Minimax Estimation of Information Measures

Jiantao Jiao
Stanford University
jiantao@stanford.edu

Kartik Venkat
Stanford University
kvenkat@stanford.edu

Yanjun Han
Tsinghua University
hanyj11@mails.tsinghua.edu.cn

Tsachy Weissman
Stanford University
tsachy@stanford.edu

***Abstract*—We propose a general methodology for the construction and analysis of minimax estimators for functionals of discrete distributions, where the support size S is unknown and may be comparable to the number of observations n. We illustrate the merit of our approach by thoroughly analyzing non-asymptotically the performance of the resulting schemes for estimating two important information measures: the entropy $H(P) = \sum_{i=1}^{S} -p_i \ln p_i$ and $F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, \alpha > 0$. We obtain the minimax L_2 risks for estimating these functionals up to a universal constant. In particular, we demonstrate that our estimator achieves the optimal sample complexity $n \gg S/\ln S$ for entropy estimation. We also demonstrate that the sample complexity for estimating $F_\alpha(P), 0 < \alpha < 1$ is $n \gg S^{1/\alpha}/\ln S$, which can be achieved by our estimator and not by the popular plug-in Maximum Likelihood Estimator (MLE). For $1 < \alpha < 3/2$, we show the minimax L_2 rate for estimating $F_\alpha(P)$ is $(n \ln n)^{-2(\alpha-1)}$ regardless of the support size, while the exact L_2 rate for the MLE is $n^{-2(\alpha-1)}$. For all the above cases, the behavior of the minimax rate-optimal estimators with n samples is essentially that of the MLE with $n \ln n$ samples. Finally, we highlight the practical advantages of our schemes for the estimation of entropy and mutual information.**

I. Introduction

One of the key tasks of information theory is to characterize fundamental limits of operational problems by means of information measures, namely, functionals of probability distributions or conditional distributions (channels). Among the most fundamental of such functionals is the entropy [1],

$$H(P) \triangleq \sum_{i=1}^{S} -p_i \ln p_i. \tag{1}$$

Another widely applicable information measure which we shall consider in detail is the functional $F_\alpha(P)$:

$$F_\alpha(P) \triangleq \sum_{i=1}^{S} p_i^\alpha, \alpha > 0. \tag{2}$$

The significance of functional $F_\alpha(P)$ can be seen via the connection $H_\alpha(P) = \frac{\ln F_\alpha(P)}{1-\alpha}$, where $H_\alpha(P)$ is the Rényi entropy [2], which also emerges in operational roles in information theory [3] [4]. In addition to their prominent operational roles in the traditional realms of information theory, information measures such as the ones above have found numerous applications, among other fields, in statistics and machine learning, biology, neuroscience, image processing, linguistics, secrecy, ecology, physics, finance, etc. In most real-world inferential applications, the true underlying distribution that generates the data is unknown. Thus many statistical modeling, signal processing and machine learning tasks rest upon data-driven procedures for accurately estimating information measures.

A. Problem formulation

Given n independent samples from an unknown discrete probability distribution $P = (p_1, p_2, \ldots, p_S)$, with *unknown* support size S, consider the problem of estimating a functional of the distribution of the form:

$$F(P) = \sum_{i=1}^{S} f(p_i), \tag{3}$$

where $f : [0, 1] \to \mathbb{R}$ is analytic[1] at $(0, 1]$, and $f(0) = 0$. Note that this includes the information measures discussed above. Denote by $\mathcal{M}_S$ all discrete distributions with support size S. Regarding the task of estimating functional $F(P)$, the L_2 risk of an arbitrary estimator $\hat{F}$, which is a Borel measurable function of the observations, is defined as $\mathbb{E}_P \left(F(P) - \hat{F}\right)^2$, where the expectation is taken with respect to the distribution P that generates the observations used by $\hat{F}$. The L_2 risk is a function of both the *unknown* distribution P and the estimator $\hat{F}$, and our goal is to minimize this risk. Since P is unknown, we cannot directly minimize it, but if we want to do well no matter what the true distribution P is, we may want to adopt the minimax criterion [5], and try to minimize the *maximum risk*

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(F(P) - \hat{F}\right)^2. \tag{4}$$

The estimator that minimizes the maximum risk above is called the *minimax* estimator, and the corresponding risk is called the *minimax* risk. The exact solution to this problem for general $F(P)$ seems intractable. Indeed, it corresponds to a high dimensional non-convex optimization problem, which in general does not admit an efficient solution. In this work, our goal is to design *minimax rate-optimal* estimators whose maximum risk is equal to the minimax risk up to a universal multiplicative constant.

Notation: Given non-negative sequences a_γ, b_γ, we use the notation $a_\gamma \lesssim b_\gamma$ to denote that there exists a universal constant C such that $\sup_\gamma \frac{a_\gamma}{b_\gamma} \leq C$. Notation $a_\gamma \asymp b_\gamma$ is equivalent to $a_\gamma \lesssim b_\gamma$ and $b_\gamma \lesssim a_\gamma$. Notation $a_\gamma \ll b_\gamma$ means $\limsup_\gamma \frac{a_\gamma}{b_\gamma} = 0$.

[1] A function f is analytic at a point x_0 if and only if its Taylor series about x_0 converges to f in some neighborhood of x_0.

B. A brief history

A natural estimator for functionals of the form (3) is the maximum likelihood estimator (MLE), or plug-in estimator, which simply evaluates $F(P_n)$, where P_n is the empirical distribution of the data. How well does the MLE perform? Interestingly, if we focus on n i.i.d. observations from a distribution with support size S, then the problem of estimating $F(P)$ becomes amenable to using the MLE when S is fixed, and the number of observations $n \to \infty$. This maximum likelihood estimator is *asymptotically efficient* [6, Chapter 8]. Thus, the MLE is unbeatable in the asymptotic regime, where S is fixed, and n grows without bound. However the apparent optimality of the MLE is misleading when S might be growing with n, where it can be shown to be *strictly sub-optimal* [7]. Hence, it is an intriguing question to investigate the optimal estimator given finitely many samples.

The literature hinted that it is possible to come up with *consistent* entropy estimators that only require $n \ll S$ samples. The earliest indication appeared in Paninski [8], in which he showed that there exists a consistent entropy estimator that requires only sublinear samples, but only an existential proof based on the Stone–Weierstrass theorem was provided. It was therefore a breakthrough when Valiant and Valiant [9] introduced the first explicit entropy estimator requiring a sublinear number of samples. In [9], they showed that $n \gg S/\ln S$ samples are both necessary and sufficient to estimate the entropy of a discrete distribution. However, the entropy estimators based on linear programming proposed in Valiant and Valiant [9], [10] have not been shown to achieve the minimax risk. Moreover, the scheme of [9] can only be applied to functionals that are Lipschitz continuous with respect to a Wasserstein metric, which can be roughly understood as those functionals that are "smoother" than entropy. Notably, this does not include the functional $F_\alpha, \alpha < 1$ and other interesting nonsmooth functionals of distributions.

Conceivably, there is a fundamental connection between the smoothness of a functional, and the hardness of estimating it. The ideal solution to this problem would be systematic and capture this trade-off for nearly every functional. This motivates our present work, in which we provide a general framework and procedure for minimax estimation of functionals with non-asymptotic performance guarantees.

II. Main results

A. Our estimators

Our main goal in this work is to present a general approach to the construction of minimax rate-optimal estimators for functionals of discrete distributions. To illustrate our approach, we describe and analyze explicit constructions for the specific cases of entropy $H(P)$ and $F_\alpha(P)$. Our estimators are agnostic with respect to the support size S, and achieve the minimax L_2 rates.

Our approach is to tackle the estimation problem separately for the cases of "small" values of p and "large" values of p (for both $H(P)$ and $F_\alpha(P)$ estimation), corresponding respectively to treating regions where the functional is "nonsmooth" and "smooth" in different ways. In the nonsmooth region, we rely on the best polynomial approximation of the function f by employing an unbiased estimator for this approximation. The part pertaining to the smooth region is estimated by a bias-corrected maximum likelihood estimator. We apply this procedure coordinate-wise based on the empirical distribution of each observed symbol, and finally sum the respective estimates. The best polynomial approximation for a function $f(x)$ on domain A with order no more than K is defined as

$$P_K^*(x) \triangleq \operatorname*{argmin}_{P \in \mathsf{poly}_K} \max_{x \in A} |f(x) - P(x)|, \tag{5}$$

where poly_K is the collection of polynomials with order at most K on A.

We now look at the specific cases of entropy and $F_\alpha(P)$ separately. For the entropy, after we obtain the empirical distribution P_n, for each coordinate $P_n(i)$, if $P_n(i) \ll \ln n/n$, we (i) compute the best polynomial approximation for $-p_i \ln p_i$ in the regime $0 \le p_i \ll \ln n/n$, (ii) use the unbiased estimators for integer powers p_i^k to estimate the corresponding terms in the polynomial approximation for $-p_i \ln p_i$ up to order $K_n \sim \ln n$, and (iii) use that polynomial as an estimate for $-p_i \ln p_i$. If $P_n(i) \gg \ln n/n$, we use the estimator $-P_n(i) \ln P_n(i) + \frac{1}{2n}$ to estimate $-p_i \ln p_i$. Then, we add the estimators corresponding to each coordinate. Our estimator for $F_\alpha(P)$ is very similar to that of entropy, with the only difference that we conduct polynomial approximation for x^α with order $K_n \sim \ln n$, and use the estimator $\left(1 + \frac{\alpha(1-\alpha)}{2nP_n(i)}\right) P_n^\alpha(i)$ when $P_n(i) \gg \ln n/n$.

Figure 1 demonstrates the estimators for $H(P)$ and $F_\alpha(P)$ pictorially. Where $\hat{p}_i = P_n(i)$ denotes the empirical frequency of i-th symbol.

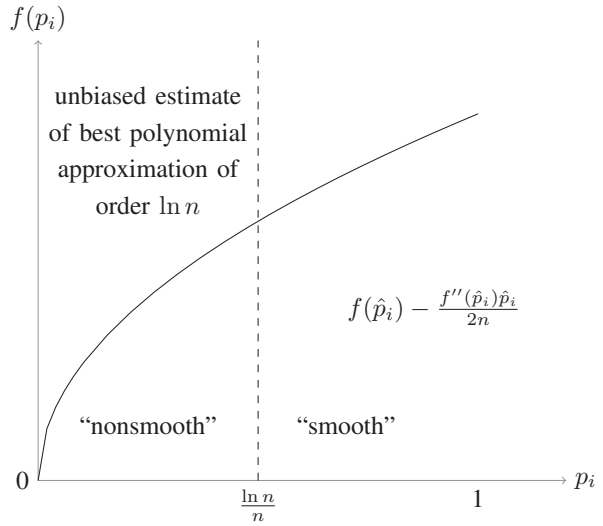

Fig. 1: Pictorial explanation of our estimators.

We remark that our estimator is both conceptually and algorithmically simple, with complexity linear in the number of samples n. Indeed, the only non-trivial computation required

is the best polynomial approximation for functions, which is data-independent and can be done *offline* before obtaining any samples.

B. Performance guarantees

Simple as our estimators are to describe and implement, they can be shown to be "optimal" in a very strong sense.

Remark: In our analysis, we consider the "Poissonized" observation model [11, Pg. 508]. In the Poisson setting, we first draw a Poisson random number $N \sim \mathsf{Poi}(n)$, and then conduct the sampling N times. Consequently the observed number of occurrences for each symbol are independent. We can show that the minimax risks under the original multinomial model and the Poisson model are essentially the same [7].

We have the following characterization of the minimax risk for entropy estimation.

Theorem 1. *Suppose $n \gtrsim \frac{S}{\ln S}$. Then the minimax risk of estimating entropy $H(P)$ satisfies*

$$\inf_{\hat{H}} \sup_{P \in \mathcal{M}_S} E_P \left(\hat{H} - H(P)\right)^2 \asymp \frac{S^2}{(n \ln n)^2} + \frac{(\ln S)^2}{n}. \quad (6)$$

Our estimator achieves this bound without knowledge of the support size S under the Poisson model.

The following is an immediate consequence of Theorem 1.

Corollary 1. *For our entropy estimator, the maximum L_2 risk vanishes provided $n \gg \frac{S}{\ln S}$. Moreover, if $n \lesssim \frac{S}{\ln S}$, then the maximum risk of any estimator for entropy will be bounded from zero.*

Corollary 1 is consistent with [9], where it was shown that one must have $n \gg \frac{S}{\ln S}$ for estimating the entropy. Recently, Wu and Yang [12] independently applied the idea of best polynomial approximation to entropy estimation, and obtained its minimax L_2 rates under a stricter stipulation ($\ln n \lesssim \ln S$). The minimax lower bound part of Theorem 1 follows from Wu and Yang [12]. We also remark that, unlike the estimator we propose, the estimator in Wu and Yang [12] relies on knowledge of the support size S, which generally may not be known. Figure 2 demonstrates the performance of our proposed entropy estimator in comparison with the MLE.

We now consider the functional $F_\alpha(P), 0 < \alpha < 1$.

Theorem 2. *Suppose $n \gtrsim \frac{S^{1/\alpha}}{\ln S}$ when we estimate $F_\alpha(P), 0 < \alpha < 1$. Then we have the following characterizations of the minimax risk.*

1) *$0 < \alpha \le 1/2$. If we also have $\ln n \lesssim \ln S$, then*

$$\inf_{\hat{F}_\alpha} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{F}_\alpha - F_\alpha(P)\right)^2 \asymp \frac{S^2}{(n \ln n)^{2\alpha}}. \quad (7)$$

2) *$1/2 < \alpha < 1$.*

$$\inf_{\hat{F}_\alpha} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{F}_\alpha - F_\alpha(P)\right)^2 \asymp \frac{S^2}{(n \ln n)^{2\alpha}} + \frac{S^{2-2\alpha}}{n}. \quad (8)$$

Our estimators $\hat{F}_\alpha$ achieves this bound without knowledge of the support size S under the Poisson model.

One immediate corollary of Theorem 2 is the following.

Corollary 2. *For our estimators of F_α, the maximum L_2 risk vanishes provided $n \gg \frac{S^{1/\alpha}}{\ln S}, 0 < \alpha < 1$. Moreover, if $n \lesssim \frac{S^{1/\alpha}}{\ln S}$, then the maximum risk of any estimator for F_α will be bounded away from zero.*

The minimax lower bound we present in Theorem 2 significantly improves on Paninski's lower bound in [8], which states that if $n \lesssim S^{1/\alpha - 1}$, then the maximum L_2 risk of any estimator for $F_\alpha(P), 0 < \alpha < 1$, is bounded away from zero.

The next two theorems correspond to estimation of $F_\alpha(P)$, $\alpha > 1$, where consistent estimation is possible even when the support size is countably infinite.

Theorem 3. *Under the Poissonized model, our estimator $\hat{F}_\alpha$ satisfies, for $1 < \alpha < \frac{3}{2}$,*

$$\sup_{P \in \cup_S \mathcal{M}_S} \mathbb{E}_P \left(\hat{F}_\alpha - F_\alpha(P)\right)^2 \lesssim \frac{1}{(n \ln n)^{2(\alpha-1)}}. \quad (9)$$

In other words, our estimator $\hat{F}_\alpha, 1 < \alpha < 3/2$ achieves an L_2 convergence rate of $(n \ln n)^{-2(\alpha-1)}$ regardless of the support size. This also turns out to be the minimax rate as shown by the following lower bound result.

Theorem 4. *There exists $c > 0$, such that if $S \ge cn \ln n$, then for $1 < \alpha < \frac{3}{2}$,*

$$\inf_{\hat{F}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{F} - F_\alpha(P)\right)^2 \gtrsim \frac{1}{(n \ln n)^{2(\alpha-1)}}, \quad (10)$$

where the infimum is taken over all possible estimators $\hat{F}$.

For $\alpha > 3/2$, the MLE achieves the parametric rate $1/n$ for the mean squared error, which is optimal.

C. Discussion of main results

Table I summarizes the minimax L_2 rates and the L_2 convergence rates of the MLE in estimating $F_\alpha(P), \alpha > 0$ and $H(P)$. When the L_2 rates have two terms, the first and second terms represent respectively the contributions of the bias and the variance. When there is a single term, only the dominant term is retained. Conditions for these results are presented in parentheses.

From a sample complexity perspective (i.e. how should the number of samples n scale with the support size S to achieve consistent estimation), Table I directly implies the results stated in Table II.

Our work (including the companion paper [13]) are the first to obtain the minimax rates, minimax rate-optimal estimators, and the maximum risk of MLE for estimating $F_\alpha(P), 0 < \alpha < 3/2$, and entropy $H(P)$ in the most comprehensive regimes of (S, n). Evident from Table I is the fact that the MLE cannot achieve the minimax risk for estimation of $H(P)$, and $F_\alpha(P)$ when $0 < \alpha < 3/2$. In these cases, our estimators have performance with n samples essentially the same of the MLE with $n \ln n$ samples, and it is the best possible. In other words, the minimax rate-optimal scheme *enlarge* the effective sample size from n to $n \ln n$. Furthermore, all the improvements we

	Minimax L_2 rates	L_2 rates of MLE
$H(P)$	$\frac{S^2}{(n\ln n)^2}+\frac{\ln^2 S}{n}$ $\left(n \gtrsim \frac{S}{\ln S}\right)$ (Thm. 1, [12])	$\frac{S^2}{n^2}+\frac{\ln^2 S}{n}$ $(n \gtrsim S)$ [13]
$F_\alpha(P), 0<\alpha\le\frac{1}{2}$	$\frac{S^2}{(n\ln n)^{2\alpha}}$ $\left(n \gtrsim S^{1/\alpha}/\ln S, \ln n \lesssim \ln S\right)$ (Thm. 2)	$\frac{S^2}{n^{2\alpha}}$ $\left(n \gtrsim S^{1/\alpha}\right)$ [13]
$F_\alpha(P), \frac{1}{2}<\alpha<1$	$\frac{S^2}{(n\ln n)^{2\alpha}}+\frac{S^{2-2\alpha}}{n}$ $\left(n \gtrsim S^{1/\alpha}/\ln S\right)$ (Thm. 2)	$\frac{S^2}{n^{2\alpha}}+\frac{S^{2-2\alpha}}{n}$ $\left(n \gtrsim S^{1/\alpha}\right)$ [13]
$F_\alpha(P), 1<\alpha<\frac{3}{2}$	$(n\ln n)^{-2(\alpha-1)}$ $(S \gtrsim n\ln n)$ (Thm. 3,4)	$n^{-2(\alpha-1)}$ $(S \gtrsim n)$ [13]
$F_\alpha(P), \alpha\ge\frac{3}{2}$	n^{-1} [13]	n^{-1}

TABLE I: Summary of results in this paper and the companion [13]

	MLE	Minimax rate-optimal
$H(P)$	$n \gg S$	$n \gg S/\ln S$
$F_\alpha(P), 0<\alpha<1$	$n \gg S^{1/\alpha}$	$n \gg S^{1/\alpha}/\ln S$
$F_\alpha(P), \alpha>1$	$n \gg 1$	$n \gg 1$

TABLE II: The number of samples needed to achieve consistent estimation

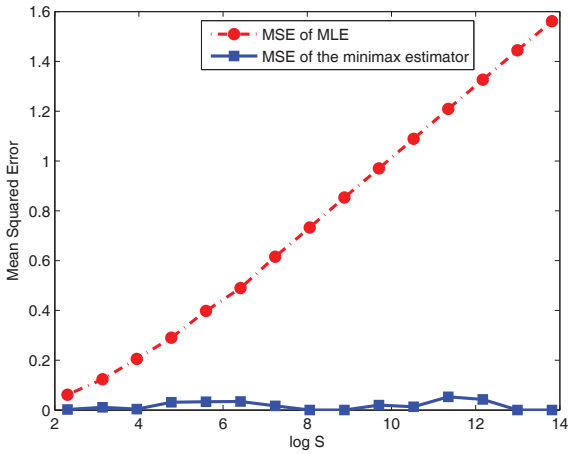

Fig. 2: The empirical mean squared error (MSE) of our estimator and the MLE along sequence $n = 5S/\ln S$. For each data point, we computer the MSE via 20 Monte Carlo simulations of sampling n times from uniform distributions with support size S. The horizontal line is $\ln S$.

have are in the bias, which is the dominating factor in the risk. This observation suggests a simple way to obtain the minimax L_2 rates from the L_2 rates of the MLE. One just needs to find the bias term in the expression of MLE L_2 rates, and replace the term n by $n\ln n$. This simple rule has deep connections with the rationale behind the construction and analysis of our estimators.

Some readers may be concerned that the minimax decision theoretic framework we adopt is too pessimistic. In some sense, it characterizes the worst case performance over all possible distributions $P \in \mathcal{M}_S$, and it would be disappointing if our estimator fails to behave optimally for distributions lying in a strict subset of $\mathcal{M}_S$ not including the worst case distribution. We remark that our estimators can be shown to be optimal in a much stronger sense. The statistics community usually uses the *adaptive* estimation framework to alleviate the pessimism of minimaxity [14]. Specifically, one specifies a nested sequence of subsets of $\mathcal{M}_S$, and try to construct an estimator that achieves simultaneously the minimax risks over each of the subsets without knowing which subset the parameter P belongs to. It was shown recently by another companion paper [15] that along the nested subsets $\mathcal{M}_S(H) = \{P : H(P) \le H\}$, our estimator (without knowing H) simultaneously achieves the minimax rates over $P \in \mathcal{M}_S(H)$ for all $H \le \ln S$. Most surprisingly, the performance of our estimator with n samples is still essentially that of the MLE with $n\ln n$ samples over every set $\mathcal{M}_S(H)$, further strengthening the advantages of our estimator in practice.

It is instructive to consider our results in the context of the intriguing connections and differences between three important problems in information theory: entropy estimation, estimating a discrete distribution under relative entropy loss, and minimax redundancy in compressing i.i.d. sources. Table III summarizes the known results and conveys several important messages. First, in the asymptotic regime, there is a logarithmic factor between the redundancy of the compression problem on one hand, and the distribution estimation problem on the other. Indeed, since compression requires use of a coding distribution Q that does not depend on the data, the redundancy of compression will definitely be larger than the risk under relative entropy in estimating the distribution. However, in the large alphabet setting, the problems are equally difficult - the phase transition of vanishing risk for both compression and distribution estimation happen when n is linear in the support size S.

Second, the large alphabet setting shows that estimation of entropy is considerably easier than both estimating the corresponding distribution, or compressing the source. There have been interesting developments in the large alphabet setting in information theory, cf. [20]–[24]. One of the implications of Table III is that the approach of entropy estimation via compression can be sub-optimal in certain regimes.

Recently, [25] has studied the complexity of estimating Rényi entropy $H_\alpha(P)$. Interestingly, although we have the relation $H_\alpha(P) = \frac{\ln F_\alpha(P)}{1-\alpha}$, the phase transitions for $H_\alpha(P)$ and $F_\alpha(P)$ can be quite different for $\alpha > 1$.

	entropy estimation	estimation of distribution	compression with blocklength n
S fixed	MSE $\sim \frac{\mathsf{Var}(-\ln P(X))}{n}$ [11]	$\inf_{\hat{P}} \sup_P \mathbb{E}D(P_X \Vert \hat{P}_X) \sim \frac{S-1}{2n}$ [16], [17]	$\min_Q \sup_P \frac{1}{n} D(P_{X^n} \Vert Q_{X^n}) \sim \frac{S-1}{2n} \ln n$ [18]
large S	$n \gg S/\ln S$ [9]	$n \gg S$ [19]	$n \gg S$ [20], [21]

TABLE III: Comparison of difficulties in entropy estimation, estimation of distribution, and data compression under classical asymptotics and high dimensional asymptotics

III. Applications: mutual information estimation

As central as it is in information theory, mutual information [1] has been adopted and widely used in a variety of other disciplines. Recently, in [26], the present authors highlight the applicability of the aforementioned methodology to statistical problems beyond functional estimation, and show that it can yield substantial gains. For example, we demonstrate that for learning tree-structured graphical models, our approach achieves a significant reduction of the required data size compared with the classical Chow–Liu algorithm [27], which is an implementation of the MLE, to achieve the same accuracy. The key step in improving the Chow–Liu algorithm is to replace the empirical mutual information with the estimator for mutual information induced via our entropy estimator. This estimator is shown to achieve the minimax L_2 rates for mutual information. Further, applying the same replacement approach to classical Bayesian network classification, the resulting classifiers uniformly outperform the previous classifiers on 26 widely used datasets.

IV. concluding remarks

The functional estimation problem examined here has been studied extensively in the fields of statistics, information theory, computer science, physics, economics, psychology, and several other disciplines. We implore the reader to refer to the full version of this paper [7] for comprehensive discussions pertaining to the motivation, related literature, proofs, methodologies, and additional results which we could not delve into here due to space limitations. Also absent from the present discussion are deep mathematical ideas from approximation theory that have allowed us to develop the necessary tools for the construction and analysis of essentially minimax estimators for a wide class of functionals.

References

[1] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[2] A. Rényi, "On measures of entropy and information," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961, pp. 547–561.

[3] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *Information Theory, IEEE Transactions on*, vol. 41, no. 1, pp. 26–34, 1995.

[4] T. A. Courtade and S. Verdú, "Cumulant generating function of codeword lengths in optimal lossless compression," in *Proc. of IEEE International Symposium on Information Theory (ISIT)*, 2014.

[5] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer, 1998, vol. 31.

[6] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[7] 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.

[8] L. Paninski, "Estimating entropy on m bins given fewer than m samples," *Information Theory, IEEE Transactions on*, vol. 50, no. 9, pp. 2200–2203, 2004.

[9] 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 43rd annual ACM symposium on Theory of computing*. ACM, 2011, pp. 685–694.

[10] P. Valiant and G. Valiant, "Estimating the unseen: improved estimators for entropy and other properties," in *Advances in Neural Information Processing Systems*, 2013, pp. 2157–2165.

[11] L. Le Cam, *Asymptotic methods in statistical decision theory*. Springer, 1986.

[12] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *arXiv preprint arXiv:1407.0381*, 2014.

[13] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Maximum likelihood estimation of functionals of discrete distributions," *arXiv preprint arXiv:1406.6959*, 2014.

[14] T. T. Cai, "Minimax and adaptive inference in nonparametric function estimation," *Statistical Science*, vol. 27, no. 1, pp. 31–50, 2012.

[15] Y. Han, J. Jiao, and T. Weissman, "Adaptive estimation of Shannon entropy," *in preparation*.

[16] N. Chentsov, *Statisticheskie reshayushchie pravila i optimalnye vyvody*. Nauka, Moskva (Engl. transl.: 1982, Statistical decision rules and optimal inference, American Mathematical Society, Providence), 1972.

[17] D. Braess and H. Dette, "The asymptotic minimax risk for the estimation of constrained binomial and multinomial probabilities," *Sankhyā: The Indian Journal of Statistics*, pp. 707–732, 2004.

[18] J. Rissanen, "Stochastic complexity and modeling," *The Annals of Statistics*, pp. 1080–1100, 1986.

[19] L. Paninski, "Variational minimax estimation of discrete distributions under KL loss," in *Advances in Neural Information Processing Systems*, 2004, pp. 1033–1040.

[20] A. Orlitsky and N. P. Santhanam, "Speaking of infinity," *Information Theory, IEEE Transactions on*, vol. 50, no. 10, pp. 2215–2230, 2004.

[21] W. Szpankowski and M. J. Weinberger, "Minimax pointwise redundancy for memoryless models over large alphabets," *Information Theory, IEEE Transactions on*, vol. 58, no. 7, pp. 4094–4104, 2012.

[22] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Probability estimation in the rare-events regime," *Information Theory, IEEE Transactions on*, vol. 57, no. 6, pp. 3207–3229, 2011.

[23] M. I. Ohannessian, V. Y. Tan, and M. A. Dahleh, "Canonical estimation in a rare-events regime," in *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*. IEEE, 2011, pp. 1840–1847.

[24] X. Yang and A. Barron, "Large alphabet coding and prediction through poissonization and tilting," in *The Sixth Workshop on Information Theoretic Methods in Science and Engineering, Tokyo*, 2013.

[25] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi, "The complexity of estimating Rényi entropy," in *SODA*, 2015.

[26] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Beyond maximum likelihood: from theory to practice," *arXiv preprint arXiv:1409.7458*, 2014.

[27] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *Information Theory, IEEE Transactions on*, vol. 14, no. 3, pp. 462–467, 1968.