\
&\qquad\quad \left. + \frac{\partial^2 f}{\partial x_1 \partial x_2}(\xi)(\widehat{p}_h(x) - p(x))(\widehat{q}_h(x) - q(x))\right| \\
&\leq C_f\left(|B_p(x)| + |B_q(x)| + \mathbb{E}\left[\widehat{p}_h(x) - p(x)\right]^2 + \mathbb{E}\left[\widehat{q}_h(x) - q(x)\right]^2 + |B_p(x)B_q(x))|\right),
\end{aligned}$$

where the last line follows from the triangle inequality and (2.9). Thus, using (2.10),

$$\begin{aligned}
|\mathbb{E}D_\alpha(\widehat{p}_h\|\widehat{q}_h) - D_\alpha(p\|q)| &= \left|\frac{1}{\alpha-1}\left(\mathbb{E}\log\int_{\mathcal{X}} f(\widehat{p}_h(x),\widehat{q}_h(x))\,dx - \log\int_{\mathcal{X}} f(p(x),q(x))\,dx\right)\right| \\
&\leq \frac{C_{\log}}{|\alpha-1|}\int_{\mathcal{X}} |\mathbb{E}f(\widehat{p}_h(x),\widehat{q}_h(x)) - f(p(x),q(x))\,dx| \\
&\leq \frac{C_f C_{\log}}{|\alpha-1|}\int_{\mathcal{X}} |B_p(x)| + |B_q(x)| + \mathbb{E}\left[\widehat{p}_h(x) - p(x)\right]^2 \\
&+ \mathbb{E}\left[\widehat{q}_h(x) - q(x)\right]^2 + |B_p(x)B_q(x)|\ dx.
\end{aligned}$$

By Hölder's Inequality, we then have

$$\begin{aligned}
|\mathbb{E}D_\alpha(\widehat{p}_h\|\widehat{q}_h) - D_\alpha(p\|q)| \leq \frac{C_f C_{\log}}{|\alpha-1|\kappa_1}&\left(\sqrt{\int_{\mathcal{X}} B_p^2(x)\,dx} + \sqrt{\int_{\mathcal{X}} B_q^2(x)\,dx}\right. \\
&+ \int_{\mathcal{X}} \mathbb{E}\left[\widehat{p}_h(x) - p(x)\right]^2 + \mathbb{E}\left[\widehat{q}_h(x) - q(x)\right]^2\ dx \\
&\left. + \sqrt{\int_{\mathcal{X}} B_p^2(x)\,dx\int_{\mathcal{X}} B_q^2(x)\,dx}\right).
\end{aligned}$$

Applying the Bias Lemma (2.13) and a standard result in kernel density estimation (see, for example, Propositions 1.1 and 1.2 of [38]) gives

$$\begin{aligned}
|\mathbb{E}D_\alpha(\widehat{p}_h\|\widehat{q}_h) - D_\alpha(p\|q)| &\leq (C_2 + C_3)\,h^\beta + C_2 h^{2\beta} + \kappa_2\frac{\|K\|_1^d}{nh^d} \\
&\leq C\left(h^\beta + h^{2\beta} + \frac{1}{nh^d}\right), \qquad (2.18)
\end{aligned}$$

for some $C > 0$ not depending on n or h. ■

2.2.7 Variance Bound

The main tool in proving our exponential bound is McDiarmid's Inequality [24] (also known as the method of bounded differences), a special case of Azuma's Inequality:

McDiarmid's Inequality: Suppose $X_1, \cdots, X_n$ are independent random varibles and a function $f : \mathbb{R}^n \to \mathbb{R}$ has the property

$$\sup_{i\in[n]} \sup_{x_1,\ldots,x_n,x_i'} f(x_1, \cdots, x_n) - f(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_n) \leq C,$$

for some $C \in \mathbb{R}$ (which may depend on n). That is, the change in the function value when changing any input is bounded by a constant. Then, for all $\varepsilon > 0$,

$$\mathbb{P}(|f(X_1, \cdots, X_n) - \mathbb{E}f(X_1, \cdots, X_n)| \geq \varepsilon) \leq 2\exp\left(-\frac{2\varepsilon^2}{nC^2}\right). \qquad (2.19)$$

The observation underlying our use of McDiarmid's inequality is that the change in our estimate when changing one sample is bounded by twice the mass of the kernel times a constant over n, since changing one sample amounts to moving one instance of the mirrored kernel in our density estimate. In particular, the C we plug into McDiarmid's Inequality decays as n^{-1}.

The proof proceeds as follows. Consider i.i.d. samples $x^1, \ldots, x^n \sim p, y^1, \ldots, y^n \sim q$. In anticipation of using McDiarmid's Inequality, let $\widehat{p}_h'(x)$ denote our kernel density estimate with the sample x^j replaced by $(x^j)'$. By the Logarithm Bound (2.10),

$$\begin{aligned}
&|D_\alpha(\widehat{p}_h \| \widehat{q}_h) - D_\alpha(\widehat{p}_h' \| \widehat{q}_h')| \\
&= \frac{1}{|\alpha - 1|}\left|\log\left(\int_{\mathcal{X}} f(\widehat{p}_h(x), \widehat{q}_h(x))\,dx\right) - \log\left(\int_{\mathcal{X}} f(\widehat{p}_h'(x), \widehat{q}_h(x))\,dx\right)\right| \\
&\leq \frac{C_{\log}}{|\alpha - 1|}\int_{\mathcal{X}} |f(\widehat{p}_h(x), \widehat{q}_h(x)) - f(\widehat{p}_h'(x), \widehat{q}_h(x))|\,dx.
\end{aligned}$$

Then, applying the Mean Value Theorem followed by the bound (2.9) on f's derivatives gives, for some $\xi : \mathcal{X} \to \mathbb{R}^2$ on the line segment between $(\widehat{p}_h, \widehat{q}_h)$ and (p, q),

$$\begin{aligned}
|D_\alpha(\widehat{p}_h \| \widehat{q}_h) - D_\alpha(\widehat{p}_h' \| \widehat{q}_h')| &\leq \frac{C_{\log}}{|\alpha - 1|}\int_{\mathcal{X}} \left|\frac{\partial f}{\partial x_1}(\xi(x))(\widehat{p}_h(x) - \widehat{p}_h'(x))\right|\,dx \\
&\leq \frac{C_f C_{\log}}{|\alpha - 1|}\int_{\mathcal{X}} |\widehat{p}_h(x) - \widehat{p}_h'(x)|\,dx.
\end{aligned}$$

Expanding $\widehat{p}_h$ as per its construction gives

$$\begin{aligned}
|D_\alpha(\widehat{p}_h \| \widehat{q}_h) - D_\alpha(\widehat{p}'_h \| \widehat{q}'_h)| &\leq \frac{C_f C_{\log}}{|\alpha - 1|} \int_{\mathcal{X}} |\widetilde{p}_h(x) - \widetilde{p}'_h(x)| \; dx \\
&\leq \frac{C_f C_{\log}}{|\alpha - 1| n h^d} \sum_{S \in \mathcal{S}} \int_{\mathcal{X}} \left| K_S(x, x^j) - K_S(x, (x^j)') \right| \; dx \\
&\leq \frac{2 C_f C_{\log}}{|\alpha - 1| n h^d} \sup_{y \in \mathcal{X}} \sum_{S \in \mathcal{S}} \int_{\mathcal{X}} |K_S(x, y)| \; dx = \frac{2 C_f C_{\log}}{|\alpha - 1| n} \|K\|_1^d,
\end{aligned}$$

where the last line follows from the triangle inequality and (2.12). An identical proof holds if we vary some y^i rather than x^i. Thus, since we have $2n$ independent samples, McDiarmid's Inequality gives the bound,

$$\mathbb{P}\left(|D_\alpha(\widehat{p}_h, \widehat{q}_h) - \mathbb{E} D_\alpha(\widehat{p}_h, \widehat{q}_h)| > \varepsilon \right) \leq 2 \exp\left(-\frac{C^2 \varepsilon^2 n}{\|K\|_1^{2d}} \right),$$

$$\text{where} \qquad C = \frac{|\alpha - 1|}{2 C_f C_{\log}} \tag{2.20}$$

depends only on κ and α (see Inequalities (2.10) and (2.9) for the exact dependence) . ■

2.3 General Density Functionals

Having shown the desired results for the case of Rényi-α divergence, we are now interested in showing similar results for a larger class of density functionals. Some functionals of interest include other Rényi-α quantities, related Shannon and Tsallis-α quanitities, and $\mathcal{L}_p$-norms and metrics, and the various divergences discussed in Section 1.1.

2.3.1 Problem Statement

For given dimensions $d_1, \dots, d_k \geq 1$, consider random vectors $X_1, \dots, X_k$ on the unit cubes $\mathcal{X}_i := [0,1]^{d_i}$ distributed according to densities $p_i : \mathcal{X}_i \to \mathbb{R}$ (for $i \in \{1, \dots, k\}$). For an appropriately smooth $f : \mathbb{R}^k \to \mathbb{R}$ we are interested in using random sample of n

i.i.d. points from the distribution of each X_i to estimate the functional

$$F(p_1, \dots, p_k) = \int_{\mathcal{X}} f(p_1(x), \dots, p_k(x))\, dx.$$

where $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$.

2.3.2 Estimator

For a given bandwidth h, we use the mirrored kernel density estimator $\hat{p}_i$ to estimate each p_i, and then estimate $F(p_1, \dots, p_k)$ by

$$F(\hat{p}_1, \dots, \hat{p}_k) := \int_{\mathcal{X}} f(\hat{p}_1(x), \dots, \hat{p}_k(x))\, dx.$$

2.3.3 Main Result

If each p_i is in the bounded Hölder class $\Sigma(\beta, L, r, d_i)$ with vanishing boundary derivatives, f is twice continuously differentiable, and the kernel $K : \mathbb{R} \to \mathbb{R}$ has order ℓ and is supported in $[-1, 1]$,

$$|F(p_1, \dots, p_k) - \mathbb{E}F(\hat{p}_1, \dots, \hat{p}_k)| \leq C \left(h^\beta + h^{2\beta} + \frac{1}{nh^d} \right)$$

for some $C \in \mathbb{R}$ not depending on n or h.

On the other hand without any conditions on p_i, if $f : [\kappa_1, \kappa_2] \to \mathbb{R}$ is Lipschitz continuous with constant C_f and $K \in \mathcal{L}_1$, then

$$P\left(|F(\hat{p}) - \mathbb{E}F(\hat{p})| > \varepsilon\right) \leq 2 \exp\left(-\frac{2\varepsilon^2 n}{C_V^2} \right).$$

2.3.4 Bias Bound

Assumptions

We assume that each p_i is in the bounded Hölder class $\Sigma(\beta, L, r, d_i)$ with vanishing boundary derivatives, and that $f : p_1(\mathcal{X}_1) \times \cdots \times p_k(\mathcal{X}_k) \to \mathbb{R}$ is twice continuously differentiable, with first and second partial derivatives bounded in magnitude by $C_f \in \mathbb{R}$.

We also assume the kernel $K : \mathbb{R} \to \mathbb{R}$ has bounded support $[-1, 1]$ and is of order ℓ.

Proof of Bias Bound

By Taylor's Theorem, $\forall x \in \mathcal{X}$, for some $\xi : \mathcal{X} \to \mathbb{R}$ on the line segment between $\widetilde{p}_h(x)$ and $p(x)$,

$$\begin{aligned}
&|\mathbb{E} f((\widetilde{p}_h)_1(x), \dots, (\widetilde{p}_h)_k(x)) - f(p_1(x), \dots, p_k(x))| \\
&= \left| \mathbb{E} \nabla f(p(x)) \cdot (\widetilde{p}_h(x) - p(x)) + \frac{1}{2} (\widetilde{p}_h(x) - p(x))^T H(f)(\xi)(\widetilde{p}_h(x) - p(x)) \right| \\
&\leq C_f \left(\sum_{i=1}^{k} |B_{p_i}(x)| + \sum_{i<j\leq k} |B_{p_i}(x)||B_{p_j}(x)| + \sum_{i \leq k} \mathbb{E} \left[\widetilde{p}_h(x) - p(x) \right]^2 \right)
\end{aligned}$$

Hence, applying Hölder's Inequality,

$$\begin{aligned}
|\mathbb{E} F(\widetilde{p}_h) - F(p)| &\leq \int_{\mathcal{X}} |\mathbb{E} f(\widetilde{p}_h(x)) - f(p(x))| \, dx \\
&\leq C_f \int_{\mathcal{X}} \sum_{i=1}^{k} |B_{p_i}(x)| + \sum_{i<j\leq k} |B_{p_i}(x)||B_{p_j}(x)| + \sum_{i\leq k} B_{p_i}^2(x) \, dx \\
&\leq C_f \sum_{i=1}^{k} \left(\sqrt{\int_{\mathcal{X}} B_p^2(x) \, dx} + \int_{\mathcal{X}} \mathbb{E}[\widetilde{p}_h(x) - p(x)]^2 \, dx \right) \\
&\quad + C_f \sum_{i<j\leq k} \sqrt{\int_{\mathcal{X}} B_{p_i}^2(x) \int_{\mathcal{X}} B_{p_j}^2(x) \, dx}.
\end{aligned}$$

Applying the Bias Lemma (Inequality 2.13) and standard results in kernel density estimation (see, for example, [38]) gives

$$\begin{aligned}
|\mathbb{E} F((\widetilde{p}_h)_1, \dots, (\widetilde{p}_h)_k) - F(p_1, \dots, p_k)| &\leq (C_2 + C_3)^2 \left(k^2 h^\beta + k h^{2\beta} \right) + k \kappa_2 \frac{\|K\|_1^d}{n h^d} \\
&\leq C \left(h^\beta + h^{2\beta} + \frac{1}{n h^d} \right)
\end{aligned}$$

for some $C > 0$ not depending on n or h, where $d = \max\{d_1, \dots, d_k\}$.

2.3.5 Variance Bound

Assumptions

We assume that f is Lipschitz continuous with constant C_f in the 1-norm on $p_1(\mathcal{X}_1) \times \cdots \times p_k(\mathcal{X}_k)$, and that $K \in \mathcal{L}_1$.

Proof of Variance Bound

Consider i.i.d. samples $x^1, \dots, x^n \sim p$. In anticipation of using McDiarmid's Inequality (recall 2.19), let $\hat{p}'$ denote our kernel density estimate with the sample x^i replaced by a new sample $(x^i)'$. By the Mean Value Theorem

$$\begin{aligned}
&|F(\hat{p}) - F(\hat{p}')| \leq C_f \|\hat{p} - \hat{p}'\|_1 \leq C_f \left(\sum_{j=1}^{k} \int_{\mathcal{X}_j} |\hat{p}_j(x) - \hat{p}_j(x)| \, dx_j \right) \\
&\leq \frac{C_f}{nh} \left(\sum_{j=1}^{k} \int_{\mathcal{X}_j} \left| K_{d_j}\left(\frac{x_j - x_j^i}{h} \right) - K_{d_j}\left(\frac{x_j - (x_j^i)'}{h} \right) \right| \, dx_j \right) \\
&\leq \frac{C_f}{n} \left(\sum_{j=1}^{k} \int_{\mathcal{X}_j} \left| K_{d_j}\left(x_j - x_j^i \right) - K_{d_j}\left(x_j - (x_j^i)' \right) \right| \, dx_j \right) \\
&\leq \frac{2C_f}{n} \sum_{j=1}^{k} \|K\|_1^{d_j} =: \frac{C_V}{n},
\end{aligned}$$

so that McDiarmid's Inequality gives

$$P\left(|F(\hat{p}) - \mathbb{E}F(\hat{p})| > \varepsilon \right) \leq 2 \exp\left(-\frac{2\varepsilon^2}{nC_V^2/n^2} \right) = 2 \exp\left(-\frac{2\varepsilon^2 n}{C_V^2} \right).$$

Thus,

$$\begin{aligned}
\mathbb{V}[F(\hat{p}_1, \cdots, \hat{p}_k)] &= \mathbb{E}\left[\left(F(\hat{p}_1, \cdots, \hat{p}_k) - \mathbb{E}F(\hat{p}_1, \cdots, \hat{p}_k) \right)^2 \right] \\
&= \int_0^\infty \mathbb{P}\left(F(\hat{p}_1, \cdots, \hat{p}_k) - \mathbb{E}F(\hat{p}_1, \cdots, \hat{p}_k))^2 > \varepsilon \right) \, d\varepsilon \\
&\leq \int_0^\infty 2 \exp\left(-\frac{2\varepsilon n}{C_V^2} \right) \, d\varepsilon = \frac{C_V^2}{n}.
\end{aligned}$$

Optimizing over h (so $h \asymp n^{\frac{1}{d+\beta}}$) gives a mean squared error of

$$\mathbb{E}\left[(F(\widehat{p}_h) - F(p))^2\right] = \frac{C_V^2}{n} + C_B^2\left(h^\beta + h^{2\beta} + \frac{1}{nh^d}\right)^2 \in O\left(n^{-1} + n^{-\frac{2\beta}{d+\beta}}\right).$$

which is in $O(n^{-1})$ if $\beta \geq d$ and in $O\left(n^{-\frac{2\beta}{d+\beta}}\right)$ otherwise. This asymptotic rate is consistent with previous bounds in density functional estimation [4, 35].

2.4 Application to Rényi-α Conditional Mutual Information

We now apply our general result to the case of Renyi-α Conditional Mutual Information, a important because of its applicability to conditional independence testing (see Section 3.2 for further discussion).

2.4.1 Problem Statement

For given dimensions $d_x, d_y, d_z \geq 1$, consider random vectors X, Y, and Z distributed on unit cubes $\mathcal{X} := [0,1]^{d_x}, \mathcal{Y} := [0,1]^{d_y}$ and $\mathcal{Z} := [0,1]^{d_z}$ according to a joint density $P_{X,Y,Z} : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$. For a given $\alpha \in (0,1) \cup (1,\infty)$, we are interested in using a random sample of $4n$ i.i.d. points from P to estimate the Renyi-α conditional mutual information of X and Y given Z:

$$I_\alpha(X;Y|Z) = \int_{\mathcal{Z}} P(z) \int_{\mathcal{X}\times\mathcal{Y}} \left(\frac{P(x,y,z)}{P(z)}\right)^\alpha \left(\frac{P(x,z)P(y,z)}{P^2(z)}\right)^{1-\alpha} d(x,y)\,dz$$

where $P(z), P(x,y), P(x,z)$, and $P(y,z)$ denote the respective marginal distributions.

2.4.2 Assumptions

Density Assumptions: We assume the joint density $P_{X,Y,Z}(x,y,z)$ is in the Hölder class $\Sigma(\beta, L, r, d_x + d_y + d_z)$ with vanishing boundary derivatives, and, furthermore, there exists $\kappa = (\kappa_1, \kappa_2) \in (0,\infty)^2$ with $\kappa_1 \leq P_{X,Y,Z} \leq \kappa_2$.

Kernel Assumptions: We assume the kernel $K : \mathbb{R} \to \mathbb{R}$ has bounded support $[-1,1]$ and is of order ℓ, as defined in (2.7).

2.4.3 Proof

In order to apply the Mean Value Theorem, define lower and upper bounds, respectively, to the argument of the logarithm.

$$\begin{aligned}
& \kappa_* := \kappa_1^\alpha \min\left\{\kappa_1^{2(1-\alpha)}\kappa_2^{\alpha-2}, \kappa_2^{2(1-\alpha)}\kappa_2^{\alpha-2}, \kappa_2^{2(1-\alpha)}\kappa_1^{\alpha-2}\right\} = (\kappa_1/\kappa_2)^{\max\{2-\alpha,\alpha,2(\alpha-1)\}} \\
& \quad \leq \int_{\mathcal{X}\times\mathcal{Y}} P^\alpha(x,y,z)\left(P(x,z)P(y,z)\right)^{1-\alpha} P^{\alpha-2}(z)\, d(x,y) \\
\text{and} \quad & \kappa^* := \kappa_*^{-1} = (\kappa_2/\kappa_1)^{\max\{2-\alpha,\alpha,2(\alpha-1)\}} \\
& \quad \geq \int_{\mathcal{X}\times\mathcal{Y}} P^\alpha(x,y,z)\left(P(x,z)P(y,z)\right)^{1-\alpha} P^{\alpha-2}(z)\, d(x,y)
\end{aligned}$$

and define $f_\alpha : [\kappa_1,\kappa_2]^4 \to \mathbb{R}$ by $f_\alpha(w,x,y,z) := (w^\alpha (xy)^{1-\alpha} z^{\alpha-2})$, noting that f is Lipschitz on this domain. Applying the Mean Value Theorem to the logarithm,

$$\begin{aligned}
& |1-\alpha||\hat{I}_\alpha(X;Y|Z) - \hat{I}'_\alpha(X;Y|Z)| \\
& \leq \frac{\kappa_2}{\kappa_*}\int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} \Big| P^\alpha(x,y,z)\left(P(x,z)P(y,z)\right)^{1-\alpha} P^{\alpha-2}(z) \\
& - \left(P'(x,y,z)\right)^\alpha \left(P'(x,z)P'(y,z)\right)^{1-\alpha} \left(P'(z)\right)^{\alpha-2}\Big|\, d(x,y,z) + \log(\kappa^*)\int_{\mathcal{Z}} |P(z) - P'(z)|\, dz \\
& = \kappa_2\kappa^* \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} |f(P(x,y,z),P(x,z),P(y,z),P(z)) - f(P'(x,y,z),P'(x,z),P'(y,z),P'(z))|\, d(x,y,z) \\
& + \log(\kappa^*)\int_{\mathcal{Z}} |P(z) - P'(z)|\, dz = \kappa_2\kappa^*|F(P) - F(P')| + \log(\kappa^*)|G(P) - G(P')| \\
& \leq \kappa_2\kappa^* \frac{8C_f\|K\|_1^{d_x+d_y+d_z}}{n} + \log(\kappa^*)\frac{2\|K\|_1^{d_z}}{n} = Cn^{-1},
\end{aligned}$$

for $C := 2\|K\|_1^{d_z}\left(4\kappa_2\kappa^* C_f \|K\|_1^{d_x+d_y} + \log(\kappa^*)\right)$. Hence, McDiarmid's Inequality gives

$$P\left(|I(\hat{P}) - \mathbb{E}I(\hat{P})| > \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2 n}{C^2}\right),$$

which in turn gives the Variance Bound $\mathbb{V}[I(\hat{P})] \leq C_V^2 n^{-1}$, giving the Mean Square Error bound

$$\mathbb{E}\left[(F(\widehat{p}_h) - F(p))^2\right] = C_V^2 n^{-1} + C_B^2\left(h^\beta + h^{2\beta} + \frac{1}{nh^d}\right)^2 \in O\left(n^{-1} + n^{-\frac{2\beta}{d+\beta}}\right),$$

(where $d = d_x + d_y + d_z$) which is in $O(n^{-1})$ if $\beta \geq d$ and in $O\left(n^{-\frac{2\beta}{d+\beta}}\right)$ otherwise. [4, 35].

2.5 Comments regarding the assumptions

Suppose a p_i is γ times continuously differentiable for a positive integer γ. Since $\mathcal{X}$ is compact, the γ-order derivatives of p_i are bounded. Hence, since $\mathcal{X}$ is convex, the $(\gamma - 1)$-order derivatives of p_i are Lipschitz, by the Mean Value Theorem. Consequently, any degree of continuous differentiability suffices for the Hölder condition.

The existence of an upper bound κ_2 is trivial, since each p_i is continuous and $\mathcal{X}$ is compact. The existence of a positive lower bound κ_1 for (some) p_i's is in some cases quite a natural assumption. For example, in the case of Rényi-α divergence, it is natural to assume that the reference density (our q) is bounded below almost everywhere as otherwise the Rényi-α divergence may be infinite. In other cases, the existence of κ_1 is a technical necessity due to certain singularities at 0 (for example, the Logarithm Bound (2.10)). In some cases (including the important special case of Rényi-α entropy (i.e., Rényi-α divergence with respect to the uniform distribution $q = 1$)), the assumption of κ_1 for p can still be dropped via an argument using Jensen's Inequality (when $\alpha > 1$, this argument requires the domain to have finite measure; see the next paragraph).

Understanding the choice of a seemingly well-behaved (in particular, bounded) domain such as the unit cube $[0, 1]^d$ is important for understanding the ramifications of this work. Unlike many problems in analysis, non-parametric estimation problems can be significantly more difficult on bounded domains, because of lack of data near the boundary, which typically leads to boundary bias. [9] However, when working with, for example, information-theoretic density functionals, it can be important to establish lower bounds on the density in question, which, of course, cannot be done if the domain has infinite Lebesgue measure, since the density must have unit mass. Indeed, it is possible that the combined necessity and difficulty of working on bounded domains is a major reason for the lack of general results concerning non-parametric estimation of information-theoretic density functionals;

[9] [13] suggests how, given bounds on the mean and variance of the random variables, our results can be extended to the case of an unbounded domain in a straightforward manner.

certainly, it is the reason for one of the restrictive and artificial assumptions we make: that the density in question has derivatives vanishing at the boundary of the domain.

The assumed Hölder condition, in combination with Taylor's Theorem, lies at the core of our bounds of the density functional estimator's bias. It is straightforward to adapt these assumptions and our results to those Sobolev spaces which embed into Hölder spaces. However, based on previous work (see Section 1.3 of [38]) using Fourier analysis to bound the error of Kernel Density Estimators over the Sobolev spaces $H^\ell = W^{\ell,2}$, it seems likely that an interesting new family of bias bounds might hold over these spaces. This may be a worthwhile topic for future work on the subject, although the generalizing to the multidimensional case may be somewhat more complicated here.

The proof of the bias bound also depends critically on the assumption that the kernel is of order ℓ. When $\beta > 2$, this, implies that K is non-positive on a subset of $[-1, 1]$ of positive measure. In combination with the assumption that $\int_{-1}^{1} K(x)\,dx = 1$, this implies $\|K\|_1 > 1$. Since the final bounds on both bias and variance include $\|K\|_1^d$ terms, both bounds increase exponentially in the dimension d. One possible solution is to modify the product kernel in such a manner that its L_1-norm is not exponential in the dimension. Using a radial kernel, for example, might reduce this somewhat, but the dependence would still be exponential in d. Another perhaps more promising solution is to consider Fourier analysis proofs over Sobolev spaces, as discussed above. The basic forms of these proofs require only that K have mean 0 each $K, p_i \in \mathcal{L}_2$.

Chapter 3

Conclusions and Future Work

3.1 Conclusion

In this paper we derived a finite sample exponential concentration bound for a consistent, nonparametric density functional estimator. To the best of our knowledge this is the first such exponential concentration bound for Renyi divergence.

3.2 Future Work

One of the primary motivations for studying conditional mutual information estimation is in determining conditional independence of two variables given a third variable or family of variables. This is important, for example, in determining graph structure in graphical models. Given three variables, X, Y, and Z, we would like to determine whether knowing

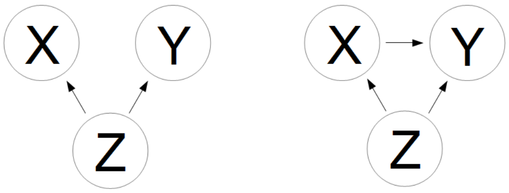

Figure 3-1: Two possible graphs of dependence between the variable X,Y, and Z. Our results suggest a hypothesis test for distinguishing the two.

the value of Z is sufficient to explain away any dependence between X and Y, in order to differentiate between the two graphical models illustrated in Figure 3-1. Hence, a useful extension of this work would be to establish a hypothesis test for conditional independence. This could be performed by estimating conditional mutual information using our estimator, and then computing an appropriate confidence interval about that estimate using the error bounds we derive. X and Y would then be considered conditionally independent given Z (at a particular confidence level) if and only if this confidence interval contains 0.

Appendix A

Experimental Results

A.1 Experiment

We used our estimator to estimate the Rényi α-divergence between two normal distributions in $\mathbb{R}^3$ restricted to the unit cube. In particular, for

$$\vec{\mu}_1 = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.3 \end{bmatrix}, \quad \vec{\mu}_2 = \begin{bmatrix} 0.7 \\ 0.7 \\ 0.7 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.2 \end{bmatrix},$$

$p = \mathcal{N}(\vec{\mu}_1, \Sigma), q = \mathcal{N}(\vec{\mu}_2, \Sigma)$. For each $n \in \{1, 2, 5, 10, 50, 100, 500, 1000, 2000, 5000\}$, n data points were sampled according to each distribution and constrained (via rejection sampling) to lie within $[0, 1]^3$. Our estimator was computed from these samples, for $\alpha = 0.8$, using the Epanechnikov Kernel $K(u) = \frac{3}{4}(1 - u^2)$ on $[-1, 1]$, with bandwidth $h = 0.25$. The true α-divergence was computed directly according to its definition on the (renormalized) distributions on $[0, 1]^3$. The bias and variance of our estimator were then computed in the usual manner based on 100 trials. Figure A-1 shows the error and variance of our estimator for each n.

We also compared our estimator's empirical error to our theoretical bound. Since the distributions used are infinitely differentiable, $\beta = \infty$, and so the estimator's mean squared error should converge as $O(n^{-1})$. An appropriate constant multiple was computed from

(2.20), (2.16), and (2.18). The resulting bound is also shown in Figure A-1.

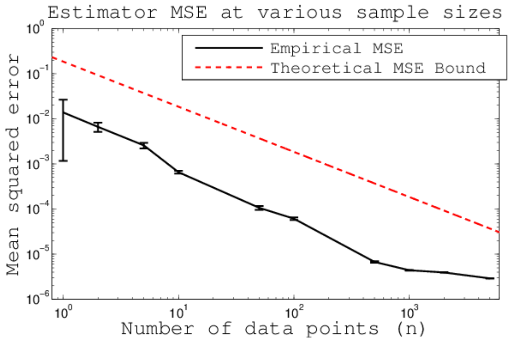

Figure A-1: Log-log plot of mean squared error (computed over 100 trials) of our estimator for various sample sizes n, alongside our theoretical bound. Error bars indicate standard deviation of estimator over 100 trials.

Appendix B

A Note on The Special Case $\alpha \to 1$

Suppose $\mathcal{X} \subseteq \mathbb{R}^n$ has finite Lebesgue measure $0 < \mu(\mathcal{X}) < \infty$, and let $p : \mathcal{X} \to [0, \infty)$ be a probability density on $\mathcal{X}$. Assume p is continuous and $\exists \kappa_1, \kappa_2 \in (0, \infty)$ with $\kappa_1 \leq p(x) \leq \kappa_2, \forall x \in \mathcal{X}$. Define the Shannon entropy

$$H(p) := -\int_{\mathcal{X}} p(x) \log p(x)\, dx \in [-\infty, +\infty] \tag{B.1}$$

and, for $\alpha \in (0, \infty) \backslash \{1\}$, define Rényi-$\alpha$ entropy

$$H_\alpha(p) := \frac{1}{1-\alpha} \log \int_{\mathcal{X}} p^\alpha(x)\, dx. \tag{B.2}$$

It is well-known that $H_\alpha(p) \to H(p)$ as $\alpha \to 1$. Indeed, applying l'Hospital's rule, differentiating under the integral sign, [1] and noting $\int_{\mathcal{X}} p(x) dx = 1$,

$$\begin{aligned}
\lim_{\alpha \to 1} \frac{\log \int_{\mathcal{X}} p^\alpha(x)\, dx}{1-\alpha} &= -\lim_{\alpha \to 1} \frac{d}{d\alpha} \log \int_{\mathcal{X}} p^\alpha(x)\, dx = -\lim_{\alpha \to 1} \frac{\frac{d}{d\alpha} \int_{\mathcal{X}} p^\alpha(x)\, dx}{\int_{\mathcal{X}} p^\alpha(x)\, dx} \\
&= -\lim_{\alpha \to 1} \frac{\int_{\mathcal{X}} p^\alpha(x) \log p(x)\, dx}{\int_{\mathcal{X}} p^\alpha(x)\, dx} = -\int_{\mathcal{X}} p(x) \log p(x)\, dx = H(p).
\end{aligned}$$

(since p is bounded, the convergence in α is uniform, and hence the integrals converge).

Under similar assumptions, we can similarly show that, as $\alpha \to 1$, Rényi-α divergence

[1] This can be justified rigorously using the facts that $\mathcal{X}$ has finite measure and the function $(\alpha, x) \mapsto p^\alpha(x)$ is continuously differentiable in α and continuous in x.

converges to Kullback–Leibler divergence, Rényi-α Conditional Mutual Information converges to Shannon Conditional Mutual Information, etc. Shannon quantities are uniquely useful in application because of theorems in Source Coding and Channel Coding for which they were first studied, and also because of the many algebraic relationships between them. However, they can be difficult to study analytically due to the explosive behavior of their $\log$ terms (in particular, their resulting sensitivity to low probability outcomes). Because of this, and also because of a results about Rényi-α quantities for other values of α (typically, $\alpha \in \{0, 1/2, 2, \infty\}$), it is common to study the analytic properties of Rényi-α quantities.

However, in order to extend estimation error bounds for Rényi-α quantities to error bounds for Shannon quanitites, it is often necessary to understand the rate of convergence as $\alpha \to 1$. Understanding this is not straightforward, because the convergence is shown using l'Hospital's rule, with the denominator of $1 - \alpha$ vanishing, and consequently, it is difficult to find rate in the literature on the convergence of Rényi-α quantities as $\alpha \to 1$. Here, we make an observation for the case of Rényi-α entropy, which may suggest new means of bounding the error of approximating this limit in general.

B.1 Main Idea

The difficulty in proving a rate for the convergence of $H_\alpha(p) \to H(p)$ as $\alpha \to 1$ appears to be due to the vanishing $1-\alpha$ term in the denominator. Letting $g_p : (0, \infty) \to \mathbb{R}$ defined by

$$g_p(\alpha) := \log \int_{\mathcal{X}} p^\alpha(x)\, dx,$$

observe that

$$H(p) = \lim_{\alpha \to 1} H_\alpha(p) = -\lim_{\alpha \to 1} \frac{g_p(\alpha) - g_p(1)}{\alpha - 1} = -g_p'(\alpha)\bigg|_{\alpha=1}.$$

Since g_p is very smooth near $\alpha = 1$, we show that two estimates of g_p can be used to estimate the value of g' using a simple secant approximation. Furthermore, $g_p(\alpha)$ is naturally estimated as part of many $H_\alpha(p)$ estimators, with error bounds not depending on α, and so this result complements such estimators of Rényi-α entropy.

B.2 Proof:

Since we are bounding the error of approximating a first derivative, it is natural to bound the second derivative:

Lemma:

$$\left|g_p''(\alpha)\right| \leq B := 2\kappa^2 \left(\frac{\kappa_2}{\kappa_1}\right)^{2\alpha}.$$

Proof: By differentiation under the integral sign and the quotient rule, noting that

$$\frac{d}{d\alpha} p^\alpha(x) = p^\alpha(x) \log p(x) \quad \forall x \in \mathcal{X},$$

$$\begin{aligned} \left|g_p''(\alpha)\right| &= \left| \frac{d}{d\alpha} \frac{\int_{\mathcal{X}} p^\alpha(x) \log p(x)\, dx}{\int_{\mathcal{X}} p^\alpha(x)\, dx} \right| \\ &= \left| \frac{\left(\int_{\mathcal{X}} p^\alpha(x) \log^2 p(x)\, dx\right) \left(\int_{\mathcal{X}} p^\alpha(x)\, dx\right) - \left(\int_{\mathcal{X}} p^\alpha(x) \log p(x)\, dx\right)^2}{\left(\int_{\mathcal{X}} p^\alpha(x)\, dx\right)^2} \right| \\ &\leq \frac{2\kappa_2^{2\alpha} \kappa^2 \mu(\mathcal{X})^2}{\kappa_1^{2\alpha} \mu(\mathcal{X})^2} = 2\kappa^2 \left(\frac{\kappa_2}{\kappa_1}\right)^{2\alpha}. \end{aligned}$$

By virtue of this lemma, it is easy to bound the error of approximating $g_p'(1)$ by a secant.

Main Result: $\forall h \in (0, 1)$,

$$\left| \frac{g(1 + h) - g(1 - h)}{2h} - g'(1) \right| \leq Bh.$$

Proof: Applying the Lemma and the Mean Value Theorem repeatedly,

$$g'(1) - Bh \leq \inf_{\alpha \in (1-h, 1+h)} g'(x) \leq \frac{g(1 + h) - g(1 - h)}{2h} \leq \sup_{\alpha \in (1-h, 1+h)} g'(x) \leq g'(1) + Bh.$$

Bibliography

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

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

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

[4] L. Birge and P. Massart. Estimation of integral functions of a density. *The Annals of Statistics*, 23:11–29, 1995.

[5] Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schlkopf, and Alex J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[6] L. Bregman. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:79–86, 1967.

[7] Barry Chai, Dirk B. Walther, Diane M. Beck, and Li Fei-Fei. Exploring functional connectivity of the human brain using multivariate information analysis. In *NIPS*, 2009.

[8] I. Csiszár. Information-type measures of differences of probability distributions and indirect observations. *Studia Sci. Math. Hungarica*, 2:299–318, 1967.

[9] 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*, 17:277–297, 2005.

[10] M. Gupta and S. Srivastava. Parametric bayesian estimation of differential entropy and relative entropy. *Entropy*, 12:818–843, 2010.

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

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

[13] Jean Honorio and Tommi Jaakkola. Two-sided exponential concentration bounds for bayes error rate and shannon entropy. In *ICML*, pages 459–467, 2013.

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

[15] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

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

[17] B. Laurent. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24:659–681, 1996.

[18] Erik G. Learned-Miller and John W. Fisher. ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003.

[19] N. Leonenko and L. Pronzato. Correction of 'a class of Rényi information estimators for mulitidimensional densities' Ann. Statist., 36(2008) 2153-2182, 2010.

[20] Nikolai Leonenko, Luc Pronzato, and Vippal Savani. A class of Rényi information estimators for multidimensional densities. *Annals of Statistics*, 36(5):2153–2182, 2008.

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

[22] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37:145151, 1991.

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

[24] Colin McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141:148–188, 1989.

[25] 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, To appear.*, 2010.

[26] J. Oliva, B. Poczos, and J. Schneider. Distribution to distribution regression. In *International Conference on Machine Learning (ICML)*, 2013.

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

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

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

[30] B. Poczos and J. Schneider. On the estimation of alpha-divergences. In *International Conference on AI and Statistics (AISTATS)*, volume 15 of *JMLR Workshop and Conference Proceedings*, pages 609–617, 2011.

[31] B. Poczos, L. Xiong, D. Sutherland, and J. Schneider. Nonparametric kernel estimators for image classification. In *25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[32] A. Rényi. On measures of entropy and information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press*, 1961.

[33] A. Rényi. *Probability Theory*. North-Holland Publishing Company, Amsterdam, 1970.

[34] Caifeng Shan, Shaogang Gong, and Peter W. Mcowan. Conditional mutual information based boosting for facial expression recognition. In *British Machine Vision Conference (BMVC)*, 2005.

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

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

[37] Z. Szabó, B. Póczos, and A. Lőrincz. Undercomplete blind subspace deconvolution. *Journal of Machine Learning Research*, 8:1063–1095, 2007.

[38] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[39] T. Villmann and S. Haase. Mathematical aspects of divergence based vector quantization using Frechet-derivatives, 2010. University of Applied Sciences Mittweida.

[40] Qinq Wang, Sanjeev R. Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5), 2009.

[41] Eric Wolsztynski, Eric Thierry, and Luc Pronzato. Minimum-entropy estimation in semi-parametric models. *Signal Process.*, 85(5):937–949, 2005.