Minimax Estimation of Functionals of Discrete Distributions

Jiantao Jiao, Student Member, IEEE, Kartik Venkat, Student Member, IEEE, Yanjun Han, Student Member, IEEE, and Tsachy Weissman, Fellow, IEEE.

Abstract

We propose a general framework for the construction and analysis of minimax estimators for a wide class of functionals of discrete distributions, where the alphabet size $S$ is unknown and may be scaling with the number of observations $n$. We treat the respective regions where the functional is “nonsmooth” and “smooth” separately. In the “nonsmooth” regime, we apply an unbiased estimator for the best polynomial approximation of the functional whereas, in the “smooth” regime, we apply a bias-corrected version of the Maximum Likelihood Estimator (MLE).

We illustrate the merit of this approach by thoroughly analyzing 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 $L_2$ rates for estimating these functionals. In particular, we demonstrate that our estimator achieves the optimal sample complexity $n = \Theta(S/\ln S)$ for entropy estimation. We also demonstrate that the sample complexity for estimating $F_\alpha(P), 0 < \alpha < 1$ is $\Theta(S^{1/\alpha}/\ln S)$, which can be achieved by our estimator but not the 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 alphabet size, while the exact $L_2$ rate for the MLE is $n^{-2(\alpha-1)}$. For all the above cases, the behavior of the optimal estimators with $n$ samples is essentially that of the MLE with $n \ln \ln n$ samples.

We highlight the practical advantages of our schemes for the estimation of entropy and mutual information. We compare our performance with the popular MLE and with the order-optimal entropy estimator of Valiant and Valiant. As we illustrate with a few experiments, our approach reduces running time and boosts the accuracy.

Index Terms

Mean squared error, entropy estimation, nonsmooth functional estimation, maximum likelihood estimator, approximation theory, minimax lower bound, polynomial approximation, minimax-optimality, high dimensional statistics, Rényi entropy

I. INTRODUCTION AND MAIN RESULTS

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:

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

where $f : [0, 1] \rightarrow \mathbb{R}$ is a continuous function, and $f(0) = 0$. Among the most fundamental of such functionals is the entropy $H(P)$,

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

which is of the form $H$ with $f(x) = -x \ln x$. Another well motivated example of the form $H$ is the functional $F_\alpha(P)$, defined as

$$F_\alpha(P) = \sum_{i=1}^{S} p_i^{\alpha}.$$ 

In 1961, Rényi [2] generalized the Shannon entropy and obtained the Rényi entropy of order $\alpha$, which is closely related to $F_\alpha(P)$:

$$H_\alpha(P) = \frac{1}{1-\alpha} \ln \left( \sum_{i=1}^{S} p_i^{\alpha} \right), \quad \alpha \geq 0, \alpha \neq 1.$$ 

Like Shannon entropy, Rényi entropy is an important information measure emerging in an increasing variety of disciplines such as ecology (as an index of diversity [3]), quantum information (as a measure of entanglement [4]), and information theory and statistics (as the generalized cutoff rate in source block coding [5] and fundamental limits in coding problems [6], [7]).
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 the form \( \tilde{F} \) under \( L_2 \) loss. To illustrate our approach, we focus on and describe explicit constructions for the specific cases of entropy \( \hat{H}(P) \) and \( F_\alpha(P) \), from which the construction for any other functional of the form \( \tilde{F} \) will be clear. Our estimators for each of these two functionals are agnostic with respect to the alphabet size \( S \), and achieve the minimax \( L_2 \) rates.

Our approach is to tackle the estimation problem separately for the cases of “small \( p \)” and “large \( p \)” in \( H(P) \) and \( F_\alpha(P) \) estimation, corresponding to treating regions where the functional is nonsmooth and smooth in different ways. As we describe in detail in the sections to follow, where we give a full account of our estimators, 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.

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 \( 0 \leq p_i \ll \ln n/n \), (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 \( K_n \ll \ln n \), 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 \( 1 - P_n(i) \ln P_n(i) + \frac{1}{2p_i^2} \) 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 \( x^\alpha \) with order \( K_n \ll \ln n \), and use the estimator \( \left( 1 + \frac{\alpha (1-\alpha)}{2 \ln p_n(i)} \right) P_n(i) \) when \( P_n(i) \gg \ln n/n \).

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. We show that even this best polynomial approximation step can be performed very efficiently using well developed machinery from approximation theory.

B. Main results

Simple as our estimators are to describe and implement, they have strong performance guarantees. For the analysis of our schemes, we consider, without loss of generality (cf. Lemma 16), the “Poissonized” observation model [8, Pg. 508]. Under this model, we first draw a Poisson random number \( N \sim \text{Poi}(\ln(n)) \), and then conduct the sampling \( N \) times. Consequently the observed number of occurrences for each symbol are independent.

We use the notation \( a \preceq b \) to denote that there exists a universal constant \( C \) such that \( \sup_{n} \frac{a}{b} \leq C \). Let \( \mathcal{M}_S \) denote the space of distributions with support size \( S \).

Theorem 1. Under the Poissonized model, our estimator \( \hat{H} \) satisfies

\[
\sup_{P \in \mathcal{M}_S} E_P \left( \hat{H} - H(P) \right)^2 \leq \frac{S^2}{(n \ln n)^2} + \frac{S(\ln n)^4}{n^{2-\epsilon}} + \frac{(\ln S)^2}{n},
\]

for all \( \epsilon > 0 \).

The following is an immediate consequence of Theorem 1.

Corollary 1. For the estimator \( \hat{H} \) in Theorem 1, the maximum \( L_2 \) risk vanishes provided \( n = \omega \left( \frac{S}{\ln S} \right) \).

Evidently, the estimator from Theorem 1 is order-optimal in the number of samples required for consistent estimation, since it was shown in [9] that one must have \( n = \omega \left( \frac{S}{\ln S} \right) \) for estimating the entropy. Wu and Yang [10] independently applied the idea of best polynomial approximation to entropy estimation, and obtained its minimax \( L_2 \) rates. We remark that, unlike the estimator we propose, the estimator in Wu and Yang [10] relies on knowledge of the alphabet size \( S \), which generally may not be known.

For the functional \( F_\alpha(P), 0 < \alpha < 1 \), we have the following.

Theorem 2. Under the Poissonized model, our estimator \( \hat{F}_\alpha \) satisfies

\[
\sup_{P \in \mathcal{M}_S} E_P \left( \hat{F}_\alpha - F_\alpha(P) \right)^2 \leq \begin{cases} 
\frac{S^2}{(n \ln n)^{2\alpha}} & 0 < \alpha \leq 1/2, \\
\frac{S^2}{(n \ln n)\alpha} + \frac{S(\ln n)^{2+2\alpha}}{n^{2\alpha-1}} & 1/2 < \alpha < 1
\end{cases}
\]

for all \( \epsilon > 0 \).

Corollary 2. For the estimator \( \hat{F}_\alpha \) in Theorem 2, the maximum \( L_2 \) risk vanishes provided \( n = \omega \left( \frac{S^{1/\alpha}}{\ln S} \right) \), \( 0 < \alpha < 1 \).

The following minimax lower bound implies that our estimator in Theorem 2 attains the optimal sample complexity.
Theorem 3. For any fixed positive constant c, let \( n = \frac{S^{2+/\alpha}}{\ln S} \). Then,

\[
c^{2\alpha} \cdot \liminf_{n \to \infty} \sup_{P \in M_s} \mathbb{E}_P \left( \hat{F} - F_\alpha(P) \right)^2 \geq D > 0, \quad 0 < \alpha < 1,
\]

where \( D > 0 \) is a constant that does not depend on \( c \), and the infimum is taken over all possible estimators \( \hat{F} \).

This minimax lower bound significantly improves on Paninski’s lower bound in [11], which states that if \( n = O(S^{1/\alpha - 1}) \), then the maximum \( L_2 \) risk of any estimator for \( F_\alpha(P) \), \( 0 < \alpha < 1 \), is bounded from zero. Evidently, the number of samples required by our estimator in Corollary 2 matches the minimax lower bound shown in Theorem 3.4

Moreover, comparing the dependency on \( c \) in Theorem 3 with that implied by Theorem 2, we know that our estimator in Theorem 2 is also minimax optimal within a constant multiplicative factor. The same is true of our estimator for the entropy, since it achieves the optimal dependency on \( c \), also achieved by Valiant and Valiant [12] and Wu and Yang [10]. We remark that the entropy estimators based on linear programming in Valiant and Valiant [9, 13] have not been shown to achieve the optimal dependency on \( c \).

The next two theorems correspond to estimation of \( F_\alpha(P) \), \( \alpha > 1 \).

Theorem 4. Under the Poissonized model, our estimator \( \hat{F}_\alpha \) satisfies

\[
\limsup_{n \to \infty} (n \ln n)^{2(\alpha - 1)} \cdot \sup_{S \in M_s} \sup_{P \in M_s} \mathbb{E}_P \left( \hat{F}_\alpha - F_\alpha(P) \right)^2 < \infty, \quad 1 < \alpha < \frac{3}{2}
\]

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 alphabet size. This also turns out to be the minimax rate as shown by the following result.

Theorem 5. If \( S = cn \ln n, c > 0 \), then

\[
\liminf_{n \to \infty} (n \ln n)^{2(\alpha - 1)} \cdot \inf_{F \in M_s} \sup_{P \in M_s} \mathbb{E}_P \left( \hat{F} - F_\alpha(P) \right)^2 > 0, \quad 1 < \alpha < \frac{3}{2},
\]

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

In a companion paper [14], a non-asymptotic analysis of the Maximum Likelihood Estimator (MLE) is presented, wherein it is shown that the exact \( L_2 \) rate of convergence for the MLE of \( F_\alpha(P) \) is \( n^{-2(\alpha - 1)} \) regardless of the alphabet size, when \( 1 < \alpha < 3/2 \). Hence, Theorem 4 implies that although the MLE requires constant sample complexity, it is strictly rate sub-optimal. It was shown in [14] that when \( \alpha \geq 3/2 \), the MLE achieves the optimal \( L_2 \) convergence rate \( n^{-1} \).

The following table 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.

| \( H(P) \) | Minimax \( L_2 \) rates | \( L_2 \) rates of MLE |
|-----------|-----------------|-----------------|
| \( F_\alpha(P), 0 < \alpha \leq 1/2 \) | \( \frac{S^2}{(n \ln n)^{\alpha}} + \ln^2 S \) (n = \( \Omega(S/\ln S) \), \( \ln n \leq \ln S \)) (Thm. 10) | \( \frac{S^2}{(n \ln n)^{\alpha}} + \ln^2 S \) (n = \( \Omega(S) \)) [14] |
| \( F_\alpha(P), 1/2 < \alpha < 1 \) | \( \frac{S^2}{(n \ln n)^{\alpha}} + \frac{S^{2-\alpha}}{n} \) (n = \( \Omega(S^{1/\alpha}/\ln S) \), \( \ln n \leq \ln S \)) (Thm. 10) | \( \frac{S^2}{(n \ln n)^{\alpha}} + \frac{S^{2-\alpha}}{n} \) (n = \( \Omega(S^{1/\alpha}) \)) [14] |
| \( F_\alpha(P), 1 < \alpha < 3/2 \) | \( n^{-\frac{2(\alpha - 1)}{2}} \) (n = \( O(S/\ln S) \)) (Thm. 15) | \( n^{-2(\alpha - 1)} \) (n = \( O(S) \)) [14] |
| \( F_\alpha(P), \alpha \geq 3/2 \) | \( n^{-1} \) [14] | \( n^{-1} \) |

**TABLE I:** Summary of results in this paper and the companion [14]

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, there exist strictly better estimators whose performance with \( n \) samples is essentially the same as that of the MLE with \( n \ln n \) samples. In other words, the optimal estimators *enlarge* the effective sample size by a logarithmic factor.

2Note that the only two results in Table I that are not proved explicitly are the characterizations of the variance bound \( \frac{S^2}{(n \ln n)^{\alpha}} \cdot \frac{1}{2} < \alpha < 1 \) as the minimax variance, and as the the variance of the MLE. However, these follow respectively from the standard information bound such as [15], and the well-known asymptotic efficiency of the MLE [18].

In a previous version of the manuscript, there is a \( \sqrt{\ln S} \) gap between our minimax lower bound and the achievability in Corollary 2. Partially inspired by Wu and Yang [10], we modified the proof of Theorem 3 by using an argument similar to that in [18] Thm. 4, thereby closing the gap.

3Note that the only two results in Table I that are not proved explicitly are the characterizations of the variance bound \( \frac{S^2}{(n \ln n)^{\alpha}} \cdot \frac{1}{2} < \alpha < 1 \) as the minimax variance, and as the the variance of the MLE. However, these follow respectively from the standard information bound such as [15], and the well-known asymptotic efficiency of the MLE [18].
C. Motivation

Existing theory proves inadequate for addressing the problem of estimating functionals of probability distributions. A natural estimator for functionals of the form $(1)$ 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 $f \in C^1(0,1)$ and we focus on $n$ i.i.d. observations from a distribution with alphabet size $S$, then the problem of estimating $F(P)$ is trivial under classical asymptotics where $S$ is fixed, and the number of observations $n \to \infty$. This maximum likelihood estimator is asymptotically efficient $(16$, Thm. 8.11, Lemma 8.14) in the sense of the Hájek convolution theorem $(17)$ and the Hájek–Le Cam local asymptotic minimax theorem $(18)$. It is therefore not surprising to encounter the following quote from the introduction of Wyner and Foster $(19)$ who considered entropy estimation:

“The plug-in estimate is universal and optimal not only for finite alphabet i.i.d. sources but also for finite alphabet, finite memory sources. On the other hand, practically as well as theoretically, these problems are of little interest.”

In light of this, is it fair to say that the entropy estimation problem is solved in the finite alphabet setting? It was observed in Paninski $(20)$ that the maximum of $\text{Var}(\ln P(X))$ over distributions with support size $S$ is of order $(\ln S)^2$ (a tight bound is also given by Lemma $(15)$ in the appendix). Since classical asymptotics (with the Delta method $(16$, Chap. 3)) show that

$$E_P(H(P_n) - H(P))^2 \sim \frac{\text{Var}(\ln P(X))}{n}, \quad n \gg 1,$$

a naive interpretation of $(10)$ might be that it suffices to take $n = \omega((\ln S)^2)$ samples to guarantee the consistency of $H(P_n)$. Such interpretation, however, would be blatantly wrong. It was already observed in Paninski $(20)$ that if $n = O(S^{1-\delta})$, $\delta > 0$, then the maximum $L_2$ risk of any entropy estimator would be unbounded as $S \to \infty$.

This apparent discrepancy shows that $(10)$ is not valid when $S$ might be growing with $n$, and it is of utmost importance to obtain risk bounds for estimators of entropy and other functionals of distributions in the latter regime. Indeed, in the modern era of “big data”, we often encounter situations where the alphabet size is comparable to, or much larger than, the number of observations. If we trace the progress on entropy estimation in the non-asymptotic regime, we find several thrusts in various communities, including, for example, the Miller–Madow bias-corrected estimator and its variants $(21)$–$(23)$, the jackknifed estimator $(24)$, the shrinkage estimator $(25)$, the Bayes estimator under various priors $(26)$, $(27)$, the coverage adjusted estimator $(28)$, the Best Upper Bound (BUB) estimator $(20)$, the B-Splines estimator $(29)$, etc.

However, there has been relatively little theoretical understanding of how the estimators mentioned above behave in the regime where $S$ is comparable to or even larger than $n$. To this effect, Paninski $(20)$ showed that the MLE, the Miller–Madow estimator, and the jackknifed estimator, all fail to consistently estimate entropy when the number of samples is linear in the alphabet size. In other words, the worst case risk for all these estimators is bounded away from zero if the sample size $n$ is linear in the alphabet size $S$. It was, however, pointed out in Paninski $(11)$ 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 $n = \Theta(S/\ln S)$ samples are both necessary and sufficient to estimate the entropy of a discrete distribution. The family of schemes they presented extends to several other symmetric functionals of discrete distributions and is of relevance to the current discussion. Readers are referred to Valiant’s thesis $(30)$ for a comprehensive treatment. We note, however, that the functionals for which the techniques of $(9)$ can be applied are limited to those 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 Rényi entropy of order $\alpha < 1$ and other interesting nonsmooth functionals of distributions. Also, it is not clear whether these techniques generally lead to minimax optimal estimators for functionals that are “smoother” than entropy. Further, Valiant and Valiant $(9)$ focused on estimators that are close to the correct value with high probability, which don’t directly translate to risk bounds on the performance of these estimators under certain loss functions.

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. Such a comprehensive view of functional estimation has yet to be realized. George Pólya $(31)$ commented that “the more general problem may be easier to solve than the special problem”. This motivates our present work, in which we provide a general framework and procedure for minimax estimation of functionals with non-asymptotic performance guarantees. In specializing our procedure to various interesting functionals, we obtain the minimax $L_2$ rates for estimating $H(P)$ and $F_\alpha(P)$, $\alpha > 0$.

D. General principles of nonsmooth functional estimation

1) Exploiting prior knowledge: One of the most basic ideas in constructing statistical procedures is to exploit the prior knowledge about the structure of the problem. To frame our general procedure for functional estimation in this context, we first briefly review some milestones in the development of statistics. The seminal work of Stein $(32)$ revealed the famous Stein’s phenomenon that uniformly minimum variance unbiased estimators could be inadmissible, i.e. the MLE $\hat{\theta}_{\text{MLE}} = Y$ in
normal model $Y \sim \mathcal{N}(\theta, I_p)$ can be uniformly outperformed. The key idea in Stein’s estimator involves shrinkage of the MLE. Following this rationale, Donoho and Johnstone [33] proposed the soft-thresholding estimator to estimate the normal mean given that we know a priori that the mean $\theta$ lies in an $\ell_p$ ball, $p \in (0, \infty)$. Later, Donoho and Johnstone [34] applied this idea to nonparametric estimation in Besov spaces, and obtained the famous wavelet shrinkage estimator for denoising. Interestingly, if we assume no prior knowledge about $\theta$, then the MLE $\hat{\theta}_{MLE} = Y$ is also minimax over $\mathbb{R}^p$, which has considerably larger risk than the shrinkage estimator when $\theta$ is indeed small. This demonstrates that prior knowledge can reduce the risk in statistical estimation.

The success of compressed sensing proposed by Candès and Tao [35] and Donoho [36] is another example of exploiting prior knowledge. For any linear inverse problem, if the dimension of the unknowns $p$ is much larger than that of the observations $n$, then it is generally impossible to estimate the unknown. However, if we know a priori that the unknown is sparse, then it is possible to construct efficient statistical procedures to exploit the sparsity and conduct inference even when $p \gg n$. The problem of matrix completion [37] and the literature related to compressed sensing [38] provide numerous examples where various kinds of prior knowledge are exploited.

2) A general procedure for nonsmooth functional estimation: While our main focus in this work is on estimating functionals of distributions, we note that our procedures and approach are applicable to more general problems. In this discussion, we do not restrict ourselves to probability functional estimation, but instead consider estimating functionals of a parameter of distributions, we note that our procedures and approach are applicable to more general problems. In this discussion, we do not restrict ourselves to probability functional estimation, but instead consider estimating functionals of a parameter $\theta$ of distributions, we note that our procedures and approach are applicable to more general problems.

### Classify Regime

1) **Choosing $\Delta_n$ and $K_n$**

There are two components to the $L_2$ risk of an estimator - the bias and the variance:

$$
\text{Risk} = \text{Bias}^2 + \text{Variance}
$$

A key feature for any good estimator is that it should have a good balance between bias and variance for all $\theta \in \Theta$. In our case, we want to control the bias and variance for both the “smooth” and “nonsmooth” regimes. Controlling the variance, it so happens, is not technically very challenging with the well-known tools from measure concentration. Indeed, as was illuminated by Donoho in [39], one of the blessings of high dimensionality is the “concentration of measure” [40] phenomenon, which allows one to control the fluctuations of an estimator. The most challenging part of estimation in high-dimensional problems, such as the current setting, is that a large bias will lead us to concentrate around a wrong point, leading to a large risk. Indeed, there arises a need to address the problem of bias control in both the “smooth” and “nonsmooth” regimes. It is the behavior of the bias in both these regimes that dictates the choices of the parameters.
and the estimators in our general estimation recipe. Hence, in order to tune parameters $\Delta_n$ and $K_n$ in an optimal way, it is necessary to understand the bias of our statistical procedure, which relies on a tight characterization of the best approximation error with an arbitrary order $K_n$. Fortunately, modern approximation theory serves this purpose well, with various profound results developed over the last century. Ever since Karl Weierstrass showed in 1885 [41] that any continuous real-valued functions on a compact interval could be uniformly approximated via algebraic and trigonometric polynomials, there has been great interest in studying the best approximation error rate $E_n[f]_A$:

$$E_n[f]_A = \inf_{P \in \text{poly}_n} \sup_{x \in A} |f(x) - P(x)|,$$

where poly$_n$ is the collection of polynomials with order at most $n$ on $A$. Quantifying $E_n[f]_A$ and obtaining the polynomial that achieves it turned out to be extremely challenging. Remez [42] in 1934 proposed an efficient algorithm for computing the best polynomial approximation, and it was recently implemented and highly optimized in Matlab by the Chebfun team [43], [44]. Regarding the theoretical understanding of $E_n[f]_A$, de la Vallée-Poussin, Bernstein, Ibragimov, Markov, Kolmogorov and others have made significant contributions, and it is still an active research area. Among others, Bernstein [45], [46] and Ibragimov [47] showed various exact limiting results for some important classes of functions like $|x|^p$ and $|x|^n \ln |x|^n$. For example, we have

**Theorem 6.** [46] The following limit exists for all $p > 0$:

$$\lim_{n \to \infty} n^p E_n[|x|^p]_{[-1,1]} = \mu(p),$$

where $\mu(p)$ is a constant bounded as

$$\frac{\Gamma(p)}{\pi} \sin \frac{\pi p}{2} \left(1 - \frac{1}{p - 1}\right) \leq \mu(p) \leq \frac{\Gamma(p)}{\pi} \sin \frac{\pi p}{2},$$

where $\Gamma(\cdot)$ denotes the Gamma function.

Regarding bounds on $E_n[f]$ for any finite $n$, Korneichuk [48, Chap. 6] provides a comprehensive study. For a comprehensive treatment of modern approximation theory, DeVore and Lorentz [49], Ditzian and Totik [50] provide excellent references. Using the extensive machinery of approximation theory, the risk bounds obtained lend natural choices towards selection of the parameters $\Delta_n$ and $K_n$ in our general recipe specialized to probability functional estimation. Regarding the choice of $\Delta_n$, according to classical asymptotics [16] and the asymptotic efficiency of $F(\hat{\theta}_n)$, $\Delta_n$ should satisfy $\lim_{n \to \infty} \Delta_n = 0$. Concretely, for the probability functional estimation problem, we will demonstrate that the correct order is $\Delta_n \sim (\ln n)/n$. Regarding the order $K_n$ of the polynomial approximation, it should grow to infinity as $n \to \infty$, but much slower than $n$. We will show that for the probability functional estimation problem, the order should be roughly $K_n \sim \ln n$.

2) **Estimating $F(\theta)$ in the “smooth” regime**

We demonstrate that, perhaps somewhat surprisingly, $F(\hat{\theta}_n)$ is generally not optimal even in the “smooth” regime, and a slight modification is needed. We propose to conduct the first-order bias correction [51] for $F(\hat{\theta}_n)$ in the “smooth” regime. Let us illustrate the rationale assuming $\theta \in \mathbb{R}^1$ and some additional regularity conditions. Since $\hat{\theta}_n$ is unbiased for $\theta$, for $\theta \notin \Theta_0$, it follows from a Taylor expansion that the bias of $F(\hat{\theta}_n)$ is

$$EF(\hat{\theta}_n) - F(\theta) = \frac{1}{2} F''(\theta) \text{Var}_\theta(\hat{\theta}_n) + O\left(\frac{1}{n^2}\right),$$

where $\text{Var}_\theta(X)$ is the variance of the random variable $X$ under probability law $P_\theta$, specified by $\theta$. We define the first-order bias-corrected estimator $F^c(\hat{\theta}_n)$:

$$F^c(\hat{\theta}_n) \triangleq F(\hat{\theta}_n) - \frac{F''(\hat{\theta}_n)}{2} \text{Var}_\hat{\theta}_n(\hat{\theta}_n).$$

In general, we utilize the first-order bias-corrected estimator in the “smooth” regime.

It is worth mentioning that our recipe is fundamentally different from the shrinkage idea. The rationale behind shrinkage is to significantly reduce the variance at the expense of slightly increasing the bias. However, it has long been observed in the literature on entropy estimation that the bias dominates the $L_2$ risk. Hence, our recipe is complementary to the idea of shrinkage: we significantly reduce the bias at the expense of slightly increasing the variance. Thus, estimation of nonsmooth functionals operates at the other end of the bias-variance trade-off, and can be thought of in this sense as a dual of shrinkage based estimation. In some sense, we show that the two main approaches in improving over MLE in statistical estimation are shrinkage and the method of approximation, which are dual to each other, complementing Efron [52] who mentioned shrinkage as an impressive improvement over the MLE.

Our general recipe has precedents in the literature. Nemirovski [53] pioneered the usage of approximation theory in functional estimation. Later, Lepski, Nemirovski, and Spokoiny [54] considered estimating the $L_1$ norm of a regression function, and
utilized trigonometric approximation. Valiant and Valiant [9] observed that the MLE performs well when enough samples are available for entropy estimation. Cai and Low [55] used best polynomial approximation to estimate the $\ell_1$ norm of a normal mean, and Vinck et al. [56] applied Taylor polynomial approximation to entropy estimation.

E. Discussion and significance of the main results

Through the lens of the general recipe for nonsmooth functional estimation, we now review and discuss the implications of our main results.

For the entropy estimation problem, Theorem 1 and its proof illustrate the rationale of our general recipe: we significantly reduce the bias at the expense of slightly increasing the variance. The proof of Theorem 1 shows that the squared bias of our estimator $H$ corresponds to the first term $c_2S^2/n$ in (5), and the variance corresponds to the next two terms. Using techniques developed in [14], we show that the squared bias of the MLE is of scale $(S/n)^2$ when $S/n$ is not too large, which shows that our estimator has a $(\ln n)^2$ improvement of the squared bias over the MLE. However, the variance of our estimator is slightly larger than that of the MLE, but this increment is a higher order term that can be neglected under mild growth conditions.

Since our estimator $H$ is essentially equivalent to the MLE $H(P_n)$ when we are in the “smooth” regime, it is also asymptotically efficient under classical asymptotics. Indeed, (5) is consistent with and easily recovers classical asymptotics whence the first two terms would be higher order terms compared to the third one.

More significantly, moving beyond classical asymptotics, Corollary 1 demonstrates that our estimator achieves the optimal $n = \Theta(S/\ln S)$ scaling of measurements with alphabet size established in [9], and we remark that for the MLE $H(P_n)$, the phase transition is at $\Theta(S)$ rather than $\Theta(S/\ln S)$. In other words, if $n = \omega(S)$, then the maximum $L_2$ risk of $H(P_n)$ vanishes, but it is bounded from zero if $n = cS$, for $c > 0$ constant. Paninski [11] first observed this fact and [14] provides a comprehensive rigorous treatment.

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 II summarizes the known results.

| Entropy estimation | Estimation of distribution | Compression with blocklength $n$ |
|--------------------|----------------------------|---------------------------------|
| $S$ fixed          | $\text{MSE} \sim \frac{\text{Var}(\ln P(X))}{n}$ [8] | $\inf_P \sup_x \mathbb{E}D(P_X||\hat{P}_X) \sim \frac{S-1}{2n}$ [57], [58] |
| $n = \Theta(S/\ln S)$ [9] | $n = \Theta(S)$ [60] | $\min_Q \sup_P \frac{1}{n}D(P_X||Q_{X^n}) \sim \frac{S-1}{2n} \ln n$ [59] |

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

Table II conveys several important messages. First, in the asymptotic regime, there is a logarithmic factor between the redundancy of the compression and distribution estimation problems. 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 alphabet size $S$.

Second, the large alphabet setting shows that estimation of entropy is considerably easier than both estimating the corresponding distribution, and compression. It is somewhat surprising and enlightening, since there has been a well-received tradition to apply data compression techniques to estimate entropy, even beyond the information theory community, e.g. [63], whereas one of the implications of Table II is that the approach of entropy estimation via compression can be highly sub-optimal in large alphabet regimes.

The estimation of $F_\alpha(P)$ is another example demonstrating the usefulness of our general recipe. Note that, since $F_\alpha(P), 0 < \alpha < 1$, is not Lipschitz with respect to the Wasserstein distance considered by Valiant and Valiant [9], their achievability technique does not apply here. Again, we can show that our estimator outperforms the maximum likelihood estimator. It is shown in [14] that if $n = cS^{1/\alpha}$, where $c > 0$ is a constant, then the maximum $L_2$ risk of $F_\alpha(P_n)$ is bounded away from zero. On the other hand, our results imply, for example, if we are interested in estimating the functional $\sum_{i=1}^S \sqrt{P_i}$, then it is necessary and sufficient to consider a sample size of $\omega(S^{2/\alpha})$. To our knowledge, this is the first consistent estimation result for functionals of this form in high dimensions in the literature.

For estimation of $F_\alpha(P), \alpha > 1$, it was shown in the companion paper [14] that the MLE achieves universal $L_2$ convergence rate $\max\{n^{-2(\alpha-1)}, n^{-1}\}$, which implies that it suffices to take $n = \omega(1)$ samples for the MLE to have vanishing worst-case $L_2$ risk. If we plot the phase transitions of $\ln n/\ln S$ for estimating $F_\alpha(P)$ using $F_\alpha(P_n)$ with respect to $\alpha$, we obtain Figure 1.
depending on whether $\alpha > 1$ or $\alpha < 1$.

However, certain questions remain unanswered. For example, it is not clear, for estimation requires a number of measurements super-linear or constant in the size of the alphabet according to whether $\alpha < 1$ or $\alpha > 1$.

Combining Table II and Figure 1 leads to the interesting observation that, in high dimensional asymptotics, estimating a distribution itself for various values of $\alpha > 1$ can be achieved using $\omega(S^{1/\alpha}/\ln S)$ samples. Since moment generating function for random variable $E$ can be interpreted as the moment generating function for random variable $\eta(X)$ as

$$F_{\alpha}(P) = \mathbb{E}_P \left[ e^{(1-\alpha)\eta(X)} \right].$$

It is shown in Valiant and Valiant [9] that the distribution of $\eta(X)$ can be estimated using $\omega(S^{1/\alpha}/\ln S)$ samples. Since moment generating functions can determine the distribution under some conditions, it is indeed plausible to see that the problem of estimating $F_{\alpha}(P)$, or the moment generating function of $\eta(X)$, is either easier or harder than estimating the distribution of $\eta(X)$ itself for various values of $\alpha$.

We now briefly shift our focus towards estimation of Rényi entropy $H_{\alpha}(P)$, which is closely related to the functional $F_{\alpha}(P)$ via $H_{\alpha}(P) = \frac{\ln F_{\alpha}(P)}{\ln S}$. Acharya et al. [67] considered the estimation of $H_{\alpha}(P)$, and demonstrated that the sample complexity for estimating $H_{\alpha}(P)$ may exhibit a different behavior than that of estimating $F_{\alpha}(P)$ for certain values of $\alpha$. By a partial application of results from the current paper, [67] showed that for $0 < \alpha < 1$, the sample complexity for estimating $H_{\alpha}(P)$ is between $\omega(S^{1/(1-\alpha)})$, $\forall \eta > 0$ and $O(S^{1/\alpha}/\ln S)$. It was also shown in [67] that for $\alpha > 1$, $\alpha \in \mathbb{Z}^+$, the sample complexity is $\Theta(S^{1-1/\alpha})$, which can be achieved by a bias-corrected MLE. Finally, for non-integer $\alpha > 1$, [67] showed that the sample complexity for estimating $H_{\alpha}(P)$ is between $\omega(S^{1-\eta})$, $\forall \eta > 0$ and $O(S)$, and that it suffices to take $n = O(S)$ samples for the MLE to be consistent. However, certain questions remain unanswered. For example, it is not clear, for $\alpha > 1$, $\alpha \notin \mathbb{Z}^+$, whether the MLE indeed requires $\omega(S)$ samples, or whether there exist estimators that can consistently estimate $H_{\alpha}(P)$ with sublinear samples. We provide partial answers to these questions below by focusing on the case when $1 < \alpha < 3/2$. First, we show in Theorem 4 that simply plugging in the novel estimator $F_{\alpha}$ from Theorem 3 to the definition of $H_{\alpha}(P)$ results in an estimator that needs at most $n = O(S/\ln S)$ samples when $1 < \alpha < 3/2$. 

Fig. 1: For any point above the thick curve, consistent estimation of $F_{\alpha}(P)$ is achieved using MLE $F_{\alpha}(P_n)$ [14]. Our estimator slightly improves over MLE to achieve the optimal $\Theta(S^{1/\alpha}/\ln S)$ sample complexity when $0 < \alpha < 1$. For the regime $0 < \alpha < 1$ below the thick curve, Theorem 3 shows that no estimator can have vanishing maximum $L_2$ risk.

We observe a sharp phase transition at $\alpha = 1$, as the sample size requirement shifts from $n = \omega(S^{1/\alpha}/\ln S)$ to $n = \omega(1)$, depending on whether $\alpha$ is in the left or right neighborhood of 1, respectively. Hence, $\alpha = 1$ is a critical point in that consistent estimation requires a number of measurements super-linear or constant in the size of the alphabet according to whether $\alpha < 1$ or $\alpha > 1$.

Combining Table II and Figure 1 leads to the interesting observation that, in high dimensional asymptotics, estimating a functional of a distribution could be easier (e.g. $H(P)$, $F_{\alpha}(P)$, $\alpha > 1$) or harder (e.g. $F_{\alpha}(P)$, $0 < \alpha < 1$) than estimating the distribution itself. This observation taps into another interesting interpretation of the functional $F_{\alpha}(P)$. In information theory, the random variable $\eta(X) = \ln \frac{1}{P(x)}$ is known as the information density, and plays important roles in characterizing higher order fundamental limits of coding problems [65], [66]. The functional $F_{\alpha}(P)$ can be interpreted as the moment generating function for random variable $\eta(X)$ as

$$F_{\alpha}(P) = \mathbb{E}_P \left[ e^{(1-\alpha)\eta(x)} \right].$$

It is shown in Valiant and Valiant [9] that the distribution of $\eta(X)$ can be estimated using $\omega(S/\ln S)$ samples. Since moment generating functions can determine the distribution under some conditions, it is indeed plausible to see that the problem of estimating $F_{\alpha}(P)$, or the moment generating function of $\eta(X)$, is either easier or harder than estimating the distribution of $\eta(X)$ itself for various values of $\alpha$. 

We now briefly shift our focus towards estimation of Rényi entropy $H_{\alpha}(P)$, which is closely related to the functional $F_{\alpha}(P)$ via $H_{\alpha}(P) = \frac{\ln F_{\alpha}(P)}{\ln S}$. Acharya et al. [67] considered the estimation of $H_{\alpha}(P)$, and demonstrated that the sample complexity for estimating $H_{\alpha}(P)$ may exhibit a different behavior than that of estimating $F_{\alpha}(P)$ for certain values of $\alpha$. By a partial application of results from the current paper, [67] showed that for $0 < \alpha < 1$, the sample complexity for estimating $H_{\alpha}(P)$ is between $\omega(S^{1/(1-\alpha)})$, $\forall \eta > 0$ and $O(S^{1/\alpha}/\ln S)$. It was also shown in [67] that for $\alpha > 1$, $\alpha \in \mathbb{Z}^+$, the sample complexity is $\Theta(S^{1-1/\alpha})$, which can be achieved by a bias-corrected MLE. Finally, for non-integer $\alpha > 1$, [67] showed that the sample complexity for estimating $H_{\alpha}(P)$ is between $\omega(S^{1-\eta})$, $\forall \eta > 0$ and $O(S)$, and that it suffices to take $n = O(S)$ samples for the MLE to be consistent. However, certain questions remain unanswered. For example, it is not clear, for $\alpha > 1$, $\alpha \notin \mathbb{Z}^+$, whether the MLE indeed requires $\omega(S)$ samples, or whether there exist estimators that can consistently estimate $H_{\alpha}(P)$ with sublinear samples. We provide partial answers to these questions below by focusing on the case when $1 < \alpha < 3/2$. First, we show in Theorem 4 that simply plugging in the novel estimator $F_{\alpha}$ from Theorem 3 to the definition of $H_{\alpha}(P)$ results in an estimator that needs at most $n = O(S/\ln S)$ samples when $1 < \alpha < 3/2$. 

Fig. 1: For any point above the thick curve, consistent estimation of $F_{\alpha}(P)$ is achieved using MLE $F_{\alpha}(P_n)$ [14]. Our estimator slightly improves over MLE to achieve the optimal $\Theta(S^{1/\alpha}/\ln S)$ sample complexity when $0 < \alpha < 1$. For the regime $0 < \alpha < 1$ below the thick curve, Theorem 3 shows that no estimator can have vanishing maximum $L_2$ risk.
Theorem 7. For any $\alpha \in (1,3/2)$ and any $\delta > 0, \epsilon \in (0,1)$, there exists a constant $c = c_\alpha(\delta, \epsilon) > 0$ such that,

$$\limsup_{S \to \infty} \sup_{P \in \mathcal{M}_S, n \geq \frac{c S}{\alpha}} \mathbb{P} \left( \left| \frac{\ln \hat{F}_n}{1-\alpha} - H_\alpha(P) \right| \geq \delta \right) \leq \epsilon,$$

(17)

where $\hat{F}_n$ is the estimator from Theorem 4.

In words, with high probability $(\ln \hat{F}_n)/(1-\alpha)$ is close to the Rényi entropy provided $n = \omega(S/\ln S)$. In contrast, the MLE requires $n = \omega(S)$ samples for estimating $H_\alpha(P)$, $1 < \alpha < \frac{3}{2}$, as is implied by the following theorem.

Theorem 8. For any $\alpha \in (1,3/2)$ and any constant $c > 0$, there exist some $\delta = \delta_\alpha(c) > 0$ such that the MLE $H_\alpha(P_n)$ satisfies

$$\liminf_{n \to \infty} \inf_{S \geq n/\epsilon} \sup_{P \in \mathcal{M}_S} \mathbb{P} \left( |H_\alpha(P_n) - H_\alpha(P)| \geq \delta \right) = 1,$$

(18)

where $P_n$ is the MLE of $P$.

To conclude this discussion, we conjecture that plugging in our optimal estimators for $F_\alpha(P)$ into the definition of $H_\alpha(P)$ results in minimax rate-optimal estimators for $H_\alpha(P)$ for all $\alpha > 0$.

F. Related work under alternative frameworks

The problem of estimating functionals of probability distributions has a long history. In particular, the problem of entropy estimation has attracted attention from various communities, including information theory, statistics, psychology, computer science, neuroscience, and physics, to name a few. Different communities have focused on different aspects of this problem.

In the information theory community, following the seminal work by Shannon [68], the focus has been on estimating entropy rates of general stationary ergodic processes with fixed (usually small) alphabet sizes. Outside of the favored binary alphabet, printed English contributed the other interesting example of alphabet size 27 (including the “space”). Cover and King [69] gave an overview of the entropy rate estimation literature until 1978. Soon after the appearance of universal data compression algorithms proposed by Ziv and Lempel [70], [71], the information theory community started applying these ideas in entropy rate estimation, e.g. Wyner and Ziv [72], and Kontoyiannis et al. [73]. Verdú [74] provides an overview of universal estimation of information measures until 2005. Jiao et al. [75] constructed a general framework for applying data compression algorithms to establish near-optimal estimators for information rates, with a focus on directed information.

Statisticians have traditionally favored i.i.d. observation models. Due to the triviality of the finite-alphabet regime in classical asymptotic statistics, the focus has shifted to the countably-infinite alphabet case, and on constructing efficient nonparametric estimators for the differential entropy. Antos and Kontoyiannis [76], Wyner and Foster [19], Vu, Yu, and Kass [77], Zhang [78], contributed to the countably-infinite alphabet situation. For nonparametric estimation of differential entropy, the readers are referred to [80], [81], [82], [83], and [84]. Beirlant et al. [85], Wang, Kulkarni, and Verdú [86] provide overviews.

Recently, the seminal work of Říčky et al. [61], [87] ignited interest in the estimation and compression of large alphabet sources. Wagner, Viswanath, and Kulkarni [88] constructed a framework for studying probability estimation in a rare event regime, based on which Ohannessian et al. [89] proposed a methodology for probability functional estimation. Szpankowski and Weinberger [62] calculated the precise minimax redundancy incurred in compressing i.i.d. large alphabet sources. Yang and Barron [90] proposed coding techniques via Poissonization and tilting in the large alphabet regime.

G. Remaining content

The rest of the paper is organized as follows. Section II details the construction of our estimators $\hat{H}$ and $\hat{F}_\alpha$ and their analysis. We present our general approach for proving minimax lower bounds and apply it to Theorem 3 and 5 in Section III. Section IV presents a few experiments comparing the performance of our entropy estimator with that of Valiant and Valiant [9], [13]. Complete proofs of the rest of theorems and lemmas are provided in the appendices.

II. ESTIMATOR CONSTRUCTION AND ANALYSIS

Throughout our analysis, we utilize the Poisson sampling model, which is equivalent to having a $S$-dimensional random vector $Z$ such that each component $Z_i$ in $Z$ has distribution $\text{Poi}(np_i)$, and all coordinates of $Z$ are independent. For simplicity of analysis, we conduct the classical “splitting” operation [91] on the Poisson random vector $Z$, and obtain two independent identically distributed random vectors $X = [X_1, X_2, \ldots, X_S]^T$, $Y = [Y_1, Y_2, \ldots, Y_S]^T$, such that each component $X_i$ in $X$ has distribution $\text{Poi}(np_i/2)$, and all coordinates in $X$ are independent. For each coordinate $i$, the splitting process generates a random variable $T_i$ such that $T_i | Z \sim \text{B}(Z_i, 1/2)$, and assign $X_i = T_i, Y_i = Z_i - T_i$. All the random variables $\{T_i : 1 \leq i \leq S\}$ are conditionally independent given our observation $Z$. 


For simplicity, we re-define $n/2$ as $n$, and denote
\[
\hat{p}_{i,1} = \frac{X_i}{n}, \hat{p}_{i,2} = \frac{Y_i}{n}, \Delta = \frac{c_1 \ln n}{n}, K = c_2 \ln n, t = \frac{\Delta}{4},
\]
where $c_1, c_2$ are positive parameters to be specified later. Note that $\Delta, K, t$ are functions of $n$, where we omit the subscript $n$ for brevity. We remark that the “splitting” operation is used to simplify the analysis, and is not performed in the experiments.

We demonstrate our analysis techniques via the proof of Theorem 2 and 4, and note that similar techniques allow us to establish Theorem 1. Our estimator $\hat{F}_\alpha, \alpha > 0$ is constructed as follows.
\[
\hat{F}_\alpha \triangleq \sum_{i=1}^{S} [L_\alpha(\hat{p}_{i,1}) \mathbb{1}(\hat{p}_{i,2} \leq 2\Delta) + U_\alpha(\hat{p}_{i,1}) \mathbb{1}(\hat{p}_{i,2} > 2\Delta)],
\]
where
\[
L_\alpha(x) \triangleq \min \{S_{K,\alpha}(x), 1\},
\]
\[
S_{K,\alpha}(x) \triangleq \sum_{k=1}^{K} g_k,\alpha (4\Delta)^{-k+\alpha} \prod_{r=0}^{k-1} (x - r/n),
\]
\[
U_\alpha(x) \triangleq I_n(x) \left(1 + \frac{\alpha(1 - \alpha)}{2nx} \right)^{x/\alpha}.
\]
It is evident from the construction that the function $L_\alpha(\cdot)$ (means “lower part”) is the sophisticated estimator we construct to reduce the bias in the “nonsmooth” regime, and the function $U_\alpha(\cdot)$ (means “upper part”) is just the bias-corrected MLE with an interpolation function $I_n(\cdot)$ to make the function $U_\alpha(\cdot)$ smooth. Indeed, when $0 < \alpha < 1$, were it not for the interpolation function, $U_\alpha(x)$ would be unbounded for $x$ close to zero. Note that $L_\alpha(x)$ and $U_\alpha(x)$ are dependent on $n$. We omit this dependence in notation for brevity.

The following lemma characterizes the properties of the function $g(x; a)$ appearing in the definition of $I_n(x)$:

**Lemma 1.** For the function $g(x; a)$ on $[0, a]$ defined as follows,
\[
g(x; a) \triangleq 126 \left(\frac{x}{a}\right)^5 - 420 \left(\frac{x}{a}\right)^6 + 540 \left(\frac{x}{a}\right)^7 - 315 \left(\frac{x}{a}\right)^8 + 70 \left(\frac{x}{a}\right)^9,
\]
we have the following properties:
\[
g(0; a) = 0, \quad g^{(i)}(0; a) = 0, 1 \leq i \leq 4
\]
\[
g(a; a) = 1, \quad g^{(i)}(a; a) = 0, 1 \leq i \leq 4
\]
The function $g(x; 1)$ is depicted in Figure 2.

The usage of the interpolation function was partially inspired by Valiant and Valiant [12], but we are the first to construct it explicitly.
Lemma 1 implies that \( I_n(x) \in C^4[0,1] \). The coefficients \( g_{k,\alpha} \), \( 0 \leq k \leq K \) are coefficients of the best polynomial approximation of \( x^\alpha \) over \( [0,1] \) up to degree \( K \), i.e.,

\[
\sum_{k=0}^{K} g_{k,\alpha} x^k = \arg \min_{g(x) \in \text{poly}_K} \sup_{x \in [0,1]} |g(x) - x^\alpha|,
\]

where \( \text{poly}_K \) denotes the set of algebraic polynomials up to order \( K \). Note that in general \( g_{k,\alpha} \) depends on \( K \), which we do not make explicit for brevity.

Similarly, we define our estimator for entropy \( H(P) \) as

\[
\hat{H} \triangleq \sum_{i=1}^{S} \left[ L_H(\hat{p}_{i,1}) \hat{1}(\hat{p}_{i,2} \leq 2\Delta) + U_H(\hat{p}_{i,1}) \hat{1}(\hat{p}_{i,2} > 2\Delta) \right],
\]

where

\[
L_H(x) \triangleq \min \{ S_{K,H}(x), 1 \}
\]

\[
S_{K,H}(x) \triangleq \sum_{k=1}^{K} g_{k,\alpha} (4\Delta)^{-k+1} \prod_{r=0}^{k-1} (x - r/n)
\]

\[
U_H(x) \triangleq I_n(x) \left( -x \ln x + \frac{1}{2n} \right) \cdot
\]

The coefficients \( \{g_{k,H}\}_{1 \leq k \leq K} \) are defined as follows. We first define

\[
\sum_{k=0}^{K} r_{k,H} x^k = \arg \min_{g(x) \in \text{poly}_K} \sup_{x \in [0,1]} |g(x) - (-x \ln x)|
\]

and then define

\[
g_{k,H} = r_{k,H}, 2 \leq k \leq K, g_{1,H} = r_{1,H} - \ln(4\Delta).
\]

The next two lemmas shows that the estimators \( U_\alpha(x) \), \( U_H(x) \) have nice bias and variance properties when the true probability \( p \) is not too small.

**Lemma 2.** If \( nX \sim \text{Poi}(np), p \geq \Delta \), then for \( c_1 \ln n \geq 1 \),

\[
|\mathbb{E}U_\alpha(X) - p^\alpha| \leq \left\{ \begin{array}{ll}
\frac{17}{n^{\alpha(c_1 \ln n)^2 - \alpha}} + \frac{8310(1+\alpha)}{\alpha(2-\alpha)} p^{\alpha} n^{-c_1/8} & 0 < \alpha < 3/2 \\
\frac{24}{n^{2\alpha(c_1 \ln n)^2 - \alpha}} + \frac{576}{\alpha} p^{2\alpha} n^{-c_1/8} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} & 0 < \alpha \leq 1/2 \\
\frac{14p^{\alpha - 1}}{n} + \frac{576}{\alpha} p^{2\alpha} n^{-c_1/8} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} + \frac{8}{n^{c_1(c_1 \ln n)^2 - \alpha}} & 1/2 < \alpha < 1 \\
\frac{202p^{\alpha}}{n} + \frac{8}{n^2} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} + \frac{120}{\alpha} p^{2\alpha} n^{-c_1/8} & 1 < \alpha < 3/2
\end{array} \right.
\]

\[
\text{Var}(U_\alpha(X)) \leq \left\{ \begin{array}{ll}
\frac{24}{n^{2\alpha(c_1 \ln n)^2 - \alpha}} + \frac{576}{\alpha} p^{2\alpha} n^{-c_1/8} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} & 0 < \alpha \leq 1/2 \\
\frac{14p^{\alpha - 1}}{n} + \frac{576}{\alpha} p^{2\alpha} n^{-c_1/8} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} + \frac{8}{n^{c_1(c_1 \ln n)^2 - \alpha}} & 1/2 < \alpha < 1 \\
\frac{202p^{\alpha}}{n} + \frac{8}{n^2} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} + \frac{120}{\alpha} p^{2\alpha} n^{-c_1/8} & 1 < \alpha < 3/2
\end{array} \right.
\]
Lemma 3. If $nX \sim \text{Poi}(np), p \geq \Delta,$

\[
|E(U_H(X) + p \ln p)| \leq \frac{3}{c_1 n \ln n} + \frac{2}{3(c_1 \ln n)^2 n} + 8024 (p \ln(1/p) + 2p) n^{-c_1/8}
\]

(37)

\[
\text{Var}(U_H(X)) \leq 2p(\ln p - \ln 2)^2/n + 54p^2 [2(\ln p)^2 - 2 \ln p + 3] n^{-c_1/8} + \left(\frac{1}{n} + 60 (p \ln(1/p) + 2p) n^{-c_1/8}\right)^2
\]

+ 2 \left( p \ln(1/p) + \frac{1}{2n} \right) \left( \frac{1}{n} + 60 (p \ln(1/p) + 2p) n^{-c_1/8} \right).
\]

(38)

The following three lemmas characterize the performance of $S_{K,\alpha}(X)$ and $S_{K,H}(X).$

Lemma 4. If $nX \sim \text{Poi}(np), p \leq 4\Delta, \alpha > 0,$ we have

\[
|E(S_{K,\alpha}(X) - p^\alpha| \leq \frac{c_3}{(n \ln n)^{\alpha}}.
\]

(39)

and for $n$ large enough, we can take $c_3 = \frac{2p(2\alpha)c_1^2}{c_2^2},$ where $c_3$ is the constant appearing in Lemma 19. If we also have $c_2 \leq 4c_1,$ then

\[
E(S_{K,\alpha}^2(X) \leq n^{8c_2 \ln 2} \left( \frac{4c_1 \ln n}{n^{2\alpha}} \right)^2.
\]

(40)

For the entropy, if $p \leq 4\Delta,$ we have

\[
|E(S_{K,H}(X) + p \ln p| \leq \frac{C}{n \ln n}.
\]

(41)

When $n$ is large enough, $C$ can be taken to be $\frac{4c_1 c_2^2 (2)}{c_2^2},$ which is given in Lemma 20. If we also have $c_2 \leq 4c_1,$ then

\[
E(S_{K,H}^2(X) \leq n^{8c_2 \ln 2} \left( \frac{4c_1 \ln n}{n^2} \right)^4.
\]

(42)

Lemma 5. If $nX \sim \text{Poi}(np), p \leq \frac{4c_1}{c_2 n \ln n}, 1 < \alpha < 3/2,$ then for $c_2 < 4c_1 \leq c_2 \ln n,$

\[
|E(S_{K,\alpha}(X) - p^\alpha| \leq p^\alpha + D_1 \left( \frac{4c_1}{c_2 n \ln n} \right)^{\alpha-1} p
\]

(43)

\[
E(S_{K,\alpha}^2(X) \leq n^{10c_2 \ln 2} \left( \frac{4c_1 \ln n}{n^{2\alpha+1}} \right)^2.
\]

(44)

where $D_1$ is a positive constant given in Lemma 17.

Lemma 6. If $nX \sim \text{Poi}(np), \frac{4c_1}{c_2 n \ln n} < p \leq 4\Delta, 1 < \alpha < 3/2,$ then for $c_2 < 4c_1,$

\[
|E(S_{K,\alpha}(X) - p^\alpha| \leq 6 \left( \frac{\pi^2 c_1}{c_2 n \ln n} \right)^\alpha
\]

(45)

\[
E(S_{K,\alpha}^2(X) \leq n^{8c_2 \ln 2} \left( \frac{4c_1 \ln n}{n^{2\alpha+2}} \right)^2.
\]

(46)

With the machinery established in Lemma 2 3 4 5 and 6, we are now ready to bound the bias and variance of each summand in our estimators. Define,

\[
\xi = \xi(X,Y) = L_\alpha(X) \mathbb{I}(Y \leq 2\Delta) + U_\alpha(X) \mathbb{I}(Y > 2\Delta),
\]

(47)

where $nX \overset{D}{=} nY \sim \text{Poi}(np),$ and $X$ is independent of $Y.$ Apparently, we have

\[
\hat{F}_\alpha = \sum_{i=1}^S \xi(\hat{p}_{i,1}, \hat{p}_{i,2}),
\]

(48)

and each of the $S$ summands are independent. Hence, it suffices to analyze the bias and variance of $\xi(X,Y)$ thoroughly for all values of $p$ in order to obtain a risk bound for $\hat{F}_\alpha.$ We break this into three different regimes. In the first case when $p \leq \Delta,$ we shall show that the estimator essentially behaves like $L_\alpha(X),$ which is a good estimator when $p$ is small. In the second case when $\Delta \leq p \leq 4\Delta,$ we show that our estimator uses either $L_\alpha(X)$ or $U_\alpha(X),$ which are both good estimators in this case. In the last case $p \geq 4\Delta,$ we show that our estimator behaves essentially like $U_\alpha(X),$ which has good properties when $p$ is not too small.

We denote $B(\xi) \triangleq E(\xi(X,Y) - p^\alpha$ as the bias of $\xi.$

Lemma 7. Suppose $0 < \alpha < 1, 0 < c_1 = 16(\alpha + \delta), 0 < 8c_2 \ln 2 = \epsilon < \alpha, \delta > 0.$ Then,
1) when $p \leq \Delta$,
\[
|B(\xi)| \leq \frac{1}{(n \ln n)^\alpha}, \quad \text{Var}(\xi) \leq \frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}}.
\] (49) (50)

2) when $\Delta \leq p \leq 4\Delta$,
\[
|B(\xi)| \leq \frac{1}{(n \ln n)^\alpha}, \quad \text{Var}(\xi) \leq \begin{cases} 
\frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}} & 0 < \alpha \leq 1/2, \\
\frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}} + \frac{\epsilon^2 - 1}{n} & 1/2 < \alpha < 1.
\end{cases}
\] (51) (52)

3) when $p > 4\Delta$,
\[
|B(\xi)| \leq \frac{1}{n^{\alpha}(\ln n)^{2-\alpha}}, \quad \text{Var}(\xi) \leq \begin{cases} 
\frac{1}{n^{\alpha}(\ln n)^{2+\alpha}} & 0 < \alpha \leq 1/2, \\
\frac{1}{n^{\alpha}(\ln n)^{2+\alpha}} + \frac{\epsilon^2 - 1}{n} & 1/2 < \alpha < 1.
\end{cases}
\] (53) (54)

Now the result of Theorem 2 follows easily from Lemma 7. We have
\[
|\text{Bias}(\hat{F}_\alpha)| \leq \sum_{i=1}^{S} |B(\xi(\hat{p}_{i,1}, \hat{p}_{i,2}))| 
\leq \sum_{i=1}^{S} \frac{1}{(n \ln n)^\alpha} 
\leq \frac{S}{(n \ln n)^\alpha},
\] (55) (56) (57)

and
\[
\text{Var}(\hat{F}_\alpha) = \sum_{i=1}^{S} \text{Var}(\xi(\hat{p}_{i,1}, \hat{p}_{i,2})) 
\leq \sum_{i=1}^{S} \begin{cases} 
\frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}} & 0 < \alpha \leq 1/2, \\
\frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}} + \frac{\epsilon^2 - 1}{n} & 1/2 < \alpha < 1.
\end{cases}
\] (58) (59)

Here we have used the fact that $x^{2\alpha-1}$ is a concave function when $1/2 < \alpha < 1$.

Combining the bias and variance bounds, we have
\[
\sup_{P \in \mathcal{M}_S} \sum_{i=1}^{S} p_i^{2\alpha-1} = S(1/S)^{2\alpha-1} = S^{2-2\alpha},
\] (60) (61) (62)

since $x^{2\alpha-1}$ is a concave function when $1/2 < \alpha < 1$.

The proof of Theorem 1 is essentially the same as that for Theorem 2, with the only differences being replacing Lemma 2 with Lemma 3, applying the entropy part of Lemma 4 and Lemma 15. The proof of Theorem 3 is slightly more involved, and we need to split the analysis in four different regimes.
Lemma 8. Suppose $1 < \alpha < 3/2$. Setting $c_1 = 16(\alpha + \delta)$, $0 < 10c_2 \ln 2 = \epsilon < 2\alpha - 2$, $\delta > 0$, we have the following bounds on $|B(\xi)|$ and $\text{Var}(\xi)$.

1) when $p \leq \frac{4c_1}{c_2^2 n \ln n}$,

\begin{align*}
|B(\xi)| & \leq p^\alpha + \frac{p}{(n \ln n)^{\alpha - 1}}, \\
\text{Var}(\xi) & \leq \frac{(\ln n)^{2\alpha + 2} p}{n^{2\alpha - 1 - \epsilon}}.
\end{align*}

(64)

(65)

2) when $\frac{4c_1}{c_2^2 n \ln n} < p \leq \Delta$,

\begin{align*}
|B(\xi)| & \leq \frac{1}{(n \ln n)^\alpha}, \\
\text{Var}(\xi) & \leq \frac{(\ln n)^{2\alpha + 2}}{n^{2\alpha - \epsilon}}.
\end{align*}

(66)

(67)

3) when $p > \Delta$,

\begin{align*}
|B(\xi)| & \leq \frac{1}{n^\alpha (\ln n)^{2 - \alpha}}, \\
\text{Var}(\xi) & \leq \frac{1}{n^2} + \frac{p}{n}.
\end{align*}

(68)

(69)

Now the result of Theorem [4] follows easily from Lemma [8]. Firstly, the total bias can be bounded by

\begin{equation}
|\text{Bias}(\hat{F}_\alpha)| \leq \sum_{i=1}^{S} |B(\xi(\hat{\beta}_{i,1}, \hat{\beta}_{i,2}))|
\end{equation}

\begin{align*}
&= \sum_{i: p_i \leq \frac{4c_1}{c_2^2 n \ln n}} |B(\xi(\hat{\beta}_{i,1}, \hat{\beta}_{i,2}))| + \sum_{i: \frac{4c_1}{c_2^2 n \ln n} < p_i \leq \Delta} |B(\xi(\hat{\beta}_{i,1}, \hat{\beta}_{i,2}))| + \sum_{i: p_i > \Delta} |B(\xi(\hat{\beta}_{i,1}, \hat{\beta}_{i,2}))| \\
&\leq \sum_{i: p_i \leq \frac{4c_1}{c_2^2 n \ln n}} \left( p^\alpha + \frac{p}{(n \ln n)^{\alpha - 1}} \right) + \sum_{i: \frac{4c_1}{c_2^2 n \ln n} < p_i \leq \Delta} \frac{1}{(n \ln n)^\alpha} + \sum_{i: p_i > \Delta} \frac{1}{n^\alpha (\ln n)^{2 - \alpha}} \\
&\leq \left( \frac{4c_1}{c_2^2 n \ln n} \right)^{\alpha - 1} + \frac{1}{(n \ln n)^{\alpha - 1}} + \frac{1}{(n \ln n)^{\alpha}} \cdot \frac{c_2^2 n \ln n}{4c_1} + \frac{1}{n^\alpha (\ln n)^{2 - \alpha}} \cdot \frac{n}{c_1 \ln n}
\end{align*}

\begin{equation}
\leq \frac{1}{(n \ln n)^{\alpha - 1}}.
\end{equation}

(70)

(71)

(72)

(73)

Secondly, the total variance is bounded by

\begin{equation}
\text{Var}(\hat{F}_\alpha) = \sum_{i=1}^{S} \text{Var}(\xi(\hat{\beta}_{i,1}, \hat{\beta}_{i,2}))
\end{equation}

\begin{align*}
&= \sum_{i: p_i \leq \frac{4c_1}{c_2^2 n \ln n}} \text{Var}(\xi(\hat{\beta}_{i,1}, \hat{\beta}_{i,2})) + \sum_{i: \frac{4c_1}{c_2^2 n \ln n} < p_i \leq \Delta} \text{Var}(\xi(\hat{\beta}_{i,1}, \hat{\beta}_{i,2})) + \sum_{i: p_i > \Delta} \text{Var}(\xi(\hat{\beta}_{i,1}, \hat{\beta}_{i,2})) \\
&\leq \sum_{i: p_i \leq \frac{4c_1}{c_2^2 n \ln n}} \frac{(\ln n)^{2\alpha + 2} p}{n^{2\alpha - 1 - \epsilon}} + \sum_{i: \frac{4c_1}{c_2^2 n \ln n} < p_i \leq \Delta} \frac{(\ln n)^{2\alpha + 2}}{n^{2\alpha - \epsilon}} + \sum_{i: p_i > \Delta} \left( \frac{p}{n} + \frac{1}{n^2} \right) \\
&\leq \frac{(\ln n)^{2\alpha + 2}}{n^{2\alpha - 1 - \epsilon}} + \frac{(\ln n)^{2\alpha + 2}}{n^{2\alpha - \epsilon}} \cdot \frac{c_2^2 n \ln n}{4c_1} + \left( \frac{1}{n} + \frac{1}{n^2} \right) \cdot \frac{n}{c_1 \ln n}
\end{align*}

\begin{equation}
\leq \frac{(\ln n)^{2\alpha + 3}}{n^{2\alpha - 1 - \epsilon}} + \frac{1}{n}
\end{equation}

\begin{equation}
\leq \frac{1}{(n \ln n)^{2\alpha - 2}}.
\end{equation}

(75)

(76)

(77)

(78)

(79)

(80)

Combining the bias and variance bounds, we have

\begin{equation}
\sup_P \mathbb{E}_P \left( \frac{\hat{F}_\alpha - F_\alpha(P)}{2} \right)^2 = (\text{Bias}(\hat{F}_\alpha))^2 + \text{Var}(\hat{F}_\alpha) \leq \frac{1}{(n \ln n)^{2\alpha - 2}}, \quad 1 < \alpha < \frac{3}{2}.
\end{equation}

which completes the proof of Theorem [4].
III. Minimax Lower Bounds for Estimating $F_\alpha(P), 0 < \alpha < 3/2$

The key lemma we will employ in the proof of Theorem [3] is the so-called method of two fuzzy hypotheses presented in Tsybakov [92]. Below we briefly review this general minimax lower bound.

Suppose we observe a random vector $Z \in (\mathcal{Z}, \mathcal{A})$ which has distribution $P_\theta$ where $\theta \in \Theta$. Let $\sigma_0$ and $\sigma_1$ be two prior distributions supported on $\Theta$. Let $\hat{T} = \hat{T}(Z)$ be an arbitrary estimator of a function $T(\theta)$ based on $Z$. We have the following general minimax lower bound.

Lemma 9. [92 Thm. 2.15] Suppose there exist $\zeta \in \mathbb{R}, s > 0, 0 \leq \beta_0, \beta_1 < 1$ such that

\[
\begin{align*}
\sigma_0(\theta : T(\theta) \leq \zeta - s) &\geq 1 - \beta_0, \\
\sigma_1(\theta : T(\theta) \geq \zeta + s) &\geq 1 - \beta_1.
\end{align*}
\]

If $V(F_1, F_0) \leq \eta < 1$, then

\[
\inf \sup \mathbb{P}_\theta \left( |\hat{T} - T(\theta)| \geq s \right) \geq \frac{1 - \eta - \beta_0 - \beta_1}{2}.
\]

Here $V(P, Q)$ is the total variation distance between two probability measures $P, Q$ on the measurable space $(\mathcal{Z}, \mathcal{A})$. Concretely, we have

\[
V(P, Q) \triangleq \sup_{A \in \mathcal{A}} |P(A) - Q(A)| = \frac{1}{2} \int |p - q| d\nu,
\]

where $p = \frac{dP}{d\nu}, q = \frac{dQ}{d\nu}$, and $\nu$ is a dominating measure so that $P \ll \nu, Q \ll \nu$.

A. Case $0 < \alpha < 1$

Towards establishing the minimax lower bound, we construct the two fuzzy hypotheses required by Lemma 9. This type of construction is well-known in approximation theory [49], and is applied in statistical problems in [54] and [55].

Lemma 10. For any given positive integer $L > 0$, there exists two probability measures $\nu_0$ and $\nu_1$ on $[0, 1]$ that satisfy the following conditions:

1) $\int t^i \nu_1(dt) = \int t^i \nu_0(dt)$, for $l = 0, 1, 2, \ldots, L$;
2) $\int t^\alpha \nu_1(dt) - \int t^\alpha \nu_0(dt) = 2E_L[x^\alpha]|_{0,1]}$

where $E_L[x^\alpha]|_{0,1]}$ is the distance in the uniform norm on $[0, 1]$ from the function $f(x) = x^\alpha$ to the space $\text{poly}_L$ of polynomials of no more than degree $L$.

According to Lemma 17 we have

\[
\lim_{L \to \infty} L^{2\alpha} E_L[x^\alpha]|_{0,1]} = \frac{\mu(2\alpha)}{2^{2\alpha}}.
\]

Now we start the proof of Theorem 3 in earnest. Since we have assumed that $n = c \frac{\ln n}{\ln S}$, we have

\[
S \sim \left( \frac{\alpha}{C} \right)^\alpha n^\alpha (\ln n)^\alpha.
\]

Define

\[
M = d_1 \frac{\ln n}{n}, \quad L = d_2 \ln n, \quad S' = S - 1,
\]

where $d_1, d_2$ are positive constants (not depending on $n$) that will be determined later. Without loss of generality we assume that $d_2 \ln n$ is always a positive integer.

For a given integer $L$, let $\nu_0$ and $\nu_1$ be the two probability measures possessing the properties given in Lemma 10. Let $g(x) = Mx$ and let $\mu_i$ be the measures on $[0, 1]$ defined by $\mu_i(A) = \nu_i(g^{-1}(A))$ for $i = 0, 1$. It follows from Lemma 10 that:

1) $\int t^i \mu_1(dt) = \int t^i \mu_0(dt)$, for $l = 0, 1, 2, \ldots, L$;
2) $\int t^\alpha \mu_1(dt) - \int t^\alpha \mu_0(dt) = 2M^\alpha E_L[x^\alpha]|_{0,1]}$

Let $\mu_1^{S'}$ and $\mu_0^{S'}$ be the product priors $\mu_i^{S'} = \prod_{j=1}^{S'} \mu_i$. We assign these priors to the length-$S'$ vector $(p_1, p_2, \ldots, p_{S'})$. Under $\mu_0^{S'}$ or $\mu_1^{S'}$, we have almost surely

\[
\sum_{i=1}^{S'} p_i \leq S'M \sim d_1 \left( \frac{\alpha}{C} \right)^\alpha \left( \frac{(\ln n)^{\alpha+1}}{n^{1-\alpha}} \right) \ll 1,
\]

hence

\[
p_S^0 \geq \left( 1 - O \left( \frac{(\ln n)^{\alpha+1}}{n^{1-\alpha}} \right) \right)^\alpha \sim 1, \quad n \to \infty.
\]
We decompose $F_\alpha(P)$ as
\[ F_\alpha(P) = F_{\hat{\alpha}}(P) + p_0^\alpha, \] (91)
where
\[ F_{\hat{\alpha}}(P) = \sum_{i=1}^{s'} p_i^\alpha. \] (92)

We will first show that Theorem 3 holds when we replace $F_\alpha(P)$ by $F_{\hat{\alpha}}(P)$, and then argue that this lower bound also holds for $F_\alpha(P)$. Indeed, if Theorem 3 is true for $F_{\hat{\alpha}}(P)$, and there exists an estimator $\hat{F}$ for $F_{\hat{\alpha}}(P)$ such that when $n = c_{\ln S}^{-\alpha}$, the maximum risk of $\hat{F}$ converges to zero, then we can construct an estimator $\tilde{F} - 1$ for estimating $F_{\hat{\alpha}}(P)$ with vanishing maximum $L_2$ risk when $n = c_{\ln S}^{-\alpha}$. It then violates the assumption that Theorem 3 is true for $F_\alpha(P)$.

For $Y|p \sim \text{Poi}(np), p \sim \mu_0$, we denote the marginal distribution of $Y$ by $F_{0,M}(y)$, whose pmf can be computed as
\[ F_{0,M}(y) = \int \frac{e^{-np}(np)^y}{y!} \mu_0(dp). \] (93)

We define $F_{1,M}(y)$ in a similar fashion.

**Lemma 11.** The following bounds are true if $d_1 = 1, d_2 = 10c$:
\[ E_{\mu_0^\alpha} F_\alpha(P) - E_{\mu_0^\alpha} F_{\hat{\alpha}}(P) = 2 \left( \frac{\alpha}{c} \right)^\alpha \frac{\mu(2\alpha)d_1^\alpha}{(2d_2)^{2\alpha}} (1 + o(1)) > 0, \] (94)
\[ \text{Var}_{\mu_0^\alpha} (F_\alpha(P)) \leq \left( \frac{\alpha d_1^2}{c} \right)^\alpha \frac{(\ln n)^{3\alpha}}{n^\alpha} \to 0, \quad j = 0, 1, \] (95)
\[ V(F_{1,M}, F_{0,M}) = \frac{1}{2} \sum_{y=0}^{\infty} |F_{1,M}(y) - F_{0,M}(y)| \leq \frac{1}{n^\alpha}. \] (96)

Now setting
\[ \sigma_j = \mu_j^{s'}, j = 0, 1, \]
\[ \theta = (p_1, p_2, \ldots, p_{S-1}), \]
\[ T(\theta) = F_\alpha(P), \]
\[ s = \frac{1}{2} \left( \frac{\alpha}{c} \right)^\alpha \frac{\mu(2\alpha)d_1^\alpha}{(2d_2)^{2\alpha}} \]
\[ \zeta = E_{\mu_0^\alpha} F_\alpha(P) + 2s \]
in Lemma 9, it follows from Chebyshev’s inequality that
\[ \sigma_0(F_\alpha(P) > \zeta - s) = \sigma_0(F_\alpha(P) - E_{\mu_0^\alpha} F_\alpha(P) > s) \leq \frac{\text{Var}_{\sigma_0}(F_\alpha(P))}{s^2} = \beta_0 \to 0, \] (97)
and
\[ \sigma_1(F_\alpha(P) < \zeta + s) = \sigma_1(F_\alpha(P) - E_{\mu_0^\alpha} F_\alpha(P) < -s) \leq \frac{\text{Var}_{\sigma_1}(F_\alpha(P))}{s^2} = \beta_1 \to 0. \] (98)

Also, it follows from the general fact that $V(\prod_{i=1}^{\alpha} P_i, \prod_{i=1}^{\alpha} Q_i) \leq \sum_{i=1}^{\alpha} V(P_i, Q_i)$ (which follows easily from a coupling argument [93]) that
\[ \eta \leq \frac{S'}{n^\alpha} = O(n^{-5}) \to 0. \] (99)

Applying Lemma 9, we have
\[ \inf_F \sup_{i' \neq i} \mathbb{P} \left( |\hat{F} - F_\alpha| \geq s \right) \geq \frac{1}{2}, \quad n \to \infty. \] (100)

According to Markov’s inequality, we have
\[ \inf_F \sup_{i' \neq i} \mathbb{E} \left( \hat{F} - F_\alpha \right)^2 \geq \frac{1}{2} s^2 = \frac{D}{c^{2\alpha}} > 0, \quad n \to \infty. \] (101)

According to the equivalence argument between $F_\alpha(P)$ and $F_\alpha(P)$, we know that
\[ c^{2\alpha} \lim_{n \to \infty} \inf_F \sup_{i' \in M_S} \mathbb{E} \left( \hat{F} - F_\alpha \right)^2 \geq D > 0, \] (102)
where $D > 0$ is a constant that only depends on $\alpha$. 

Lemma 12. For any $0 < \eta < 1$ and positive integer $L > 0$, there exist two probability measures $\nu_0$ and $\nu_1$ on $[\eta, 1]$ such that

1) $\int t^l \nu_1(dt) = \int t^l \nu_0(dt)$, for all $l = 0, 1, 2, \cdots, L$;
2) $\int t^{n-1} \nu_1(dt) - \int t^{n-1} \nu_0(dt) = 2E_L[x^{n-1}][\eta, 1],$

where $E_L[x^{\beta}][\eta, 1]$ is the distance in the uniform norm on $[\eta, 1]$ from the function $f(x) = x^\beta$ to the space spanned by $\{1, x, \cdots, x^L\}$.

Based on Lemma 12, two new measures $\tilde{\nu}_0$, $\tilde{\nu}_1$ can be constructed as follows: for $i = 0, 1$, the restriction of $\tilde{\nu}_i$ on $[\eta, 1]$ is absolutely continuous with respect to $\nu_i$, with the Radon-Nikodym derivative given by

$$
\frac{d\tilde{\nu}_i}{d\nu_i}(t) = \frac{\eta}{t}, \quad t \in [\eta, 1],
$$

and $\tilde{\nu}_i(\{0\}) = 1 - \tilde{\nu}_i([\eta, 1]) \geq 0$. Hence, $\tilde{\nu}_0$, $\tilde{\nu}_1$ are both probability measures on $[0, 1]$, with the following properties

1) $\int t^l \tilde{\nu}_1(dt) = \int t^l \tilde{\nu}_0(dt) = \eta$;
2) $\int t^l \tilde{\nu}_1(dt) = \int t^l \tilde{\nu}_0(dt)$, for all $l = 2, \cdots, L + 1$;
3) $\int t^n \tilde{\nu}_1(dt) - \int t^n \tilde{\nu}_0(dt) = 2\eta E_L[x^{n-1}][\eta, 1]$.

The following lemma characterizes the properties of $E_L[x^{\beta}][\eta, 1]$.

Lemma 13. For $0 < \beta < 1/2$, there exists a universal positive constant $D$ such that

$$
\liminf_{L \to \infty} L^{2\beta} E_L[x^{\beta}][\eta(DL^{-2}, 1)] > 0.
$$

Define

$$
L = d_2 \ln n, \quad \eta = \frac{1}{(DL)^2}, \quad M = \frac{d_1}{S\eta} = \frac{d_1d_2^2D^2 \ln n}{cn},
$$

with universal positive constants $d_1, d_2$ to be determined later. Without loss of generality we assume that $d_2 \ln n$ is always a positive integer. By the choice of $\eta$ we know that

$$
\liminf_{n \to \infty} (\ln n)^{2(\alpha - 1)} E_L[x^{\alpha-1}][\eta, 1] > 0.
$$

Let $g(x) = Mx$ and let $\mu_i$ be the measures on $[0, M]$ defined by $\mu_i(A) = \tilde{\nu}_i(g^{-1}(A))$ for $i = 0, 1$. It then follows that

1) $\int t^l \mu_1(dt) = \int t^l \mu_0(dt) = d_1/S$;
2) $\int t^l \mu_1(dt) = \int t^l \mu_0(dt)$, for all $l = 2, \cdots, L + 1$;
3) $\int t^n \mu_1(dt) - \int t^n \mu_0(dt) = 2\eta M^\alpha E_L[x^{\alpha-1}][\eta, 1]$.

Let $\mu_0^\beta$ and $\mu_1^\beta$ be product priors which we assign to the length-$S$ vector $P = (p_1, p_2, \cdots, S)$. Note that $P$ may not be a probability distribution, we consider the set of approximate probability vectors

$$
\mathcal{M}_S(\gamma) \triangleq \left\{ P : \sum_{i=1}^S p_i - d_1 \leq \frac{1}{(\ln n)^\gamma} \right\},
$$

with universal constant $\gamma > 0$ to be specified later, and further define under the Poissonized model,

$$
R_P(S, n, \gamma) \triangleq \inf_{\tilde{F} \in \mathcal{M}_S(\gamma)} \sup_{P \in \mathcal{M}_S(\gamma)} \mathbb{E}_P[\tilde{F} - F^\alpha(P)]^2.
$$

The equivalence of $R_P(S, n, \infty)$ and $R_P(S, n, \gamma)$ is established in the following lemma.

Lemma 14. For any $S, n \in \mathbb{N}$ and $\gamma > 0$, we have

$$
R_P(S, \frac{d_1n}{4}, \infty) \geq \frac{1}{4d_1^2} R_P(S, n, \gamma) - \frac{1}{2d_1^2} \exp\left(-\frac{d_1n}{8} \right) - \frac{2M^{2\alpha - 2}}{d_1^2 (\ln n)^{2\gamma}}.
$$

In light of Lemma 14, it suffices to consider $R_P(S, n, \gamma)$ to give a lower bound of $R_P(S, n, \infty)$. Denote

$$
\chi \triangleq \mathbb{E}_{\mu_0^\beta} F_{\alpha}(P) - \mathbb{E}_{\mu_0^\beta} F_{\alpha}(P) = 2\eta M^\alpha E_L[x^{\alpha-1}][\eta, 1] \cdot S = 2d_1 M^{\alpha-1} E_L[x^{\alpha-1}][\eta, 1],
$$

and

$$
E_i \triangleq \mathcal{M}_S(\gamma) \bigcap \left\{ P : |F_{\alpha}(P) - \mathbb{E}_{\mu_i^\beta} F_{\alpha}(P)| \leq \frac{\chi}{4} \right\}, \quad i = 0, 1.
$$
Applying Chebyshev’s inequality and the union bound yields that
\[
\mu_i^S[(E_i)^c] \leq \mu_i^S \left[ \left| \sum_{i=1}^{S} p_i - d_1 \right| > \frac{1}{(\ln n)^\gamma} \right] + \mu_i^S \left[ |F_\alpha(P) - \mathbb{E}_{\mu_i^S} F_\alpha(P)| > \frac{\chi}{4} \right]
\]
(112)
\[
\leq (\ln n)^{2\gamma} \sum_{i=1}^{S} \text{Var}_{\mu_i^S}[p_i] + \frac{16}{\chi^2} \text{Var}_{\mu_i^S}[F_\alpha(P)]
\]
(113)
\[
\leq (\ln n)^{2\gamma} \frac{SM^2}{n} + \frac{16}{\chi^2} \frac{SM^{2\alpha}}{n}
\]
(114)
\[
= \frac{d_1^2 d_2 D^4 (\ln n)^{2\gamma+3}}{cn} + \frac{4d_1^2 D_4 (\ln n)^3}{cn(E_L[x^{\alpha-1}]_{[0,1]})^2}
\]
(115)
\[
\rightarrow 0 \text{ as } n \rightarrow \infty,
\]
(116)
where (116) follows from (106). Denote by \( \pi_i \) the conditional distribution defined as
\[
\pi_i(A) = \frac{\mu_i^S(E_i \cap A)}{\mu_i^S(E_i)}, \quad i = 0, 1.
\]
(117)
Now consider \( \pi_0, \pi_1 \) as two priors and \( F_0, F_1 \) as the corresponding marginal distributions. By setting
\[
\zeta = \mathbb{E}_{\mu_0^S} F_\alpha(P) + \frac{\chi}{2}, \quad s = \frac{\chi}{4}, \quad d_1 = \frac{c}{(10eD)^2}, \quad d_2 = 10e, \quad \gamma = 2\alpha,
\]
(118)
we have \( \beta_0 = \beta_1 = 0 \). The total variational distance is then upper bounded by
\[
V(F_0, F_1) \leq V(F_0, G_0) + V(G_0, G_1) + V(G_1, F_1)
\]
(119)
\[
\leq \mu_0^S[(E_0)^c] + V(G_0, G_1) + \mu_1^S[(E_1)^c]
\]
(120)
\[
\leq \mu_0^S[(E_0)^c] + \frac{S}{n^\gamma} + \mu_1^S[(E_1)^c]
\]
(121)
\[
\rightarrow 0,
\]
(122)
where \( G_i \) is the marginal probability under prior \( \mu_i^S \). (121) is given by Lemma 11 and (122) follows from (116). Hence, it follows from Lemma 9 and Markov’s inequality that
\[
R_P(S, n, \gamma) \geq s^2 \inf_{F \in \mathcal{M}_S(\gamma)} \sup_{P \in \mathcal{M}} \mathbb{P} \left( |\hat{F} - F_\alpha(\theta)| \geq s \right)
\]
(123)
\[
\geq 1 - \frac{32}{(\ln n)^2 \chi^2}
\]
(124)
\[
= \frac{d_1^2 (1 - V(F_0, F_1))}{8} \left( \frac{\ln n}{n} \right)^{2(\alpha-1)} (E_L[x^{\alpha-1}]_{[0,1]})^2.
\]
(125)
Now it follows from (106) and Lemma 14 that
\[
\liminf_{n \rightarrow \infty} (n \ln n)^{2(\alpha-1)} \cdot \inf_{F \in \mathcal{M}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P |\hat{F} - F_\alpha(P)|^2
\]
(126)
\[
\geq \liminf_{n \rightarrow \infty} (n \ln n)^{2(\alpha-1)} \cdot \left[ \frac{1}{4d_1^2} R_P(S, \frac{4n}{d_1}, \gamma) - \frac{1}{2d_1^2 \exp \left( \frac{4n}{8} \right)} - \frac{2}{d_1^2 \exp \left( \frac{4n}{8} \right)} \left( \frac{\ln(4n/d_1)}{4n/d_1} \right)^{2\alpha-2} \right]
\]
(127)
\[
\geq \liminf_{n \rightarrow \infty} \left[ \frac{1 - V(F_0, F_1)}{2\alpha+1} \left( \frac{\ln n}{n} \right)^{2\alpha-2} E_L[x^{\alpha-1}]_{[0,1]} \right]^2 - \frac{(n \ln n)^{2(\alpha-1)}}{2d_1^2 \exp \left( \frac{n}{2} \right)} - \frac{2^{5-4\alpha}}{d_1^2 (\ln n)^4} \right] \geq 0.
\]
(128)
(129)

IV. Experiments

As mentioned in the Introduction, the implementation of our algorithm is extremely efficient and has linear complexity with respect to the sample size \( n \), independent of the alphabet size. The only overhead that deserves special mention is the computation of the best polynomial approximation, which is performed via the Remez algorithm [42]. The Chebfun team [43] provides a highly optimized implementation of the Remez algorithm in Matlab [44]. In numerical analysis, the convergence of an algorithm is called quadratic if the error \( e_m \) after the \( m \)-th computation satisfies \( e_m \leq C \alpha^{-2m} \) for some \( C > 0 \) and \( 0 < \alpha < 1 \). Under some assumptions about the function to approximate, one can prove [49, Pg. 96] the quadratic convergence of the Remez algorithm. Empirical experiments partially validate the efficiency of the Remez algorithm, which computes order 500 best polynomial approximation for \( -x \ln x, x \in [0,1] \) in a fraction of a second on a Thinkpad X220 laptop. Considering the fact that the order of approximation we conduct is logarithmic in \( n \), our estimator requires very modest computation.
We emphasize that although the value of constants $c_1, c_2$ required in Lemma 7 lead to rather poor constants in the bias and variance bounds, the practical performance could be much better than what the theoretical bounds guarantee. It is due to the fact that we keep on using worst case upper bounds in the analysis. Practically, experimentation shows that $c_1 \in [0.1, 0.5], c_2 = 0.7$ results in very effective entropy estimation. In our experiments, we do not conduct “splitting” and lose half of the samples, and we evaluate our estimator on the multinomial rather than the Poisson sampling model required for the analysis.

Our experiments show that in practice our estimator is amenable to rather tight confidence intervals, despite its somewhat involved nature. Noting that the bias of an estimator is usually harder to estimate than the variance [94, Chap. 10], it is desirable to obtain a tight theoretical bound on the bias, and to apply Bootstrap [95] to construct confidence intervals. The bias estimates provided in Lemma 4 are quite tight, and with the practical value of constants $c_1, c_2$, they lead to very good confidence intervals. The idea of decreasing the bias at the expense of increasing variance to obtain good confidence intervals also appears in [96].

Given the extensive literature on entropy estimation, we demonstrate the efficacy of our general recipe by detailing a few experiments for that problem.

A. Convergence properties along $n = c \frac{S}{\ln S}$

First, we demonstrate that if we choose $n = c \frac{S}{\ln S}$ and take $S \to \infty$, then the MSE of our estimator is bounded, whereas that of the maximum likelihood estimator goes to infinity. In fact, our analysis of the MLE in [14] and Paninski [20] showed that along the sequence $n = c \frac{S}{\ln S}$, the supremum risk of MLE grows as $(\ln S)^2$ when $S/n$ is relatively small, and grows as $(\ln \ln S)^2$ when $S/n$ is very large. As we now see, the experiments validate the theory.

We choose $c = 8$, and sample 30 points equally spaced in a logarithmic scale from $10^{0.5}$ to $10^6$ as candidates for alphabet size $S$. For each alphabet size $S$, we take $n = 8S/\ln S$ samples from a uniform distribution with alphabet size $S$, and do 10 Monte Carlo simulations to obtain the empirical MSE. The result is demonstrated in Figure 3.
Figure 3: The empirical MSE of our estimator and the MLE along sequence $n = 8S/\ln S$, where $S$ is sampled equally spaced logarithmically from $10^{0.5}$ to $10^6$. The horizontal line is $S$, and the vertical line is the MSE.

Figure 3 demonstrates that indeed along the sequence $n = 8S/\ln S$, the MSE of our estimator stays bounded by 0.031. However, the MSE of the MLE grows unboundedly.

It deserves mentioning that when $S = 10^6$, the entropy associated with the uniform distribution over $S$ elements is $\ln S = 13.8155$. It is evident from Figure 3 that the MSE of the MLE is roughly 0.8, but the MSE of our estimator is uniformly bounded by 0.031 for all $S$ in the experiment.

B. Comparison of MLE, our estimator, and Valiant and Valiant [13]

Recently, Valiant and Valiant [13] provided a modification of [9] to estimate entropy, and demonstrated its superior empirical performance via comparison with various existing algorithms, even with the algorithm proposed in Valiant and Valiant [9]. Hence, it is most informative to compare our algorithm with that of [13]. In our experiments, we downloaded and used the Matlab implementation of the estimator in [13], with default parameters.

1) Data rich regime: $S \ll n$: We first experiment in the regime $S \ll n$, which is an “easy” regime where even the MLE is known to perform very well. However, the estimator in [13] exhibits peculiar behavior. We conduct 10 Monte Carlo simulations of estimation based on $n = 10000$ observations from a uniform distribution over an alphabet of size $S = 200$. The outputs of each Monte Carlo iteration are exhibited in Figure 4.
It is quite clear that over the 10 Monte Carlo iterations, our algorithm performs quite well and is stable, the MLE is stable but its average value is far from the true entropy, but the estimator in [13] is oscillating quite wildly around some point which is also far from the true entropy. We experimented on other distributions such as the Zipf, with similar empirical findings.

We remark that the estimator in [13] has substantially longer running time than ours in the data rich regime. The total running time of our estimator in 10 Monte Carlo simulations is 0.09s, whereas the one in [13] takes 10.5s to complete the 10 simulations.

2) Data sparse regime: $S \gg n$: This is the regime where the conventional approaches such as MLE fail. We fix $S = 20000$, and sample $n = 10000$ times from a uniform distribution with $S$ elements, i.e., the number of observations is half the size of the alphabet. The outputs of MLE, our estimator, and the estimator in [13] in 10 Monte Carlo simulations are exhibited in Figure 5.
Figure 5 shows that the MLE is stable, but is far from the true entropy. Both our estimator and that of [13] perform quite well. Interestingly, with the same sample size $n = 10000$, the estimator in [13] runs much faster than in the data rich regime, with a total running time 0.96s. However, it is still slower than our estimator, which takes 0.1s to complete the 10 simulations.

3) Estimation of mutual information: One functional of particular significance in various applications is the mutual information $I(X;Y)$, but it cannot be directly expressed in the form of (1). Indeed, we have

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \ln \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = \sum_{x,y} P_{XY}(x,y) \ln \frac{P_{XY}(x,y)}{(\sum_y P_{XY}(x,y))(\sum_x P_{XY}(x,y))}. \ (130)$$

However, one can easily show that if $X,Y$ both take values in alphabets of size $S$, then the sample complexity for estimating $I(X;Y)$ is $\Theta(S^2/\ln S)$, rather than $\Theta(S^2)$ required by the MLE. Applying our entropy estimator in the following way results in an order-optimal mutual information estimator. We represent

$$I(X;Y) = H(X) + H(Y) - H(XY), \ (131)$$

where $H(XY)$ is the entropy associated with the joint distribution $P_{XY}$, and use our entropy estimator to estimate each term. As was exhibited in previous experiments, in the data rich regime, MLE is better than the estimator in [13], and in the data sparse regime, [13] is better than the MLE, and in both regimes our estimators are doing well uniformly. However, in mutual information estimation, the estimators of $H(X)$ and $H(Y)$ may be operating in the data rich regime, but that of $H(XY)$ in the data sparse regime. Conceivably, in this situation neither the MLE nor [13] would perform well, but our estimator is expected to have good performance.

In order to investigate this intuition, we fix $S = 200, n = 20000$, and generate two random variables $X,Y$ both with alphabet size $S$ as follows. We first randomly generate the marginal distribution $P_X(i), 1 \leq i \leq S$, where for each $i$ we choose an independent random variable distributed as Beta(0.6, 0.5), and we normalize at the end to make $P_X$ a distribution. We pass $X$ through a transition channel to obtain $Y$, such that $Y = X$ with probability 0.5, and $Y$ takes all other $S - 1$ values with equal probability $0.5/(S - 1)$. We conduct 10 Monte Carlo simulations, and the results are exhibited in Figure 6.
It is clear from Figure 6 that both the MLE and the estimator in [13] suffer from large bias and/or large variance, but our estimator is quite robust and accurate. At the same time, the estimator in [13] has considerably longer running time than our estimator. It takes 159.6s to complete the 10 simulations, whereas ours requires 0.25s.

We have experimented with other distributions such as the Zipf, as well as randomly generated distributions, with similar results. In summary, we observe that

1) the performance of the MLE is always quite stable, but usually concentrates at some point away from the true functional value;
2) the estimator in [13] performs quite well in the data sparse regime $S \gg n$, but performs worse than the MLE in the data rich regime $S \ll n$, which is undesirable in applications such as mutual information estimation and situations where the alphabet size $S$ is unknown;
3) our estimator has stable performance, linear complexity, high accuracy, and the potential of admitting tight confidence intervals.

V. ACKNOWLEDGMENTS

We are grateful to Gregory Valiant for introducing the entropy estimation problem to us, which motivated this work. We thank many approximation theorists for very helpful discussions, in particular, Dany Leviatan, Kirill Kopotun, Feng Dai, Volodymyr Andriyevskyy, Gancho Tachev, Radu Paltanea, Paul Nevai, Doron Lubinsky, Dingxuan Zhou, and Allan Pinkus. We thank Marcelo Weinberger for bringing up the question of comparing the difficulty between entropy estimation and data compression, and Thomas Courtade for proposing to view $F_{\alpha}(P)$ as moment generating functions of the information density random variable and suggesting some tricks in the proof of Lemma 15. We thank Adel Javanmard for discussions related to constructing confidence intervals for high dimensional statistical problems, and Liam Paninski for interesting discussions related to Paninski [20]. Finally, we thank Dmitri Serguei Pavlichin for translating articles from Bernstein’s collected works, whose English versions are unavailable.

APPENDIX A

Auxiliary Lemmas

Lemma 15. If the support of distribution $P$ is of size $S, S \geq 56$, then

$$\text{Var}(-\ln P(X)) \leq \frac{3}{4} (\ln S)^2.$$ (132)
The next lemma relates the minimax risk under the Poissonized model and that under the Multinomial model. We define the minimax risk for Multinomial model with \( n \) observations on alphabet size \( S \) for estimating functional \( F \) as

\[
R(S, n) \triangleq \inf_F \sup_{P \in \mathcal{M}_S} E_{\text{Multinomial}} \left( \hat{F} - F(P) \right)^2,
\]

and the counterpart for the Poissonized model as

\[
R_P(S, n) \triangleq \inf_F \sup_{P \in \mathcal{M}_S} E_{\text{Poisson}} \left( \hat{F} - F(P) \right)^2.
\]

The next lemma is an extension of Wu and Yang [10], Eq. (9).

**Lemma 16.** The minimax risks under the Poissonized model and the Multinomial model are related via the following inequalities:

\[
R_P(S, 2n) - e^{-n/4} \sup_{P \in \mathcal{M}_S} |F(P)|^2 \leq R(S, n) \leq 2R_P(S, n/2).
\]

The following lemma characterizes the best polynomial approximation error of \( x^\alpha \) over \([0, 1]\) in a very precise sense. Concretely, denoting the best polynomial approximation error with order at most \( n \) for function \( f \) as \( E_n[f] \), we have the following lemma.

**Lemma 17.** For any positive integer \( n \) and any \( \alpha > 0 \), we have

\[
E_n[x^\alpha]_{[0, 1]} \leq \left( \frac{\pi}{2n} \right)^{2\alpha},
\]

Moreover, the following limit exists:

\[
\lim_{n \to \infty} n^{2\alpha} E_n[x^\alpha]_{[0, 1]} = \frac{\mu(2\alpha)}{2^{2\alpha}},
\]

where \( \mu(p) \triangleq \lim_{n \to \infty} n^n E_n[|x|^p]_{[-1, 1]}, p > 0 \) is the Bernstein function introduced by [46].

Denote the best polynomial approximation of \( x^\alpha \), \( \alpha > 0 \) to the \( n \)-th degree by \( \sum_{k=0}^n g_{k,\alpha} x^k \), and define \( R_{n,\alpha}(x) \triangleq \sum_{k=1}^n g_{k,\alpha} x^k \). For \( 1 < \alpha < 3/2 \), we have the norm bound

\[
\max_{0 \leq x \leq 1} |R_{n,\alpha}(x) - x^\alpha| \leq \left( \frac{\pi}{n} \right)^{2\alpha},
\]

and the pointwise bound

\[
|R_{n,\alpha}(x)| \leq D_1 x \frac{n}{n^{2(\alpha - 1)}}, \quad \forall x \in \left[ 0, \frac{1}{n^2} \right],
\]

where \( D_1 > 0 \) is a universal positive constant. Furthermore, for any \( \alpha > 0 \), we have

\[
|g_{k,\alpha}| \leq 2^{3n}, \quad |g_{k,H}| \leq 2^{3n}, \quad k = 1, 2, \ldots, n,
\]

where the coefficients \( g_{k,H} \) are defined in (34).

Although the Bernstein function \( \mu(p) \) seems hard to analyze, we can compute it fairly easily using well-developed machinery in numerical analysis. For example, [47] showed the following bound on \( \mu(1) \) using analytical methods:

\[
0.2801685460\ldots \leq \mu(1) \leq 0.2801733791,
\]

but we can easily obtain it numerically in the Chebfun system [43] using polynomial approximation order roughly 100.

We have the following result by Ibragimov [47].

**Lemma 18.** The following limits exists:

\[
\lim_{n \to \infty} n^2 E_n[-x \ln x]_{[0, 1]} = \frac{\nu_1(2)}{2} < \frac{1}{2}.
\]

The function \( \nu_1(p) \) was introduced by Ibragimov [47] as the following limit for \( p \) positive even integer and \( m \) positive integer:

\[
\lim_{n \to \infty} \frac{n^p}{(\ln n)^{m-p}} E_n[|x|^p \ln^m |x|]_{[-1, 1]} = \nu_1(p).
\]

This Lemma follows from Ibragimov [47], Thm. 95]. Note that Ibragimov [47] contained a small mistake where the limit of \( n^2 E_n[(1-x) \ln(1-x)]_{[-1, 1]} \) was wrongly computed to be \( 4 \nu_1(2) \), but it is supposed to be \( \nu_1(2) \). Using numerical computation provided by the Chebfun [43] toolbox, we obtain that

\[
\nu_1(2) \approx 0.453,
\]
and this asymptotic result starts to be very accurate even for small order of polynomials such as 5.

The following two lemmas characterize the approximation error of $x^\alpha$ and $-x \ln x$ when $x$ is small.

**Lemma 19.** For all $x \in [0, 4\Delta]$, the following bound holds:

$$\left| \sum_{k=1}^{K} g_{k,\alpha} \Delta^{-k+\alpha} x^k - x^\alpha \right| \leq \left( \frac{\pi}{2} \right)^{2\alpha} \frac{2(4\Delta)^{\alpha}}{K^{2\alpha}} = \frac{c_3}{(n \ln n)^\alpha},$$

where $c_3 = 2 \left( \frac{\pi^2 c_1}{c_2} \right)^\alpha$. When $n$ is large enough, we could take

$$c_3 = \frac{2 \mu(2\alpha)c_1^\alpha}{c_2^\alpha},$$

where the function $\mu(\cdot)$ is the Bernstein function introduced in Theorem 6.

**Lemma 20.** For all $x \in [0, 4\Delta]$, there exists a constant $C > 0$ such that

$$\left| \sum_{k=1}^{K} g_{k,H}(4\Delta)^{-k+1} x^k + x \ln x \right| \leq C \frac{n \ln n}{n \ln n}.$$

Moreover, when $n$ is large enough, we could take $C$ to be

$$C = \frac{4c_1 \nu_1(2)}{c_2^2} \approx \frac{1.81 c_1}{c_2^2},$$

where the function $\nu_1(p)$ is introduced in Lemma 18.

According to Lemma 18, the asymptotic result $C \approx \frac{1.81 c_1}{c_2^2}$ starts to become very accurate even from very small values of $K$ such as 5. The following lemma gives some tails bounds for Poisson random variables.

**Lemma 21.** If $X \sim \text{Poi}(\lambda)$, then for any $\delta > 0$, we have

$$\Pr(X \geq (1 + \delta)\lambda) \leq \left( \frac{e^{\delta}}{(1 + \delta)^{1+\delta}} \right)^\lambda \leq e^{-\delta^2/2}. \quad (149)$$

$$\Pr(X \leq (1 - \delta)\lambda) \leq \left( \frac{e^{-\delta}}{(1 - \delta)^{1-\delta}} \right)^\lambda \leq e^{-\delta^2/2}. \quad (150)$$

Next lemma gives an upper bound on the $k$-th moment of a Poisson random variable.

**Lemma 22.** Let $X \sim \text{Poi}(\lambda)$, $k$ be an positive integer. Taking $M = \max\{\lambda, k\}$, we have

$$\mathbb{E}X^k \leq (2M)^k. \quad (151)$$

The next two lemmas from Cai and Low [55] are simple facts we will utilize in the analysis of our estimators.

**Lemma 23.** [55] Lemma 4] Suppose $\mathbb{I}(A)$ is an indicator random variable independent of $X$ and $Y$, then

$$\text{Var}(X\mathbb{I}(A) + Y\mathbb{I}(A^c)) = \text{Var}(X)\text{P}(A) + \text{Var}(Y)\text{P}(A^c) + (\text{EX} - \text{EY})^2\text{P}(A)\text{P}(A^c). \quad (152)$$

**Lemma 24.** [55] Lemma 5] For any two random variables $X$ and $Y$,

$$\text{Var}(\min\{X, Y\}) \leq \text{Var}(X) + \text{Var}(Y). \quad (153)$$

In particular, for any random variable $X$ and any constant $C$,

$$\text{Var}(\min\{X, C\}) \leq \text{Var}(X). \quad (154)$$

**APPENDIX B**

**PROOF OF THEOREM 7** AND MAIN LEMMAS

**A. Proof of Theorem 7**

The convexity of $x^\alpha, \alpha > 1$ yields

$$F_\alpha(P) \geq \sum_{i=1}^{S} \left( \frac{1}{S} \right)^\alpha = S^{1-\alpha}, \quad (155)$$

where $S = \sum_{i=1}^{n} P_i$. The convex function $F_\alpha(P)$ is an upper bound for the sum of the powers of the probabilities $P_i$.
hence for any $\delta > 0$, 
\[
\left\{ \mathbf{Z} : \frac{\ln \hat{F}_\alpha(Z)}{1-\alpha} - H_\alpha \geq \delta \right\} \subseteq \left\{ \mathbf{Z} : \left| \hat{F}_\alpha(Z) - F_\alpha \right| \geq \left(1 - e^{(1-\alpha)\delta} \right) F_\alpha \right\} 
\]
\[
\subseteq \left\{ \mathbf{Z} : \left| \hat{F}_\alpha(Z) - F_\alpha \right| \geq \left(1 - e^{(1-\alpha)\delta} \right) S^{1-\alpha} \right\}. 
\]

Theorem 4 implies that there exists a constant $0 < C_\alpha < \infty$ such that 
\[
\sup_P \mathbb{E}_P \left( \left| \hat{F}_\alpha - F_\alpha \right| \right)^2 \leq \frac{C_\alpha}{(n \ln n)^{2\alpha-2}}, 
\]
where the supremum is taken over all discrete distributions supported on countably infinite alphabet. Using this estimator and applying Chebychev’s inequality, 
\[
\sup_P \mathbb{P} \left( \left| \frac{\ln \hat{F}_\alpha(Z)}{1-\alpha} - H_\alpha \right| \geq \delta \right) \leq \sup_P \mathbb{P} \left( \left| \hat{F}_\alpha - F_\alpha \right| \geq \left(1 - e^{(1-\alpha)\delta} \right) S^{1-\alpha} \right) 
\]
\[
\leq \frac{\sup_P \mathbb{E}_P \left| \hat{F}_\alpha - F_\alpha \right|^2}{(1 - e^{(1-\alpha)\delta})^2 S^{2-2\alpha}} \leq \frac{C_\alpha}{(1 - e^{(1-\alpha)\delta})^2} \left( \frac{S}{n \ln n} \right)^{2\alpha-2}. 
\]

The proof is finished by choosing 
\[
c_\alpha(\delta, \epsilon) = \left( \frac{(1 - e^{(1-\alpha)\delta})^2}{C_\alpha} \right)^{-\frac{1}{2\alpha-2}}. 
\]

**B. Proof of Theorem 8**

Since the central limit theorem claims that 
\[
\frac{\text{Poi}(\lambda) - \lambda}{\sqrt{\lambda}} \sim \mathcal{N}(0, 1) 
\]
as $\lambda \to \infty$, there exists $\lambda_0 > 0$ such that 
\[
\mathbb{P}(\text{Poi}(\lambda) > \lambda + 1) \geq \frac{1}{3}, \quad \forall \lambda \geq \lambda_0. 
\]
Denoting $c_m = \max\{c, \lambda_0\}$, we set $S_0 = \lfloor \frac{n}{c_m} \rfloor \leq \lfloor \frac{n}{\epsilon} \rfloor \leq S$ and consider the distribution $P = (1/S_0, 1/S_0, \ldots, 1/S_0, 0, 0, \ldots, 0)$, then $H_\alpha(P) = \ln S_0$. Under the Poissonized model $n \hat{p}_i \sim \text{Poi}(np_i)$, $1 \leq i \leq S$, we have $\hat{p}_i = 0$ for $i > S_0$, and 
\[
p \triangleq \mathbb{P} \left( \hat{p}_i > \frac{c_m + 1}{n} \right) \geq \frac{1}{3}, \quad \forall i = 1, 2, \cdots, S_0. 
\]
Defining 
\[
N = \sum_{i=1}^{S_0} \mathbb{1} \left( \hat{p}_i > \frac{c_m + 1}{n} \right), 
\]
then the random variable $N$ follows a Binomial distribution $N \sim \text{B}(S_0, p)$, and by the central limit theorem again we have 
\[
\lim_{n \to \infty} \mathbb{P} \left( N \geq \frac{S_0}{6} \right) = 1. 
\]
Given $\eta \triangleq N/S_0 \geq 1/6$, it follows from the convexity of $x^\alpha, \alpha > 1$ that 
\[
\sum_{i=1}^{S} \hat{p}_i^\alpha = \sum_{1 \leq i \leq S_0; \hat{p}_i > \frac{c_m + 1}{n}} \hat{p}_i^\alpha + \sum_{1 \leq i \leq S_0; \hat{p}_i \leq \frac{c_m + 1}{n}} \hat{p}_i^\alpha \geq \eta S_0 \cdot \left( \frac{M}{S_0} \right)^\alpha + (S_0 - \eta S_0) \cdot \left( \frac{1 - \eta M}{S_0 - \eta S_0} \right)^\alpha \triangleq S_0^{1-\alpha} \cdot f(\eta, M), 
\]

(168)
where
\[ \frac{1}{\eta} \geq \frac{S_0}{N} \geq M \triangleq S_0 \cdot \frac{1}{N} \sum_{1 \leq i \leq S_0; \tilde{p}_i > \frac{c_m + 1}{n}} \tilde{p}_i \geq S_0 \cdot \frac{c_m + 1}{n} \geq 1 + \frac{1}{c_m}. \] (169)

It can be easily checked that
\[ f(x, y) = xy^\alpha + \frac{(1 - xy)^\alpha}{(1 - x)^\alpha - 1} \] (170)
\[ \frac{\partial f}{\partial y} = \alpha xy^{\alpha - 1} - \frac{\alpha x(1 - xy)^{\alpha - 1}}{(1 - x)^{\alpha - 1}} = \frac{\alpha x}{(1 - x)^{\alpha - 1}} \cdot ((y - xy)^{\alpha - 1} - (1 - xy)^{\alpha - 1}) > 0, \quad 0 < x < xy \leq 1. \] (171)

Hence, due to $0 < 1/6 \leq \eta < 1 < M \leq 1/\eta$, we conclude that $f(\eta, M) \geq f(\eta, 1 + c_m^{-1}) > f(\eta, 1) = 1$. Since $f(\eta, 1 + c_m^{-1})$ is continuous with respect to $\eta \in [1/6, c_m/(c_m + 1)]$, we have
\[ \tilde{H}_n(P_n) \leq \ln S_0 - \frac{\ln f(\eta, M)}{\alpha - 1} \leq \ln S_0 - \frac{\ln f(\eta, 1 + c_m^{-1})}{\alpha - 1} \leq \ln S_0 - \frac{\ln \min_{\eta \in [1/6, c_m/(c_m + 1)]} f(\eta, 1 + c_m^{-1})}{\alpha - 1} < \ln S_0. \] (172)

Then the proof is completed by choosing
\[ \delta_\alpha(c) = \frac{\min_{\eta \in [1/6, c_m/(c_m + 1)]} \ln f(\eta, 1 + c_m^{-1})}{\alpha - 1} > 0. \] (173)

### C. Proof of Lemma [2]

For $p \geq \Delta$, we do Taylor expansion of $U_\alpha(x)$ around $x = p$. We have
\[ U_\alpha(x) = U_\alpha(p) + U_\alpha'(p)(x - p) + \frac{1}{2} U_\alpha''(p)(x - p)^2 + \frac{1}{6} U_\alpha'''(p)(x - p)^3 + R(x; p), \] (174)
where the remainder term enjoys the following representations:
\[ R(x; p) = \frac{1}{6} \int_p^x (x - u)^3 U_\alpha^{(4)}(u) du = \frac{U_\alpha^{(4)}(\xi_x)}{24}(x - p)^4, \quad \xi_x \in [\min\{x, p\}, \max\{x, p\}] \] (175)

The first remainder is called the integral representation of Taylor series remainders, and the second remainder is called the Lagrange remainder.

Since $p \geq \Delta$, we know that
\[ U_\alpha'(p) = \alpha p^{\alpha - 1} + \frac{\alpha(1 - \alpha)}{2n} (\alpha - 1)p^{\alpha - 2}, \] (176)
\[ U_\alpha''(p) = \alpha(\alpha - 1)p^{\alpha - 2} + \frac{\alpha(1 - \alpha)(\alpha - 1)(\alpha - 2)}{2n} p^{\alpha - 3}, \] (177)
\[ U_\alpha^{(3)}(p) = \alpha(\alpha - 1)(\alpha - 2)p^{\alpha - 3} + \frac{\alpha(1 - \alpha)(\alpha - 1)(\alpha - 2)(\alpha - 3)}{2n} p^{\alpha - 4}, \] (178)
\[ U_\alpha^{(4)}(p) = \alpha(\alpha - 1)(\alpha - 2)(\alpha - 3)p^{\alpha - 4} + \frac{\alpha(1 - \alpha)(\alpha - 1)(\alpha - 2)(\alpha - 3)(\alpha - 4)}{2n} p^{\alpha - 5}. \] (179)

Replacing $x$ by random variable $X$ in (174), where $nX \sim \text{Poi}(np), p \geq \Delta$, and taking expectations on both sides, we have
\[ \mathbb{E}U_\alpha(X) = U_\alpha(p) + \frac{1}{2} U_\alpha''(p) \frac{p}{n} + \frac{1}{6} U_\alpha'''(p) \frac{p^2}{n^2} + \mathbb{E}[R(X; p)] \] \[ = p^\alpha + \frac{\alpha(\alpha - 1)(\alpha - 2)(\alpha - 3)}{12n^3} p^{\alpha - 2} - \frac{\alpha(1 - \alpha)^2(2 - \alpha)(3 - \alpha)}{12n^3} p^{\alpha - 3} + \mathbb{E}[R(X; p)] \] (180)
where we have used the fact that if $nX \sim \text{Poi}(np)$, then $\mathbb{E}(X - p)^2 = \frac{p}{n}, \mathbb{E}(X - p)^3 = \frac{p^2}{n^2}$.

Since the representation of $R(x; p)$ involves $U_\alpha^{(4)}(\xi_x)$, it would be helpful to obtain some estimates of $U_\alpha^{(4)}(x)$ over $[0, 1]$.

Denoting $U_\alpha(x) = I_n(x)f(x)$, where $f(x) = x^\alpha + \frac{\alpha(1 - \alpha)}{2n} x^{\alpha - 1}$, we have
\[ U_\alpha^{(4)}(x) = I_n^{(4)} f + 4I_n^{(3)} f^{(1)} + 6I_n^{(2)} f^{(2)} + 4I_n^{(1)} f^{(3)} + I_n f^{(4)}. \] (182)

Hence, it suffices to bound each term in (182) separately.

For $x \in [0, t], U_\alpha(x) \equiv 0$, so we do not need to consider this regime. For $x \in [t, 1], U_\alpha(x) = f(x)$, hence
\[ |U_\alpha^{(4)}(x)| = |f^{(4)}(x)| = \alpha(\alpha - 1)(\alpha - 2)(\alpha - 3)x^{\alpha - 4} + \frac{\alpha(1 - \alpha)(\alpha - 1)(\alpha - 2)(\alpha - 3)(\alpha - 4)}{2n} x^{\alpha - 5}, \] (183)
which implies that for $x \geq 2t$,
\[
\sup_{z \in [x,1]} |U_\alpha^{(4)}(z)| \leq 6x^{\alpha-4} + \frac{12}{n} x^{\alpha-5}.
\] (184)

Finally we consider $x \in (t,2t)$. Denoting $y = x - t$, the derivatives of $I_n(x)$ for $x \in (t,2t)$ are as follows:

\[
I_n'(x) = \frac{630y^4(t-y)^4}{t^9}
\] (185)

\[
I_n''(x) = \frac{2520y^4(t-2y)(t-y)^3}{t^9}
\] (186)

\[
I_n^{(3)}(x) = \frac{2520y^2(t-y)^2(3t^2 - 14ty + 14y^2)}{t^9}
\] (187)

\[
I_n^{(4)}(x) = \frac{15120(t - 2y)(t - y)(t^2 - 7ty + 7y^2)}{t^9}
\] (188)

Considering the fact that $y/t \in [0,1]$, we can maximize $|I_n^{(i)}(x)|$ over $x \in (t,2t)$ for $1 \leq i \leq 4$. With the help of Mathematica \[\text{[98]}, we could show that for $x \in (t,2t)$,

\[
|I_n'(x)| \leq \frac{4}{t}
\] (189)

\[
|I_n''(x)| \leq \frac{20}{t^2}
\] (190)

\[
|I_n^{(3)}(x)| \leq \frac{100}{t^3}
\] (191)

\[
|I_n^{(4)}(x)| \leq \frac{1000}{t^4}.
\] (192)

Plugging these upper bounds in (182), we know for $x \in (t,2t)$

\[
|U_\alpha^{(4)}(x)| \leq \frac{1000}{t^4}t^{\alpha} + \frac{4 \times 100}{t^3}t^{\alpha-1} + 6 \times \frac{20}{t^2}t^{\alpha-2} + 4 \times \frac{4}{t}t^{\alpha-3} + 6t^{\alpha-4} \leq 1558t^{\alpha-4} \leq 1558(x/2)^{\alpha-4} \leq 24928x^{\alpha-4}.
\] (193)

Now we proceed to upper bound $\mathbb{E}[|R(X;p)|], p \geq \Delta$. We consider the following two cases:

1) Case 1: $x \geq p/2$. In this case,
\[
|R(x;p)| = \left| \frac{U_\alpha^{(4)}(\xi_x)(x-p)^4}{24} \right| \leq \sup_{x \in [p/2,1]} |U_\alpha^{(4)}(x)(x-p)^4| \leq \left( \frac{6(p/2)^{\alpha-4} + \frac{12}{n}(p/2)^{\alpha-5}}{24} \right) \frac{(x-p)^4}{24}.
\] (194)

2) Case 2: $0 \leq x < p/2$. In this case, denoting $y = \max\{x,\Delta/4\},$
\[
|R(x;p)| \leq \frac{1}{6} \int_y^p (u-x)^3 |U_\alpha^{(4)}(u)| du
\] (195)

\[
\leq \frac{1}{6} \int_y^p (u-x)^3 24928u^{\alpha-4} du
\] (196)

\[
\leq 8310 \left( \frac{1}{\alpha - 2} \right) \frac{3}{\alpha - 1} (p^{\alpha-2} - y^{\alpha-2}) - \frac{x^3}{\alpha - 3} (p^{\alpha-3} - y^{\alpha-3})
\] (199)

Now we have
\[
\mathbb{E}[|R(X;p)|] = \mathbb{E}[|R(X;p)| \mathbb{I}(X \geq p/2)] + \mathbb{E}[|R(X;p)| \mathbb{I}(X < p/2)] = B_1 + B_2.
\] (204)
For the term $B_1$, we have
\[
B_1 = E[|R(X;p)|1(X \geq p/2)] \leq \left(6(p/2)^{\alpha-4} + \frac{12}{n}(p/2)^{\alpha-5}\right) E[(X-p)^4]/24 \leq \left(\frac{1}{4}(p/2)^{\alpha-4} + \frac{1}{2n}(p/2)^{\alpha-5}\right) \left(\frac{p}{n^3} + \frac{3p^2}{n^2}\right),
\]
where we have used the fact that if $nX \sim \text{Poi}(np)$, then $E(X-p)^4 = (np + 3np^2)/n^4$.

For the term $B_2$, we have
\[
B_2 = E[|R(X;p)|1(X < p/2)] \leq \frac{8310(1+\alpha)}{\alpha(2-\alpha)} p^\alpha P(nX < np/2).
\]
Applying Lemma 21 we have
\[
B_2 \leq \frac{8310(1+\alpha)}{\alpha(2-\alpha)} p^\alpha e^{-np/8} \leq \frac{8310(1+\alpha)}{\alpha(2-\alpha)} p^\alpha n^{-c_1/8}.
\]
Hence, we have
\[
E[R(X;p)] \leq E[|R(X;p)|] \leq \left(\frac{1}{4}(p/2)^{\alpha-4} + \frac{1}{2n}(p/2)^{\alpha-5}\right) \left(\frac{p}{n^3} + \frac{3p^2}{n^2}\right) + \frac{8310(1+\alpha)}{\alpha(2-\alpha)} p^\alpha n^{-c_1/8}.
\]
Plugging this into (181), we have for $p \geq \Delta$,
\[
|\mathbb{E}U_\alpha(X) - p^\alpha| \leq \frac{\alpha(\alpha-1)(\alpha-2)(5-3\alpha)}{12n^2} p^{\alpha-2} + \left(\frac{1}{4}(p/2)^{\alpha-4} + \frac{1}{2n}(p/2)^{\alpha-5}\right) \left(\frac{p}{n^3} + \frac{3p^2}{n^2}\right) + \frac{8310(1+\alpha)}{\alpha(2-\alpha)} p^\alpha n^{-c_1/8}
\]
\[
\leq \frac{17p^{\alpha-2}}{n^2} + \frac{8310(1+\alpha)}{\alpha(2-\alpha)} p^\alpha n^{-c_1/8}
\]
\[
= \frac{17}{\alpha^2(\alpha-2)^2} + \frac{8310(1+\alpha)}{\alpha(2-\alpha)} p^\alpha n^{-c_1/8}.
\]
For the upper bound on the variance $\text{Var}(U_\alpha(X))$, denoting $f(p) = p^\alpha + \frac{(1-\alpha)}{2n} p^{\alpha-1}$, for $p \geq \Delta$, we have
\[
\text{Var}(U_\alpha(X)) = E[U_\alpha^2(X) - (E[U_\alpha(X)])^2
\]
\[
= E[U_\alpha^2(X) - f^2(p) + f^2(p) - (E[U_\alpha(X)])^2
\]
\[
\leq [E[U_\alpha^2(X) - f^2(p)] + |f^2(p) - (E[U_\alpha(X)] - f(p) + f(p))^2]
\]
\[
= [E[U_\alpha^2(X) - f^2(p)] + |(E[U_\alpha(X)] - f(p))^2 + 2f(p)(E[U_\alpha(X)] - f(p))|
\]
\[
\leq [E[U_\alpha^2(X) - f^2(p)] + |E[U_\alpha(X)] - f(p)]^2 + 2f(p)|E[U_\alpha(X)] - f(p)|.
\]
Hence, it suffices to obtain bounds on $[E[U_\alpha^2(X) - f^2(p)]$ and $[E[U_\alpha(X)] - f(p)]$. Denoting $r(x) = U_\alpha(x)$, we know that $r(x) \in C^4[0,1]$, and it follows from Taylor's formula and the integral representation of the remainder term that
\[
r(X) = f^2(p) + r^2(p)(X-p) + R_1(X;p),
\]
\[
R_1(X;p) = \int_p^X (X-u)r''(u)du = \frac{1}{2}r''(\eta_X)(X-p)^2, \quad \eta_X \in [\min\{X,p\}, \max\{X,p\}].
\]
Similarly, we have
\[
U_\alpha(X) = f(p) + f'(p)(X-p) + R_2(X;p),
\]
\[
R_2(X;p) = \int_p^X (X-u)U_\alpha''(u)du = \frac{1}{2}U_\alpha''(\nu_X)(X-p)^2, \quad \nu_X \in [\min\{X,p\}, \max\{X,p\}].
\]
Taking expectation on both sides with respect to $X$, where $nX \sim \text{Poi}(np)$, $p \geq \Delta$, we have
\[
|E[U_\alpha^2(X) - f^2(p)| = |E[R_1(X;p)].
\]
Similarly, we have
\[
|E[U_\alpha(X) - f(p)] = |E[R_2(X;p)].
\]
As we did for function $U_\alpha(x)$, now we give some upper estimates for $|r''(x)|$ over $[0,1]$. Over regime $[0,t]$, $r(x) \equiv 0$, so we ignore this regime. Over regime $[2t,1]$, since $U_\alpha(x) = f(x)$, $f(x) = x^\alpha + \frac{(1-\alpha)}{2n} x^{\alpha-1}$, we have
\[
r'(x) = 2ff',
\]
\[
r''(x) = 2(f')^2 + 2ff''.
\]
Hence, for $x \geq 2t$, 

$$
\sup_{z \in [x, 1]} |r''(z)| \leq 4x^{2\alpha - 2}. 
$$

(225)

$$
\sup_{z \in [x, 1]} |U''_n(z)| \leq x^{\alpha - 2}. 
$$

(226)

Over regime $[t, 2t]$, we have 

$$
r'(x) = 2f f' I_n^2 + 2I_n I'_n f^2 
$$

(227)

$$
r''(x) = 2 \left( (f')^2 I_n^2 + 2f f' I_n I'_n + (I_n^2 f^2 + I_n I'_n f^2 + 2f f' I_n I'_n) \right). 
$$

(228)

Hence, we have for $x \in [t, 2t]$, 

$$
|r''(x)| \leq 2 \left( t^{2\alpha - 2} + t^{2\alpha - 2} + 2t^{2\alpha - 1} - \frac{4}{t} \right) 
$$

(229)

$$
\leq 108t^{2\alpha - 2} 
$$

(230)

$$
\leq 108(x/2)^{2\alpha - 2} 
$$

(231)

$$
= 432x^{2\alpha - 2} 
$$

(232)

Also, over regime $[t, 2t]$, 

$$
U''_n(x) = I_n'' f + I_n f'' + 2I_n f', 
$$

(233)

hence for $x \in [t, 2t]$, 

$$
|U''_n(x)| \leq \frac{20}{t^2} t^{\alpha} + t^{\alpha - 2} + 2\frac{4}{t} t^{\alpha - 1} \leq 30t^{\alpha - 2} \leq 30(x/2)^{\alpha - 2} \leq 120x^{\alpha - 2}. 
$$

(234)

Now we are in the position to bound $|\mathbb{E}R_1(X;p)|$ and $|\mathbb{E}R_2(X;p)|$.

We have 

$$
|\mathbb{E}R_1(X;p)| \leq \mathbb{E}|R_1(X;p)| 
$$

(235)

$$
= \mathbb{E}[|R_1(X;p)\mathbb{I}(X \geq p/2)|] + \mathbb{E}[R_1(X;p)\mathbb{I}(X < p/2)]] 
$$

(236)

$$
\leq \mathbb{E} \left[ \frac{1}{2} (4p/2)^{2\alpha - 2}(X - p)^2 \right] + \mathbb{E}[R_1(X;p)\mathbb{I}(X < p/2)] 
$$

(237)

$$
= \frac{8p^{2\alpha - 1}}{n} + \sup_{x \leq p/2} |R_1(x;p)|\mathbb{P}(nX < np/2)) 
$$

(238)

$$
\leq \frac{8p^{2\alpha - 1}}{n} + \sup_{x \leq p/2} |R_1(x;p)|n^{-\epsilon_1 / 8}, 
$$

(239)

where in the last step we have applied Lemma [21].

Regarding $\sup_{x \leq p/2} |R_1(x;p)|$, for any $x \leq p/2$, denoting $y = \max\{x, \Delta / 4\}$, we have 

$$
R_1(x;p) = \int_x^p (u - x)r''(u)du 
$$

(240)

$$
\leq \int_y^p (u - x)432u^{2\alpha - 2}du 
$$

(241)

$$
\leq 432 \int_y^p u^{2\alpha - 1}du 
$$

(242)

$$
= \frac{432}{2\alpha} (p^{2\alpha} - y^{2\alpha}) 
$$

(243)

$$
\leq \frac{432}{2\alpha} p^{2\alpha} 
$$

(244)

$$
\leq \frac{216}{\alpha} p^{2\alpha}. 
$$

(245)

Hence, we have 

$$
|\mathbb{E}R_1(X;p)| \leq \frac{8p^{2\alpha - 1}}{n} + \frac{216}{\alpha} p^{2\alpha} n^{-\epsilon_1 / 8}. 
$$

(246)

Analogously, we obtain the following bound for $|\mathbb{E}R_2(X;p)|$: 

$$
|\mathbb{E}R_2(X;p)| \leq \frac{2p^{2\alpha - 1}}{n} + \frac{120}{\alpha} p^{2\alpha} n^{-\epsilon_1 / 8}. 
$$

(247)
Plugging these estimates of $|R_1(X; p)|$ and $|R_2(X; p)|$ into (216), we have for $p \geq \Delta, c_1 \ln n \geq 1$,

$$\text{Var}(U_{\alpha}(X)) \leq \frac{8p^{2\alpha-1}}{n} + \frac{216}{\alpha} p^{2\alpha} n^{-c_1/8} + \left(2 \frac{p^{\alpha-1}}{n} + \frac{120}{\alpha} p^{\alpha} n^{-c_1/8}\right)^2 + 2f(p) \left(2 \frac{p^{\alpha-1}}{n} + \frac{120}{\alpha} p^{\alpha} n^{-c_1/8}\right).$$

(248)

We need to distinguish two cases: $0 < \alpha \leq 1/2$, and $1/2 < \alpha < 1$.

1) $0 < \alpha \leq 1/2$: in this case, we have

$$\text{Var}(U_{\alpha}(X)) \leq \frac{8p^{2\alpha-1}}{n} + \frac{216}{\alpha} p^{2\alpha} n^{-c_1/8} + \left(2 \frac{p^{\alpha-1}}{n} + \frac{120}{\alpha} p^{\alpha} n^{-c_1/8}\right)^2 + 2f(p) \left(2 \frac{p^{\alpha-1}}{n} + \frac{120}{\alpha} p^{\alpha} n^{-c_1/8}\right)$$

$$\leq \frac{8}{n^{2\alpha}(c_1 \ln n)^{1-2\alpha}} + \frac{216}{\alpha} p^{2\alpha} n^{-c_1/8} + 2 \left(\frac{4p^{2\alpha-2}}{n^2} + \frac{14400}{\alpha^2} p^{2\alpha} n^{-c_1/4}\right)$$

$$+ 2\alpha \left(1 + \frac{1}{8\alpha c_1 \ln n}\right) \left(2 \frac{p^{\alpha-1}}{n} + \frac{120}{\alpha} p^{\alpha} n^{-c_1/8}\right)$$

$$\leq \frac{16}{n^{2\alpha}(c_1 \ln n)^{1-2\alpha}} + \frac{8}{n^{2\alpha}(c_1 \ln n)^{1-2\alpha}} + \frac{576}{\alpha^2} p^{2\alpha} n^{-c_1/8} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4}$$

$$\leq \frac{24}{n^{2\alpha}(c_1 \ln n)^{1-2\alpha}} + \frac{576}{\alpha^2} p^{2\alpha} n^{-c_1/8} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4}.$$

(249)

(250)

(251)

(252)

(253)

2) $1/2 < \alpha < 1$: in this case, we have

$$\text{Var}(U_{\alpha}(X)) \leq \frac{8p^{2\alpha-1}}{n} + \frac{216}{\alpha} p^{2\alpha} n^{-c_1/8} + \left(2 \frac{p^{\alpha-1}}{n} + \frac{120}{\alpha} p^{\alpha} n^{-c_1/8}\right)^2 + 2f(p) \left(2 \frac{p^{\alpha-1}}{n} + \frac{120}{\alpha} p^{\alpha} n^{-c_1/8}\right)$$

$$\leq \frac{8p^{2\alpha-1}}{n} + \frac{216}{\alpha} p^{2\alpha} n^{-c_1/8} + \frac{8p^{2\alpha-2}}{n^2} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} + 3\alpha \left(2 \frac{p^{\alpha-1}}{n} + \frac{120}{\alpha} p^{\alpha} n^{-c_1/8}\right)$$

$$\leq \frac{14p^{2\alpha-1}}{n} + \frac{576}{\alpha^2} p^{2\alpha} n^{-c_1/8} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} + \frac{8}{n^{2\alpha}(c_1 \ln n)^{1-2\alpha-2\alpha}}.$$

(254)

(255)

(256)

For $1 < \alpha < 3/2$, following the same procedures, we obtain some upper bounds on $|r''(x)|$ and $|U''_{\alpha}(x)|$. Over regime $[0, t]$, we have $r'(x) = U''_{\alpha}(x) = 0$, so we have $r''(x) = U''_{\alpha}(x) = 0$. For $x$ in $[2t, 1]$, we have $r''(x) = |U''_{\alpha}(x)| = |I''_{\alpha} f + 2I''_{\alpha} f' + I''_{\alpha} f''| \leq \frac{20}{t^2} x^\alpha + 2\alpha x^{\alpha-2} \leq 120 x^{\alpha-2}$.

(257)

(258)

Over regime $[t, 2t]$, we have

$$|r''(x)| = 2 |f'' I_n^2 + f f' I_n^2 + 2 f f' I_n I_n' + (I_n')^2 f^2 + I_n I_n' f^2 + 2 f f' I_n I_n'|$$

$$\leq 2 \left(\alpha x^{2\alpha-2} + \alpha x^{2\alpha-2} + 20 x^{2\alpha-1} \cdot \frac{4}{t^2} x^\alpha + 20 x^{2\alpha-1} \cdot \frac{4}{t^2} x^{2\alpha} + 2 \alpha x^{2\alpha-2} \cdot \frac{4}{t^2}\right)$$

$$\leq 400 x^{2\alpha-2},$$

(259)

(260)

(261)

and

$$|U''_{\alpha}(x)| = |I''_{\alpha} f + 2I''_{\alpha} f' + I''_{\alpha} f''| \leq \frac{20}{t^2} x^\alpha + 2\alpha x^{\alpha-2} \cdot \frac{4}{t^2} x^{2\alpha} + \alpha x^{\alpha-2} \leq 120 x^{\alpha-2},$$

(262)

where we have used the inequality

$$|I_n(x)| \leq 1, \quad |I_n'(x)| \leq \frac{4}{t^2}, \quad |I_n''(x)| \leq \frac{400}{t^4}, \quad \forall x \in [t, 2t].$$

(263)

Noting that we have obtained a norm bound for $|r''(x)|$ over all regimes expressed as

$$|r''(x)| \leq 400 x^{2\alpha-2} \leq 400, \quad \forall x \in [0, 1],$$

(264)

we have the upper bound

$$|\mathbb{E} R_1(X; p)| \leq \mathbb{E} |R_1(X; p)| \leq \frac{1}{2} \mathbb{E} |f''(\eta X)(X - p)|^2 \leq 200 \mathbb{E} |(X - p)|^2 = \frac{200p^2}{n}.$$

(265)

For the upper bound of $|\mathbb{E} R_2(X; p)|$, we first consider the upper bound of $|R_2(x; p)|$ when $x \leq p/2$. Denoting $y = \max\{x, \Delta/4\}$, we have

$$R_2(x; p) = \int_x^p (u - x) U''_{\alpha}(u) du \leq \int_y^p (u - x) 120 u^{\alpha-2} du \leq \int_y^p 120 u^{\alpha-1} du \leq \frac{120}{\alpha} p^\alpha,$$

(266)
then

$$|\mathbb{E}R_2(X; p)| \leq \mathbb{E}|R_2(X; p)|$$

(267)

$$= \mathbb{E}|R_2(X; p)| \mathbb{1}(X \geq p/2) + \mathbb{E}|R_2(X; p)| \mathbb{1}(X < p/2)|$$

(268)

$$\leq \mathbb{E} \left\{ \frac{1}{2} \cdot 2 \left( \frac{p}{2} \right)^{n-2} (X - p)^2 \right\} + \sup_{x < p/2} |R_2(x; p)| \cdot \mathbb{P} \left( nX < \frac{nD}{2} \right)$$

(269)

$$\leq \frac{2p^{\alpha-1}}{n} + \frac{120}{\alpha} p^\alpha n^{-c_1/8},$$

(270)

where in the last step we have applied Lemma 21. Plugging in the upper bound of $|\mathbb{E}R_1(X; p)|$ and $|\mathbb{E}R_2(X; p)|$ together, we know when $1 < \alpha < 3/2$,

$$\text{Var}(U_\alpha(X)) \leq \frac{200p}{n} + \left( \frac{2p^{\alpha-1}}{n} + \frac{120}{\alpha} p^\alpha n^{-c_1/8} \right)^2 + f(p) \left( \frac{2p^{\alpha-1}}{n} + \frac{120}{\alpha} p^\alpha n^{-c_1/8} \right)$$

(271)

$$\leq \frac{200p}{n} + 2 \left( \frac{4p^{2(\alpha-1)}}{n^2} + \frac{14400}{\alpha^2} p^{2\alpha n^{-c_1/4}} \right) + p^\alpha \left( \frac{2p^{\alpha-1}}{n} + \frac{120}{\alpha} p^\alpha n^{-c_1/8} \right)$$

(272)

$$\leq \frac{202p}{n} + \frac{8}{n^2} + \frac{28800}{\alpha^2} p^{2\alpha n^{-c_1/4}} + \frac{120}{\alpha} p^\alpha n^{-c_1/8}.$$ 

(273)

### D. Proof of Lemma 3

We have

$$U_H(x) = I_n(x) \left( -x \ln x + \frac{1}{2n} \right) = I_n(x) f(x),$$

(274)

where $f(x) = -x \ln x + 1/(2n)$.

For $p \geq \Delta$, we do Taylor expansion of $U_H(x)$ around $x = p$. We have

$$U_H(x) = U_H(p) + U'_H(p)(x - p) + \frac{1}{2} U''_H(p)(x - p)^2 + \frac{1}{6} U'''_H(p)(x - p)^3 + R(x; p),$$

(275)

where the remainder term enjoys the following representations:

$$R(x; p) = \frac{1}{6} \int_p^x (x - u)^3 U''_H(u) du = \frac{U''_H(\xi_x)}{24} (x - p)^4, \quad \xi_x \in [\min\{x, p\}, \max\{x, p\}]$$

(276)

The first remainder is called the integral representation of Taylor series remainders, and the second remainder is called the Lagrange remainder.

Since $p \geq \Delta$, we know that

$$U'_H(p) = -\ln p - 1$$

(277)

$$U''_H(p) = -1/p$$

(278)

$$U'''_H(p) = 1/p^2$$

(279)

$$U''''_H(p) = -2/p^3$$

(280)

Replacing $x$ by random variable $X$ in (275), where $nX \sim \text{Poi}(np), p \geq \Delta$, and taking expectations on both sides, we have

$$\mathbb{E}U_H(X) = U_H(p) + \frac{1}{2} U''_H(p) \frac{p}{n} + \frac{1}{6} U'''_H(p) \frac{p}{n^2} + \mathbb{E}[R(X; p)]$$

(281)

$$= -x \ln x + \frac{1}{6pn^2} + \mathbb{E}[R(X; p)]$$

(282)

where we have used the fact that if $nX \sim \text{Poi}(np)$, then $\mathbb{E}(X - p)^2 = \frac{p}{n}, \mathbb{E}(X - p)^3 = \frac{p}{n^2}$.

Since the representation of $R(x; p)$ involves $U''''_H(\xi_x)$, it would be helpful to obtain some estimates of $U''''_H(x)$ over $[0, 1]$. We have

$$U''''_H(x) = I''''_n f + 4T''_n f^{(1)} + 6I''_n f^{(2)} + 4I'_{(n)} f^{(3)} + I_n f^{(4)}.$$ 

(283)

Hence, it suffices to bound each term in (283) separately.

For $x \in [0, t]$, $U_H(x) \equiv 0$, so we do not need to consider this regime. For $x \in [2t, 1]$, $U_H(x) = f(x)$, hence

$$|U''''_H(x)| = |f^{(4)}(x)| = 2/x^3,$$

(284)

which implies that for $x \geq 2t$,

$$\sup_{x \in [x, 1]} |U''''_H(x)| \leq 2/x^3.$$ 

(285)
Finally we consider \( x \in (t, 2t) \). Denoting \( y = x - t \), the derivatives of \( I_n(x) \) for \( x \in (t, 2t) \) are as follows:

\[
I'_n(x) = \frac{630y^4(t - y)^4}{t^9},
\]

(286)

\[
I''_n(x) = \frac{2520y^3(t - 2y)(t - y)^3}{t^9},
\]

(287)

\[
I^{(3)}_n(x) = \frac{2520y^2(t - y)^2(3t^2 - 14ty + 14y^2)}{t^9},
\]

(288)

\[
I^{(4)}_n(x) = \frac{15120y(t - 2y)(t - y)(t^2 - 7ty + 7y^2)}{t^9}.
\]

(289)

Considering the fact that \( y/t \in [0, 1] \), we can maximize \( |I^{(i)}_n(x)| \) over \( x \in (t, 2t) \) for \( 1 \leq i \leq 4 \). With the help of Mathematica \([98]\), we could show that for \( x \in (t, 2t) \),

\[
|I'_n(x)| \leq \frac{4}{t},
\]

(290)

\[
|I''_n(x)| \leq \frac{20}{t^2},
\]

(291)

\[
|I^{(3)}_n(x)| \leq \frac{100}{t^3},
\]

(292)

\[
|I^{(4)}_n(x)| \leq \frac{1000}{t^4}.
\]

(293)

Plugging these upper bounds in (283), we know for \( x \in (t, 2t) \)

\[
|U^{(4)}_H(x)| \leq \frac{1000}{t^4} t \ln(1/t) + \frac{4 \times 100}{t^3} \ln(1/t) + 6 \times \frac{20}{t^2} 1/t + 4 \times \frac{4}{t} 1/t^2 + 2/t^3 \leq 1538 \ln(1/t) \frac{t^3}{t^3} \leq 1538 \ln(2/x) \frac{x^3}{x^3} \leq 12304 \frac{1 + \ln(1/x)}{x^3}.
\]

(294)

Now we proceed to upper bound \( |E[R(X; p)]|, p \geq \Delta \). We consider the following two cases:

1) Case 1: \( x \geq p/2 \). In this case,

\[
|R(x; p)| = \left| U^{(4)}_H(x) \right| = \frac{1}{24} (x - p)^4 \leq \sup_{x \in [p/2, 1]} \left| U^{(4)}_H(x) \right| \frac{2}{(p/2)^4} \frac{2}{24} = \frac{2(x - p)^4}{3p^4}.
\]

(295)

2) Case 2: \( 0 \leq x < p/2 \). In this case, denoting \( y = \max\{x, \Delta/4\} \),

\[
|R(x; p)| \leq \frac{1}{6} \int_y^p (u - x)^4 |U^{(4)}_H(u)| du
\]

(296)

\[
\leq \frac{1}{6} \int_y^p (u - x)^3 12304 \frac{1 + \ln(1/u)}{u^3} du
\]

(297)

\[
\leq 2051 \int_y^p (u - x)^3 (1 + \ln(1/u)) \frac{u^3}{u^3} du
\]

(298)

\[
= 2051 \int_y^p (u^3 - 3ux^2u + 3x^2u - x^3)(1 + \ln(1/u)) du
\]

(299)

\[
\leq 2051 \int_y^p (u^3 + 3x^2u)(1 + \ln(1/u)) du
\]

(300)

\[
= 2051 \int_y^p \left(1 + \frac{3x^2}{u^2}\right)(1 + \ln(1/u)) du
\]

(301)

\[
\leq 8204 \int_y^p (1 + \ln(1/u)) du
\]

(302)

\[
\leq 8204 p(\ln(1/p) + 2).
\]

(303)

Now we have

\[
E[|R(X; p)|] = E[|R(X; p)|1(X \geq p/2)] + E[|R(X; p)|1(X < p/2)] = B_1 + B_2.
\]

(304)

For the term \( B_1 \), we have

\[
B_1 = E[|R(X; p)|1(X \geq p/2)] \leq E \left[ \frac{2(X - p)^4}{3p^3} \right] = \frac{2}{3p^2 n^3} + \frac{2}{pn^2}.
\]

(305)

where we have used the fact that if \( nX \sim \text{Poi}(np) \), then \( E(X - p)^4 = (np + 3n^2 p^2)/n^4 \).
For the term $B_2$, we have

$$B_2 = \mathbb{E}[|R(X;p)|^2] \leq 8024p (\ln(1/p) + 2) \mathbb{P}(nX < np/2).$$

(306)

Applying Lemma 21, we have

$$B_2 \leq 8024 \left( p \ln(1/p) + 2p \right) e^{-np/8} = 8024 \left( p \ln(1/p) + 2p \right) n^{-c_1/8}. $$

(307)

Hence, we have

$$\mathbb{E}[R(X;p)] \leq \mathbb{E}[|R(X;p)|] \leq \frac{2}{3p^n} + \frac{2}{p n^2} + 8024 \left( p \ln(1/p) + 2p \right) n^{-c_1/8}. $$

(308)

Plugging this into (282), we have for $p \geq \Delta$,

$$|\mathbb{E}U_H(X) + p \ln(p)| \leq \frac{1}{6p^2} + \frac{2}{3p^n} + \frac{2}{p n^2} + 8024 \left( p \ln(1/p) + 2p \right) n^{-c_1/8} $$

(309)

$$\leq \frac{3}{p n^2} + \frac{2}{3p^n} + 8024 \left( p \ln(1/p) + 2p \right) n^{-c_1/8} $$

(310)

$$\leq \frac{3}{c_1 n \ln(n)} + \frac{2}{3(c_1 \ln(n))^2} + 8024 \left( p \ln(1/p) + 2p \right) n^{-c_1/8}. $$

(311)

For the upper bound on the variance $\text{Var}(U_H(X))$, recalling that $f(p) = -x \ln x + \frac{1}{p^n}$, for $p \geq \Delta$, we have

$$\text{Var}(U_H(X)) = \mathbb{E}U_H^2(X) - (\mathbb{E}U_H(X))^2$$

(312)

$$= \mathbb{E}U_H^2(X) - f^2(p) + f^2(p) - (\mathbb{E}U_H(X))^2 $$

(313)

$$\leq |\mathbb{E}U_H^2(X) - f^2(p)| + |f^2(p) - (\mathbb{E}U_H(X) - f(p))^2| $$

(314)

$$\leq |\mathbb{E}U_H^2(X) - f^2(p)| + |\mathbb{E}U_H^2(X) - f(p)|^2 + 2f(p)\mathbb{E}(U_H(X) - f(p)) $$

(315)

$$\leq |\mathbb{E}U_H^2(X) - f^2(p)| + |\mathbb{E}U_H^2(X) - f(p)|^2 + 2f(p)|\mathbb{E}U_H(X) - f(p)|. $$

(316)

Hence, it suffices to obtain bounds on $|\mathbb{E}U_H^2(X) - f^2(p)|$ and $\mathbb{E}U_H^2(X) - f(p)$]. Denoting $r(x) = U_H^2(x)$, we know that $r(x) \in C^4[0,1]$, and it follows from Taylor’s formula and the integral representation of the remainder term that

$$r(X) = f^2(p) + r'(p)(X - p) + R_1(X;p),$$

(317)

$$R_1(X;p) = \int_p^X (X - u) r''(u) du = \frac{1}{2} r''(\eta_X)(X - p)^2, \quad \eta_X \in \min\{X,p\}, \max\{X,p\}. $$

(318)

Similarly, we have

$$U_H(X) = f(p) + f'(p)(X - p) + R_2(X;p),$$

(319)

$$R_2(X;p) = \int_p^X (X - u) U_H''(u) du = \frac{1}{2} U_H''(\nu_X)(X - p)^2, \quad \nu_X \in \min\{X,p\}, \max\{X,p\}. $$

(320)

Taking expectation on both sides with respect to $X$, where $nX \sim \text{Poi}(np), p \geq \Delta$, we have

$$|\mathbb{E}U_H^2(X) - f^2(p)| = |\mathbb{E}R_1(X;p)|. $$

(321)

Similarly, we have

$$|\mathbb{E}U_H(X) - f(p)| = |\mathbb{E}R_2(X;p)|. $$

(322)

As we did for function $U_H(x)$, now we give some upper estimates for $|r''(x)|$ over $[0,1]$. Over regime $[0,t]$, $r(x) \equiv 0$, so we ignore this regime. Over regime $[2t,1]$, since $U_H(x) = f(x)$, we have

$$r'(x) = 2f f',$$

(323)

$$r''(x) = 2(f')^2 + 2f f''.$$ 

(324)

Hence, for $x \geq 2t$,

$$\sup_{x \in [x,1]} |r''(x)| \leq 4(\ln x)^2. $$

(325)

$$\sup_{x \in [x,1]} |U_H''(x)| \leq 1/x. $$

(326)

Over regime $[t,2t]$, we have

$$r'(x) = 2f f' I_n^2 + 2I_n f' f,$$

(327)

$$r''(x) = 2 ((f')^2 I_n^2 + 2f f' f' I_n^2 + (I_n')^2 f^2 + I_n f_n f^2 + 2f f' I_n I_n') .$$

(328)
Hence, we have for $x \in [t, 2t]$,
\[
|r''(x)| \leq 2 \left( 4t^2 + (\ln t)^2 + 2(t \ln(1/t))(\ln(1/t)) \frac{4}{t} + (4/t^2) t^2 (\ln t)^2 + 2(t \ln(1/t)) (\ln(1/t)) \frac{4}{t} \right)
\]
\[
\leq 108(\ln t)^2
\]
\[
\leq 108(\ln(x/2))^2
\]
\[
= 216((\ln x)^2 + 1),
\]
where we have used the fact that $|\ln 2| \approx 0.69 < 1$. Also, over regime $[t, 2t],$
\[
U''_H(x) = I''_n f + I_n f'' + 2I'_n f',
\]
hence for $x \in [t, 2t],$
\[
|U''_H(x)| \leq \frac{20}{t^2} (t \ln(1/t)) + \frac{1}{t^4} \ln(1/t) \leq \frac{30}{x^2} \ln(2/x) \leq \frac{60}{x} (\ln(1/x) + 1).
\]
Now we are in the position to bound $|\mathbb{E}R_1(X; p)|$ and $|\mathbb{E}R_2(X; p)|$. We have
\[
|\mathbb{E}R_1(X; p)| \leq \mathbb{E}|R_1(X; p)|
\]
\[
= \mathbb{E}[|R_1(X; p)| (X \geq p/2)] + \mathbb{E}[R_1(X; p) | (X < p/2)]
\]
\[
\leq \mathbb{E} \left( \frac{1}{2} x \ln(2/p) \right)^2 + \mathbb{E}[R_1(X; p) | (X < p/2)]
\]
\[
= 2p(\ln p - \ln 2)^2/n + \sup_{x \leq p/2} |R_1(x; p)| \mathbb{P}(nX < np/2)
\]
\[
= 2p(\ln p - \ln 2)^2/n + \sup_{x \leq p/2} |R_1(x; p)| n^{-c_1/8},
\]
where in the last step we have applied Lemma 21.
Regarding $\sup_{x \leq p/2} |R_1(x; p)|$, for any $x \leq p/2$, denoting $y = \max\{x, \Delta/4\}$, we have
\[
R_1(x; p) = \int_{x}^{p} (u - x) r''(u) du
\]
\[
\leq \int_{y}^{p} (u - x) 216((\ln u)^2 + 1) du
\]
\[
\leq 216 \int_{y}^{p} u ((\ln u)^2 + 1) du
\]
\[
\leq 54p^2 (2(\ln p)^2 - 2 \ln p + 3).
\]
Hence, we have
\[
|\mathbb{E}R_1(X; p)| \leq 2p(\ln p - \ln 2)^2/n + 54p^2 |2(\ln p)^2 - 2 \ln p + 1| n^{-c_1/8}.
\]
Analogously, we obtain the following bound for $|\mathbb{E}R_2(X; p)|$:
\[
|\mathbb{E}R_2(X; p)| \leq \frac{1}{n} + 60 (p \ln(1/p) + 2p) n^{-c_1/8}.
\]
Plugging these estimates of $|\mathbb{E}R_1(X; p)|$ and $|\mathbb{E}R_2(X; p)|$ into (316), we have for $p \geq \Delta$,
\[
\text{Var}(U_H(X)) \leq 2p(\ln p - \ln 2)^2/n + 54p^2 |2(\ln p)^2 - 2 \ln p + 3| n^{-c_1/8} + \left( \frac{1}{n} + 60 (p \ln(1/p) + 2p) n^{-c_1/8} \right)^2
\]
\[
+ 2 \left( p \ln(1/p) + \frac{1}{2n} \right) \left( \frac{1}{n} + 60 (p \ln(1/p) + 2p) n^{-c_1/8} \right).
\]
E. Proof of Lemma 4
We first bound the bias term. It follows from differentiating the moment generating function of the Poisson distribution that if $X \sim \text{Poi}(\lambda)$, then
\[
\mathbb{E}X(X - 1) \ldots (X - r + 1) = \lambda^r,
\]
for any $r$ positive integer.
Then, we know that for \( nX \sim \text{Poi}(np) \),

\[
\mathbb{E}S_{K,\alpha}(X) = \sum_{k=1}^{K} g_{k,\alpha}(4\Delta)^{-k+\alpha} p^k.
\] (347)

Applying Lemma 19, we know that for all \( p \leq 4\Delta \),

\[
|\mathbb{E}S_{K,\alpha}(X) - p^\alpha| \leq \frac{e_3}{(n \ln n)^{\alpha}}.
\] (348)

Now we bound the second moment of \( S_{K,\alpha}(X) \). Denote

\[
E_{k,n}(x) = \prod_{r=0}^{k-1} (x - r/n),
\] (349)

we have

\[
\mathbb{E}S_{K,\alpha}^2(X) \leq \left( \sum_{k=1}^{K} |g_{k,\alpha}|(4\Delta)^{-k+\alpha} \left( \mathbb{E}E_{k,n}^2(X) \right)^{1/2} \right)^2
\] (350)

\[
\leq 2^{6K} \left( \sum_{k=1}^{K} (4\Delta)^{-k+\alpha} \left( \mathbb{E}E_{k,n}^2(X) \right)^{1/2} \right)^2
\] (351)

Here we have used Lemma 17.

Since \( K \leq 4n\Delta \), applying Lemma 22,

\[
\mathbb{E}E_{k,n}^2(X) = \frac{1}{n^{2k}} \mathbb{E} \prod_{r=0}^{k-1} (nX - r)^2
\] (352)

\[
\leq \frac{1}{n^{2k}} \mathbb{E} \prod_{r=0}^{k-1} (nX)^2
\] (353)

\[
= \frac{1}{n^{2k}} \mathbb{E}(nX)^{2k}
\] (354)

\[
\leq \frac{1}{n^{2k}} (8c_1 \ln n)^{2k}
\] (355)

\[
= \left( \frac{8c_1 \ln n}{n} \right)^{2k},
\] (356)

we know

\[
\mathbb{E}S_{K,\alpha}^2(X) \leq 2^{6K} \left( \sum_{k=1}^{K} (4\Delta)^{-k+\alpha} \left( \frac{8c_1 \ln n}{n} \right)^{k} \right)^2
\] (357)

\[
\leq 2^{6K} 2^{2K} \left( \sum_{k=1}^{K} (4\Delta)^{-k+\alpha} (4\Delta)^{k} \right)^2
\] (358)

\[
\leq 2^{6K} 2^{2K} \left( \sum_{k=1}^{K} (4\Delta)^{\alpha} \right)^2
\] (359)

\[
= 2^{8K} K^2 (4\Delta)^{2\alpha}
\] (360)

\[
\leq n^{8c_2 \ln 2} (c_2 \ln 2)^{2}(4c_1 \ln n)^{2\alpha}
\] (361)

\[
\leq n^{8c_2 \ln 2} (4c_1 \ln n)^{2+2\alpha}/n^{2\alpha}
\] (362)

The proof for the \( S_{K,H}(x) \) case is essentially the same as that for \( S_{K,\alpha}(x) \) via replacing \( \alpha \) by 1 and applying Lemma 20 rather than Lemma 19.
F. Proof of Lemma 5

We first bound the bias term. It follows from the property of Poisson distribution that

\[ \mathbb{E} S_{K,\alpha}(X) = \sum_{k=1}^{K} g_{k,\alpha}(4\Delta)^{-k+\alpha} p^k. \]  

(363)

Since \( p/4\Delta \leq 1/K^2 \), it follows from a variation of the pointwise bound in Lemma 17 that

\[ |\mathbb{E} S_{K,\alpha}(X)| \leq D_1 (4\Delta)^\alpha \cdot \frac{p}{4\Delta} = D_1 \left( \frac{4c_1}{c_2^2 n \ln n} \right)^{\alpha-1} p, \]

(364)

which completes the proof of the first part of Lemma 5. For the variance, denote

\[ E_{k,n}(x) = \prod_{r=0}^{k-1} \left( x - \frac{r}{n} \right), \]

(365)

we have

\[ \mathbb{E} E_{k,n}^2(X) = \frac{1}{n^{2k}} \prod_{r=0}^{k-1} (nX - r)^2 \leq \frac{1}{n^{2k}} \mathbb{E}[nX]^{2k} \]

(366)

\[ \leq \frac{1}{n^{2k}} \sum_{i=1}^{2k} \binom{2k}{i} (np)^i \]

(367)

\[ \leq \frac{1}{n^{2k}} \sum_{i=1}^{2k} \binom{2k}{i} (2k-1)^{2k-i} (np)^i \]

(368)

\[ \leq \frac{1}{n^{2k}} \sum_{i=1}^{2k} \binom{2k}{i} (2k)^{2k-i} np \]

(369)

\[ \leq \frac{(2k+1)^{2k}}{n^{2k-1}}, \]

(370)

where \( \binom{k}{i} \) is the Stirling numbers of the second kind, and we have used the inequality

\[ \binom{k}{i} \leq \binom{k}{i} i^{k-i}. \]

(371)

Hence, we can bound the second moment of \( S_{K,\alpha}(X) \) as

\[ \mathbb{E} S_{K,\alpha}^2(X) \leq \left( \sum_{k=1}^{K} |g_{k,\alpha}|(4\Delta)^{-k+\alpha} \left( \mathbb{E} E_{k,n}^2(X) \right)^{1/2} \right)^2 \]

(372)

\[ \leq 26K \left( \sum_{k=1}^{K} (4\Delta)^{-k+\alpha} \left( \frac{2k+1}{n^{k-2}} \right)^{2k} \right) \]

(373)

\[ \leq 2^{10K} \left( \sum_{k=1}^{K} \left( \frac{4c_1 \ln n}{n} \right)^{-k+\alpha} K^k \sqrt{p} \right)^2 \]

(374)

\[ \leq \frac{2^{10K} (4c_1 \ln n)^{2\alpha}}{n^{2\alpha-1}} \left( \sum_{k=1}^{K} \frac{c_2}{4c_1} \right)^k \]

(375)

\[ \leq 2^{10c_2 \ln 2} \left( \frac{4c_1 \ln n}{n^{2\alpha-1}} \right)^{2\alpha} K^2 \frac{2^p}{n^{2\alpha-1}} \]

(376)

\[ \leq 2^{10c_2 \ln 2} \left( \frac{4c_1 \ln n}{n^{2\alpha-1}} \right)^{2\alpha+2} p \]

(377)

given \( c_2 < 4c_1 \), where we have used Lemma 17.
G. Proof of Lemma 6

The bias bound follows from (363) and a variation of the norm bound in Lemma 17, and the variance bound is given by Lemma 4.

H. Proof of Lemma 7

We apply Lemma 23 and Lemma 24 to calculate the bias and variance of $\xi$.

1) Case 1: $p \leq \Delta$

**Claim:** when $p \leq \Delta$, we have

$$|B(\xi)| \leq \frac{1}{(n \ln n)^\alpha}$$

$$\text{Var}(\xi) \leq \frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}}.$$

Now we prove this claim. In this regime, we write $L_\alpha(X) = S_{K,\alpha} - (S_{K,\alpha}(X) - 1) I(S_{K,\alpha}(X) \geq 1)$. We have

$$|B(\xi)| = \left| E S_{K,\alpha}(X) P(Y \leq 2\Delta) - [E (S_{K,\alpha}(X) - 1) I(S_{K,\alpha} \geq 1)] P(Y \leq 2\Delta) + E U_\alpha(X) P(Y > 2\Delta) - p^\alpha \right|$$

$$= \left| E S_{K,\alpha}(X) - p^\alpha - [E (S_{K,\alpha}(X) - 1) I(S_{K,\alpha} \geq 1)] P(Y \leq 2\Delta) + (E U_\alpha(X) - E S_{K,\alpha}(X)) P(Y > 2\Delta) \right|$$

$$\leq |E S_{K,\alpha}(X) - p^\alpha| + E (S_{K,\alpha}(X) - 1) I(S_{K,\alpha} \geq 1) |(E U_\alpha(X) + |E S_{K,\alpha}(X)|)| P(Y > 2\Delta)$$

$$\equiv B_1 + B_2 + B_3.$$

Now we bound $B_1, B_2, B_3$ separately. It follows from Lemma 4 that

$$B_1 = |E S_{K,\alpha}(X) - p^\alpha| \leq \frac{c_3}{(n \ln n)^\alpha} \leq \frac{1}{(n \ln n)^\alpha}.$$

Now consider $B_2$. Note that for any random variable $Z$ and any constant $\lambda > 0$,

$$E(Z 1(Z \geq \lambda)) \leq \lambda^{-1} E(Z^2 1(X \geq \lambda)) \leq \lambda^{-1} E Z^2.$$

Hence, we have

$$B_2 = E(S_{K,\alpha}(X) - 1) I(S_{K,\alpha} \geq 1) \leq E S_{K,\alpha}(X) I(S_{K,\alpha} \geq 1) \leq E S_{K,\alpha} \leq n^{c_2} \ln 2 \frac{(4c_1 \ln n)^{2+2\alpha}}{n^{2\alpha}} \leq \frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}}$$

where we used Lemma 4 in the last step.

Now we deal with $B_3$. We have

$$|E S_{K,\alpha}(X)| \leq p^\alpha + \frac{c_3}{(n \ln n)^\alpha} \leq \left( \frac{c_1 \ln n}{n} \right)^\alpha + \frac{c_3}{(n \ln n)^\alpha}$$

$$\text{E}[U_\alpha(X)] \leq \sup_{x \in [0,1]} |U_\alpha(x)| \leq 1 + \frac{\alpha(1-\alpha)}{2c_1 \ln n} \leq 1 + \frac{1}{8c_1 \ln n}$$

$$P(Y \geq 2\Delta) = P(n Y \geq 2n \Delta) \leq (e/4)^{c_1 \ln n} = n^{-c_1 \ln (4/e)},$$

where we have used Lemma 4 and Lemma 21. Thus, we have

$$B_3 = (|E S_{K,\alpha}(X)| + E[U_\alpha(X)]) P(Y \geq 2\Delta) \leq n^{-c_1 \ln (4/e)}.$$

To sum up, we have the following bound on $|B(\xi)|$:

$$|B(\xi)| \leq \frac{1}{(n \ln n)^\alpha} + \frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}} + n^{-16\alpha} \leq \frac{1}{(n \ln n)^\alpha}.$$
We now consider the variance. It follows from Lemma 23 and Lemma 24 that

\[ \text{Var}(\xi) \leq \text{Var}(S_{K,\alpha}(X)) + \text{Var}(U_{\alpha}(X)) + (\mathbb{E}L_{\alpha}(X) - \mathbb{E}U_{\alpha}(X))^2 \mathbb{P}(Y > 2\Delta) \]  
\[ \leq \mathbb{E}S_{K,\alpha}^2(X) + (\mathbb{E}U_{\alpha}^2(X) + 1 + 2\mathbb{E}U_{\alpha}(X)) \mathbb{P}(Y > 2\Delta) \]  
\[ \leq n^{8c_2\ln 2} \left(\frac{4c_1\ln n}{n^{2\alpha}}\right)^{2+2\alpha} \left(1 + \frac{1}{8c_1\ln n}\right)^2 n^{-c_1(4/\epsilon)} \]  
\[ \leq \frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}} + n^{-8\alpha} \]  
\[ \leq \frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}} \] (396)

2) Case 2: \( \Delta \leq p \leq 4\Delta \).

**Claim:** when \( \Delta \leq p \leq 4\Delta \), we have

\[ |B(\xi)| \leq \frac{1}{(n\ln n)^\alpha} \] (397)
\[ \text{Var}(\xi) \leq \begin{cases} 0 < \alpha \leq 1/2 \\ \frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}} + \frac{\epsilon^{2\alpha-1}}{\epsilon} & 1/2 < \alpha < 1 \end{cases} \] (398)

Now we prove this claim. In this case,

\[ |B(\xi)| = \left| (\mathbb{E}L_{\alpha}(X) - p^\alpha)\mathbb{P}(Y \leq 2\Delta) + (\mathbb{E}U_{\alpha}(X) - p^\alpha)\mathbb{P}(Y > 2\Delta) \right| \] (399)
\[ \leq |\mathbb{E}L_{\alpha}(X) - p^\alpha| + |\mathbb{E}U_{\alpha}(X) - p^\alpha| \] (400)
\[ \leq |\mathbb{E}S_{K,\alpha}(X) - p^\alpha| + \mathbb{E}(S_{K,\alpha}(X) - 1)\mathbb{I}(S_{K,\alpha} \geq 1) + |\mathbb{E}U_{\alpha}(X) - p^\alpha| \] (401)
\[ \equiv B_1 + B_2 + B_3. \] (402)

It follows from Lemma 4 that

\[ B_1 = |\mathbb{E}S_{K,\alpha}(X) - p^\alpha| \leq \frac{c_3}{(n\ln n)^\alpha} \leq \frac{1}{(n\ln n)^\alpha}. \] (403)

As in 386, we have

\[ B_2 = \mathbb{E}(S_{K,\alpha}(X) - 1)\mathbb{I}(S_{K,\alpha} \geq 1) \leq n^{8c_2\ln 2} \left(\frac{4c_1\ln n}{n^{2\alpha}}\right)^{2+2\alpha} \leq \frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}}. \] (404)

Regarding \( B_3 \), applying Lemma 2 we have

\[ B_3 \leq \frac{17}{n^\alpha(c_1\ln n)^{2-\alpha}} + \frac{8310(1+\alpha)}{\alpha(2-\alpha)} p^\alpha n^{-c_1/8} \leq \frac{1}{n^\alpha(\ln n)^{2-\alpha}} + n^{-2\alpha}. \] (405)

To sum up, we have

\[ |B(\xi)| \leq \frac{1}{(n\ln n)^\alpha} + \frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}} + \frac{1}{n^\alpha(\ln n)^{2-\alpha}} + n^{-2\alpha} \leq \frac{1}{(n\ln n)^\alpha}. \] (406)

For the variance, we have

\[ \text{Var}(\xi) \leq \text{Var}(S_{K,\alpha}(X)) + \text{Var}(U_{\alpha}(X)) + (\mathbb{E}L_{\alpha}(X) - \mathbb{E}U_{\alpha}(X))^2. \] (407)

Applying Lemma 4 we have

\[ \text{Var}(S_{K,\alpha}(X)) \leq n^{8c_2\ln 2} \left(\frac{4c_1\ln n}{n^{2\alpha}}\right)^{2+2\alpha} \leq \frac{(\ln n)^{2+2\alpha}}{n^{2\alpha-\epsilon}}. \] (408)

Lemma 2 implies that

\[ \text{Var}(U_{\alpha}(X)) \leq \begin{cases} \frac{24}{n^\alpha(c_1\ln n)^{1-2\alpha}} + \frac{576}{\alpha} p^\alpha n^{-c_1/8} + \frac{28800}{\alpha^2} p^2\alpha n^{-c_1/4} & 0 < \alpha \leq 1/2 \\ \frac{14c_2\ln 2}{n^{2\alpha-1}} + \frac{576}{\alpha} p^2\alpha n^{-c_1/8} + \frac{28800}{\alpha^2} p^2\alpha n^{-c_1/4} + \frac{8}{n^{2\alpha}(c_1\ln n)^{2-2\alpha}} & 1/2 < \alpha < 1 \end{cases} \] (409)
Regarding \((E L_\alpha(X) - E U_\alpha(X))^2\), we have
\[
(E L_\alpha(X) - E U_\alpha(X))^2 \leq \left| |E S_{K,\alpha}(X) - p^\alpha| + E(S_{K,\alpha}(X) - 1)I(S_{K,\alpha} \geq 1) + |E U_\alpha(X) - p^\alpha|\right|^2
\]
\[
\leq \left[ \frac{c_3}{(n \ln n)^{\alpha}} + n^{8c_2 \ln 2} \left( \frac{4c_1 \ln n}{n^{2\alpha}} \right)^{2+2\alpha} + \frac{17}{n^{\alpha} (c_1 \ln n)^{2-\alpha}} + \frac{8310(1 + \alpha)}{\alpha(2 - \alpha)} p^\alpha n^{-c_1/8} \right]^2
\]
\[
\geq \frac{1}{(n \ln n)^{2\alpha}}.
\]

To sum up, we have
\[
\text{Var}(\xi) \leq \begin{cases} 
\frac{(\ln n)^{2+2\alpha}}{n^{2\alpha}} & 0 < \alpha \leq 1/2 \\
\frac{1}{n^{2\alpha} (c_1 \ln n)^{2-\alpha}} + \frac{p^{2\alpha - 1}}{n} & 1/2 < \alpha < 1
\end{cases}
\]

3) Case 3: \(p > 4\Delta\).
Claim: when \(p > 4\Delta\), we have
\[
|B(\xi)| \leq \frac{1}{n^{2\alpha} (c_1 \ln n)^{2-\alpha}} + \frac{p^{2\alpha - 1}}{n} \quad 0 < \alpha \leq 1/2
\]
\[
\text{Var}(\xi) \leq \begin{cases} 
\frac{1}{n^{2\alpha} (c_1 \ln n)^{2-\alpha}} + \frac{p^{2\alpha - 1}}{n} & 1/2 < \alpha < 1
\end{cases}
\]

Now we prove this claim. In this case,
\[
|B(\xi)| \leq |E U_\alpha(X) - p^\alpha| + (|E L_\alpha(X)| + p^\alpha) \mathbb{P}(Y \leq 2\Delta)
\]
\[
\leq \frac{17}{n^{\alpha}(c_1 \ln n)^{2-\alpha}} + \frac{8310(1 + \alpha)}{\alpha(2 - \alpha)} p^\alpha n^{-c_1/8} + 2\mathbb{P}(Y \leq 2\Delta)
\]
\[
\leq \frac{17}{n^{\alpha}(c_1 \ln n)^{2-\alpha}} + \frac{8310(1 + \alpha)}{\alpha(2 - \alpha)} p^\alpha n^{-c_1/8} + 2\mathbb{P}(n Y \leq 2n\Delta)
\]
\[
\leq \frac{17}{n^{\alpha}(c_1 \ln n)^{2-\alpha}} + \frac{8310(1 + \alpha)}{\alpha(2 - \alpha)} p^\alpha n^{-c_1/8} + 2e^{-c_1/2\ln n}
\]
\[
= \frac{17}{n^{\alpha}(c_1 \ln n)^{2-\alpha}} + \frac{8310(1 + \alpha)}{\alpha(2 - \alpha)} p^\alpha n^{-c_1/8} + 2n^{-c_1/2}
\]
\[
\leq \frac{1}{n^{\alpha}(c_1 \ln n)^{2-\alpha}}.
\]

Regarding the variance, we have
\[
\text{Var}(\xi) \leq \text{Var}(U_\alpha(X)) + (\text{Var}(L_\alpha(X)) + (E L_\alpha(X) - E U_\alpha(X))^2) \mathbb{P}(Y \leq 2\Delta)
\]
\[
\leq \begin{cases} 
\frac{24}{n^{\alpha}(c_1 \ln n)^{2-\alpha}} + \frac{576}{14p^{2\alpha - 1}} p^{2\alpha} n^{-c_1/8} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} & 0 < \alpha \leq 1/2 \\
\frac{1}{n^{\alpha}(c_1 \ln n)^{2-\alpha}} + \frac{576}{14p^{2\alpha - 1}} p^{2\alpha} n^{-c_1/8} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} + \frac{8}{n^{\alpha}(c_1 \ln n)^{2-\alpha}} & 1/2 < \alpha < 1
\end{cases}
\]
\[
\leq \begin{cases} 
\frac{24}{n^{\alpha}(c_1 \ln n)^{2-\alpha}} + \frac{576}{14p^{2\alpha - 1}} p^{2\alpha} n^{-c_1/8} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} & 0 < \alpha \leq 1/2 \\
\frac{1}{n^{\alpha}(c_1 \ln n)^{2-\alpha}} + \frac{576}{14p^{2\alpha - 1}} p^{2\alpha} n^{-c_1/8} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} + \frac{8}{n^{\alpha}(c_1 \ln n)^{2-\alpha}} & 1/2 < \alpha < 1
\end{cases}
\]
\[
\leq \begin{cases} 
\frac{1}{n^{\alpha}(c_1 \ln n)^{2-\alpha}} + \frac{p^{2\alpha - 1}}{n} & 0 < \alpha \leq 1/2 \\
\frac{1}{n^{\alpha}(c_1 \ln n)^{2-\alpha}} + \frac{p^{2\alpha - 1}}{n} & 1/2 < \alpha < 1
\end{cases}
\]

I. Proof of Lemma 8
We use Lemma 7, Lemma 9, and Lemma 10 to compute the bias and variance of \(\xi\). We distinguish four cases.

1) Case 1: \(p \leq \frac{4\Delta}{c_1 \ln n}\).
In this regime, we write \(L_\alpha(X) = S_{K,\alpha}(X) - (S_{K,\alpha}(X) - 1)I(S_{K,\alpha}(X) \geq 1)\) and bound the bias as
\[
|B(\xi)| = |E S_{K,\alpha}(X)\mathbb{P}(Y \leq 2\Delta) - (E S_{K,\alpha}(X) - 1)I(S_{K,\alpha}(X) \geq 1)\mathbb{P}(Y \leq 2\Delta) + E U_\alpha(X)\mathbb{P}(Y > 2\Delta) - p^\alpha|
\]
\[
\leq |E S_{K,\alpha}(X) - p^\alpha| + E(S_{K,\alpha}(X) - 1)I(S_{K,\alpha}(X) \geq 1) + (|E U_\alpha(X)| + |E S_{K,\alpha}(X)\mathbb{P}(Y > 2\Delta))
\]
\[
\equiv B_1 + B_2 + B_3.
\]
Now we bound $B_1, B_2, B_3$ separately. It follows from Lemma 5 that

\[ B_1 = |\mathbb{E} S_{K, \alpha}(X) - p^\alpha| \leq p^\alpha + D_1 \left( \frac{4c_1}{c_2^2 n \ln n} \right) \alpha^{-1} p \leq p^\alpha + \frac{p}{(n \ln n)^{\alpha-1}}, \]

\[ B_2 = \mathbb{E} (S_{K, \alpha}(X) - 1) \mathbb{1} (S_{K, \alpha}(X) \geq 1) \leq \mathbb{E} S_{K, \alpha}^2 \leq n^{10c_2 \ln 2} \frac{4c_1 \ln n}{n^{2\alpha-1}} \leq (\ln n)^{2\alpha+2} p \cdot n^{-2\alpha-1-\epsilon}. \]

Now we consider $B_3$. Since

\[ |\mathbb{E} U_\alpha(X)| \leq |\mathbb{E} X^\alpha| \leq \frac{1}{n^\alpha} |\mathbb{E} (nX)^\alpha| \leq \frac{1}{n^\alpha} |\mathbb{E} (nX)^2| = \frac{np + (np)^2}{n^\alpha} \leq \frac{p}{n^\alpha-1} \]

\[ |\mathbb{E} S_{K, \alpha}(X)| \leq D_1 \left( \frac{4c_1}{c_2^2 n \ln n} \right) \alpha^{-1} p \leq \frac{p}{(n \ln n)^{\alpha-1}} \]

\[ \mathbb{P}(Y > 2\Delta) = \mathbb{P}(nY > 2n\Delta) \leq \left( \frac{enp}{2c_1 \ln n} \right)^{2c_1 \ln n} \leq \left( \frac{2e}{c_2^2 \ln n} \right)^{2c_1 \ln n} \leq n^{-2c_1 \ln n}, \]

where we have used Lemma 21 and the pointwise bound in Lemma 17 thus

\[ B_3 = (|\mathbb{E} U_\alpha(X)| + |\mathbb{E} S_{K, \alpha}(X)|) \mathbb{P}(Y > 2\Delta) \leq pn^{-2c_1 \ln n}. \]

To sum up, we have the following bound on $|B(\xi)|$:

\[ |B(\xi)| \leq p^\alpha + \frac{p}{(n \ln n)^{\alpha-1}} + \frac{(\ln n)^{2\alpha+2} p}{n^{2\alpha-1-\epsilon}} + pn^{-2c_1 \ln n} \leq p^\alpha + \frac{p}{(n \ln n)^{\alpha-1}}. \]

As for variance, it follows from Lemma 23 and Lemma 24 that

\[ \text{Var}(\xi) \leq \text{Var}(S_{K, \alpha}(X)) + \text{Var}(U_\alpha) \mathbb{P}(Y > 2\Delta) + (\mathbb{E} L_\alpha(X) - \mathbb{E} U_\alpha(X))^2 \mathbb{P}(Y > 2\Delta) \]

\[ \leq \mathbb{E} S_{K, \alpha}^2 \leq 2(\mathbb{E} L_\alpha^2(X) + \mathbb{E} U_\alpha^2(X)) \mathbb{P}(Y > 2\Delta) \]

\[ \leq \left( \frac{(\ln n)^{2\alpha+2} p}{n^{2\alpha-1-\epsilon}} + 2 \left( \frac{(\ln n)^{2\alpha+2} p}{n^{2\alpha-1-\epsilon}} + \frac{p}{n^{2\alpha-1}} \right) \right) n^{-2c_1 \ln n} \]

\[ \leq \left( \frac{(\ln n)^{2\alpha+2} p}{n^{2\alpha-1-\epsilon}} \right), \]

where we have used the fact that when $p \leq \frac{4c_1}{c_2^2 n \ln n}$,

\[ |\mathbb{E} U_\alpha^2(X)| \leq |\mathbb{E} X^{2\alpha}| \leq \frac{1}{n^{2\alpha}} |\mathbb{E} (nX)^{2\alpha}| \leq \frac{1}{n^{2\alpha}} |\mathbb{E} (nX)^3| = \frac{(np)^3 + 3(np)^2 + np}{n^{2\alpha}} \leq \frac{p}{n^{2\alpha-1}}. \]

2) Case $\frac{4c_1}{c_2^2 n \ln n} < p \leq \Delta$.

To bound the bias, we use the same definition of $B_1, B_2, B_3$ as in Case 1. It follows from Lemma 6 that

\[ B_1 = |\mathbb{E} S_{K, \alpha}(X) - p^\alpha| \leq 6 \left( \frac{\pi^2 c_1}{c_2^2 n \ln n} \right) \alpha \leq \frac{1}{(n \ln n)^{\alpha-1}}, \]

\[ B_2 = \mathbb{E} (S_{K, \alpha}(X) - 1) \mathbb{1} (S_{K, \alpha}(X) \geq 1) \leq \mathbb{E} S_{K, \alpha}^2 \leq n^{8c_2 \ln 2} \frac{4c_1 \ln n}{n^{2\alpha}} \leq \frac{(\ln n)^{2\alpha+2}}{n^{2\alpha-\epsilon}}. \]

To deal with $B_3$, we use

\[ |\mathbb{E} U_\alpha(X)| \leq \sup_{x \in [0, 1]} |U_\alpha(x)| = 1 \]

\[ |\mathbb{E} S_{K, \alpha}(X)| \leq |\mathbb{E} S_{K, \alpha}(X) - p^\alpha| + p^\alpha \leq 6 \left( \frac{\pi^2 c_1}{c_2^2 n \ln n} \right) \alpha \leq \left( \frac{\ln n}{n} \right)^\alpha \]

\[ \mathbb{P}(Y > 2\Delta) = \mathbb{P}(nY > 2n\Delta) \leq \left( \frac{e}{4} \right)^{\frac{c_1 \ln n}{n}} \leq n^{-c_1 \ln (4/\epsilon)}, \]

to obtain that

\[ B_3 = (|\mathbb{E} U_\alpha(X)| + |\mathbb{E} S_{K, \alpha}(X)|) \mathbb{P}(Y > 2\Delta) \leq n^{-c_1 \ln (4/\epsilon)}. \]

Hence, the bias is upper bounded by

\[ |B(\xi)| \leq \frac{1}{(n \ln n)^{\alpha}} + \frac{(\ln n)^{2\alpha+2}}{n^{2\alpha-\epsilon}} + n^{-c_1 \ln (4/\epsilon)} \leq \frac{1}{(n \ln n)^{\alpha}}. \]
Similar to the analysis in Case 1, the variance is upper bounded by
\[
\text{Var}(\xi) \leq \text{Var}(S_{K, \alpha}(X)) + \text{Var}(U_{\alpha}(X))\mathbb{P}(Y > 2\Delta) + \left(\mathbb{E}L_{\alpha}(X) - \mathbb{E}U_{\alpha}(X)\right)^2\mathbb{P}(Y > 2\Delta)
\]
(448)
\[
\leq \mathbb{E}S_{K, \alpha}^2(X) + 2(\mathbb{E}L_{\alpha}^2(X) + \mathbb{E}U_{\alpha}^2(X))\mathbb{P}(Y > 2\Delta)
\]
(449)
\[
\leq \left(\frac{\ln n}{n^{2\alpha - \epsilon}}\right)^{2\alpha + 2} + 2\left(1 + 1\right)n^{-c_1\ln(4/e)}
\]
(450)
\[
\leq \left(\frac{\ln n}{n^{2\alpha - \epsilon}}\right)^{2\alpha + 2}.
\]
(451)

3) Case 3: \(\Delta < p < 4\Delta\).

In this case,
\[
|B(\xi)| = |(\mathbb{E}L_{\alpha}(X) - p^\alpha)\mathbb{P}(Y \leq 2\Delta) + (\mathbb{E}U_{\alpha}(X) - p^\alpha)\mathbb{P}(Y > 2\Delta)|
\]
(452)
\[
\leq |\mathbb{E}L_{\alpha}(X) - p^\alpha| + |\mathbb{E}U_{\alpha}(X) - p^\alpha|
\]
(453)
\[
\leq |\mathbb{E}S_{K, \alpha}(X) - p^\alpha| + \mathbb{E}(S_{K, \alpha}(X) - 1)\mathbb{I}(S_{K, \alpha}(X) \geq 1) + |\mathbb{E}U_{\alpha}(X) - p^\alpha|
\]
(454)
\[
\equiv B_1 + B_2 + B_3.
\]
(455)

It follows from Lemma 5 and Lemma 2 that
\[
B_1 = |\mathbb{E}S_{K, \alpha}(X) - p^\alpha| \leq 6\left(\frac{\pi^2c_1}{c_1^2\ln n}\right)^\alpha \leq \frac{1}{(n\ln n)^\alpha},
\]
(456)
\[
B_2 = \mathbb{E}(S_{K, \alpha}(X) - 1)\mathbb{I}(S_{K, \alpha}(X) \geq 1) \leq \mathbb{E}S_{K, \alpha}^2 \leq n^{8c_2\ln 2} \frac{(4c_1\ln n)^{2\alpha + 2}}{n^2} \leq \frac{(\ln n)^{2\alpha + 2}}{n^{2\alpha - \epsilon}},
\]
(457)
\[
B_3 = |\mathbb{E}U_{\alpha}(X) - p^\alpha| \leq \frac{17}{n^\alpha(c_1\ln n)^{2-\alpha}} + \frac{8310(1 + \alpha)}{\alpha(2 - \alpha)}p^\alpha n^{-c_1/8} \leq \frac{1}{n^\alpha(\ln n)^{2-\alpha}}.
\]
(458)

To sum up, the total bias is upper bounded by
\[
|B(\xi)| \leq \frac{1}{(n\ln n)^\alpha} + \frac{(\ln n)^{2\alpha + 2}}{n^{2\alpha - \epsilon}} + \frac{1}{n^\alpha(\ln n)^{2-\alpha}} \leq \frac{1}{n^\alpha(\ln n)^{2-\alpha}}.
\]
(459)

For the variance, we have
\[
\text{Var}(\xi) \leq \text{Var}(S_{K, \alpha}(X)) + \text{Var}(U_{\alpha}(X)) + (\mathbb{E}L_{\alpha}(X) - \mathbb{E}U_{\alpha}(X))^2
\]
(460)
\[
\leq \mathbb{E}S_{K, \alpha}^2(X) + \text{Var}(U_{\alpha}(X)) + (\mathbb{E}L_{\alpha}(X) - \mathbb{E}U_{\alpha}(X))^2
\]
(461)
\[
\leq \left(\frac{\ln n}{n^{2\alpha - \epsilon}}\right)^{2\alpha + 2} + \left(\frac{p}{n} + \frac{1}{n^2}\right) + \frac{1}{n^\alpha(\ln n)^{2-2\alpha}}
\]
(462)
\[
\leq \frac{1}{n^2} + \frac{p}{n},
\]
(463)

where we have used Lemma 4 and
\[
|\mathbb{E}L_{\alpha}(X) - \mathbb{E}U_{\alpha}(X)| \leq |\mathbb{E}S_{K, \alpha}(X) - p^\alpha| + \mathbb{E}(S_{K, \alpha}(X) - 1)\mathbb{I}(S_{K, \alpha}(X) \geq 1) + |\mathbb{E}U_{\alpha}(X) - p^\alpha|
\]
(464)
\[
= B_1 + B_2 + B_3 + 1
\]
(465)
\[
\leq \frac{1}{n^\alpha(\ln n)^{2-\alpha}}.
\]
(466)

4) Case 4: \(p \geq 4\Delta\).

In this case, the bias is upper bounded by
\[
|B(\xi)| \leq |\mathbb{E}U_{\alpha}(X) - p^\alpha| + (|\mathbb{E}L_{\alpha}(X)| + p^\alpha)\mathbb{P}(Y \leq 2\Delta)
\]
(467)
\[
\leq \frac{17}{n^\alpha(c_1\ln n)^{2-\alpha}} + \frac{8310(1 + \alpha)}{\alpha(2 - \alpha)}p^\alpha n^{-c_1/8} + 2\mathbb{P}(nY \leq 2n\Delta)
\]
(468)
\[
\leq \frac{17}{n^\alpha(c_1\ln n)^{2-\alpha}} + \frac{8310(1 + \alpha)}{\alpha(2 - \alpha)}p^\alpha n^{-c_1/8} + 2e^{-c_1\ln n/2}
\]
(469)
\[
= \frac{17}{n^\alpha(c_1\ln n)^{2-\alpha}} + \frac{8310(1 + \alpha)}{\alpha(2 - \alpha)}p^\alpha n^{-c_1/8} + 2n^{-c_1/2}
\]
(470)
\[
\leq \frac{1}{n^\alpha(\ln n)^{2-\alpha}}.
\]
(471)
where we have used Lemma [21] to bound \( P(nY \leq 2n\Delta) \). The variance is then upper bounded by

\[
\text{Var}(\xi) \leq \text{Var}(U_\alpha(X)) + (\text{Var}(L_\alpha(X)) + (\mathbb{E}L_\alpha(X) - \mathbb{E}U_\alpha(X))^2) P(nY \leq 2n\Delta) \quad (472)
\]

\[
\leq \frac{202p}{n} + \frac{8}{n^2} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} + \frac{120}{\alpha} p^{2\alpha} n^{-c_1/8} + 2P(nY \leq 2n\Delta) \quad (473)
\]

\[
\leq \frac{202p}{n} + \frac{8}{n^2} + \frac{28800}{\alpha^2} p^{2\alpha} n^{-c_1/4} + \frac{120}{\alpha} p^{2\alpha} n^{-c_1/8} + 2n^{-c_1/2} \quad (474)
\]

\[
\leq \frac{p}{n} + \frac{1}{n^2}, \quad (475)
\]

where we have used Lemma 2.

\[\]

**J. Proof of Lemma 10**

The existence of the two prior distributions \( \nu_0 \) and \( \nu_1 \) follows directly from a standard functional analysis argument proposed by Lepski, Nemirovski, and Spokoiny [54], and elaborated in best polynomial approximation by Cai and Low [55, Lemma 1]. It suffices to replace the interval \([-1, 1]\) with \([0, 1]\) and the function \(|x|\) with \(x^\alpha\) in the proof of Lemma 1 in [55].

\[\]

**K. Proof of Lemma 11**

We compute the difference of the expectations as follows,

\[
\mathbb{E}_{\mu_0} F_{\alpha}(P) - \mathbb{E}_{\mu_0} F_{\alpha}(P) = \sum_{i=1}^{s'} (\mathbb{E}_{\mu_i} p_i^\alpha - \mathbb{E}_{\mu_0} p_i^\alpha) \quad (476)
\]

\[
= 2S'M^\alpha E\{x^\alpha\}_{[0,1]} \quad (477)
\]

\[
= 2 \left( \frac{\alpha}{c} \right)^\alpha n^\alpha (\ln n)^\alpha \left( \frac{d_1 \ln n}{n} \right)^\alpha \mu(2\alpha) \frac{1}{2^{2\alpha} (d_2 \ln n)^{2\alpha}} (1 + o(1)) \quad (478)
\]

\[
= 2 \left( \frac{\alpha}{c} \right)^\alpha \mu(2\alpha) \frac{d_1^\alpha}{(2d_2)^{2\alpha}} (1 + o(1)), \quad (479)
\]

where \(\mu(2\alpha)\) is the constant given by Lemma [17].

We also have bounds for the variance:

\[
\text{Var}_{\mu_j} (F_\alpha(P)) = \mathbb{E}_{\mu_j} \left( F_\alpha(P) - \mathbb{E}_{\mu_j} F_\alpha(P) \right)^2 \quad (480)
\]

\[
= \sum_{i=1}^{s'} \mathbb{E}_{\mu_i} (p_i^\alpha - \mathbb{E}_{\mu_i} p_i^\alpha)^2 \quad (481)
\]

\[
\leq S' \mathbb{E}_{\mu_j} p_i^{2\alpha} \quad (482)
\]

\[
\leq S' M^{2\alpha} \quad (483)
\]

\[
\leq \left( \frac{\alpha d_1^2}{c} \right)^\alpha (\ln n)^{3\alpha} n^{-\alpha} \to 0, \quad j = 0, 1. \quad (484)
\]

For any integer \( y \geq 0 \),

\[
F_{1,M}(y) - F_{0,M}(y) = \int \frac{e^{-np(ny)}}{y!} \left( \mu_1(dp) - \mu_0(dp) \right) \quad (485)
\]

\[
= \int \sum_{i=0}^{\infty} \frac{(-1)^i (ny)^{i+y}}{i! y!} \left( \mu_1(dp) - \mu_0(dp) \right), \quad (486)
\]

where we have used the Taylor expansion of \( e^{-x} \).

Now we proceed to bound the total variation distance between the marginal distributions under two priors \( \mu_0, \mu_1 \).

\[
\sum_{y=0}^{\infty} |F_{1,M}(y) - F_{0,M}(y)| = \sum_{y=0}^{d_1 \ln n} |F_{1,M}(y) - F_{0,M}(y)| + \sum_{y=d_2 \ln n}^{\infty} |F_{1,M}(y) - F_{0,M}(y)| \quad (487)
\]

\[
\triangleq D_1 + D_2. \quad (488)
\]
Note that we take \( d_1 = 1, d_2 = 10e \) in the assumption. We bound \( D_2 \) in the following way:

\[
D_2 = \sum_{y \geq \frac{d_2 \ln n}{2}} e^{-np} \frac{(np)^y}{y!} (\mu_1(dp) - \mu_0(dp))
\]

(489)

\[
\leq P \left( \text{Poi}(np) > \frac{d_2 \ln n}{2} \right)
\]

(490)

\[
\leq P \left( \text{Poi}(d_2 \ln n) > \frac{d_2 \ln n}{2} \right)
\]

(491)

\[
\leq \left( \frac{e^{5e-1}}{(5e)^{d_2 \ln n}} \right) \ln n
\]

(492)

\[
= \frac{1}{n^{1+5e \ln 5}} \leq \frac{1}{n^{12}},
\]

(493)

where in the fourth step we have applied Lemma 21.

We bound \( D_1 \) as follows:

\[
D_1 = \sum_{y=0}^{\frac{d_1 \ln n}{2}} |F_{1,M}(y) - F_{0,M}(y)|
\]

(494)

\[
= \sum_{y=0}^{\frac{d_2 \ln n}{2}} \frac{1}{y!} \left| \sum_{i=0}^{\infty} \frac{(-1)^i}{i!} (np)^{i+y} (\mu_1(dp) - \mu_0(dp)) \right|
\]

(495)

\[
= \sum_{y=0}^{\frac{d_2 \ln n}{2}} \frac{1}{y!} \left| \sum_{i=d_2 \ln n - y}^{\infty} \frac{(-1)^i}{i!} (np)^{i+y} (\mu_1(dp) - \mu_0(dp)) \right|
\]

(496)

\[
\leq \sum_{y=0}^{\frac{d_2 \ln n}{2}} \frac{1}{y!} \left| \sum_{i=d_2 \ln n - y}^{\infty} \frac{(d_1 \ln n)^{i+y}}{i!} \right|
\]

(497)

\[
= \sum_{y=0}^{\frac{d_2 \ln n}{2}} \frac{(d_1 \ln n)^y}{y!} \sum_{i=d_2 \ln n - y}^{\infty} \frac{(d_1 \ln n)^i}{i!},
\]

(498)

where in the third step we have used the fact that \( \mu_1 \) and \( \mu_0 \) have matching moments up to order \( d_2 \ln n \). The Lagrangian remainder for Taylor series of \( e^x, x > 0 \) shows that

\[
\sum_{j>m} \frac{x^j}{j!} = e^x \frac{x^{m+1}}{(m+1)!},
\]

(499)

where \( 0 \leq \xi \leq x \). Applying this result, we have

\[
D_1 \leq \sum_{y=0}^{\frac{d_2 \ln n}{2}} \frac{(d_1 \ln n)^y}{y!} \sum_{i=d_2 \ln n - y}^{d_1 \ln n} \frac{(d_1 \ln n)^i}{i!}
\]

(500)

\[
\leq \sum_{y=0}^{\frac{d_2 \ln n}{2}} \frac{(d_1 \ln n)^y}{y!} \sum_{i=d_2 \ln n - y}^{d_1 \ln n} \frac{(d_1 \ln n)^i}{i!} e^{d_1 \ln n - (d_2 \ln n - y + 1)}
\]

(501)

\[
\leq \sum_{y=0}^{\frac{d_2 \ln n}{2}} \frac{(d_1 \ln n)^y}{y!} e^{d_1 \ln n - (d_2 \ln n - y + 1)} \frac{(d_1 \ln n)^i}{i!} e^{d_2 \ln n - y + 1}
\]

(502)

\[
\leq e^{d_1 \ln n + d_2 \ln n + 1} \left( \frac{d_1 \ln n}{d_2 \ln n + 1} \right)^{d_2 \ln n + 1} \sum_{y=0}^{\frac{d_2 \ln n}{2}} \frac{(d_1 \ln n)^y}{y!}
\]

(503)

\[
\leq e^{2d_1 \ln n + d_2 \ln n + 1} \left( \frac{d_1 \ln n}{d_2 \ln n + 1} \right)^{d_2 \ln n + 1} \sum_{y=0}^{\frac{d_2 \ln n}{2}} \frac{(d_1 \ln n)^y}{y!}
\]

(504)

where in the third step we have used the fact that \( n! \geq \left( \frac{n}{e} \right)^n \).
Plugging \( d_1 = 1, d_2 = 10e \) in, we have
\[
D_1 \leq \frac{1}{5n^6}. \tag{505}
\]

Combining bounds on \( D_1 \) and \( D_2 \) together, we have
\[
V(F_{1, M}, F_{0, M}) = \frac{1}{2} \sum_{y=0}^{\infty} |F_{1, M}(y) - F_{0, M}(y)| \leq \frac{1}{2} (D_1 + D_2) \leq \frac{1}{n^6}. \tag{506}
\]

L. Proof of Lemma 13

For \( \eta \in (0, 1/2) \), define
\[
f_\eta(x) = \left( \frac{1 - \eta}{2} x + \frac{1 + \eta}{2} \right)^\beta,
\]
then \( E_L[f_\eta][-1,1] = E_L[x^\beta]|_{\eta,1} \). For \( \varphi(x) = \sqrt{1 - x^2} \), denote the second-order Ditzian-Totik modulus of smoothness by
\[
\omega_\varphi^2(f, t) \triangleq \sup \left\{ \left| f(u) + f(v) - 2f\left( \frac{u + v}{2} \right) \right| : u, v \in [-1, 1], |u - v| \leq 2t \varphi\left( \frac{u + v}{2} \right) \right\},
\]
then it is straightforward to obtain that
\[
\omega_\varphi^2(f_\eta, n^{-1}) = \left| \eta^\beta + \left( \eta + \frac{2(1 - \eta)}{n^2 + 1} \right)^\beta \right|, \quad \forall n \leq (4\eta)^\frac{2}{\beta} - 1. \tag{509}
\]

It follows directly from (509) that
\[
\frac{2 \cdot 5^\beta - 4^\beta - 6^\beta}{(2n)^{2\beta}} \leq \omega_\varphi^2(f_\eta, n^{-1}) \leq \frac{1}{n^{2\beta}}, \quad n \leq \min\left\{ \frac{1}{\sqrt{\eta}}, (4\eta)^\frac{2}{\beta} - 1 \right\}. \tag{510}
\]

The relationship between \( \omega_\varphi^2(f_\eta, n^{-1}) \) and \( E_n[f_\eta][-1,1] \) was shown in [50] Thm. 7.2.1, Thm. 7.2.4] that there exists two universal positive constants \( M_1, M_2 \) such that
\[
E_n[f_\eta][-1,1] \leq M_1 \omega_\varphi^2(f_\eta, n^{-1}), \tag{511}
\]
\[
\frac{1}{n^2} \sum_{k=0}^{n} (k+1) E_k[f_\eta][-1,1] \geq M_2 \omega_\varphi^2(f_\eta, n^{-1}). \tag{512}
\]

Applying (511) and (512) and setting the approximation order \( N = DL \) with a positive constant \( D \) to be specified later, then given \( \eta = 1/N^2 \), the non-increasing property of \( E_n[f_\eta][-1,1] \) with respect to \( n \) yields that
\[
E_L[f_\eta][-1,1] \geq \frac{1}{N-L} \sum_{k=L+1}^{N} E_k[f_\eta][-1,1] \tag{513}
\]
\[
\geq \frac{1}{N^2} \sum_{k=L+1}^{N} (k+1) E_k[f_\eta][-1,1] \tag{514}
\]
\[
\geq M_2 \omega_\varphi^2(f_\eta, N^{-1}) - \frac{E_0[f_\eta][-1,1]}{N^2} - \frac{1}{N^2} \sum_{k=1}^{L} (k+1) E_k[f_\eta][-1,1] \tag{515}
\]
\[
\geq M_2 \frac{(2 \cdot 5^\beta - 4^\beta - 6^\beta)}{(2N)^{2\beta}} - \frac{1}{(DL)^{2\beta}} - \frac{M_1}{N} \sum_{k=1}^{L} \omega_\varphi^2(f_\eta, k^{-1}) \tag{516}
\]
\[
\geq M_2 \frac{(2 \cdot 5^\beta - 4^\beta - 6^\beta)}{(2DL)^{2\beta}} - \frac{1}{D^2L^{2\beta}} - \frac{M_1}{DL} \int_{0}^{L} \frac{dx}{x^{2\beta}} \tag{517}
\]
\[
= L^{-2\beta} \left[ M_2 \frac{(2 \cdot 5^\beta - 4^\beta - 6^\beta)}{(2D)^{2\beta}} - \frac{1}{D^2} - \frac{M_1}{D(1 - 2\beta)} \right]. \tag{519}
\]

Due to \( 0 < 2\beta < 1 \), for a sufficiently large universal constant \( D \) we can obtain that
\[
\lim_{L \to \infty} \inf_{L^2\beta} E_L[x^\beta]|_{\eta,1} = \lim_{L \to \infty} \inf_{L^2\beta} E_L[f_\eta][-1,1] > 0. \tag{520}
\]
M. Proof of Lemma 14

Fix \( \delta > 0 \). Let \( \hat{F}(Z) \) be a near-minimax estimator of \( F_\alpha(P) \) under the Multinomial model. The estimator \( \hat{F}(Z) \) obtains the number of samples \( n \) from observation \( Z \). By definition, we have

\[
\sup_{P \in \mathcal{M}_S} E_{\text{Multinomial}}|\hat{F}(Z) - F_\alpha(P)|^2 < R(S, n) + \delta,
\]

where \( R(S, n) \) is the minimax \( L_2 \) risk under the Multinomial model. Given \( P \in \mathcal{M}_S(\gamma) \), let \( Z = [Z_1, \cdots, Z_S]^T \) with \( Z_i \sim \text{Poi}(np_i) \) and let \( n' = \sum_{i=1}^{S} Z_i \sim \text{Poi}(n \sum_{i=1}^{S} p_i) \), we use the estimator \( d_1 \cdot \hat{F}(Z) \) to estimate \( F_\alpha(P) \).

The triangle inequality gives

\[
\frac{1}{2} E_P|d_1^\alpha \cdot \hat{F}(Z) - F_\alpha(P)|^2 \leq E_P \left[ \left| d_1^\alpha \cdot \hat{F}(Z) - d_1^\alpha F_\alpha \left( \frac{P}{\sum_{i=1}^{S} p_i} \right) \right|^2 + \left| \left( \sum_{i=1}^{S} p_i \right)^\alpha - d_1^\alpha \right|^2 F_\alpha^2 \left( \frac{P}{\sum_{i=1}^{S} p_i} \right) \right] \leq \frac{1}{2} E_P \left[ \left( P - d_1 n \right) \cdot \hat{F}(Z) - P \right]^2 \leq d_1^{2\alpha} \sum_{m=0}^{\infty} R(S, m) \mathbb{P}(n' = m) + 4M^{2\alpha - 2} \left( \frac{\ln n}{2\gamma} \right)^{2\gamma} \leq d_1^{2\alpha} R(S, \frac{d_1 n}{2}) \mathbb{P}(n' \geq \frac{d_1 n}{2}) + \mathbb{P}(n' \leq \frac{d_1 n}{2}) + \frac{4M^{2\alpha - 2} \left( \frac{\ln n}{2\gamma} \right)^{2\gamma}}{8} \leq d_1^{2\alpha} R(S, \frac{d_1 n}{2}) + \exp(-\frac{d_1 n}{8}) + \frac{4M^{2\alpha - 2} \left( \frac{\ln n}{2\gamma} \right)^{2\gamma}}{8},
\]

where we have used the fact that conditioned on \( n' = m \), \( Z \sim \text{Multinomial}(m, P) \) and the last step follows from Lemma 21.

The proof is completed by the arbitrariness of \( \delta \) and Lemma 16.

APPENDIX C
PROOF OF AUXILIARY LEMMAS

A. Proof of Lemma 7

We obtain the polynomial \( g(x; a) \) via the Hermite interpolation formula. Concretely, the following WolframAlpha (http://www.wolframalpha.com/) command will give us \( g(x; a) \):

\[
\text{InterpolatingPolynomial}([\{0,0,0,0,0\}, \{a,1,0,0,0\}], x)
\]

B. Proof of Lemma 15

For brevity, denote \( \text{Var}(-\ln P(X)) \) as \( V(P) \), we have

\[
V(P) = \sum_{i=1}^{S} p_i (\ln p_i)^2 - \left( \sum_{i=1}^{S} p_i \ln p_i \right)^2.
\]

We construct the Lagrangian:

\[
\mathcal{L} = \sum_{i=1}^{S} p_i (\ln p_i)^2 - \left( \sum_{i=1}^{S} p_i \ln p_i \right)^2 + \lambda \left( \sum_{i=1}^{S} p_i - 1 \right).
\]

Taking derivatives with respect to \( p_i \), we obtain

\[
\frac{\partial \mathcal{L}}{\partial p_i} = (\ln p_i)^2 + p_i (2 \ln p_i) \cdot 1 = 2 \sum_{i=1}^{S} p_i \ln p_i \left(1 + \ln p_i \right) + \lambda = 0,
\]

It is equivalent to

\[
(\ln p_i)^2 + 2 \ln p_i + 2H(P)(1 + \ln p_i) + \lambda = 0,
\]

Note that it is a quadratic form for \( \ln p_i \) with the same coefficients. Solving for \( \ln p_i \), we obtain that

\[
\ln p_i = -H(P) \pm \sqrt{1 + H^2(P) - \lambda}.
\]

It implies that components of the maximum achieving distribution can only take two values. Assume \( p_i \in \{q_1, q_2\}, \forall i \). Suppose \( q_1 \) appears \( k \) times, we have

\[
kq_1 + (S-k)q_2 = 1.
\]
Now we compute the functional

\[
V(P) = kq_1(\ln q_1)^2 + (S-k)q_2(\ln q_2)^2 - (kq_1 \ln q_1 + (S-k)q_2 \ln q_2)^2
= kq_1(\ln q_1)^2 + (1 - kq_1)(\ln q_2)^2 - k^2q_2(\ln q_1)^2 - (1 - kq_1)^2(\ln q_2)^2 - 2kq_1(1 - kq_1)(\ln q_1)(\ln q_2)
= kq_1(1 - kq_1)\left(\ln \frac{q_2}{q_1}\right)^2.
\]  

(533)

(534)

(535)

Since \( q_2 = \frac{1 - kq_1}{S - k} \), we have

\[
V(P) = kq_1(1 - kq_1)\left(\ln \frac{1 - kq_1}{S q_1 - kq_1}\right)^2.
\]  

(536)

Denote \( x = kq_1, y = k/S \), we have

\[
V(P) = x(1 - x) \left(\ln \frac{1 - x}{x} - \ln \frac{1 - y}{y}\right)^2.
\]  

(537)

Fixing \( x \), we see \( V(P) \) is a monotone function of \( y \). Without loss of generality, by symmetry we assume \( x \leq 1/2 \). Then, the maximum achieving \( y = \frac{S - 1}{S} \), and \( V(P) \) as a function of \( x \) is

\[
V(P) = x(1 - x) \left(\ln \frac{1 - x}{x} + \ln(S - 1)\right)^2.
\]  

(538)

Taking derivatives with respect to \( x \), ignoring the minimum achieving \( x \), we obtain the following equation for maximum achieving value of \( x \), which is denoted as \( x_1 \):

\[
(1 - 2x_1) \left(\ln \frac{1}{x_1} - 1\right) + \ln(S - 1) = 2.
\]  

(539)

Denoting \( S - 1 \) by \( m \), we have

\[
(1 - 2x_1) \ln \frac{m(1 - x_1)}{x_1} = 2,
\]  

(540)

which is equivalent to

\[
(2 - 2x_1) \ln \frac{m(1 - x_1)}{x_1} = 2 + \ln \frac{m(1 - x_1)}{x_1}.
\]  

(541)

Multiplying both sides by \( \frac{x_1}{2} \ln \frac{m(1 - x_1)}{x_1} \), we obtain

\[
V_{\max} = x_1(1 - x_1) \left(\ln \frac{m(1 - x_1)}{x_1}\right)^2 = x_1 \left(\ln \frac{m(1 - x_1)}{x_1} + \frac{1}{2} \left(\ln \frac{m(1 - x_1)}{x_1}\right)^2\right).
\]  

(542)

Note that for \( x \in (0, 1/2) \),

\[
\ln \frac{m(1 - x)}{x} \in [\ln m, \infty),
\]  

(543)

and if \( S \geq 4 \), we have \( \ln m = \ln(S - 1) > 1 \).

Using the bound \( z \leq z^2, z \geq 1 \), we have

\[
V_{\max} \leq \frac{3}{2} x_1 \left(\ln \frac{m(1 - x_1)}{x_1}\right)^2.
\]  

(544)

Taking derivatives with respect to \( z \) for \( z \left(\ln \frac{m(1 - z)}{z}\right)^2 \), \( z \in (0, 1/2] \), we have

\[
\frac{d}{dz} \left( z \left(\ln \frac{m(1 - z)}{z}\right)^2 \right) = \ln \left(\frac{m(1 - z)}{z}\right) \times \left(\ln \frac{m(1 - z)}{z} + \frac{2}{z - 1}\right) \geq \ln \left(\frac{m(1 - z)}{z}\right) \times (\ln m - 4),
\]  

(545)

which is always nonnegative if \( m \geq e^3 \), i.e., \( S \geq 56 \). Hence, we know that when \( S \geq 56 \), the function \( z \left(\ln \frac{m(1 - z)}{z}\right)^2 \) is an increasing function of \( z \) for \( z \in (0, 1/2] \), thus achieves its maximum at \( z = 1/2 \).

Then, we obtain

\[
V_{\max} \leq \frac{3}{4} (\ln m)^2 \leq \frac{3}{4} (\ln S)^2.
\]  

(546)
C. Proof of Lemma 17

Applying [100] Thm. 5.6, we know

\[ R_P(S, n) = \sum_{k \geq 0} R(S, k)p(Poi(n) = k). \] (547)

Note that the function \( R(S, n) \) is decreasing in \( n \) since we have taken the infimum over all possible estimators. We prove the first inequality as follows.

\[ R_P(S, 2n) = \sum_{k \geq 0} R(S, k)p(Poi(2n) = k) \] (548)
\[ \leq \sum_{k=0}^{n-1} R(S, k)p(Poi(2n) = k) + \sum_{k=n}^{\infty} R(S, k)p(Poi(2n) = k) \] (549)
\[ \leq \sup_{P \in M_S} |F(P)|^2e^{-n/4} + R(S, n), \] (550)

where we have applied Lemma 21 in the last step.

Regarding the second inequality, we have

\[ R_P(S, n/2) = \sum_{k \geq 0} R(S, k)p(Poi(n/2) = k) \] (552)
\[ \geq R(S, n)p(Poi(n/2) \leq n) \] (553)
\[ \geq \frac{1}{2} R(S, n). \] (554)

D. Proof of Lemma 17

We first show the limiting result. Defining \( y^2 = x \), we know

\[ E_n[x^{\alpha}]_{[0,1]} = E_{2n}[y^{2\alpha}]_{[-1,1]} \] (555)

Applying Theorem 6 to our settings, for any \( \alpha > 0 \) we have

\[ \lim_{n \to \infty} n^{2\alpha} E_n[x^{\alpha}]_{[0,1]} = \lim_{n \to \infty} n^{2\alpha} E_{2n}[y^{2\alpha}]_{[-1,1]} \] (556)
\[ = \frac{1}{2^{2\alpha}} \lim_{n \to \infty} (2n)^{2\alpha} E_{2n}[y^{2\alpha}]_{[-1,1]} \] (557)
\[ = \frac{\mu(2\alpha)}{2^{2\alpha}}. \] (558)

Korneichuk [48] Sec. 6.2.5] showed the inequality

\[ E_n[f] \leq \omega \left( f, \frac{\pi}{n+1} \right) \] (559)

for all \( f \in C[-1, 1] \), where \( \omega(f, \delta) \) is the first order modulus of smoothness, defined as

\[ \omega(f, \delta) \triangleq \sup\{|f(x) - f(x + \delta)| : x \in [-1, 1], x + \delta \in [-1, 1]\}. \] (560)

Bernstein [101] Pg. 171] showed

\[ E_{n+1}[f(x)] \leq \frac{\pi}{2(n+1)} E_n[f'(x)], \] (561)

where \( f \in C([-1, 1]). \)

For \( 0 < \alpha \leq 1/2, \omega(y^{2\alpha}, \delta) \leq \delta^{2\alpha}, \delta \leq 2 \), hence we know

\[ E_n[x^{\alpha}]_{[0,1]} = E_{2n}[y^{2\alpha}]_{[-1,1]} \leq \left( \frac{\pi}{2n} \right)^{2\alpha}. \] (562)

For \( 1/2 < \alpha < 1 \), noting that

\[ (y^2)^{\alpha} = 2\alpha y^{2\alpha-1}, \] (563)

and that

\[ \omega(2\alpha y^{2\alpha-1}, \delta) \leq 2\alpha \delta^{2\alpha-1}, \delta \leq 2, \] (564)
we know for $1/2 < \alpha < 1$,
\[ E_{2n}[y^{2\alpha}][-1,1] \leq \frac{\pi}{2(2n)} E_{2n-1}[2\alpha y^{2\alpha-1}][-1,1] \leq 2\alpha \frac{\pi}{4n} \left( \frac{\pi}{2n} \right)^{2\alpha - 1} = \alpha \left( \frac{\pi}{2n} \right)^{2\alpha} \leq \left( \frac{\pi}{2n} \right)^{2\alpha}. \] (565)

For $1 < \alpha < 3/2$, by defining $y^2 = x$ we know that
\[ E_n[x^n][0,1] = E_{2n}[y^{2\alpha}][-1,1] \leq \frac{\pi^2}{4n} E_{2n-1}[2\alpha y^{2\alpha-1}][-1,1] \leq \frac{\pi^2}{4n(2n - 1)} E_{2n-2}[2\alpha(2\alpha - 1)y^{2\alpha-2}][-1,1] \] (566)
and
\[ \omega \left( 2\alpha(2\alpha - 1)y^{2\alpha-2}, \frac{1}{2n - 1} \right) = \frac{2\alpha(2\alpha - 1)\pi^{2\alpha-2}}{(2n - 1)^{2\alpha-2}}. \] (567)

Hence we have
\[ E_n[x^n][0,1] \leq \frac{\pi^2}{4n(2n - 1)} \cdot \frac{2\alpha(2\alpha - 1)\pi^{2\alpha-2}}{(2n - 1)^{2\alpha-2}} < \alpha(2\alpha - 1) \left( \frac{\pi}{2n - 1} \right)^{2\alpha} < 3 \left( \frac{\pi}{n} \right)^{2\alpha}. \] (568)

Plugging in $x = 0$ yields $|g_{0,\alpha}| < 3(\pi/n)^{2\alpha}$, hence
\[ \max_{0 \leq x \leq 1} |R_{n,\alpha}(x) - x^n| \leq E_n[x^n][0,1] + |g_{0,\alpha}| \leq 6 \left( \frac{\pi}{n} \right)^{2\alpha}. \] (569)

Moreover, it has been shown in [49, Pg. 207] that
\[ \max_{0 \leq x \leq 1} |R'_{n,\alpha}(x) - (x^n)'| \leq D : E_n[(x^n)'][0,1] \leq D\alpha \left( \frac{\pi}{2n} \right)^{2\alpha-1}, \] (570)
where $D > 0$ is a positive universal constant, and the last inequality follows directly from (562). Hence, when $x \leq n^{-2}$, we have
\[ |R'_{n,\alpha}(x)| \leq |R'_{n,\alpha}(x) - (x^n)'| + \alpha x^{n-1} \leq D\alpha \left( \frac{\pi}{2n} \right)^{2\alpha-1} + \frac{\alpha}{n^{2(\alpha-1)}} \leq \frac{D_1}{n^{2(\alpha-1)}}, \] (571)
then integrating on $x$ yields the pointwise bound
\[ |R_{n,\alpha}(x)| \leq |R_{n,\alpha}(0)| + \int_0^x |R'_{n,\alpha}(t)| dt \leq 0 + \int_0^\infty \frac{D_1}{n^{2(\alpha-1)}} dt = \frac{D_1}{n^{2(\alpha-1)}}, \quad \forall x \in \left[ 0, \frac{1}{n^2} \right]. \] (572)

In order to bound the coefficients of best polynomial approximations, we need the following result by Qazi and Rahman [102, Thm. E] on the maximal coefficients of polynomials on a finite interval.

**Lemma 25.** Let $p_n(x) = \sum_{\mu=0}^n a_\mu x^\mu$ be a polynomial of degree at most $n$ such that $\|p_n(x)\| \leq 1$ for $x \in [-1,1]$. Then, $|a_{n-2\mu}|$ is bounded above by the modulus of the corresponding coefficient of $T_n$ for $\mu = 0, 1, \ldots, [n/2]$, and $|a_{n-1-2\mu}|$ is bounded above by the modulus of the corresponding coefficient of $T_{n-1}$ for $\mu = 0, 1, \ldots, [(n-1)/2]$.

It is shown in Cai and Low [55, Lemma 2] that all of the coefficients of Chebyshev polynomial $T_{2m}(x)$, $m \in \mathbb{Z}_+$ are upper bounded by $2^{3m}$. If we view the best polynomial approximation of $x^\alpha$ or $-x\log x$ over $[0,1]$ as the best polynomial approximation of $y^{2\alpha}$ or $-y^2 \log y^2, y^2 = x$, then we would obtain an even polynomial over interval $[-1,1]$ represented as
\[ \sum_{k=0}^n g_{k,\alpha} y^{2k} \text{ or } \sum_{k=0}^n g_{k,\alpha} y^{2k}. \] (573)

Applying Lemma 25 and equation 34, we know that for all $k \leq n$, we have
\[ |g_{k,\alpha}| \leq 2^{3n}, \quad |g_{k,\alpha}| \leq 2^{3n}. \] (574)

**E. Proof of Lemma 79**

Define $x' = \frac{x}{2\Delta} \in [0,1]$. Applying Lemma 17, we have
\[ \left| (x')^\alpha - \sum_{k=0}^K g_{k,\alpha}(x')^k \right| \leq \left( \frac{\pi}{2K} \right)^{2\alpha}. \] (575)

Multiplying both sides by $(4\Delta)^\alpha$, we have
\[ \left| \sum_{k=0}^K g_{k,\alpha}(4\Delta)^{-k+\alpha} x^k - x^\alpha \right| \leq \left( \frac{\pi}{2} \right)^{2\alpha} \frac{(4\Delta)^\alpha}{K^{2\alpha}}. \] (576)
Since \( x^\alpha = 0 \) when \( x = 0 \), taking \( x = 0 \) in Lemma \([19]\) we know
\[
g_{0,\alpha}(4\Delta)^\alpha \leq \left( \frac{\pi}{2} \right)^{2\alpha} \frac{(4\Delta)^\alpha}{K^{2\alpha}}, \tag{577}
\]
which implies that
\[
\left| \sum_{k=1}^{K} g_{k,\alpha}(4\Delta)^{-k+\alpha} x^k - x^\alpha \right| \leq 2 \left( \frac{\pi}{2} \right)^{2\alpha} \frac{(4\Delta)^\alpha}{K^{2\alpha}} = \frac{c_3}{(n \ln n)^\alpha}. \tag{578}
\]

\[\text{F. Proof of Lemma 20}\]

Define \( x' = \frac{x}{4\Delta} \), hence for \( x \in [0, 4\Delta] \), \( x' \in [0, 1] \). It follows from the best polynomial approximation result for \(-x \ln x\) on \([0, 1]\) that there exists a constant \( d > 0 \) such that for all \( x' \in [0, 1] \),
\[
\left| \sum_{k=0}^{K} r_{k,H}(x')^k - (-x' \ln x') \right| \leq \frac{d}{K^2}. \tag{579}
\]
When \( n \) is sufficiently large, we could take \( d = \frac{\nu_1(2)}{2} \). Taking \( x' = 0 \), we have
\[
r_{0,H} \leq \frac{d}{K^2}, \tag{580}
\]
hence
\[
\left| \sum_{k=1}^{K} r_{k,H}(x')^k - (-x' \ln x') \right| \leq \frac{2d}{K^2}. \tag{581}
\]
Now, multiplying both sides by \( 4\Delta \), we have
\[
\left| \sum_{k=1}^{K} g_{k,H}(4\Delta)^{-k+1} x^k + x(\ln x - \ln(4\Delta)) \right| \leq \frac{2d(4\Delta)}{K^2}. \tag{582}
\]

Since we have defined \( g_{k,H} \) as
\[
g_{k,H} = r_{k,H}, 2 \leq k \leq K, \quad g_{1,H} = r_{1,H} - \ln(4\Delta), \tag{583}
\]
we have
\[
\left| \sum_{k=1}^{K} g_{k,H}(4\Delta)^{-k+1} x^k + x \ln x \right| \leq \frac{2d(4\Delta)}{K^2} = \frac{8dc_1}{c_2^2n \ln n} = \frac{C}{n \ln n}. \tag{584}
\]
When \( n \) is sufficiently large, we could replace \( d \) by \( \nu_1(2) / 2 \), hence obtain
\[
C = \frac{4c_1\nu_1(2)}{c_2^2}. \tag{585}
\]

\[\text{G. Proof of Lemma 22}\]

We know that if \( X \sim \text{Poi}(\lambda) \), then it follows from \([103]\) that
\[
\mathbb{E}X^k = \sum_{i=1}^{k} \lambda^i \left\{ \begin{array}{c} k \\ i \end{array} \right\}, \tag{586}
\]
where \( \left\{ \begin{array}{c} k \\ i \end{array} \right\} \) is the Stirling numbers of the second kind.
Using (371), we have

\[ \mathbb{E}X^k = \sum_{i=1}^{k} \binom{k}{i} \lambda^i \]  

\[ \leq \sum_{i=1}^{k} \binom{k}{i} \lambda^{k-i} \]  

\[ \leq \sum_{i=1}^{k} M^i \binom{k}{i} M^{k-i} \]  

\[ = M^k \sum_{i=1}^{k} \binom{k}{i} \]  

\[ \leq M^k 2^k \]  

\[ = (2M)^k. \]  

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] M. O. Hill, “Diversity and evenness: a unifying notation and its consequences,” Ecology, vol. 54, no. 2, pp. 427–432, 1973.

[4] F. Franchini, A. Its, and V. Korepin, “Rényi entropy of the XY spin chain,” Journal of Mathematical Physics, vol. 41, no. 2, p. 025302, 2008.

[5] I. Csiszár, “Generalized cutoff rates and Rényi’s information measures,” Information Theory, IEEE Transactions on, vol. 41, no. 1, pp. 26–34, 1995.

[6] 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.

[7] ———, “Variable-length lossy compression and channel coding: non-asymptotic converses via cumulant generating functions,” in Proc. of IEEE International Symposium on Information Theory (ISIT), 2014.

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

[9] G. Valiant and P. Valiant, “Estimating the unseen: n/j 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] Y. Wu and P. Yang, “Minimax rates of entropy estimation on large alphabets via best polynomial approximation,” arXiv preprint arXiv:1407.0381, 2014.

[11] 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.

[12] G. Valiant and P. Valiant, “The power of linear estimators,” in Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on. IEEE, 2011, pp. 403–412.

[13] 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.

[14] J. Jiao, K. Venkat, and T. Weissman, “Non-asymptotic theory for the plug-in rule in functional estimation,” available on arXiv.

[15] R. D. Gill and B. Y. Levit, “Applications of the van trees inequality: a bayesian cramér-rao bound.” Bernoulli, pp. 59–79, 1995.

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

[17] J. Hājek, “A characterization of limiting distributions of regular estimates,” Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, vol. 14, no. 4, pp. 323–330, 1970.

[18] ———, “Local asymptotic minimax and admissibility in estimation,” in Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, vol. 1, 1972, pp. 175–194.

[19] A. J. Wyner and D. Foster, “On the lower limits of entropy estimation,” IEEE Transactions on Information Theory, submitted for publication.

[20] J. H. van der Vaart, “On the lower limits of entropy estimation,” IEEE Transactions on Information Theory, vol. 41, no. 2, p. 429–443, 2003.

[21] C. O. Daub, R. Steuer, J. Selbig, and S. Kloska, “Estimating mutual information using B-spline functions—an improved similarity measure for analysing gene expression data,” BMC bioinformatics, vol. 5, no. 1, p. 118, 2004.

[22] G. J. Valiant, “Algorithmic approaches to statistical questions,” Ph.D. dissertation, University of California, 2012.

[23] G. Pólya, How to solve it: A new aspect of mathematical method. Princeton university press, 2014.

[24] C. Stein et al., “Inadmissibility of the usual estimator for the mean of a multivariate normal distribution,” in Proceedings of the Third Berkeley symposium on mathematical statistics and probability, vol. 1, no. 399, 1956, pp. 197–206.

[25] D. L. Donoho and J. M. Johnstone, “Minimax risk over $\ell_p$-balls for $\ell_q$-error,” Probability Theory and Related Fields, vol. 99, pp. 277–303, 1994.

[26] ———, “Ideal spatial adaptation by wavelet shrinkage,” Biometrika, vol. 81, no. 3, pp. 425–455, 1994.

[27] E. J. Candès and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?” Information Theory, IEEE Transactions on, vol. 52, no. 12, pp. 5406–5425, 2006.

[28] D. L. Donoho, “Compressed sensing,” Information Theory, IEEE Transactions on, vol. 52, no. 4, pp. 1289–1306, 2006.
[86] Q. Wang, S. R. Kulkarni, and S. Verdú, “Universal estimation of information measures for analog sources,” Foundations and Trends in Communications and Information Theory, vol. 5, no. 3, pp. 265–353, 2009.

[87] A. Orlitisky, N. P. Santhanam, and J. Zhang, “Universal compression of memoryless sources over unknown alphabets,” Information Theory, IEEE Transactions on, vol. 50, no. 7, pp. 1469–1481, 2004.

[88] 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.

[89] 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.

[90] 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.

[91] A. Tsybakov, “Aggregation and high-dimensional statistics,” Lecture notes for the course given at the École dété de Probabilités in Saint-Flour, URL http://www.crest.fr/cfkfinder/userfiles/files/Pageperso/ATsybakov/Lecture_notes_SFlour.pdf, vol. 16, p. 20, 2013.

[92] ——, Introduction to Nonparametric Estimation. Springer-Verlag, 2002.

[93] T. Lindvall, Lectures on the coupling method. Courier Dover Publications, 2002.

[94] B. Efron and R. J. Tibshirani, An introduction to the bootstrap. CRC press, 1994, vol. 57.

[95] B. Efron, “Bootstrap methods: another look at the Jackknife,” The Annals of Statistics, pp. 1–26, 1979.

[96] A. Javanmard and A. Montanari, “Confidence intervals and hypothesis testing for high-dimensional regression,” arXiv preprint arXiv:1306.3171, 2013.

[97] R. S. Varga and A. J. Carpenter, “On the Bernstein conjecture in approximation theory,” Constructive Approximation, vol. 1, no. 1, pp. 333–348, 1985.

[98] R. Rennie and A. Dobson, “On Stirling numbers of the second kind,” Journal of Combinatorial Theory, vol. 7, no. 2, pp. 116–121, 1969.

[99] M. Mitzenmacher and E. Upfal, Probability and computing: Randomized algorithms and probabilistic analysis. Cambridge University Press, 2005.

[100] S. Bernstein, “Collected works: Vol 1. constructive theory of functions (1905-1930), English translation,” Atomic Energy Commission, Springfield, Va, 1958.

[101] M. Qazi and Q. Rahman, “Some coefficient estimates for polynomials on the unit interval,” Serdica Math. J, vol. 33, pp. 449–474, 2007.

[102] J. Riordan, “Moment recurrence relations for binomial, Poisson and hypergeometric frequency distributions,” The Annals of Mathematical Statistics, vol. 8, no. 2, pp. 103–111, 1937.