Estimation of Large Zipfian Distributions with Sort and Snap

Peter Matthew Jacobs[1,4], Anirban Bhattacharya[2], Debdeep Pati[3]
Lekha Patel[1,4], Jeff M. Phillips[1]
[1]University of Utah [2]Texas A&M University
[3]University of Wisconsin [4]Sandia National Laboratories

Abstract

We study the estimation of Zipfian distributions under L_1 loss, and provide near minimax optimal bounds in several regimes. Specifically, we assume observations arrive from a known alphabet, and with a known decay rate parametrizing the Zipfian, but we do not know a priori which alphabet elements have larger probability than others. We present a novel Sort and Snap estimator, which uses the empirical proportions to sort the alphabet, and then snaps them to the associated term from the Zipfian distribution. For arbitrary decay rates and smaller alphabet sizes, as well as for large decay rates and large alphabet sizes, we show an exact or minor variant of this estimator is near minimax optimal and has exponential improvement over the standard empirical proportion estimator. However, for small decay rates and larger alphabet sizes a simulation study indicates the standard empirical proportion estimator is competitive with Sort and Snap procedures. In addition to providing nearly tight bounds for important high-dimensional estimation problems, we believe the Sort and Snap estimator, and its analysis, will have independent interest.

1 INTRODUCTION

A increasingly common task in data analysis is collecting, measuring, and estimating complex and discrete data. We consider data drawn from an alphabet $[k] = \{1, 2, \ldots, k\}$, which may model IP addresses, words or subwords in natural language, cities among addresses, or first names among people. What is common is that some values are much more common than others, but the domain size k is extremely large and may be (almost) endless. We aim to model this with a discrete probability distribution of dimension k, denoted $\boldsymbol{p} = (\boldsymbol{p}(1), \boldsymbol{p}(2), \ldots, \boldsymbol{p}(k))$. The most common discrete probability distribution for modeling such patterns is the s–Zipfian, which holds if there exists a permutation function π such that for each $j \in [k]$

$$\boldsymbol{p}(\pi(j)) = \frac{j^{-s}}{\sum_{t=1}^{k} t^{-s}}.$$

Zipfian probability distributions arise naturally in natural (Dahui et al., 2005; Li, 1992; Piantadosi, 2014; Ferrer-i Cancho, 2005, 2016) and AI-generated (Diamond, 2023) languages as the proportion of words or subwords within a language or a given corpus of text. And the alphabet sizes can be very large, with over 200,000 words in English and over 80,000 sinograms in Chinese (Han et al., 2015).

We consider the problem of estimating the distribution $\boldsymbol{p}$ from a series of n samples drawn independently from a k-dimensional s-Zipfian distribution. While there are many reasonable choices for a loss function, the choice of L_1 loss is both common and well motivated in the study of estimation of high dimensional probability distributions (see Han et al. (2015); Cohen et al. (2020)): if μ, ν are two discrete distributions with a common size k alphabet, and $L_1(\mu, \nu) = \epsilon$, then for any subgroup of words A, $|\mu(A) - \nu(A)| \lesssim \epsilon$. The other extremely commonly used loss for discrete distribution estimation, Kullback-Liebler (KL), (see Paninski (2004); Orlitsky and Suresh (2015); Falahatgar et al. (2017)) only offers that if $KL(P, Q) = \epsilon$, $|\mu(A) - \nu(A)| \lesssim \sqrt{\epsilon}$. In our analysis we model the alphabet size $k = \lfloor n^{\beta} \rfloor$, as a function of n, parameterized by some $\beta > 0$; what Han et al. (2015) call the *High Dimensional Asymptotics* setting.

For an LLM example, Meta's Llama 3.0 has a vocabulary of $k = 128{,}000$ tokens, and is trained on

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

n = 15,000,000,000 (15 trillion) tokens. For data of this magnitude, can we learn the distribution of these tokens? If we know the distribution is s-Zipfian, can we improve upon the empirical estimate? If we considered a very small context size of say 2, so a vocabulary size of k^2, how does that effect our ability to learn this distribution directly, including if we know it is still s-Zipfian? While modeling Zipfian properties in learning has proven important (Chan et al., 2022a,b), these basic questions are still not well understood in the context of such high-vocabulary Zipfian distributions, especially under L_1 loss. We provide numerous results towards understanding these questions.

Prior Work The default way to estimate a discrete distribution is the *Empirical Proportion Estimator* (EPE), which provides the sample estimate for each index. The study of tight estimates of the expected absolute deviation of a binomial distribution from its mean performed by Berend and Kontorovich (2013) provides a general tool for finding the risk of the EPE under L_1 when sufficient information about the distribution is known. Cohen et al. (2020) develop a data-dependent bound for the EPE under L_1 loss suitable for use in unconstrained estimation settings. Han et al. (2015) shows that the EPE is minimax optimal for estimating a completely unconstrained probability distribution in the L_1 norm in the High Dimensional Asymptotics setting provided $\beta < 1$; when $\beta \geq 1$ (so $k \geq n$), consistent estimation is not necessarily possible in the unconstrained setting. What this work silently leaves open is if we can do better if we know the data comes from an s-Zipfian distribution. Falahatgar et al. (2017) addresses this question when the distance of interest is Kullback-Leibler (KL) and k grows at least as quickly as n and $s > 1$; in this case they show that a modification of the EPE, known as Absolute Discounting, is minimax optimal for estimating the permutation ordering of the s-Zipfian distribution under KL distance. But they do not delineate whether for more slowly increasing k (i.e $\beta < 1$), if leveraging that the probability values are known can yield estimators with improved convergence rates. And they do not present any results regarding minimax optimality under L_1 loss even in the $\beta \geq 1$ case, nor do they study $s \leq 1$ under any circumstances.

The parameter s indexes a family of distributions that have the same probability values but different order – a permutation invariant class. A question of concern in Competitive Distribution Estimation (Orlitsky and Suresh, 2015) regarding a permutation invariant class is: what is the difference between the worst case risk of a minimax estimator for this class and the worst case risk of an estimator q_n operating on this class without knowledge of the probability values? This is the *competitive regret* of q_n for the permutation invariant class. Orlitsky and Suresh (2015) prove under KL loss that regardless of the permutation invariant class considered, a variant of Good-Turing achieves competitive regret not asymptotically larger than $n^{-\frac{1}{2}}$. Hao and Orlitsky (2019b), still working with KL loss, improve this with an estimator that achieves worst case asymptotic regret of $\min(n^{-\frac{s}{s+1}}, n^{-\frac{1}{2}})$ specifically for $s > 0$ Zipfian permutation invariant classes (regardless of dimension) while also guaranteeing strong competitive regret under a wide variety of different permutation classes. However, the work on rigorous competitive regret analysis under L_1 is less well developed. Valiant and Valiant (2016) study this quantity for general permutation invariant classes, and show the existence of an estimator achieving the much weaker, asymptotically $poly(\frac{1}{\log(n)})$ magnitude regret.

There are various other works focusing on the unconstrained large discrete distribution estimation problem. Canonne et al. (2023) provides concentration of measure statements for the Laplace estimator under the KL loss. Kontorovich and Painsky (2024) provides a data dependent bound for the EPE loss under L_∞. Painsky (2023) develops confidence intervals for the distribution.

Several authors have studied estimation of the decay parameter s in low dimensional Zipfian laws (Izsák, 2006; Zörnig and Altmann, 1995), while authors (Hsu et al., 2019; Aamand et al., 2019) in the sketching communities have designed small space estimators for streams from a s-Zipfian distribution.

Our Contributions Our first main contribution is the introduction of the Sort and Snap estimator for estimating high-dimensional discrete distributions. This starts with the EPE estimator for each $j \in [k]$, and then sorts these estimated values. It uses this sorted order to estimate π, and then updates the estimates to that of the known s-Zipfian distribution, in a thresholding step we refer to as snapping. We also introduce the Truncated Sort and Snap estimator, which applies Sort and Snap to the largest estimates, but retains the EPE for the smaller values.

Second we provide new upper bounds under L_1 error for the Sort and Snap estimators. Notably, there are two regimes where Sort and Snap outperforms the EPE estimator. The first is when $\beta < \frac{1}{B(s)}$ where $B(s) = 2 + \max(s, 1)$; i.e the growth rate of $k = \lfloor n^\beta \rfloor$ is not too fast, so the alphabet is not too large (Theorem 4.2). *This surprisingly achieves exponential convergence!* The second is for Truncated Sort and Snap when $s > 2$, so the rate of decay is very fast (Theorem 4.3). Comparison with the EPE requires providing a tight characterization of its performance, which we

also derive for s-Zipfian distributions (Theorem 4.1).

Third, we provide minimax lower bounds for all these regimes. We show in the $\beta < \frac{1}{B(s)}$ regime that Sort and Snap matches the polynomial in the fastest decaying term of the lower bound precisely (Theorem 4.2) and matches the constant in the fastest decaying term of the lower bound up to a factor 16. In the $s > 2, \beta \geq \frac{1}{s+2}$ setting, we provide a minimax lower bound of $n^{-\frac{s+1}{s+2}}$ almost matching our $n^{-\frac{s}{s+2}+\tau}$ upper bound achieved by Truncated Sort and Snap (Theorem 4.3) where $\tau > 0$ is an arbitrarily small constant.

In all, this work elucidates a gap in the understanding of estimating structured discrete distributions, and introduces the Sort and Snap mechanism to close it. Before it was not possible to provide tight minimax bounds for such structured distributions in several growing k regimes, and our work explains why the standard EPE estimator fell short in the L_1 norm. This work also opens up future directions. One such direction is determining the lowest possible competitive regret in the Zipfian setting under L_1. Our results aid in this pursuit because we have characterized the s known near minimax rate in several important regimes and the EPE performance has been fully characterized; therefore EPE competitive regret calculations can be performed from our results in several cases. Other new directions include tightening a few intermediate regimes and in showing the more general applicability of this approach of snapping to one of a discrete set of options from a function class.

2 NOTATION AND DEFINITIONS

Let $\mathcal{S}_k$ denote the $k-1$ dimensional probability simplex for $k \geq 2$ and let $[k] = \{1, 2, \ldots, k\}$. For $\boldsymbol{p} \in \mathcal{S}_k$, we denote the probability of the j^{th} category as $\boldsymbol{p}(j)$. We will be studying probability *patterns* as k grows. Specifically, let

$$\mathcal{F} := \{f : \mathbb{R}^+ \to \mathbb{R}^+; \;\; f \text{ is monotonically decreasing } \}$$

and for $k \in \{1, 2, \ldots, \}$, let $\mathcal{M}_k$ be the permutation functions on $[k]$. We refer to families of probability distributions at a given dimension $k \geq 2$ generated by f as

$$\mathcal{P}_{f,k} = \{\boldsymbol{p} : \boldsymbol{p} \in \mathcal{S}_k, \exists \pi \in \mathcal{M}_k \text{ s.t } \boldsymbol{p}(j) = \frac{f(\pi(j))}{\sum_{j=1}^k f(\pi(j))}\} \tag{1}$$

This paper focuses on s-Zipfian probability patterns. That is, for $s > 0$, we study

$$f_s(x) = x^{-s}.$$

Let $H_{k,s} := \sum_{j=1}^k j^{-s}$ and for $s > 1$, $R(s) = \lim_{k\to\infty} H_{k,s}$ is the Riemann-Zeta function.

For $f \in \mathcal{F}$ and $\boldsymbol{p}^k \in \mathcal{P}_{f,k}$ (with permutation function π), consider n samples

$$Y_{1k}, Y_{2k}, \ldots, Y_{nk} \overset{iid}{\sim} \boldsymbol{p}^k.$$

That is, for $i \in \{1, 2, \ldots, n\}$, and $j \in \{1, 2, \ldots, k\}$, $\mathbb{P}(Y_{ik} = j) = \frac{f(\pi(j))}{\sum_{i=1}^k f(\pi(i))}$. For $j \in \{1, 2, \ldots, k\}$, let

$$X_{jk} = \sum_{i=1}^n \mathbb{I}(Y_{ik} = j) \tag{2}$$

be the observed count of the jth element.

Define the *inverse of the intermediate dimension*

$$B(s) := 2 + \max(1, s) = s + 2 + \max(0, 1 - s) \tag{3}$$

This will be an important function used in defining k, as a function of n, to place boundaries on where our analysis can provide significant improvement for Sort and Snap over the EPE estimator.

$X \sim Multinomial(n, \boldsymbol{p} = (p_1, p_2, \ldots, p_k))$ means X has the probability mass function resulting from dropping n balls into bins labeled $1, 2, \ldots, k$ where each ball is dropped independently and the probability a ball falls in bin j is p_j.

For functions g, f defined on all sufficiently large integers, the notation $g(n) \lesssim f(n)$ means there exists a constant $C > 0$ and a N such that for $n \geq N, g(n) \leq Cf(n)$. The notation $g(n) = o(f(n))$ means $\lim_{n\to\infty} \frac{g(n)}{f(n)} = 0$. The notation $g(n) \asymp f(n)$ means there exists a constant $C \in \mathbb{R}$ such that $\lim_{n\to\infty} \frac{g(n)}{f(n)} = C$. Note that while $k = k(n)$ will always be a function of the sample size parameter n, we often write k in place of $k(n)$.

3 SORT AND SNAP ESTIMATORS

We introduce the ***Sort and Snap*** (SS) estimator. For $f \in \mathcal{F}$, the sorted order of the counts by category are determined; then the j^{th} largest probability in laws from $\mathcal{P}_{f,k}$, $\frac{f(j)}{\sum_{i=1}^k f(i)}$, is snapped on to one category (index) with the j^{th} largest count. More formally, let $\hat{\pi} \in \mathcal{M}_k$ satisfy that index $\hat{\pi}(j)$ gives the j^{th} largest count. That is,

$$X_{\hat{\pi}(1)k} \geq X_{\hat{\pi}(2)k} \geq \cdots \geq X_{\hat{\pi}(k)k}.$$

Note $\hat{\pi}$ is not unique because of possible ties. Then the Full Sort and Snap Estimator is

$$\hat{\boldsymbol{p}}_n^k(\hat{\pi}(j)) := \frac{f(j)}{\sum_{j=1}^k f(j)}. \tag{4}$$

Next we formalize the EPE as

$$\bar{\boldsymbol{p}}_n^k(j) := \frac{X_{jk}}{n}.$$

Now the ***Truncated Sort and Snap Estimator*** (TSS) is defined as follows which uses Sort and Snap for the common categories (the head), and EPE for the uncommon (the tail). For a truncation threshold $T \in \{1, 2, \dots, \}$ TSS is defined

$$\hat{\boldsymbol{p}}^k_{n,T}(j) = \begin{cases} \hat{\boldsymbol{p}}^k_n(j) & j \in \{\hat{\pi}(1), \dots, \hat{\pi}(\min(T,k))\} \\ \bar{\boldsymbol{p}}^k_n(j) & j \in [k] - \{\hat{\pi}(1), \dots, \hat{\pi}(\min(T,k))\} \end{cases} \tag{5}$$

While these estimators are not uniquely defined (because a procedure for resolving ties is not specified), the analyses we provide for Sort and Snap procedures does not depend on the way ties are resolved.

4 ERROR BOUNDS

All mathematical results are rigorously stated in this section; full proofs are deferred to the appendix, but an overview of main analysis ideas are presented in Section 5. A pictorial summary of our results on Sort and Snap, Truncated Sorted Snap, the EPE, and minimax lower bounds is displayed in Figure 1 as a function of parameters β and s. The main takeaways are that (a) for $\beta < \frac{1}{B(s)}$ SS's exponential decay beats EPE, and is nearly tight to the minimax lower bound; (b) for $s > 2$ TSS improves upon EPE and approaches the minimax lower bound as s grows; and (c) otherwise, we leave as an open problem if a TSS or SS variant can improve upon EPE.

We first state upper and lower bounds for the performance of the EPE in estimating growing, exact s-Zipfian probability laws.

Theorem 4.1 (EPE for Exact s-Zipfian Laws)**.** *Suppose $s > 0$. Further let $0 < \beta < \infty$, $\boldsymbol{p}^k \in \mathcal{P}_{f_s,k}$ and $k = \lfloor n^\beta \rfloor$ for n sufficiently large so that $\lfloor n^\beta \rfloor \geq 2$. Also,*

$$Y_{1k}, Y_{2k}, \dots, Y_{nk} \overset{iid}{\sim} \boldsymbol{p}^k$$

and $\bar{\boldsymbol{p}}^k_n$ is the corresponding EPE. Then

$$\mathbb{E}\|\bar{\boldsymbol{p}}^k_n - \boldsymbol{p}^k\|_1 \asymp \begin{cases} n^{-\frac{1}{2}(1-\beta)} & 0 < s < 1, \beta < 1 \\ 1 & 0 < s < 1, \beta \geq 1 \\ \frac{n^{-\frac{1}{2}(1-\beta)}}{\sqrt{\log(n)}} & s = 1, \beta < 1 \\ \frac{\log(\log(n))}{\log(n)} & s = 1, \beta = 1 \\ 1 & s = 1, \beta > 1 \\ n^{-\frac{1}{2}+\beta(1-\frac{s}{2})} & 1 < s < 2, \beta < \frac{1}{s} \\ n^{\frac{1}{s}-1} & 1 < s < 2, \beta \geq \frac{1}{s} \\ \frac{\log(n)}{n^{1/2}} & s = 2 \\ n^{-\frac{1}{2}} & s > 2 \end{cases}$$

In estimation of a completely unconstrained $\boldsymbol{p}^k$ in the simplex, Han et al. (2015) shows that in the L_1 loss, the EPE is minimax optimal. But in the constrained problem of estimating a probability vector in $\mathcal{P}_{f_s,k}$, this is not always the case.

Sort And Snap Bounds The next theorem, one of the main contributions of this work, shows that provided the growth rate of k as a function of n is *not too fast*, then full Sort and Snap, but not the EPE, is a minimax optimal estimator (at least with respect to the polynomial in n and constants depending only on s and β inside the exponential). Before stating the theorem, we introduce a useful constant

$$C^*_{s,\beta} := s^2 \begin{cases} 1 - s & 0 < s < 1 \\ \frac{1}{\beta} & s = 1 \\ \frac{1}{R(s)} & s > 1. \end{cases} \tag{6}$$

Theorem 4.2 (SS Upper and Lower Bounds)**.** *Suppose $s > 0$, $0 < \beta < \frac{1}{B(s)}$ and for each n large enough so that $\lfloor n^\beta \rfloor \geq 2$, $k := \lfloor n^\beta \rfloor$ and $0 < \tau < 1 - \beta B(s)$. Then the Full Sort and Snap Estimator $\hat{\boldsymbol{p}}^k_n$ achieves that* $\sup_{\boldsymbol{p} \in \mathcal{P}_{f_s,k}} \mathbb{E}\|\hat{\boldsymbol{p}}^k_n - \boldsymbol{p}\|_1$

$$\lesssim \frac{n^{-\beta(B(s)-2)}}{\log^{\mathbb{I}(s=1)}(n)} \exp\left(-\frac{C^*_{s,\beta}(1-\tau)}{16} \frac{n^{1-\beta B(s)}}{\log^{\mathbb{I}(s=1)}(n)}\right)$$

when $0 < \beta < \frac{1}{B(s)+1}$ and $\sup_{\boldsymbol{p} \in \mathcal{P}_{f_s,k}} \mathbb{E}\|\hat{\boldsymbol{p}}^k_n - \boldsymbol{p}\|_1$

$$\lesssim \frac{n^{-\beta(B(s)-2)}}{\log^{\mathbb{I}(s=1)}(n)} h_{n,s,\beta} \exp\left(-\frac{C^*_{s,\beta}(1-\tau)}{16} \frac{n^{1-\beta B(s)}}{\log^{\mathbb{I}(s=1)}(n)}\right)$$

when $\beta < \frac{1}{B(s)}$ and $h_{n,s,\beta} = n^{2(\beta(B(s)+1)-(1-\tau))}$. Also, $\inf_{\hat{\mu}_n} \sup_{\boldsymbol{p} \in \mathcal{P}_{f_s,k}} \mathbb{E}\|\hat{\mu}_n - \boldsymbol{p}\|_1$

$$\gtrsim \frac{n^{-\beta(B(s)-1)}}{\log^{\mathbb{I}(s=1)}(n)} \exp\left(-C^*_{s,\beta}(1+\tau) \frac{n^{1-\beta B(s)}}{\log^{\mathbb{I}(s=1)}(n)}\right)$$

where $\mathcal{P}_{f_s,k}$ is the collection of all k–dimensional s–Zipifan probability distributions. And the inf *is taken over all functions of the n samples $Y_1, Y_2, \dots, Y_n \overset{iid}{\sim} \boldsymbol{p}$.*

Full Sort and Snap thus clearly outperforms the EPE in the $\beta < \frac{1}{B(s)}$ growth setting, because according to Theorem 4.1, in no regimes does the EPE achieve better than polynomial decay. Note regarding the minimax optimality of Full Sort and Snap that the polynomial inside the exponential is matched precisely in Theorem 4.2 and the constant in the exponential is mismatched only by a factor 16.

Truncated Sort And Snap Bounds The next theorem shows that a Truncation of Sort and Snap can outperform the EPE even when k grows quickly with n, so long as $s > 2$; recall for every $s > 2$, the EPE never outperforms $n^{-\frac{1}{2}}$ and for $s > 2$, $\frac{s}{s+2} > \frac{1}{2}$.

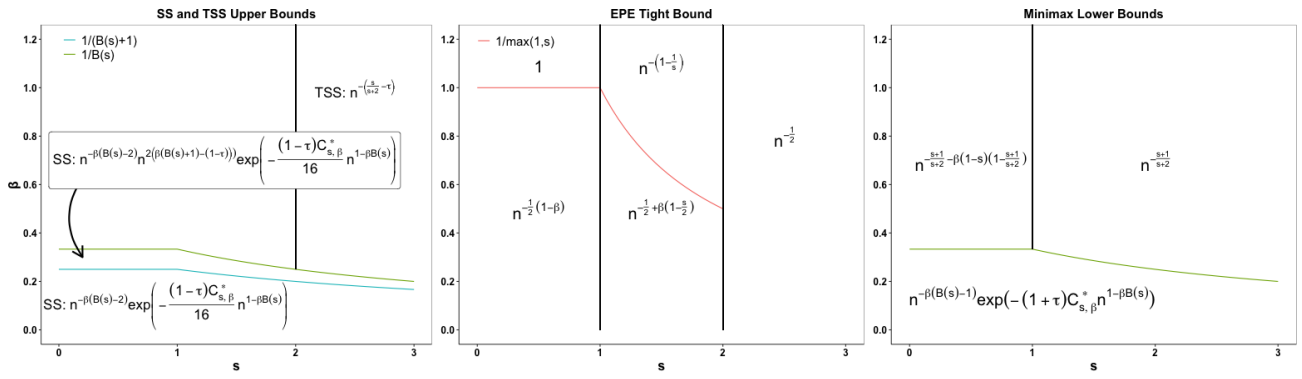

Figure 1: Bounds on the error rates as a function of n, plotted in (s, β) parameter space. The left panel shows upper bounds for SS and TSS, the middle tight bounds for EPE, and the right minimax lower bounds. The region separating curves are $\frac{1}{B(s)}$, $\frac{1}{(B(s)+1)}$, and $\frac{1}{\max(1,s)}$. Behavior at curve boundaries not displayed.

Theorem 4.3. *(TSS Upper and Lower Bounds) Suppose $s > 2$, $\beta \geq \frac{1}{B(s)}$ and $k = \lfloor n^\beta \rfloor$ for each n large enough so that $\lfloor n^\beta \rfloor \geq 2$. And let $T_{n,\epsilon} = I(k, 1, \epsilon) - 1$ (where $I(k, 1, \epsilon)$ is defined in equation 92). Then there exists $U_{s,\beta,1}, U_{s,\beta,2} > 0$ such that for any $0 < \epsilon < U_{s,\beta,1}$ and $0 < \tau < U_{s,\beta,2}$*

$$\sup_{\boldsymbol{p} \in \mathcal{P}_{f_s,k}} \mathbb{E}\|\hat{\boldsymbol{p}}^k_{n,T_{n,\epsilon}} - \boldsymbol{p}\|_1 \lesssim n^{-\frac{s}{s+2}} n^\tau \tag{7}$$

Also, for any $s > 0$ and $\beta \geq \frac{1}{B(s)}$, $\inf_{\hat{\mu}_n} \sup_{\boldsymbol{p} \in \mathcal{P}_{f_s,k}} \mathbb{E}\|\hat{\mu}_n - \boldsymbol{p}\|_1$

$$\gtrsim n^{-\frac{s+1}{s+2} - \beta \max(0,1-s)(1-\frac{s+1}{s+2})} \log^{-\mathbb{I}(s=1)\frac{1}{3}}(n) \tag{8}$$

where again the inf *is taken over all functions of the n samples $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \boldsymbol{p}$.*

Note that in Theorem 4.3, the polynomial in the lower and upper bound mismatch by a power $\frac{1}{s+2}$. Thus TSS is, up to an arbitrarily slowly increasing polynomial term, near-minimax-optimal for large s; but for s near 2 the power gap is near $\frac{1}{4}$; resolving this gap remains an open problem.

While Sort and Snap outperforms the EPE for any $s > 0$ Zipfian law that doesn't grow *too quickly* in dimension (Theorem 4.2) and for any $s > 2$ law even if the dimension grows quickly (Theorem 4.3), we do not discover a benefit to using Sort and Snap for long, flat Zipfian laws. That is, for $0 < s < 2$ and $\beta \geq \frac{1}{B(s)}$. In this case, the best upper bound we know of comes from the EPE, and our best lower bound is $n^{-\frac{s+1}{s+2}}$.

5 PROOF SKETCHES

Full proofs are contained in the appendices; here we will illustrate central ideas of Theorems 4.1, 4.2 and 4.3. The restriction to $s > 1$ in this section ensures $H_{k,s}$ is convergent, which slightly simplifies the proofs, but in the appendix we show how to handle any $s > 0$ by quantifying the growth rate of $H_{k,s}$.

5.1 On The Tight Performance Of The EPE

As discussed in Cohen et al. (2020), for $k \geq 2, \boldsymbol{p}^k \in \mathcal{P}_{f,k}$ and $Y_{1k}, Y_{2k}, \ldots, Y_{nk} \overset{iid}{\sim} \boldsymbol{p}^k$ and $\bar{\boldsymbol{p}}^k_n$ the corresponding EPE, Berend and Kontorovich (2013) shows that there exists a number $\Lambda_{n,k}(f)$ such that

$$\frac{1}{2}\Lambda_{n,k}(f) - \frac{1}{2}\frac{1}{\sqrt{n}} \leq \mathbb{E}\|\bar{\boldsymbol{p}}^k_n - \boldsymbol{p}^k\|_1 \leq 2\Lambda_{n,k}(f)$$

$\Lambda_{n,k}(f)$ is precisely the sum of the probabilities in the k dimensional probability law generated by f that are less than $\frac{1}{n}$ plus $\frac{1}{2\sqrt{n}}$ times the sum of the square roots of all the other probabilities in the k dimensional law generated by f. In the Zipfian case, $f \equiv f_s$ and the order of Λ depends on β and s; here we illustrate one case.

Lemma 5.1. *If $\beta > \frac{1}{s}$ and $k \geq 2$, $\boldsymbol{p}^k \in \mathcal{P}_{f_s,k}$ with permutation function π_k and $k = \lfloor n^\beta \rfloor$ then*

$$\mathbb{E}\|\bar{\boldsymbol{p}}^k_n - \boldsymbol{p}^k\|_1 \asymp n^{\frac{1}{s}-1} \text{ when } 1 < s < 2$$

Proof. $\frac{j^{-s}}{H_{k,s}} < n^{-1}$ if and only if $j > \left(\frac{n}{H_{k,s}}\right)^{\frac{1}{s}}$. Thus since $\beta > \frac{1}{s}$.

$$\Lambda_{n,k}(f_s) = \frac{1}{H_{k,s}} \sum_{j=\lfloor (\frac{n}{H_{k,s}})^{\frac{1}{s}} \rfloor + 1}^{k} j^{-s} + \frac{1}{2\sqrt{nH_{k,s}}} \sum_{j=1}^{\lfloor (\frac{n}{H_{k,s}})^{\frac{1}{s}} \rfloor} j^{-\frac{s}{2}}$$

Using Riemann integration to upper and lower bound the first sum and that $H_{k,s}$ converges when $s > 1$, we

have that $\sum_{j=\lfloor(\frac{n}{H_{k,s}})^{1/s}\rfloor+1}^{k} j^{-s} \asymp n^{\frac{1}{s}-1}$ when $s > 1$. By another Riemann integration argument, when $1 < s < 2$, we have that $\frac{1}{2\sqrt{nH_{k,s}}} \sum_{j=1}^{\lfloor(\frac{n}{H_{k,s}})^{\frac{1}{s}}\rfloor} j^{-\frac{s}{2}} \asymp n^{\frac{1}{s}-1}$. Thus when $1 < s < 2$, we conclude $\Lambda_{n,k}(f_s) \asymp n^{\frac{1}{s}-1}$; the lemma conclusion follows. □

In this example, the order of the two sums is the same. This happens to be the case when $1 < s < 2$ and $\beta > \frac{1}{s}$, but there are other arrangements of the values of β, s where this is not so.

5.2 A Minimax Lower Bound For Small β

Lemma 5.2. *If $0 < \beta < \frac{1}{B(s)}$ and $s > 1$, then for any $\tau > 0$*

$$\inf_{\hat{\mu}_n} \sup_{\boldsymbol{p}\in\mathcal{P}_{f_s,k}} \mathbb{E}\|\hat{\mu}_n - \boldsymbol{p}\|_1 \gtrsim$$

$$\frac{n^{-\beta(B(s)-1)}}{\log^{\mathbb{I}(s=1)}(n)} \exp\left(-C^*_{s,\beta}(1+\tau)\frac{n^{1-\beta B(s)}}{\log^{\mathbb{I}(s=1)}(n)}\right)$$

where the inf *is taken over all functions of the n samples $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \boldsymbol{p}$*

Proof. We use the Le Cam method (Tibshirani and Wasserman, 2017). Set P_{0n} to be the k dimensional Zipfian law with probabilities in decreasing order. Set P_{1n} to P_{0n} on the first $k-2$ indices, with the last 2 probabilities flipped. By Theorem 4 of Tibshirani and Wasserman (2017), the minimax lower bound is

$$\gtrsim \|P_{0n} - P_{1n}\|_1 \exp\left(-nKL(P_{0n}, P_{1n})\right).$$

Thus we need to lower bound $KL(P_{0n}, P_{1n})$ and upper bound $\|P_{0n} - P_{1n}\|_1$. First we compute the two distances. This is simple because only the last 2 entries have been flipped.

$$\begin{aligned}\|P_{0n} - P_{1n}\|_1 &= \frac{2}{H_{k,s}}|(k-1)^{-s} - k^{-s}| \\ &= \frac{2}{H_{k,s}}|f_s(k-1) - f_s(k)|\end{aligned}$$

And

$$KL(P_{0n}, P_{1n}) =$$
$$\frac{(k-1)^{-s}}{H_{k,s}}\log(\frac{(k-1)^{-s}}{k^{-s}}) + \frac{k^{-s}}{H_{k,s}}\log(\frac{k^{-s}}{(k-1)^{-s}}) =$$
$$\frac{s}{H_{k,s}}\log(\frac{k}{k-1})|f_s(k-1) - f_s(k)|.$$

Since $f_s(k-1) - f_s(k)$ occurs in both distances we need control of it from above and below. Using a two term Taylor series expansion of $f_s(k-1)$ about $f_s(k)$, and that f_s is convex so its second derivative is nonnegative, we have for some $\phi \in [k-1, k]$,

$$\begin{aligned}sf_{s+1}(k-1) &\le f_s(k-1) - f_s(k) \\ &= sf_{s+1}(k-1) + \frac{s(s+1)}{2}f_{s+2}(\phi).\end{aligned} \tag{9}$$

To handle the f_{s+2} term, we use that f_{s+2} monotonically decreases. Therefore $f_{s+2}(\phi) \le f_{s+2}(k-1)$. Now note $f_{s+2}(k-1) = o(f_s(k-1))$. Therefore, for any $\tau > 0$, there is an N so that for $n \ge N$

$$sf_{s+1}(k-1) \le f_s(k-1) - f_s(k) \le s(1+\tau)f_{s+1}(k-1) \tag{10}$$

We also need control over $H_{k,s}$ from above and below since it occurs in both distances. Since $s > 1$, $H_{k,s} \to R(s)$ (the Riemann-Zeta function). Thus for any $\tau > 0$ and N sufficiently large,

$$(1-\tau)R(s) \le H_{k,s} \le R(s)$$

The last term amongst the two distances that we have not yet controlled is $\log(\frac{k}{k-1})$. It occurs only in the numerator $KL(P_{0n}, P_{1n})$ so we just need an upper bound for it. By another Taylor series expansion, for any $\tau > 0$, there is some N such that for $n \ge N$

$$\log(\frac{k}{k-1}) \le \frac{1}{k}(1+\tau).$$

Finally, to upper bound the $KL(P_{0n}, P_{1n})$, we use the expression for KL and plug in the upper bounds for $f_s(k-1) - f_s(k)$ and $\log(\frac{k}{k-1})$ and the lower bound for $H_{k,s}$. In the KL upper bound, there is a factor above 1 and arbitrarily close to 1, $\frac{(1+\tau)^2}{1-\tau}$ appearing for $\tau > 0$. Call this $1+\tau$ for $\tau > 0$ arbitrarily small. Likewise, we use the lower bound for $f_s(k-1) - f_s(k)$ and upper bound for $H_{k,s}$ and the expression for $\|P_{0n} - P_{1n}\|_1$ to lower bound this quantity. □

5.3 Sort And Snap Upper Bound For Small β

Here we describe an approach to achieve the optimal exponential decay (up to factor 16) when $\beta < \frac{1}{B(s)}$ and $s > 1$ without achieving the optimal polynomial outside the exponential. The tighter approach in the appendix improves the lower order polynomial term in front of the exponential (and also handles $0 < s \le 1$).

Lemma 5.3. *If $\beta < \frac{1}{B(s)}$ and $s > 1$ and $k = \lfloor n^\beta \rfloor$ and $\boldsymbol{p}^k \in \mathcal{P}_{f_s,k}$ with permutation function π_k, there exists a $U_{s,\beta,3} > 0$ such that for each $0 < \tau < U_{s,\beta,3}$*

$$\mathbb{E}\|\hat{\boldsymbol{p}}^k_n - \boldsymbol{p}^k\|_1 \lesssim n^\beta \exp\left(-(1-\tau)\frac{C^*_{s,\beta}}{16}n^{1-\beta B(s)}\right)$$

Proof. We start by defining for $i \in [k]$ a radius $r_{n,i}$ so that if the distance between the probability associated with the i^{th} largest category and its respective empirical proportion is less than $r_{n,i}$ for each category $i \in [k]$, then Sort and Snap yields zero error. Denote this as event A_n. Decomposing the expectation and using an upper bound on 1–norm implies $\mathbb{E}\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1 \leq 2\mathbb{P}(A_n^C)$.

It is sufficient to set $r_{n,i}$ smaller than $\frac{1}{2}$ of the distance between i^{th} largest probability and the nearest adjacent probability. We will show the following suffices:

$$r_{n,i} := \chi_n \sqrt{\boldsymbol{p}^k(\pi_k^{-1}(i))(1-\boldsymbol{p}^k(\pi_k^{-1}(i))}$$

where $\chi_n := (1-\tau)\frac{1}{2}n^{-\frac{\beta B(s)}{2}}\sqrt{C^*_{s,\beta}}$ for $0 < \tau < 1$. Given this, using a standard Bernstein concentration of measure statement, then the union bound over $j \in [k]$, we can achieve our goal by bounding the probability of A_n^C as at most

$$2\sum_{j=1}^{k} \exp\left(-n \min\left(\frac{\chi_n^2}{4}, \frac{\chi_n\sqrt{\boldsymbol{p}^k(j)(1-\boldsymbol{p}^k(j))}}{2}\right)\right) \lesssim k\exp(-n\chi_n^2) \lesssim k\exp\left((1-\tau)^2\frac{C^*_{s,\beta}}{16}n^{1-\beta B(s)}\right)$$

where the first $\lesssim$ is since $\sqrt{\boldsymbol{p}^k(j)(1-\boldsymbol{p}^k(j))}$ is larger than this function evaluated at the smallest probability and since by definition of $B(s)$ and since $H_{k,s} \to R(s)$, one can show

$$\chi_n^2 = o\left(\chi_n\sqrt{\boldsymbol{p}^k(\pi_k^{-1}(k))(1-\boldsymbol{p}^k(\pi_k^{-1}(k))}\right).$$

What is left to argue is that for each τ in a neighborhood of 0, eventually in n, $r_{n,i}$ is smaller than $\frac{1}{2}$ the distance of the i^{th} largest probability to its nearest adjacent probability for $i \in [k]$. Since the index set is growing, we divide it into segments $S_j := \{\lfloor k^{\zeta_{j-1}}\rfloor, \ldots, \lfloor k^{\zeta_j}\rfloor\}$ where $\{\zeta_j\}_{j=0}^{\infty}$ is an increasing sequence with $\zeta_0 = 0$ and limit 1 and $r_{n,\lfloor k^{\zeta_{j-1}}\rfloor}$ is smaller than $\frac{1}{2}$ the distance between the $\lfloor k^{\zeta_j}\rfloor^{th}$ largest probability and its nearest adjacent probability for $j \geq 1$.

Assuming the existence of such a sequence, each index $i \in [k-1]$ is in some S_{j_i}. In particular, since $r_{n,i}$ decreases in i (which follows since the function $g(x) = x(1-x)$ is symmetric about $\frac{1}{2}$ on the interval $[0,1]$ and monotonically increasing on $[0,\frac{1}{2}]$), $r_{n,i}$ is smaller than the r value at the left end of the segment. By the definition of the ζ sequence, the r value at the left end of the segment is smaller than $\frac{1}{2}$ of the distance to closest adjacent probability at the right end of the segment. And since Zipfianity ensures distance to closest adjacent probability decreases as i increases, this value at the right end of the segment is smaller than this value at i.

Thus all that is left is to show the sequence

$$\zeta_j := \frac{\zeta_{j-1}\cdot s}{2\cdot(s+1)} + \frac{s+2}{2(s+1)} = \frac{s+2}{2(s+1)}\sum_{\ell=0}^{j-1}\left(\frac{s}{2(s+1)}\right)^{\ell}$$

satisfies the desired properties. First, using the geometric series formula note that $\lim_{j\to\infty}\zeta_j = 1$. Second, since $\zeta_j < 1$ always, distance to closest adjacent probability at the end of segment j is $\frac{f_s(\lfloor k^{\zeta_j}\rfloor) - f_s(\lfloor k^{\zeta_j}\rfloor+1)}{H_{k,s}}$. Using Taylor's theorem to bound this from below by $\frac{sf_{s+1}(\lfloor k^{\zeta_j}\rfloor)}{H_{k,s}}$ we have for any $0 < \tau < 1$ an N sufficiently large such that for each $j \geq 1$ and $n \geq N$

$$\begin{aligned}
\frac{f_s(\lfloor k^{\zeta_j}\rfloor) - f_s(\lfloor k^{\zeta_j}\rfloor+1)}{H_{k,s}} &\geq \frac{sf_{s+1}(\lfloor k^{\zeta_j}\rfloor)}{H_{k,s}} \\
&\geq n^{-(s+1)\beta\zeta_j}\frac{s}{R(s)} \\
&\geq n^{-\frac{\beta(s+2)}{2}}n^{-\frac{s\beta\zeta_{j-1}}{2}}\frac{s}{R(s)} \\
&\geq n^{-\frac{\beta B(s)}{2}}\frac{s}{\sqrt{R(s)}}(1-\tau)\sqrt{g(\boldsymbol{p}(\pi_k^{-1}(\lfloor k^{\zeta_{j-1}}\rfloor)))} \\
&= 2r_{n,\lfloor k^{\zeta_{j-1}}\rfloor}
\end{aligned}$$

where in the second last line we used that $B(s) = s+2$ and in the last line we used the definition of χ_n and that $C^*_{s,\beta} = \frac{s^2}{R(s)}$ when $s > 1$. This handles the ordering property for all $i \in [k-1]$. For $i = k$, one can directly show $r_{n,k}$ is smaller than $\frac{1}{2}$ of the distance to the closest adjacent probability. □

5.4 Performance Of TSS In High Dimensions

Here we explain why the stated TSS bounds of Theorem 4.3 are only for $s > 2$, not for $s > 1$.

$T_{n,\epsilon}$ of Theorem 4.3, specified in equation (92) of the appendix, satisfies $T_{n,\epsilon} \asymp n^{\frac{1}{B(s)}-\epsilon}$ where $\epsilon > 0$ is arbitrarily small. Decomposing expected 1–norm into the head (error on probabilities at least as large as the probability at the truncation index) and tail, and then using the segmentation procedure on the head, we have that when $s > 2$

$$\mathbb{E}\|\hat{\boldsymbol{p}}^k_{n,T_{n,\epsilon}}(j) - \boldsymbol{p}^k\|_1 = n^{\mathsf{C}_1}\exp(-\mathsf{C}_2 n^{\epsilon}) + \mathbb{E}\sum_{j=T_{n,\epsilon+1}}^{k}|\bar{\boldsymbol{p}}^k_n(\pi_k^{-1}(j)) - \boldsymbol{p}^k(\pi_k^{-1}(j))|.$$

$T_{n,\epsilon}$, also a function of s, is chosen arbitrarily close to the largest growing index such that use of Sort and

Snap on this head of the distribution still yields exponential decay. This leaves the EPE error in the tail. What distinguishes the heavy tailed $1 < s < 2$ case from the light tailed $s > 2$ case is that

$$\mathbb{E} \sum_{j=T_{n,\epsilon+1}}^{k} |\bar{\boldsymbol{p}}_n^k(\pi_k^{-1}(j)) - \boldsymbol{p}^k(\pi_k^{-1}(j))| \begin{cases} \asymp \mathbb{E}\|\bar{\boldsymbol{p}}_n^k - \boldsymbol{p}\|_1 & 1 < s < 2 \\ o\left(\mathbb{E}\|\bar{\boldsymbol{p}}_n^k - \boldsymbol{p}\|_1\right) & s > 2 \end{cases} \tag{11}$$

6 SIMULATIONS

We perform three simulations to demonstrate and compare performance of the EPE, TSS, and SS under various s and β. Each simulation is a Monte Carlo study: for a given n, β and s, M trials are run, drawing from the member of $\boldsymbol{p}^k \in \mathcal{P}_{f_s,k}$ possessing the identity permutation. From each of the M trials, we obtain the value of the estimator for a given estimator $A_n \in \{EPE_n, TSS_n, SS_n\}$; denote these estimates as $A_{1n}, \ldots, A_{Mn}$. For each estimate, the $\|\cdot\|_1$ distance between the estimator and truth (the member of $\mathcal{P}_{f_s,k}$ possessing the identity permutation) is computed. The Monte Carlo point estimate for $\mathbb{E}\|\boldsymbol{p}^k - A_n\|_1$ is defined as $\bar{B}_{A,n,M} := \frac{1}{M}\sum_{j=1}^{M}\|\boldsymbol{p}^k - A_{jn}\|_1$. Via the Central Limit Theorem, Delta Method, and Slutsky's theorem

$$\frac{\bar{B}_{A,n,M}}{\hat{\sigma}_M}\sqrt{M}(\log(\bar{B}_{A,n,M}) - \log(\mathbb{E}\|\boldsymbol{p}^k - A_n\|_1) \xrightarrow{\mathbb{D}} \mathcal{N}(0,1) \text{ as } M \to \infty$$

where $\hat{\sigma}_M$ is the sample standard deviation of the M Monte Carlo trials. The above equation allows us to produce approximately valid 95% confidence intervals for $\log(\mathbb{E}\|\boldsymbol{p}^k - A_n\|_1)$ for M sufficiently large. We set $M = 300$ and use sample sizes $\ln(n) \in \{5, 6, \ldots, 13, 14\}$. Simulation plots have x-axis as $\ln(n)$ and y-axis providing an approximate 95% confidence interval for $\log(\mathbb{E}\|\boldsymbol{p}^k - A_n\|_1)$. Code, data and instructions to reproduce are here: `https://github.com/jacobs269/zipfianPaper`

Simulation 1: $s = 1.05$**,** $\beta = \frac{1}{s+3}$**:** This simulation illustrates one of the main features of Theorems 4.1 and 4.2, which is that if the problem has a growth rate that does not increase too quickly with n, then Sort and Snap will achieve exponential decay while the EPE achieves only polynomial decay. The choice of $s = 1.05$ for this simulation is also motivated by Zipf's original work on the subject suggesting word frequency distributions in text often were Zipfian with s near 1 (Zipf, 1949).

If $x_n = \log(n)$ and $y_n = \log(\mathbb{E}\|\boldsymbol{p}^k - A_n\|_1)$, then Theorem 4.1 implies that with $A_n = EPE_n$, there are constants C_1, C_2 such that

$$\log(C_1) - rx_n \le y_n \le \log(C_2) - rx_n \tag{12}$$

for n sufficiently large and r the power in the rate given in Theorem 4.1. On the other hand, Theorem 4.2 implies that when $A_n = SS_n$ there are constants $r_1, r_2, C_1, C_2, C_3, C_4$ such that for n sufficiently large

$$\log(C_1) - r_1 x_n - C_3 \exp\left(x_n\left(1 - \beta(s+2)\right)\right) \le y_n \le \log(C_2) - r_2 x_n - C_4 \exp(x_n(1 - \beta(s+2))). \tag{13}$$

By Equation (12) we expect roughly linear behavior for EPE in the $\log - \log$ plots while by Equation (13) we expect a steeper slope as $\log(n)$ increases for SS. This is illustrated in the left plot in Figure 2.

Simulation 2: $s > 2$ **and** $\beta = 1$**:** This simulation compares the EPE, SS, and TSS when $s = 3$ and when $s = 5$. Using similar arguments to those explained for Simulation 1, Theorems 4.1 and 4.3 respectively suggest that asymptotically in n and on the $\log - \log$ scale, the EPE curve should be approximately linear with slope $-\frac{1}{2}$ and the TSS curve should be approximately linear with slope $-\frac{s}{s+2}$.

In particular, the EPE should not incur a rate improvement when s increases in this simulation, but TSS should. This is what we observe in Simulation 2 in the middle of Figure 2. EPE and TSS curves look roughly linear but the Ordinary Least Squares estimates of the slope for the EPE under $s = 3, 5$ are nearly identical $-.470, -.474$ while for TSS they are $-.569$ and $-.676$. The exact slope estimates are meaningful only in comparison between estimators.

Moreover, the SS curves achieve lower error than the TSS curves but the difference between TSS and SS appears to vanish as n grows. The slope estimates support the hypothesis that eventually in n, TSS will perform at least as well or possibly better than SS.

Simulation 3: $s = 1.5$ **and** $\beta = 1/s$**:** A remaining open question is whether or not the EPE retains minimax optimality when $1 < s < 2$ and $\beta > \frac{1}{s+2}$. Our hypothesis is yes, and this is supported by Simulation 3 on the right of Figure 2. We see that the EPE has the fastest decaying slope of the three methods and all three slopes are quite similar.

7 DISCUSSION AND OPEN PROBLEMS

This work provides new estimators (SS and TSS) for s-Zipfian distributions under L_1, and shows that these

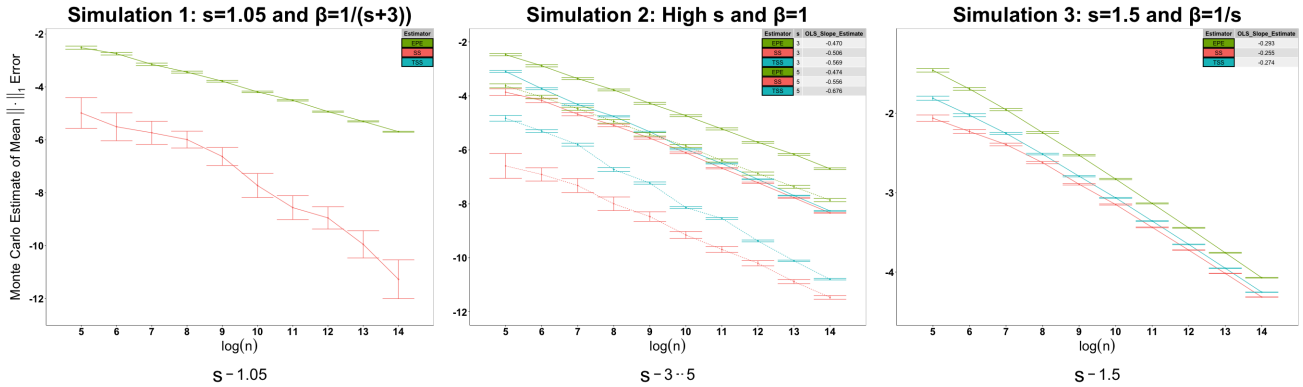

Figure 2: Simulations with $\ln(n) \in \{5, 6, \ldots, 13, 14\}$ showing $\log(\mathbb{E}\|\boldsymbol{p}^k - A_n\|_1)$ with 95% confidence intervals where $A_n \in \{EPE_n, TSS_n, SS_n\}$. For Simulations 2 and 3 the Ordinary Least Squares estimate of the slope of each line is displayed in the top right corner.

are sometimes near minimax optimal. We also provide tight analyses for the standard EPE estimator. These answer important questions about estimation for high-dimensional structured distributions, which are very relevant in the context of large language models and other featurized representation learning problems.

This work brings forward several interesting open questions, including what other distributional cases can Sort-and-Snap type estimators provide better error rates, and can we estimate distributional meta-parameters (such as s) as part of the problem. Moreover, while the main landscape of estimating this s-Zipfian is now resolved (see Figure 1), we leave open questions to completely resolve this, including: *Does the EPE remain the minimax optimal procedure when* $0 < s < 2$ *and* $\beta > \frac{1}{B(s)}$?

This is plausible because the error of the EPE in estimating largest probabilities (the $n^{\frac{1}{B(s)}}$ largest probabilities) is dominated by that of estimating smaller probabilities, *and* mis-ordering the counts of adjacent probabilities beyond the $n^{\frac{1}{B(s)}}$ point happens with constant probability for those smaller probabilities. Proving this in the affirmative would be interesting because it would demonstrate that in high-dimensional problems under L_1 loss there can arise scenarios where it does not help to use the probability values themselves, even if known.

Another consideration is that SS and TSS are examples of permutation invariant estimators in that if the categories are relabeled according to some permutation function π^*, then the probability estimate for category j under SS (or TSS) before the relabeling of categories will equal the probability estimate for category $(\pi^*)^{-1}(j)$ after the relabeling. Greenshtein and Ritov (2009) define the *permutation invariant oracle* as the permutation invariant estimator minimizing the risk when the true permutation π of the probabilities is the identity permutation. Because SS and TSS are permutation invariant, our upper bounds also upper bound the risk of the permutation invariant oracle in the s–Zipfian setting. The permutation invariant oracle offers an alternative to SS and TSS that may avoid switching from snapping to empirical proportions in TSS at a Zipfian specific dimension. This could be useful for generalizing beyond the Zipfian setting and we leave this for future work.

Finally, note that due to Berend and Kontorovich (2013), estimators that use empirical proportions or slight modifications (such as the EPE, the Good-Turing method in Orlitsky and Suresh (2015), or Absolute Discounting in Falahatgar et al. (2017)) must achieve the slower polynomial rather than exponential decay rates. The exponential decay rates of SS are due to s (and consequently the probability values) being known. In a data analysis however, it is possible s may not be known. This in combination with our regime dependent upper bounds on the minimax estimator in the s known case provides strong motivation for a more careful study of Competitive Distribution Estimation under L_1 loss in the Zipfian setting. In particular: *Will an adaptive variant of Sort and Snap that estimates* s *first, and then performs the sorting and snapping still yield an improvement over the EPE and other non multiset oracle procedures such as those presented in Orlitsky and Suresh (2015),Valiant and Valiant (2016),Falahatgar et al. (2017), Hao and Orlitsky (2019b), and Hao and Orlitsky (2019a) for estimating distributions that are* $s > 0$ *Zipfian?*

Acknowledgments

This work was supported by contract NSF CCF-2115677 and the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated

by National Technology and Engineering Solutions of Sandia, LLC, a wholly-owned subsidiary of Honeywell International, Inc., for both the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. The authors further acknowledge and thank the reviewers for helpful conversations.

References

Aamand, A., Indyk, P., and Vakilian, A. (2019). (learned) frequency estimation algorithms under zipfian distribution. *arXiv preprint arXiv:1908.05198*.

Berend, D. and Kontorovich, A. (2013). A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters*, 83(4):1254–1259.

Canonne, C. L., Sun, Z., and Suresh, A. T. (2023). Concentration bounds for discrete distribution estimation in kl divergence. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 2093–2098. IEEE.

Chan, S., Santoro, A., Lampinen, A., Wang, J., Singh, A., Richemond, P., McClelland, J., and Hill, F. (2022a). Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891.

Chan, S. C., Lampinen, A. K., Richemond, P. H., and Hill, F. (2022b). Zipfian environments for reinforcement learning. In *Conference on Lifelong Learning Agents*, pages 406–429. PMLR.

Cohen, D., Kontorovich, A., and Wolfer, G. (2020). Learning discrete distributions with infinite support. *Advances in Neural Information Processing Systems*, 33:3942–3951.

Dahui, W., Menghui, L., and Zengru, D. (2005). True reason for zipf's law in language. *Physica A: Statistical Mechanics and its Applications*, 358(2-4):545–550.

Diamond, J. (2023). " genlangs" and zipf's law: Do languages generated by chatgpt statistically look human? *arXiv preprint arXiv:2304.12191*.

Falahatgar, M., Ohannessian, M. I., Orlitsky, A., and Pichapati, V. (2017). The power of absolute discounting: all-dimensional distribution estimation. *Advances in Neural Information Processing Systems*, 30.

Ferrer-i Cancho, R. (2005). The variation of zipf's law in human language. *The European Physical Journal B-Condensed Matter and Complex Systems*, 44:249–257.

Ferrer-i Cancho, R. (2016). Compression and the origins of zipf's law for word frequencies. *Complexity*, 21(S2):409–411.

Greenshtein, E. and Ritov, Y. (2009). Asymptotic efficiency of simple decisions for the compound decision problem. *Lecture Notes-Monograph Series*, pages 266–275.

Han, Y., Jiao, J., and Weissman, T. (2015). Minimax estimation of discrete distributions under ell1 loss. *IEEE Transactions on Information Theory*, 61(11):6343–6354.

Hao, Y. and Orlitsky, A. (2019a). The broad optimality of profile maximum likelihood. *Advances in Neural Information Processing Systems*, 32.

Hao, Y. and Orlitsky, A. (2019b). Doubly-competitive distribution estimation. In *International Conference on Machine Learning*, pages 2614–2623. PMLR.

Hsu, C.-Y., Indyk, P., Katabi, D., and Vakilian, A. (2019). Learning-based frequency estimation algorithms. In *International Conference on Learning Representations*.

Izsák, F. (2006). Maximum likelihood estimation for constrained parameters of multinomial distributions—application to zipf–mandelbrot models. *Computational statistics & data analysis*, 51(3):1575–1583.

Kontorovich, A. and Painsky, A. (2024). Distribution estimation under the infinity norm. *arXiv preprint arXiv:2402.08422*.

Li, W. (1992). Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on information theory*, 38(6):1842–1845.

Orlitsky, A. and Suresh, A. T. (2015). Competitive distribution estimation: Why is good-turing good. *Advances in Neural Information Processing Systems*, 28.

Painsky, A. (2023). Large alphabet inference. *Information and Inference: A Journal of the IMA*, 12(4):3067–3086.

Paninski, L. (2004). Variational minimax estimation of discrete distributions under kl loss. *Advances in Neural Information Processing Systems*, 17.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112–1130.

Tibshirani, R. and Wasserman, L. (2017). Minimax theory.

Valiant, G. and Valiant, P. (2016). Instance optimal learning of discrete distributions. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 142–155.

Zipf, G. K. (1949). Human behavior and the principle of least effort: An introduction to human eoclogy.

Zörnig, P. and Altmann, G. (1995). Unified representation of zipf distributions. *Computational Statistics & Data Analysis*, 19(4):461–473.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes (Sample Size)]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes (see link referenced in paper)]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes (see link referenced in paper)]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes (see link referenced in paper)]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A TIGHT PERFORMANCE OF EPE

Berend and Kontorovich (2013) provide a general tight upper and lower bound for the performance of the EPE that we will apply in our setting. As discussed in Cohen et al. (2020), letting

$$\Lambda_{n,k}(f) := \sum_{j\in[k]:\frac{f(j)}{\sum_{t=1}^k f(t)}<\frac{1}{n}} \frac{f(j)}{\sum_{t=1}^k f(t)} + \frac{1}{2\sqrt{n}} \sum_{j\in[k]:\frac{f(j)}{\sum_{t=1}^k f(t)}\geq 1/n} \sqrt{\frac{f(j)}{\sum_{t=1}^k f(t)}} \tag{14}$$

Then if for $k \geq 2$, $\boldsymbol{p}^k \in \mathcal{P}_{f_s,k}$ and

$$Y_{1k}, Y_{2k}, \ldots, Y_{nk} \overset{iid}{\sim} \boldsymbol{p}^k$$

and $\bar{\boldsymbol{p}}_n^k$ is the corresponding EPE, then for $k \geq 2$

$$\frac{1}{2}\Lambda_{n,k}(f) - \frac{1}{2}\frac{1}{\sqrt{n}} \leq \mathbb{E}\|\bar{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1 \leq 2\Lambda_{n,k}(f) \tag{15}$$

This leads to the following theorem regarding the performance of the EPE for s–Zipfian laws

Theorem 4.1 (EPE for Exact s-Zipfian Laws)**.** *Suppose $s > 0$. Further let $0 < \beta < \infty$, $\boldsymbol{p}^k \in \mathcal{P}_{f_s,k}$ and $k = \lfloor n^\beta \rfloor$ for n sufficiently large so that $\lfloor n^\beta \rfloor \geq 2$. Also,*

$$Y_{1k}, Y_{2k}, \ldots, Y_{nk} \overset{iid}{\sim} \boldsymbol{p}^k$$

and $\bar{\boldsymbol{p}}_n^k$ is the corresponding EPE. Then

$$\mathbb{E}\|\bar{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1 \asymp \begin{cases} n^{-\frac{1}{2}(1-\beta)} & 0 < s < 1, \beta < 1 \\ 1 & 0 < s < 1, \beta \geq 1 \\ \frac{n^{-\frac{1}{2}(1-\beta)}}{\sqrt{\log(n)}} & s = 1, \beta < 1 \\ \frac{\log(\log(n))}{\log(n)} & s = 1, \beta = 1 \\ 1 & s = 1, \beta > 1 \\ n^{-\frac{1}{2}+\beta(1-\frac{s}{2})} & 1 < s < 2, \beta < \frac{1}{s} \\ n^{\frac{1}{s}-1} & 1 < s < 2, \beta \geq \frac{1}{s} \\ \frac{\log(n)}{n^{1/2}} & s = 2 \\ n^{-\frac{1}{2}} & s > 2 \end{cases}$$

Proof. For $j \in \mathbb{N}$, $\frac{j^{-s}}{H_{k,s}} < n^{-1}$ implies that $j > (\frac{n}{H_{k,s}})^{1/s}$. Therefore, with $f_s(x) = x^{-s}$, we have that

$$\Lambda_{n,k}(f_s) = \left(\frac{1}{H_{k,s}} \sum_{j\in\mathbb{N}:\min\left(k,(\frac{n}{H_{k,s}})^{1/s}\right)<j\leq k} j^{-s} + \frac{1}{2\sqrt{nH_{k,s}}} \sum_{j\in\mathbb{N}:1\leq j\leq\min\left(k,(\frac{n}{H_{k,s}})^{1/s}\right)} j^{-s/2} \right) = \begin{cases} \frac{1}{H_{k,s}} \sum_{j=\lceil(\frac{n}{H_{k,s}})^{1/s}\rceil}^k j^{-s} + \frac{1}{2\sqrt{nH_{k,s}}} \sum_{j=1}^{\lfloor(\frac{n}{H_{k,s}})^{1/s}\rfloor} j^{-s/2} & (\frac{n}{H_{k,s}})^{1/s} < k \\ \frac{1}{2\sqrt{H_{k,s}n}} \sum_{j=1}^k j^{-s/2} & (\frac{n}{H_{k,s}})^{1/s} \geq k \end{cases} \tag{16}$$

Observe also that for a_n an increasing sequences in n and $s' > 0$

$$\begin{cases} \frac{1}{1-s'}\left((a_n+1)^{1-s'}-1\right) & s' < 1 \\ \log(a_n+1) & s' = 1 \\ \frac{1}{s'-1}(1-(a_n+1)^{1-s'}) & s' > 1 \end{cases} = \int_1^{a_n+1} x^{-s'}dx \leq \sum_{j=1}^{a_n} j^{-s'} \leq 1 + \int_1^{a_n} x^{-s'}dx = 1 + \begin{cases} \frac{1}{1-s'}\left(a_n^{1-s'}-1\right) & s' < 1 \\ \log(a_n) & s' = 1 \\ \frac{1}{s'-1}(1-a_n^{1-s'}) & s' > 1 \end{cases} \tag{17}$$

In particular,

$$\sum_{j=1}^{a_n} j^{-s'} \asymp \begin{cases} a_n^{1-s'} & 0 < s' < 1 \\ \log(a_n) & s' = 1 \\ 1 & s' > 1 \end{cases} \tag{18}$$

Note equation 18 implies that

$$H_{k,s} \asymp \begin{cases} n^{\beta(1-s)} & 0 < s < 1 \\ \log(n) & s = 1 \\ 1 & s > 1 \end{cases} \tag{19}$$

Also, for b_n a non-decreasing sequence such that $1 \leq b_n \leq k$, we have that

$$\begin{cases} \frac{1}{1-s'}\left((k+1)^{1-s'} - b_n^{1-s'}\right) & s' < 1 \\ \log(k+1) - \log(b_n) & s' = 1 \\ \frac{1}{s'-1}\left(b_n^{1-s'} - (k+1)^{1-s'}\right) & s' > 1 \end{cases} = \int_{b_n}^{k+1} x^{-s'} dx \leq \sum_{j=b_n}^{k} j^{-s'} \leq \int_{b_n-1}^{k} x^{-s'} dx = \begin{cases} \frac{1}{1-s'}\left(k^{1-s'} - (b_n-1)^{1-s'}\right) & s' < 1 \\ \log(k) - \log(b_n - 1) & s' = 1 \\ \frac{1}{s'-1}\left((b_n-1)^{1-s'} - k^{1-s'}\right) & s' > 1 \end{cases} \tag{20}$$

Now we will handle each of the 9 cases.

Case 1: $0 < s < 1, \beta < 1$

$\beta < 1$ implies that $\frac{1-\beta(1-s)}{s} > \beta$. Using this and equation 19, we have that

$$\left(\frac{n}{H_{k,s}}\right)^{1/s} \asymp n^{\frac{1-\beta(1-s)}{s}} = \omega(k)$$

In particular for sufficiently large n, $\left(\frac{n}{H_{k,s}}\right)^{1/s} \geq k$. Thus applying equation 16 and then using equations 18 (with $a_n = k$ and $s' = \frac{s}{2}$) and 19, we have that

$$\Lambda_{n,k}(f_s) \asymp n^{-\frac{1}{2}} n^{-\frac{\beta(1-s)}{2}} n^{\beta(1-\frac{s}{2})} = n^{-\frac{1}{2}(1-\beta)} \tag{21}$$

Using this and equation 15 the case conclusion follows.

Case 2: $0 < s < 1, \beta \geq 1$

We first handle $\beta = 1$. In this case, by equation 19, $(\frac{n}{H_{k,s}})^{\frac{1}{s}} \asymp n$. By applying equation 19 and 18 (with $s' = \frac{s}{2}$ and a_n the obvious choice), we have that

$$\frac{1}{2\sqrt{nH_{k,s}}} \sum_{j=1}^{\lfloor (\frac{n}{H_{k,s}})^{\frac{1}{s}} \rfloor} j^{-s/2} \asymp n^{-\frac{1}{2}} n^{-\frac{1-s}{2}} n^{1-\frac{s}{2}} \asymp 1 \tag{22}$$

By a nearly identical argument,

$$\frac{1}{2\sqrt{nH_{k,s}}} \sum_{j=1}^{k} j^{-s/2} \asymp 1 \tag{23}$$

By equations 22 and 23 and equation 16, when $\beta = 1$

$$\Lambda_{n,k}(f_s) \gtrsim 1$$

Using this and equation 15 and that 1–norm is bounded above by a positive constant on the Simplex, the case conclusion follows for $\beta = 1$. When $\beta > 1$, $\frac{1-\beta(1-s)}{s} < \beta$. So using equation 19, we have that

$$\left(\frac{n}{H_{k,s}}\right)^{1/s} \asymp n^{\frac{1-\beta(1-s)}{s}} = o(k) \tag{24}$$

Using this and equation 16 and then equation 20 with $b_n = \lceil (\frac{n}{H_{k,s}})^{1/s} \rceil$ and $s' = s$, we have that

$$\Lambda_{n,k}(f_s) \gtrsim \frac{1}{H_{k,s}} \sum_{j=\lceil (\frac{n}{H_{k,s}})^{1/s} \rceil}^{k} j^{-s} \gtrsim n^{-\beta(1-s)} \left((k+1)^{1-s} - (\lceil (\frac{n}{H_{k,s}})^{1/s} \rceil)^{1-s} \right) \gtrsim n^{-\beta(1-s)} n^{\beta(1-s)} \gtrsim 1 \tag{25}$$

Using this and equation 15 and that 1–norm is bounded above by a positive constant on the Simplex, the case conclusion follows for $\beta > 1$

Case 3: $s = 1, \beta < 1$

By equation 19, we have that

$$\left(\frac{n}{H_{k,s}} \right)^{1/s} \asymp \frac{n}{\log(n)} = \omega(k) \tag{26}$$

Using this and equation 16 and then equation 18 (with $s' = 1/2$ and $a_n = k$), we have that

$$\Lambda_{n,k}(f_1) \asymp \frac{1}{\sqrt{nH_{k,s}}} \sum_{j=1}^{k} j^{-s/2} \asymp n^{-\frac{1}{2}} \frac{1}{\sqrt{\log(n)}} n^{\beta(1-\frac{1}{2})} \asymp \frac{n^{-\frac{1}{2}(1-\beta)}}{\sqrt{\log(n)}} \tag{27}$$

Using this and equation 15 the case conclusion follows

Case 4: $s = 1, \beta = 1$

First using equation 18 with $a_n = \lfloor (\frac{n}{H_{k,s}})^{1/s} \rfloor$ and $s' = \frac{1}{2}$ and also equation 19, we have that

$$\frac{1}{\sqrt{nH_{k,s}}} \sum_{j=1}^{\lfloor (\frac{n}{H_{k,s}})^{1/s} \rfloor} j^{-\frac{1}{2}} \asymp n^{-\frac{1}{2}} \frac{1}{\sqrt{\log(n)}} \left(\frac{n}{\log(n)} \right)^{\frac{1}{2}} \asymp \frac{1}{\log(n)} \tag{28}$$

Next, using equation 20 with $b_n = \lceil (\frac{n}{H_{k,s}})^{1/s} \rceil$ and $s' = 1$ and also equation 19, we have that

$$\frac{\log(\log(n))}{\log(n)} \lesssim \frac{1}{\log(n)} \log(\frac{k+1}{\lceil (\frac{n}{H_{k,s}})^{1/s} \rceil}) \lesssim \frac{1}{H_{k,s}} \sum_{j=\lceil (\frac{n}{H_{k,s}})^{1/s} \rceil}^{k} j^{-s} \lesssim \frac{1}{\log(n)} \log(\frac{k}{\lceil (\frac{n}{H_{k,s}})^{1/s} \rceil - 1}) \lesssim \frac{\log(\log(n))}{\log(n)} \tag{29}$$

Also, $(\frac{n}{H_{k,s}})^{1/s} = o(k)$. Using this and equations 28 and 29 and 16, we have that

$$\Lambda_{n,k}(f_1) \asymp \frac{\log(\log(n))}{\log(n)} \tag{30}$$

Using this and equation 15 the case conclusion follows

Case 5: $s = 1, \beta > 1$

Applying equation 20 with $b_n = \lceil (\frac{n}{H_{k,s}})^{1/s} \rceil$ and $s' = s$ (and again using equation 19), we have that

$$\frac{1}{H_{k,s}} \sum_{j=\lceil (\frac{n}{H_{k,s}})^{1/s} \rceil}^{k} j^{-s} \gtrsim \frac{1}{\log(n)} \left(\log(k+1) - \log(\lceil (\frac{n}{H_{k,s}})^{1/s} \rceil \right) \asymp \frac{\log(n^{\beta-1} \log(n))}{\log(n)} \asymp 1 \tag{31}$$

Also $\left(\frac{n}{H_{k,s}} \right)^{1/s} = o(k)$. Using this and equation 31, and 16, we have that

$$\Lambda_{n,k}(f_1) \gtrsim 1$$

Using this and that 1–norm on the Simplex is bounded above by a constant, and equation 15, the case conclusion follows.

Case 6: $1 < s < 2, \beta < \frac{1}{s}$

By equation 19 and since $\beta < \frac{1}{s}$

$$\left(\frac{n}{H_{k,s}} \right)^{1/s} \asymp n^{\frac{1}{s}} = \omega(k) \tag{32}$$

Also, applying equation 19 with $a_n = k$ and $s' = \frac{s}{2}$, we have that

$$\frac{1}{\sqrt{H_{k,s}n}} \sum_{j=1}^{k} j^{-\frac{s}{2}} \asymp n^{\beta(1-\frac{s}{2})} n^{-\frac{1}{2}} \asymp n^{-\frac{1}{2}+\beta(1-\frac{s}{2})} \tag{33}$$

Using this and equations 32 and 16, we have that

$$\Lambda_{n,k}(f_s) \asymp n^{-\frac{1}{2}+\beta(1-\frac{s}{2})}$$

Using this and equation 15 the case conclusion follows.

Case 7: $1 < s < 2, \beta \geq \frac{1}{s}$

First we will consider $\beta = \frac{1}{s}$. By equation 16 we have that

$$\min\left(\frac{1}{H_{k,s}} \sum_{j=\lceil (\frac{n}{H_{k,s}})^{\frac{1}{s}} \rceil}^{k} j^{-s} + \frac{1}{2\sqrt{nH_{k,s}}} \sum_{j=1}^{\lfloor (\frac{n}{H_{k,s}})^{\frac{1}{s}} \rfloor} j^{-s/2}, \frac{1}{2\sqrt{H_{k,s}n}} \sum_{j=1}^{k} j^{-s/2}\right) \lesssim \Lambda_{n,k}(f_s) \lesssim \max\left(\frac{1}{H_{k,s}} \sum_{j=\lceil (\frac{n}{H_{k,s}})^{\frac{1}{s}} \rceil}^{k} j^{-s} + \frac{1}{2\sqrt{nH_{k,s}}} \sum_{j=1}^{\lfloor (\frac{n}{H_{k,s}})^{\frac{1}{s}} \rfloor} j^{-s/2}, \frac{1}{2\sqrt{H_{k,s}n}} \sum_{j=1}^{k} j^{-s/2}\right) \tag{34}$$

Also applications of equation 19,20,and 18 yield

$$\frac{1}{H_{k,s}} \sum_{j=\lceil (\frac{n}{H_{k,s}})^{\frac{1}{s}} \rceil}^{k} j^{-s} + \frac{1}{2\sqrt{nH_{k,s}}} \sum_{j=1}^{\lfloor (\frac{n}{H_{k,s}})^{\frac{1}{s}} \rfloor} j^{-s/2} \asymp n^{\frac{1}{s}(1-s)} + n^{-\frac{1}{2}} n^{\frac{1}{s}(1-\frac{s}{2})} \asymp n^{\frac{1}{s}-1} \tag{35}$$

Also, when $\beta = \frac{1}{s}$,

$$\frac{1}{\sqrt{H_{k,s}n}} \sum_{j=1}^{k} j^{-\frac{s}{2}} \asymp n^{-\frac{1}{2}} n^{\beta(1-\frac{s}{2})} \asymp n^{\frac{1}{s}-1} \tag{36}$$

By equations 34,35,and 36, we have that when $\beta = \frac{1}{s}$,

$$\Lambda_{n,k}(f_s) \asymp n^{\frac{1}{s}-1} \tag{37}$$

Using this and equation 15 the case conclusion follows for $\beta = \frac{1}{s}$. If $\beta > \frac{1}{s}$, then

$$\left(\frac{n}{H_{k,s}}\right)^{1/s} = o(k)$$

Using this and equation 16 and equation 35, we have that

$$\Lambda_{n,k}(f_s) \asymp \frac{1}{H_{k,s}} \sum_{j=\lceil (\frac{n}{H_{k,s}})^{\frac{1}{s}} \rceil}^{k} j^{-s} + \frac{1}{2\sqrt{nH_{k,s}}} \sum_{j=1}^{\lfloor (\frac{n}{H_{k,s}})^{\frac{1}{s}} \rfloor} j^{-s/2} \asymp n^{\frac{1}{s}-1} \tag{38}$$

Using this and equation 15 the case conclusion follows for $\beta > \frac{1}{s}$.

Case 8: $s = 2, \beta > 0$

In this case, using equation 18

$$\frac{1}{\sqrt{H_{k,s}n}} \sum_{j=1}^{k} j^{-\frac{s}{2}} \asymp \frac{\log(n)}{\sqrt{n}} \tag{39}$$

and using equations 20 and 18,

$$\frac{1}{H_{k,s}} \sum_{j=\lceil (\frac{n}{H_{k,s}})^{\frac{1}{s}} \rceil}^{k} j^{-s} + \frac{1}{2\sqrt{nH_{k,s}}} \sum_{j=1}^{\lfloor (\frac{n}{H_{k,s}})^{\frac{1}{s}} \rfloor} j^{-s/2} \asymp n^{\frac{1}{2}(1-2)} + \frac{\log(n)}{\sqrt{n}} \asymp \frac{\log(n)}{\sqrt{n}} \tag{40}$$

By equations 39 and 40 and 16 and 15 the case conclusion follows.
Case 9: $s > 2, \beta > 0$
Using almost the exactly same argument as Case 8 (with the only difference being all sums are convergent), we have that

$$\Lambda_{n,k}(f_s) \asymp n^{-\frac{1}{2}} \tag{41}$$

Using this and equation 15, we have that

$$\mathbb{E}\|\hat{p}_n^k - \boldsymbol{p}^k\|_1 \lesssim n^{-\frac{1}{2}}$$

Knowing the asymptotic order of $\Lambda_{n,k}(f_s)$ is not sufficient to generate a matching lower bound in this case because the upper bound rate is $n^{-\frac{1}{2}}$. To match the upper bound, note that of course the mean 1–norm error exceeds the mean absolute error of estimating the most probable category. That is,

$$\mathbb{E}\|\bar{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1 \geq \mathbb{E}|\bar{\boldsymbol{p}}_n^k(\pi_k^{-1}(1)) - \boldsymbol{p}^k(\pi_k^{-1}(1))| \tag{42}$$

Finally, note that the most probable category has probability $\frac{1}{H_{k,s}}$, which converges to a constant as $n \to \infty$. In particular, applying line 2 of Theorem 1 of Berend and Kontorovich (2013), in conjunction with equation 42, we have that

$$\mathbb{E}\|\bar{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1 \geq \frac{1}{\sqrt{2}} n^{-\frac{1}{2}} \sqrt{\frac{1}{H_{k,s}}(1 - \frac{1}{H_{k,s}})} \gtrsim n^{-\frac{1}{2}} \tag{43}$$

Thus the upper bound is matched and the case conclusion follows. □

In the process of proving theorem 4.1 we have derived the general asymptotic order of $H_{k,s}$. Because the asymptotic order of $H_{k,s}$ plays an important role in several other proofs in this document, we state here the asymptotic order of $H_{k,s}$ as a separate lemma.

Lemma A.1. *If $k = \lfloor n^\beta \rfloor$, then for each $0 < \omega < 1$ and n sufficiently large*

$$\begin{cases} \frac{(1-\omega)}{1-s} n^{\beta(1-s)} & 0 < s < 1 \\ \beta \log(n) & s = 1 \\ (1-\omega)R(s) & s > 1 \end{cases} \leq H_{k,s} \leq \begin{cases} \frac{1}{1-s} n^{\beta(1-s)} & 0 < s < 1 \\ (1+\omega)\beta \log(n) & s = 1 \\ R(s) & s > 1 \end{cases} \tag{44}$$

Proof. Follows immediately from the top two lines of equation 17 in the proof of theorem 4.1 and that $H_{k,s} \to R(s)$ when $s > 1$ □

B PERFORMANCE OF SORT AND SNAP + MINIMAX LOWER BOUNDS

B.1 Guide

Appendix section B goes through the details of how upper (and lower bounds) are proved for Sort and Snap and truncations of Sort and Snap. Because the proof strategy is lengthy, in this section we provide both a guide to reading appendix section B, in addition to providing an outline of the proof.

With $Y_{1k}, Y_{2k}, \ldots, Y_{nk} \overset{iid}{\sim} \boldsymbol{p}^k$, and for $j \in [k]$, X_{jn} as defined in equation 2 – notating with X_{jn} instead of X_{jk} because k is a function of n – and

$$Z_{jn} = \frac{X_{jn}}{n\sqrt{\boldsymbol{p}^k(j)(1 - \boldsymbol{p}^k(j))}}$$

and $\boldsymbol{Z}_n := (Z_{1n}, Z_{2n}, \ldots, Z_{kn})$ and carefully chosen χ_n tending to zero, we will determine whether

$$\mathbb{P}\left(\|\boldsymbol{Z}_n - \mathbb{E}\boldsymbol{Z}_n\|_\infty \geq \chi_n\right) \tag{45}$$

tends to zero, and if so, how fast. If we can argue that there exists an χ_n such that the error of Sort and Snap is small when $\|\boldsymbol{Z}_n - \mathbb{E}\boldsymbol{Z}_n\|_1 < \chi_n$ and yet it is still the case that the probability in equation 45 is small, then we can decompose the expected $\|\cdot\|_1$ distance between Sort and Snap and the true probability vector on this event, and achieve reasonable bounds. This involves a search for well-balanced values of χ_n. (The proofs are slightly more complex than just decomposing on a single event (to attain a better polynomial term on the outside of the exponential in the $\beta < \frac{1}{B(s)}$ case); those aspects of the proof will be explained after explaining the single event decomposition ideas).

Towards this end, a concentration of measure statement is developed in appendix section B.3. This is lemma B.1, restated again here for convenience.

Lemma B.1. *Let $k \geq 2$. If $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} Categorical(\boldsymbol{p} = (p_1, p_2, \ldots, p_k))$ (i.e $\mathbb{P}(Y_1 = j) = p_j$ for $j \in [k]$) and*

$$\boldsymbol{X} := (\sum_{i=1}^{n} \mathbb{I}(Y_i = j))_{j=1}^{k}$$

(so that $\boldsymbol{X} \sim Multinomial(n, \boldsymbol{p} = (p_1, p_2, \ldots, p_k))$), and $Z_j = \frac{X_j}{n\sqrt{p_j(1-p_j)}}$ for $j \in [k]$ and $\boldsymbol{Z} := (Z_1, Z_2, \ldots, Z_k)$. Then for $u, C > 0$

$$\mathbb{P}\left(\|\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}\|_\infty \geq Cu + \frac{C^2u^2}{2}\frac{1}{\min_{j\in[k]}\sqrt{p_j(1-p_j)}})\right) \leq 2k\exp(-\frac{nu^2C^2}{4})$$

Section B.4 takes specific integers $I(k) \in [k]$ and determines the largest possible χ_n – now depending on $I(k)$ – such that the event

$$\|\boldsymbol{Z}_n - \mathbb{E}\boldsymbol{Z}_n\|_\infty < \chi_n \tag{46}$$

implies that the top $I(k)$ counts are in perfect order. By the language, *the top $I(k)$ counts are in perfect order*, we mean that the count associated with the category with the j^{th} largest probability is the j^{th} largest count for $j \in \{1, 2, \ldots, I(k)\}$. Note that when this is true and Sort and Snap is used for at least the top $I(k)$ largest counts, the error of Sort and Snap on the top $I(k)$ largest probabilities is zero. And for this reason, section B.4 is aptly named *Good Event Identification*.

We are in search of sufficiently large values of χ_n so that $\mathbb{P}(\|\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}\| \geq \chi_n)$ is small; yet the value of χ_n needs to be sufficiently small so that the event of equation 46 implies that the top $I(k)$ counts are in perfect order. It will be seen that if we desire more counts to be in perfect order (i.e $I(k)$ is large), we will require χ_n to be smaller. And that if $I(k)$ reaches a sufficiently large rate of growth, $n^{\frac{1}{s+2}}$, it is no longer possible to simultaneously choose χ_n large enough so that $\mathbb{P}(\|\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}\| \geq \chi_n)$ decays to zero and χ_n is small enough so that the top $I(k)$ counts are in perfect order.

B.1.1 Segmentation identifies good events

We want to determine the largest value of χ_n in equation 46 such that the top $I(k, \gamma) = \lfloor k^\gamma \rfloor - 1$ counts are in perfect order, where $0 < \gamma \leq 1$. (The reason for doing this not only when $\gamma = 1$ but also when $\gamma < 1$ has to do with achieving a better polynomial term on the outside of the exponential decay and this will be elaborated on later.) For $\boldsymbol{p} \in \mathcal{P}_{f_s,k}$ (with permutation function π_k) and $k \geq 2$ and $i \in [k]$, define

$$D_{\boldsymbol{p},i} = \min_{j\in[k]-i} |\boldsymbol{p}(i) - \boldsymbol{p}(j)|$$

$D_{\boldsymbol{p},\pi_k^{-1}(j)}$ gives the distance to the closest adjacent probability for the j^{th} largest probability in $\boldsymbol{p}$. Also, for χ_n to be chosen later, let

$$S_{\boldsymbol{p},i,n} = \chi_n\sqrt{\boldsymbol{p}(i)(1-\boldsymbol{p}(i))}$$

$S_{\boldsymbol{p},\pi_k^{-1}(j),n}$ is the width of the ball around the j^{th} largest probability that contains $\frac{X_{\pi_k^{-1}(j),n}}{n}$ under the event in equation 46. Also, let

$$B_{\boldsymbol{p},i,n} = (L_{\boldsymbol{p},i,n}, U_{\boldsymbol{p},i,n}) = (\boldsymbol{p}(i) - S_{\boldsymbol{p},i,n}, \boldsymbol{p}(i) + S_{\boldsymbol{p},i,n})$$

Under the event in equation 46, $B_{\boldsymbol{p},\pi_k^{-1}(j),n}$ is the part of $\mathbb{R}$ containing $\frac{X_{\pi_k^{-1}(j),n}}{n}$. Now suppose χ_n is set sufficiently small so that

$$S_{\boldsymbol{p},\pi_k^{-1}(j),n} \leq \frac{1}{2} D_{\boldsymbol{p},\pi_k^{-1}(j)} \text{ for each } j \in [\lfloor k^\gamma \rfloor] \tag{47}$$

Because convexity of $f_s(x) = x^{-s}$ ensures that $D_{\boldsymbol{p},\pi_k^{-1}(j)}$ is decreasing in j (see lemma B.4), the above assumption implies that $B_{\boldsymbol{p},\pi_k^{-1}(j),n}$ lies strictly above $B_{\boldsymbol{p},\pi_k^{-1}(j+1),n}$ for $j \in [\lfloor k^\gamma \rfloor - 1]$. Also, since $\boldsymbol{p} \in \mathcal{P}_{f_s,k}$, $S_{\boldsymbol{p},\pi_k^{-1}(j),n}$ and $\boldsymbol{p}(\pi_k^{-1}(j))$ are monotonically decreasing in j (see lemma B.12). In particular, for any probability that is not one of the top $\lfloor k^\gamma \rfloor$ largest, the top of its ball does not rise above the top of the ball for the $\lfloor k^\gamma \rfloor$ largest probability.

This allows us to conclude that if χ_n is set sufficiently small so that equation 47 is satisfied, then under the event in equation 46, the top $I(k,\gamma) = \lfloor k^\gamma \rfloor - 1$ counts must be in perfect order. The remainder of this section is thus devoted to answering the following question: what is the largest value of χ_n in which equation 47 is satisfied?

To this end, set

$$\chi_n = C_1 n^{-\frac{1}{2}+e} \tag{48}$$

for a C_1 and e to be determined later. First note that with this choice χ_n, any power $e \leq 0$ will yield probability not decaying to zero in the application of lemma B.1; so we are now looking for a $e > 0$ that ensures equation 47 is satisfied. To this end we construct an infinite collection of sets $R_{1e}, R_{2e}, R_{3e}, \ldots$ such that

1. For each $j \geq 0$, there is a $a(j) \leq b(j)$ such that $R_{je} = \{\pi_k^{-1}(a(j)), \pi_k^{-1}(a(j)+1), \ldots, \pi_k^{-1}(b(j))\}$
2. $a(0) = 1$ and for $j_1 \leq j_2$, $b(j_1) \leq a(j_2)$
3. For each $j \geq 0$, $S_{\boldsymbol{p},\pi_k^{-1}(a(j)),n} \leq \frac{1}{2} D_{\boldsymbol{p},\pi_k^{-1}(b(j))}$

And the functions a and b depend on the choice of e (and C_1) through condition (3). The segment R_{je} contains all the indices corresponding to at least the $a(j)$ largest probability and at most the $b(j)$ largest probability. And because $\boldsymbol{p}(\pi_k^{-1}(j))(1-\boldsymbol{p}(\pi_k^{-1}(j))$ is decreasing in j (see lemma B.12) and $D_{\boldsymbol{p},\pi_k^{-1}(j))}$ is decreasing in j (see lemma B.4), by condition (3) above, for $i \in \{a(j), a(j)+1, \ldots, b(j)\}$,

$$S_{\boldsymbol{p},\pi_k^{-1}(i),n} \leq S_{\boldsymbol{p},\pi_k^{-1}(a(j)),n} \overset{(3)}{\leq} \frac{1}{2} D_{\boldsymbol{p},\pi_k^{-1}(b(j))} \leq \frac{1}{2} D_{\boldsymbol{p},\pi_k^{-1}(i)} \tag{49}$$

Thus for any $J \geq 1$ and each of the largest $1 \leq i \leq b(J)$ probabilities, equation 47 is satisfied. What remains is to argue that there is a way to choose the sets $R_{0e}, R_{1e}, R_{2e}, \ldots$, so that $(1), (2), (3)$ are satisfied and that for every $i \in \lfloor k^\gamma \rfloor$, there is a j such that $\pi_k^{-1}(i) \in R_{je}$.

The strategy is as follows: If we set $a(j) = \lfloor k^{\zeta_j \gamma} \rfloor$ for some $0 \leq \zeta_j \leq 1$, we intend to select $b(j) = \lfloor k^{\zeta_{j+1}\gamma} \rfloor$ for some $\zeta_{j+1} \geq \zeta_j$. In particular, ζ_{j+1} is selected so that $n^{-\frac{1}{2}+e}\sqrt{\boldsymbol{p}(\pi_k^{-1}(a(j)))(1-\boldsymbol{p}(\pi_k^{-1}(a(j))))} \asymp D_{\boldsymbol{p},\pi_k^{-1}(\lfloor k^{\zeta_{j+1}\gamma} \rfloor)}$. In the proofs, constants are tracked on the left and right side of the $\asymp$ so that we can make an appropriate selection for C. This selection process yields a sequence of ζ values $\{\zeta_j\}_{j=0}^\infty$ (with $\zeta_0 = 0$) such that for $j \geq 0$,

$$a(j) = \lfloor k^{\zeta_j \gamma} \rfloor, b(j) = \lfloor k^{\zeta_{j+1}\gamma} \rfloor$$

and for $j \geq 1$

$$a(j) = b(j-1)$$

Selection of a small e will allow ζ_{j+1} to be much larger than ζ_j. And this will allow $\cup_{j=0}^\infty R_{je}$ to cover a higher number of largest probabilities. But the cost of a small e is a small χ_n which means the complement of the event of interest, $\|\boldsymbol{Z}_n - \mathbb{E}\boldsymbol{Z}_n\|_1 \geq \chi_n$, will have a high probability. A balance is struck by setting

$$e = \frac{1 - \beta\left(\gamma(s+2) + \max(0, 1-s)\right)}{2}$$

when $\beta < \frac{1}{B(s)}$ where $B(s) = 2 + \max(1, s)$. With this choice for e (formally specified in equation 66 of the appendix) and choosing $C_1 = \frac{\sqrt{C^*_{s,\beta}}}{2}$ (where $C^*_{s,\beta}$ is formally specified in equation 6), and $\{\zeta_j\}_{j=0}^{\infty}$ (formally specified in equation 73 of the appendix) [1] as

$$\zeta_0 := 0, \zeta_j := \frac{1}{\beta\gamma(s+1)}\left(\frac{1 + s\beta\gamma\zeta_{j-1}}{2} - e\right) - \frac{1}{\gamma}\max\left(0, \frac{1-s}{2(s+1)}\right) \text{ for } j \geq 1$$

we show in lemma B.3 that $\lim_{j\to\infty} \zeta_j = 1$. Specifically, this choice of e allows us to cover the $\lfloor k^\gamma \rfloor$ largest probabilities without *overshooting*. Lemma B.6 provides the full details of the argument, which yields for every $0 < \omega < 1$ some N such that for every γ in a constant sized neighborhood around 1, there is a single N such that for $n \geq N$, the event $\|\boldsymbol{Z}_n - \mathbb{E}\boldsymbol{Z}_n\|_1 < (1-\omega)Cn^{-\frac{1}{2}+e}$ implies that the top $I(k,\gamma)$ elements are in perfect order. In the precise theorem statement $I(k,\gamma)$ is defined ± 1 of the $I(k,\gamma)$ discussed in this section, but this makes no conceptual difference to the arguments explained in this section. Helper lemmas B.4 and B.5 establish non-asymptotic lower bounds on the D distance which are crucial to lemma B.6. That lemma B.6 is uniformly true for sufficiently large N across γ values in a neighborhood of 1 will play a role in the full strategy of the proof for $\beta < \frac{1}{B(s)}$, which will be detailed in the next subsection.

When $\beta \geq \frac{1}{B(s)}$, we can only establish the existence of a good event for the highest $\asymp n^{\frac{1}{B(s)}-\epsilon}$ probabilities for $\epsilon > 0$ arbitrarily small. (This is because the $n^{\frac{1}{B(s)}}$ highest probability satisfies that the order of the standard deviation of the empirical proportion for this rank is the same as the order of the D distance at this rank). This is done by setting

$$e = \frac{\epsilon}{2}\beta B(s) \tag{50}$$

as in equation 66.

Two final notes are in order. The first is that to formally argue that the complements of these good events actually have small probability using lemma B.1, we have to convert χ_n into the form $Cu_n + u_n^2\frac{C^2}{2}\frac{1}{\min_{j\in[k]}\sqrt{p_j(1-p_j)}}$. This is possible by setting $Cu_n = \frac{\chi_n}{2}$ and then arguing that $u_n^2\frac{1}{\min_{j\in[k]}\sqrt{p_j(1-p_j)}} = o(u_n)$. This argument is carried out in the $\beta < \frac{1}{B(s)}$ case in lemma B.2. For $s > 2$ and $\frac{1}{B(s)} < \beta < \frac{1}{s}$, this argument is carried out in equation 157. For $\beta \geq \frac{1}{s}$, it is no longer possible to argue that the second argument will be asymptotically dominated by u_n. This is resolved when $s > 2$ by constructing a new Multinomial distribution that groups together all categories with probabilites no larger than the $\lfloor n^{\frac{1}{s}} \rfloor$ largest probability. (See section B.7 for details). The second is that in lemma B.6, when $s = 1$, there is an additional log factor in χ_n (see equation 67 for the definition of $e_{\gamma,2}$). This is because when $s = 1$ the normalizer grows at a $\log(n)$ rate.

B.1.2 Sort and Snap upper bounds when $\beta < \frac{1}{B(s)}$

The upper bound proof for Sort and Snap in the $\beta < \frac{1}{B(s)}$ case is theorem B.7. For a γ_k growing towards 1, we use two good events. A_{1k} is the event that the top $I(k,\gamma_k) - 1$ counts are in perfect order. A_{2k} is the event that the top $I(k, \gamma = 1) = k$ counts are in perfect order. The expectation is decomposed on these events as in equation 112.

When A_{2k} occurs, Sort and Snap yields no error, so there is no contribution to the expectation. There are only two events left over. They are $A_{2k}^C \cap A_{1k}$ and $A_{2k}^C \cap A_{1k}^C$. For the latter event, the only reasonable upper bound on the 1–norm between Sort and Snap and the truth is constant sized, but $\mathbb{P}(A_{2k}^C)$ is controllable using the ideas of subsection B.1.1. For the former event, because the top $I(k,\gamma_k) - 1$ counts are in perfect order, the worst case error of Sort and Snap is not worse than $k - I(k,\gamma_k) + 1$ times the error of the count associated with the $I(k,\gamma_k)$ largest probability being the smallest count. And we again use the ideas of section B.1.1, this time to control $\mathbb{P}(A_{1k}^C)$.

[1] Note ζ_j can be simplified to $\frac{s}{2(s+1)}\zeta_{j-1} + \frac{s+2}{2(s+1)}$; the more verbose way of writing ζ_j is used to simultaneously also express ζ_j in the $\beta \geq \frac{1}{B(s)}$ case; specifically the only difference between the definition of ζ_j in the two cases is the different choice of e. See the below for the choice of e in the $\beta \geq \frac{1}{B(s)}$ case).

For $\beta < \frac{1}{B(s)}$, the Le Cam minimax lower bounds (see theorem B.8) involve flipping the last two probabilities, which yields a 1–norm error of $n^{-\beta(B(s)-1)}$. In order to match the polynomial on the outside of the exponential in this lower bound (up to the union bound factor k) [2] we need to find γ_k so that

$$\mathbb{P}(A_{2k}^C) \lesssim (\text{ Worse Case Error Of Sort and Snap under event } A_{1k} \cap A_{2k}^C) \times \mathbb{P}(A_{1k}^C) \tag{51}$$

and

$$(\text{ Worse Case Error Of Sort and Snap under event } A_{1k} \cap A_{2k}^C) \asymp n^{-\beta(B(s)-1)} \tag{52}$$

There is again a balancing act, this time based on the selection of γ_k. γ_k must be selected sufficiently small so we can satisfy criteria 51. But for a very small value of γ_k, the 1–norm error on the back end is too large and criteria 52 becomes impossible to satisfy. With the choice of γ_k given in equations 108 and 109 we strike a perfect balance for $\beta < \frac{1}{B(s)+1}$ and are able to precisely match the polynomial term (up to the union bound factor) in our Le Cam lower bounds. When $\frac{1}{B(s)+1} \leq \beta < \frac{1}{B(s)}$, the choices for γ_k satisfying criteria 51 all yield 1–norm error on the back end exceeding $n^{-\beta(B(s)-1)}$ and in the final statement of the upper bound this yields the extra non zero polynomial factor $n^{2(\beta(B(s)+1)-(1-\tau))}$ for arbitrarily small τ. This extra factor does not appear in the Le Cam lower bounds.

B.1.3 Sort and Snap upper bounds when $\beta \geq \frac{1}{B(s)}$, $s > 2$

Here we describe the main ideas of theorems B.9 (which covers $\frac{1}{s+2} \leq \beta < \frac{1}{s}$) and theorem B.10 (which covers $\beta \geq \frac{1}{s}$). The deviation between the proof strategies in these two cases is minor, so first we describe arguments that are common to both of these proofs.

Specifically, we use a truncation of Sort and Snap; namely Sort and Snap is used for the top

$$\asymp \lfloor n^{\frac{1}{s+2}-\epsilon} \rfloor$$

largest categories in the sample for ϵ arbitrarily small and then the EPE is used as the estimator for the remaining categories. The event $A_{3,k}$ indicates that the top $\lfloor n^{\frac{1}{s+2}-\epsilon} \rfloor$ counts are in perfect order. When $A_{3,k}$ is true, the only error left is the error of estimating the EPE in the tail. Also, using the Segmentation approach it is argued that $\mathbb{P}(A_{3,k}^C)$ decays at an exponential rate. What is left is to compute the EPE error for the tail probabilities. After bounding expected absolute deviation, Riemann integration is used to upper bound the error of the EPE in the tail. This leads to the upper bounds of equation 163.

Note that when the EPE error in the smallest tail such that exponential error rates are achievable in estimating the head is of the same order as the EPE error for estimating the entire probability vector (which occurs when $s \leq 2$), there is no benefit to using truncated Sort and Snap over using the EPE. This is why when $\beta \geq \frac{1}{B(s)}$ and $s \leq 2$, we do not present Sort and Snap upper bounds as a part of the main results; one may as well use the EPE on the entire probability vector in this case.

The argument in theorem B.9 used to upper bound $\mathbb{P}(A_{3,k}^C)$ does not work when $\beta \geq \frac{1}{s}$. The issue is that χ_n is not the dominantly small term in Bernstein's inequality because the $\min_j \sqrt{p_j(1-p_j)}$ is too small. For this reason, there is a separate proof for the $\beta \geq \frac{1}{s}$ case. It is provided in theorem B.10. In this case, we use a new Multinomial. It groups together as one category all of the categories that are not one of the largest $\lfloor n^{\frac{1}{s}} \rfloor - 1$ categories. Even when $s > 2$, this new category is large enough so that χ_n is now the dominating factor. The proof now proceeds similarly to theorem B.9.

B.2 A note on use of the letter N

In the proofs of appendix B, the following notation ambiguity regarding the use of the letter N is used: In a single proof, one may see the following sequence of statements:

[2] The extra factor k, due to the union bound, is likely also removable using the methods of Talagrand so that there is a completely tight match in the polynomial on the outside of the exponential, but we do not investigate that in this work

1. There exists an N such that for $n \geq N$, statement A holds
2. There exists an N such that for $n \geq N$, statement B holds
3. Now suppose $n \geq N$

The question thus arises, which value of N is being referred to in (3)? The one indicating statement A or the one indicating statement B? If the N indicating statement A is denoted N_1 and the N indicating statement B is denoted statement N_2, then both A and B are true for $n \geq N_3 = \max(N_1, N_2)$. Rather than introducing an unwieldy large collection of N_j variables in these proofs, we reuse the letter N repeatedly. The reader should keep this in mind.

B.3 Infinity norm concentration of variance adjusted Multinomial

Lemma B.1. *Let $k \geq 2$. If $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} Categorical(\boldsymbol{p} = (p_1, p_2, \ldots, p_k))$ (i.e $\mathbb{P}(Y_1 = j) = p_j$ for $j \in [k]$) and*

$$\boldsymbol{X} := (\sum_{i=1}^{n} \mathbb{I}(Y_i = j))_{j=1}^{k}$$

(so that $\boldsymbol{X} \sim Multinomial(n, \boldsymbol{p} = (p_1, p_2, \ldots, p_k))$), and $Z_j = \frac{X_j}{n\sqrt{p_j(1-p_j)}}$ for $j \in [k]$ and $\boldsymbol{Z} := (Z_1, Z_2, \ldots, Z_k)$. Then for $u, C > 0$

$$\mathbb{P}\left(\|\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}\|_\infty \geq Cu + \frac{C^2u^2}{2} \frac{1}{\min_{j\in[k]} \sqrt{p_j(1-p_j)}}) \right) \leq 2k \exp(-\frac{nu^2C^2}{4})$$

Proof. For $i \in [n]$ and $j \in [k]$, let

$$W_{ij} := \frac{1}{n\sqrt{p_j(1-p_j)}} \left(\mathbb{I}(Y_i = j) - p_j\right)$$

Then for $j \in [k]$ and $\epsilon, t > 0$, by Markov's inequality

$$\mathbb{P}\left(Z_j - \mathbb{E}Z_j \geq \epsilon\right) \leq \mathbb{E}\exp(t(Z_j - \mathbb{E}Z_j))\exp(-t\epsilon) = \mathbb{E}\exp(t\sum_{i=1}^{n} W_{ij})\exp(-t\epsilon) \tag{53}$$

Now using the taylor series form of $\exp(x)$ and that $\sum_{j=2}^{\infty} \frac{1}{j!} = e - 2$, note that for $|x| \leq 1$

$$|\exp(x) - 1 - x| \leq \sum_{j=2}^{\infty} \frac{|x|^j}{j!} \leq x^2 \sum_{j=2}^{\infty} \frac{1}{j!} = x^2(e-2) \leq x^2 \tag{54}$$

Now note that $|W_{ij}| \leq \frac{1}{n\sqrt{p_j(1-p_j)}}$. Therefore,

$$|tW_{ij}| \leq \frac{t}{n\sqrt{p_j(1-p_j)}}$$

In particular, as long as $t \leq n\sqrt{p_j(1-p_j)}$, we apply equation 54 to get that with probability 1

$$|\exp(tW_{ij}) - 1 - tW_{ij}| \leq t^2 W_{ij}^2 \tag{55}$$

Taking the expectation on both sides and then applying Jensen's inequality for the absolute value on the left hand side and then using that $\mathbb{E}W_{ij} = 0$, we have that when $t \leq n\sqrt{p_j(1-p_j)}$

$$|\mathbb{E}\exp(tW_{ij}) - 1| \leq t^2 \mathbb{V}(W_{ij}) \tag{56}$$

Using this and that $1 + z \leq \exp(z)$ for every z, we have for every $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, k\}$ and $t \leq n\sqrt{p_j(1-p_j)}$

$$\mathbb{E}\exp(tW_{ij}) \leq 1 + t^2\mathbb{V}(W_{ij}) \leq \exp(t^2 \mathbb{V}W_{ij}) \tag{57}$$

Using this and equation 53 and that $\mathbb{V}(W_{ij}) = \frac{1}{n^2}$ for each $i \in [n]$ and $j \in [k]$ we have that for $j \in [k]$ and $t \leq n\sqrt{p_j(1-p_j)}$

$$\mathbb{P}\left(Z_j - \mathbb{E}Z_j \geq \epsilon\right) \leq \prod_{i=1}^{n} \mathbb{E}\exp(tW_{ij})\exp(-t\epsilon) = \exp(-t\epsilon)\exp(\frac{t^2}{n}) = \exp(-(t\epsilon - \frac{t^2}{n})) = \exp(-g(t)) \tag{58}$$

where $g(t) = t\epsilon - \frac{t^2}{n}$. $g(t)$ is maximized at $t_1^* = \frac{n\epsilon}{2}$. Also, let $t_2^* = n\sqrt{p_j(1-p_j)}$. So when $t_1^* = \frac{n\epsilon}{2} \leq t_2^* = n\sqrt{p_j(1-p_j)}$ we have that

$$\mathbb{P}(Z_j - \mathbb{E}Z_j \geq \epsilon) \leq \exp(-g(t_1^*)) = \exp(-\frac{n\epsilon^2}{4}) \tag{59}$$

And if $t_1^* > t_2^*$, we have that

$$g(t_2^*) = t_2^*\epsilon - \frac{(t_2^*)^2}{n} = t_2^*\epsilon - t_2^*\sqrt{p_j(1-p_j)} \geq t_2^*\epsilon - \frac{\epsilon n\sqrt{p_j(1-p_j)}}{2} = t_2^*\epsilon - \frac{\epsilon}{2}t_2^* = \frac{\epsilon t_2^*}{2} \tag{60}$$

In particular, when $t_1^* > t_2^*$

$$\mathbb{P}(Z_j - \mathbb{E}Z_j \geq \epsilon) \leq \exp(-\frac{\epsilon t_2^*}{2}) = \exp(-\frac{n\epsilon}{\frac{2}{\sqrt{p_j(1-p_j)}}}) \tag{61}$$

By equations 59 and 61, we have that for $j \in [k]$

$$\mathbb{P}(Z_j - \mathbb{E}Z_j \geq \epsilon) \leq \exp(-n\min\left(\frac{\epsilon^2}{4}, \frac{\epsilon}{\frac{2}{\sqrt{p_j(1-p_j)}}}\right)) \tag{62}$$

By the same inequality holds for $-(Z_j - \mathbb{E}Z_j)$. Thus now applying the union bound, we have that for $\epsilon > 0$

$$\mathbb{P}\left(\|\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}\|_\infty \geq \epsilon\right) \leq 2\sum_{j=1}^{k}\exp(-n\min\left(\frac{\epsilon^2}{4}, \frac{\epsilon\sqrt{p_j(1-p_j)}}{2}\right)) \leq 2k\exp\left(-n\min(\frac{\epsilon^2}{4}, \frac{\epsilon}{\frac{2}{\min_{j\in[k]}\sqrt{p_j(1-p_j)}}})\right) = 2k\exp\left(-n\min(\frac{\epsilon^2}{d_2^2}, \frac{\epsilon}{d_1})\right) \tag{63}$$

where $d_1 := \frac{2}{\min_{j\in[k]}\sqrt{p_j(1-p_j)}}$ and $d_2 = 2$. Now for $a > 0$, let $\epsilon = a^2d_1 + ad_2$. Applying equation 63, we have that

$$\mathbb{P}\left(\|\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}\|_\infty \geq a^2d_1 + ad_2\right) \leq 2k\exp(-n\min\left(\frac{a^4d_1^2 + a^2d_2^2}{d_2^2}, \frac{a^2d_1 + ad_2}{d_1}\right)) \leq 2k\exp(-na^2) \tag{64}$$

Finally, set $a = \frac{Cu}{2}$ for $u > 0$ and the lemma statement follows from equation 64. □

B.4 Good event identification

This section lays the foundation for the proofs of the upper bounds for SS and TSS.

Let $s > 0$ and $\beta > 0$. We define the *inverse of the intermediate dimension*, depending on s as

$$B(s) := 2 + max(1, s) = s + 2 + \max(0, 1 - s) \tag{65}$$

The equivalent forms of $B(s)$ provided will be useful later. For each $0 < \gamma \leq 1$, we define two quantities $e_{\gamma,1}, e_{\gamma,2}$ that will index the radius of the infinity ball around the normalized multinomial. Also let $\epsilon > 0$.

$$e_{\gamma,1} := \begin{cases} \frac{\epsilon}{2}\beta B(s) & \beta \geq \frac{1}{B(s)}, s > 2 \\ \frac{1-\beta(\gamma(s+2)+\max(0,1-s))}{2} & \beta < \frac{1}{B(s)} \end{cases} \tag{66}$$

and

$$e_{\gamma,2} := \begin{cases} 0 & s \neq 1, \beta > 0 \\ \frac{1}{2} & s = 1, \beta > 0 \end{cases} \tag{67}$$

$e_{\gamma,1}, e_{\gamma,2}$ of course depend on β, s, but they are suppressed in the notation without causing ambiguities.

Lemma B.2. *Let $s > 0$ and $0 < \beta < \frac{1}{B(s)}$ and $k = \lfloor n^\beta \rfloor$ (for each n large enough so that $\lfloor n^\beta \rfloor \geq 2$). Also let $C > 0, \omega > 0$ and for $\ell \geq 2$, let $\boldsymbol{p}^\ell \in \mathcal{P}_{f_s,\ell}$. Then there exists a single N and a $0 < \Gamma_s < 1$ such that for every $1 \geq \gamma > \Gamma_s$ and $n \geq N$*

$$Cn^{-\frac{1}{2}+e_{\gamma,1}} \log(n)^{-e_{\gamma,2}} + (n^{-\frac{1}{2}+e_{\gamma,1}} \log(n)^{-e_{\gamma,2}})^2 \frac{C^2}{2} \frac{1}{\min_{j\in[k]} \sqrt{\boldsymbol{p}^k(j)(1-\boldsymbol{p}^k(j))}} \leq (1+\omega) Cn^{-\frac{1}{2}+e_{\gamma,1}} \log(n)^{-e_{\gamma,2}}$$

Proof. First note that

$$n^{-\frac{1}{2}+e_{\gamma,1}} = n^{-\frac{\beta(\gamma(s+2)+\max(0,1-s))}{2}} \tag{68}$$

and since $\boldsymbol{p}^k$ is s–Zipfian and the order of $H_{k,s}$ (lemma A.1), and the definition of $e_{\gamma,2}$ (see equation 67), we have that

$$\begin{aligned} (n^{-\frac{1}{2}+e_{\gamma,1}} \log(n)^{-e_{\gamma,2}})^2 \frac{1}{\min_{j\in[k]} \sqrt{\boldsymbol{p}^k(j)(1-\boldsymbol{p}^k(j))}} &\lesssim \\ \log(n)^{-2e_{\gamma,2}} \sqrt{H_{k,s}} n^{-\beta(\gamma(s+2)+\max(0,1-s))} k^{s/2} &\lesssim \\ \log(n)^{-2e_{\gamma,2}} n^{-\beta(\gamma(s+2)+\max(0,1-s))} \begin{cases} n^{\frac{\beta}{2}} & 0 < s < 1 \\ \sqrt{\log(n)} n^{\frac{s\beta}{2}} & s = 1 \\ n^{\frac{s\beta}{2}} & s > 1 \end{cases} &= \\ \log^{-e_{\gamma,2}}(n) n^{-\beta(\gamma(s+2)+\max(0,1-s)-\frac{\max(1,s)}{2})} & \end{aligned} \tag{69}$$

Now note that if for $\Gamma_s, \Gamma \in (0,1)$ such that $\Gamma_s > \Gamma > \frac{1}{2} + \frac{1}{2(s+2)} (\max(1,s) - \max(0,1-s)) = \frac{1}{2} + \frac{s}{2(s+2)}$, we have that

$$\begin{aligned} \max_{\gamma\in(\Gamma_s,1]} \frac{\gamma(s+2)+\max(0,1-s)}{2} &\leq \\ \frac{(s+2)+\max(0,1-s)}{2} &< \\ \Gamma(s+2)+\max(0,1-s) - \frac{\max(1,s)}{2} &\leq \\ \min_{\gamma\in(\Gamma_s,1]} \gamma(s+2)+\max(0,1-s) - \frac{\max(1,s)}{2} & \end{aligned} \tag{70}$$

Using equation 70, we have that for any $C_1 > 0$ there is a N (depending on C_1) such that for every $\gamma \in (\Gamma_s, 1)$ and $n \geq N$

$$\begin{aligned} n^{-\beta(\gamma(s+2)+\max(0,1-s)-\frac{\max(1,s)}{2})} &\leq \\ n^{-\beta\left(\Gamma(s+2)+\max(0,1-s)-\frac{\max(1,s)}{2}\right)} &\leq \\ C_1 n^{-\beta\left(\frac{(s+2)+\max(0,1-s)}{2}\right)} &\leq \\ C_1 n^{-\beta\left(\frac{\gamma(s+2)+\max(0,1-s)}{2}\right)} & \end{aligned} \tag{71}$$

Using equations 69 and 71 and setting $C_1 = \frac{2\omega}{C}$ yields an N such that for $n \geq N$ and each $\gamma \in (\Gamma_s, 1]$

$$(n^{-\frac{1}{2}+e_{\gamma,1}} \log(n)^{-e_{\gamma,2}})^2 \frac{C^2}{2} \frac{1}{\min_{j\in[k]} \sqrt{\boldsymbol{p}^k(j)(1-\boldsymbol{p}^k(j))}} \leq \log^{-e_{\gamma,2}}(n) C\omega n^{-\beta\left(\frac{\gamma(s+2)+\max(0,1-s)}{2}\right)} = C\omega \log^{-e_{\gamma,2}}(n) n^{-\frac{1}{2}+e_{\gamma,1}} \tag{72}$$

where the equality in the above line is due to equation 68.

□

Now recursively define for $\beta > 0$ and $0 < \gamma \leq 1$ and $s > 0$

$$\zeta_{0,\gamma} := 0 \text{ and for } j \geq 1, \zeta_{j,\gamma} := \frac{1}{\beta\gamma(s+1)}\left(\frac{1 + s\beta\gamma\zeta_{j-1,\gamma}}{2} - e_{\gamma,1}\right) - \frac{1}{\gamma}\max\left(0, \frac{1-s}{2(s+1)}\right) \tag{73}$$

Note $\zeta_{j,\gamma}$ only depends on γ when $\beta \geq \frac{1}{B(s)}$. This is because γ cancels when plugging in $e_{\gamma,1}$ (defined in equation 66) in the $\beta < \frac{1}{B(s)}$ case. Specifically, when $\beta < \frac{1}{B(s)}$

$$\begin{aligned}\zeta_{j,\gamma} = \frac{1}{\beta\gamma(s+1)}\left(\frac{1 + s\beta\gamma\zeta_{j-1,\gamma}}{2} - \frac{1 - \beta(\gamma(s+2) + \max(0, 1-s))}{2}\right) - \frac{1}{\gamma}\max\left(0, \frac{1-s}{2(s+1)}\right) = \\ \frac{1}{2\beta\gamma(s+1)}\left(\beta\gamma s\zeta_{j-1,\gamma} + \beta\gamma(s+2)\right) = \\ \frac{s}{2(s+1)}\zeta_{j-1,\gamma} + \frac{s+2}{2(s+1)}\end{aligned} \tag{74}$$

Lemma B.3. *Let $s > 0$ and $0 < \gamma \leq 1$ and $\beta > 0$ and $0 < \epsilon < \frac{1}{\beta B(s)}$. Also suppose either $0 < \beta < \frac{1}{B(s)}$ or $s > 2$. Then $\zeta_{j,\gamma}$ is strictly increasing in j for fixed γ and $\gamma\zeta_{j,\gamma}$ is non-decreasing in γ for fixed j and*

$$\lim_{j\to\infty}\zeta_{j,\gamma} = \begin{cases} 1 & \beta < \frac{1}{B(s)} \\ \frac{1}{\beta\gamma(s+2)} - \frac{\epsilon}{\gamma} & \beta \geq \frac{1}{B(s)}, s > 2 \end{cases}$$

Proof. We will first use induction to prove that for $j \geq 1$ for any (β, s) specified by the lemma statement,

$$\zeta_{j,\gamma} = \left(\frac{\frac{1}{2} - e_{\gamma,1}}{\beta\gamma(s+1)} - \frac{1}{\gamma}\max(0, \frac{1-s}{2(s+1)})\right)\sum_{t=0}^{j-1}(\frac{1}{2}\frac{s}{(s+1)})^t \tag{75}$$

For the base case, note that since $\zeta_{0,\gamma} = 0$, we have that

$$\zeta_{1,\gamma} = \left(\frac{\frac{1}{2} - e_{\gamma,1}}{\beta\gamma(s+1)} - \frac{1}{\gamma}\max(0, \frac{1-s}{2(s+1)})\right)\sum_{t=0}^{1-1}(\frac{1}{2}\frac{s}{(s+1)})^t$$

Suppose equation 75 is true for some $j \geq 1$. Then

$$\begin{aligned}\zeta_{j+1,\gamma} = \frac{1}{\beta\gamma(s+1)}\left(\frac{1 + s\beta\gamma\left(\left(\frac{\frac{1}{2}-e_{\gamma,1}}{\beta\gamma(s+1)} - \frac{1}{\gamma}\max(0, \frac{1-s}{2(s+1)})\right)\sum_{t=0}^{j-1}\left(\frac{1}{2}\frac{s}{(s+1)}\right)^t\right)}{2} - e_{\gamma,1}\right) - \frac{1}{\gamma}\max\left(0, \frac{1-s}{2(s+1)}\right) = \\ \frac{1}{\beta\gamma(s+1)}\left(\frac{1}{2} - e_{\gamma,1} + \frac{s}{2(s+1)}\left(\frac{1}{2} - e_{\gamma,1} - \max(0, \frac{1-s}{2})\beta\right)\sum_{t=0}^{j-1}\left(\frac{1}{2}\frac{s}{(s+1)}\right)^t\right) - \frac{1}{\gamma}\max(0, \frac{1-s}{2(s+1)}) = \\ \frac{\frac{1}{2} - e_{\gamma,1}}{\beta\gamma(s+1)} + \frac{1}{\beta\gamma(s+1)}\left(\frac{1}{2} - e_{\gamma,1} - \max(0, \frac{1-s}{2})\beta\right)\sum_{t=1}^{j}\left(\frac{1}{2}\frac{s}{(s+1)}\right)^t - \frac{1}{\gamma}\max(0, \frac{1-s}{2(s+1)}) = \\ \frac{\frac{1}{2} - e_{\gamma,1}}{\beta\gamma(s+1)} + \left(\frac{\frac{1}{2} - e_{\gamma,1}}{\beta\gamma(s+1)} - \frac{1}{\gamma}\max(0, \frac{1-s}{2(s+1)})\right)\sum_{t=1}^{j}\left(\frac{1}{2}\frac{s}{(s+1)}\right)^t - \frac{1}{\gamma}\max(0, \frac{1-s}{2(s+1)}) = \\ \left(\frac{\frac{1}{2} - e_{\gamma,1}}{\beta\gamma(s+1)} - \frac{1}{\gamma}\max(0, \frac{1-s}{2(s+1)})\right)\sum_{t=0}^{j}(\frac{1}{2}\frac{s}{(s+1)})^t\end{aligned} \tag{76}$$

Thus by induction equation 75 holds for every $j \geq 1$. Using this and that by definition $e_{\gamma,1}$ is non-increasing in γ, we conclude that for each $j \geq 1$, $\gamma\zeta_{j,\gamma}$ is non-decreasing in γ. To conclude that $\zeta_{j,\gamma}$ is strictly increasing in j it is sufficient to show that $\frac{\frac{1}{2}-e_{\gamma,1}}{\beta\gamma(s+1)} - \frac{1}{\gamma}\max(0, \frac{1-s}{2(s+1)}) > 0$. When $\beta < \frac{1}{B(s)}$, by definition of $e_{\gamma,1}$ (see equation

66), we have that

$$\begin{aligned} \frac{\frac{1}{2} - e_{\gamma,1}}{\beta\gamma(s+1)} - \frac{1}{\gamma}\max(0, \frac{1-s}{2(s+1)}) = \\ \frac{1}{\gamma(s+1)}\left(\frac{\frac{1}{2} - e_{\gamma,1}}{\beta} - \max(0, \frac{1-s}{2})\right) = \\ \frac{1}{\gamma(s+1)}\left(\frac{\beta(\gamma(s+2) + \max(0, 1-s)}{2\beta} - \max(0, \frac{1-s}{2})\right) = \\ \frac{s+2}{2(s+1)} > 0 \end{aligned} \tag{77}$$

And when $\beta \geq \frac{1}{B(s)}$ and $s > 2$, again by definition of $e_{\gamma,1}$, and now using that $\epsilon < \frac{1}{\beta B(s)}$

$$\frac{\frac{1}{2} - e_{\gamma,1}}{\beta\gamma(s+1)} - \frac{1}{\gamma}\max(0, \frac{1-s}{2(s+1)}) = \frac{1 - \epsilon\beta B(s)}{2\beta\gamma(s+1)} > 0 \tag{78}$$

By equations 77 and 78 we have conclude that whenever $\beta < \frac{1}{B(s)}$ or $s > 2$, $\zeta_{j,\gamma}$ in increasing in j. Finally, taking the limit as $j \to \infty$ in equation 75 and using the formula for a geometric series and equations 77 and 78 we conclude that

$$\lim_{j\to\infty} \zeta_{j,\gamma} = \frac{1}{1 - \frac{1}{2}\frac{s}{(s+1)}} \begin{cases} \frac{s+2}{2(s+1)} & 0 < \beta < \frac{1}{B(s)} \\ \frac{1-\epsilon\beta B(s)}{2\beta\gamma(s+1)} & \frac{1}{B(s)} \leq \beta, s > 2 \end{cases} = \begin{cases} 1 & 0 < \beta < \frac{1}{B(s)} \\ \frac{1}{\beta\gamma(s+2)} - \frac{\epsilon}{\gamma} & \frac{1}{B(s)} \leq \beta, s > 2 \end{cases} \tag{79}$$

□

Now for $k \geq 2$ and $\boldsymbol{p} \in \mathcal{S}_k$ and $i \in [k]$, define the distance to the closest adjacent probability from the probability of the i^{th} category as

$$D_{\boldsymbol{p},i} = \min_{j \in [k]-i} |\boldsymbol{p}(i) - \boldsymbol{p}(j)| \tag{80}$$

Lemma B.4. *For $s > 0$, $k \geq 2$ and $\boldsymbol{p}^k \in \mathcal{P}_{f_s,k}$ with permutation function π_k*

1. *$D_{\boldsymbol{p}^k,\pi_k^{-1}(k)} = \frac{f_s(k-1) - f_s(k)}{H_{k,s}}$ and if $i \in [k-1]$ then $D_{\boldsymbol{p}^k,\pi_k^{-1}(i)} = \frac{f_s(i) - f_s(i+1)}{H_{k,s}}$*
2. *For $i_1, i_2 \in [k]$, if $i_1 < i_2$ then $D_{\boldsymbol{p}^k,\pi_k^{-1}(i_1)} \geq D_{\boldsymbol{p}^k,\pi_k^{-1}(i_2)}$*

Proof. Recall for $i \in [k]$, $\boldsymbol{p}^k(\pi_k^{-1}(i)) = \frac{f_s(i)}{\sum_{j=1}^k f_s(j)} = \frac{f_s(i)}{H_{k,s}}$. Also since f_s is monotonically decreasing, the minima in equation 80 is realized by an adjacent index. That is, for $i \in [k]$

$$D_{\boldsymbol{p}^k,\pi_k^{-1}(i)} = \frac{1}{H_{k,s}} \begin{cases} f_s(1) - f_s(2) & i = 1 \\ \min(f_s(i-1) - f_s(i), f_s(i) - f_s(i+1)) & i \in \{2, 3, \ldots, k-1\} \\ f_s(k-1) - f_s(k) & i = k \end{cases} \tag{81}$$

For each $i \in \{2, 3, \ldots, k-1\}$, by lemma B.11 with $x_1 = i-1$ and $x_2 = i$,

$$\min(f_s(i-1) - f_s(i), f_s(i) - f_s(i+1)) = f_s(i) - f_s(i+1)$$

Using this and equation 81, we conclude (1). Specifically

$$D_{\boldsymbol{p}^k,\pi_k^{-1}(i)} = \frac{1}{H_{k,s}} \begin{cases} f_s(i) - f_s(i+1) & i \in [k-1] \\ f_s(k-1) - f_s(k) & i = k \end{cases} \tag{82}$$

Now let $i_1, i_2 \in [k-1]$ and $i_1 < i_2$. By equation 82 and using lemma B.11 with $x_1 = i_1$ and $x_2 = i_2$ (noting that $i_1 - i_2 \geq 1$), we have that

$$D_{\boldsymbol{p}^k,\pi_k^{-1}(i_1)} = \frac{f_s(i_1) - f_s(i_1+1)}{H_{k,s}} \geq \frac{f_s(i_2) - f_s(i_2+1)}{H_{k,s}} = D_{\boldsymbol{p}^k,\pi_k^{-1}(i_2)} \tag{83}$$

Also if $i_1 \in [k-2]$, then again using lemma B.11, this time with $x_1 = i_1$ and $x_2 = k-1$, we have that

$$D_{\boldsymbol{p}^k,\pi_k^{-1}(i_1)} = \frac{f_s(i_1) - f_s(i_1+1)}{H_{k,s}} \geq \frac{f_s(k-1) - f_s(k)}{H_{k,s}} = D_{\boldsymbol{p}^k,\pi_k^{-1}(k)} \tag{84}$$

And finally, note that $D_{\boldsymbol{p}^k,\pi_k^{-1}(k-1)} = D_{\boldsymbol{p}^k,\pi_k^{-1}(k)}$. Using this and equations 83 and 84 we conclude (2).

□

Lemma B.5. *Let $s > 0$ and $\beta > 0$ and $0 < \epsilon < \frac{1}{\beta B(s)}$ and $k := \lfloor n^\beta \rfloor$ (for each n large enough so that $\lfloor n^\beta \rfloor \geq 2$) and for $\ell \geq 2$, $\boldsymbol{p}^\ell \in \mathcal{P}_{f_s,\ell}$ (with permutation function denoted π_ℓ). Also, suppose either $0 < \beta < \frac{1}{B(s)}$ or $s > 2$. Then for any $0 < \omega < 1$ there is a N such that for every $1 \geq \gamma > \frac{1}{2}$ and every $j \geq 1$ and $n \geq N$*

$$D_{\boldsymbol{p}^k,\pi_k^{-1}(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor)} \geq s(1-\omega)n^{-\left(\frac{1+s\beta\gamma\zeta_{j-1,\gamma}}{2} - e_{\gamma,1}\right)} \times \begin{cases} (1-s)n^{-\frac{\beta(1-s)}{2}} & 0 < s < 1 \\ \frac{1}{\beta}\log^{-1}(n) & s = 1 \\ \frac{1}{R(s)} & s > 1 \end{cases} \tag{85}$$

Proof. By lemma B.3, $\gamma\zeta_{j,\gamma}$ is strictly increasing with limit at most 1 for fixed γ. Therefore we have that

$$\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor < k$$

for every $j \geq 0$ and $0 < \gamma \leq 1$. Thus applying lemma B.4 part (1), we have that for $j \geq 1$

$$\begin{aligned} D_{\boldsymbol{p}^k,\pi_k^{-1}(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor)} = \frac{1}{H_{k,s}}\left(f_s\left(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor\right) - f_s(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor + 1)\right) = \\ \frac{1}{H_{k,s}}\left(s(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor)^{-(s+1)} - \frac{s(s+1)}{2}\phi_j^{-(s+2)}\right) \geq \\ \frac{1}{H_{k,s}}\left(s(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor)^{-(s+1)} - \frac{s(s+1)}{2}(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor)^{-(s+2)}\right) \end{aligned} \tag{86}$$

where $\phi_j \in \left[\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor, \lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor + 1\right]$ and a second order Taylor series expansion of f_s about $\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor$ has been applied to get the second equality in the above equation while the inequality in the above equation is due to the monotonicity of f_{s+2}.

Now using that $\zeta_{0,\gamma} = 0$ and that by lemma B.3, both $\zeta_{j,\gamma}$ is strictly increasing in j for fixed γ and that $\gamma\zeta_{j,\gamma}$ is non-decreasing in γ for fixed j, we have that

$$\sup_{1 \geq \gamma > \frac{1}{2}} \max_{j \geq 1}(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor)^{-1} \leq \sup_{1 \geq \gamma > \frac{1}{2}} (\lfloor k^{\gamma\zeta_{1,\gamma}} \rfloor)^{-1} \leq \lfloor k^{\frac{1}{2}\zeta_{1,\frac{1}{2}}} \rfloor^{-1} = o(1) \tag{87}$$

Thus for $\omega_1 > 0$ there exists an N such that for all $n \geq N$

$$\sup_{1 \geq \gamma > \frac{1}{2}} \max_{j \geq 1}(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor)^{-1} \leq \frac{2\omega_1}{s+1} \tag{88}$$

By equations 86 and 88 and the definition of $\zeta_{j,\gamma}$, we have that there is a single N such that for every $1 \geq \gamma > \frac{1}{2}$ and $j \geq 1$ and $n \geq N$

$$\begin{aligned} D_{\boldsymbol{p}^k,\pi_k^{-1}(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor)} \geq \frac{1}{H_{k,s}}\left(s(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor)^{-(s+1)} - \frac{s(s+1)}{2}(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor)^{-(s+1)}(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor^{-1}\right) \geq \frac{s(1-\omega_1)}{H_{k,s}}\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor^{-(s+1)} \geq \\ \frac{s(1-\omega_1)}{H_{k,s}}n^{-\beta\gamma(s+1)\zeta_{j,\gamma}} = \\ \frac{s(1-\omega_1)}{H_{k,s}}n^{-\beta\gamma(s+1)\left(\frac{1}{\beta\gamma(s+1)}\left(\frac{1+s\beta\gamma\zeta_{j-1,\gamma}}{2} - e_{\gamma,1}\right) - \frac{1}{\gamma}\max(0,\frac{1-s}{2(s+1)})\right)} = \\ \frac{s(1-\omega_1)}{H_{k,s}}n^{-\left(\frac{1+s\beta\gamma\zeta_{j-1,\gamma}}{2} - e_{\gamma,1}\right) + \beta\max(0,\frac{1-s}{2})} \end{aligned} \tag{89}$$

By equation 89 and lemma A.1 for an upper bound on $H_{k,s}$, we conclude that for any $\omega_2 > 0$ there is a N such that for every $1 \geq \gamma > \frac{1}{2}$ and $j \geq 1$ and $n \geq N$

$$D_{\boldsymbol{p}^k, \pi_k^{-1}(\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor)} \geq s(1-\omega_2) n^{-\left(\frac{1+s\beta\gamma\zeta_{j-1,\gamma}}{2} - e_{\gamma,1}\right)} \times \begin{cases} (1-s)n^{-\frac{\beta(1-s)}{2}} & 0 < s < 1 \\ \frac{1}{\beta}\log^{-1}(n) & s = 1 \\ \frac{1}{R(s)} & s > 1 \end{cases} \tag{90}$$

□

At this point the reader is reminded of the constant, $C^*_{s,\beta}$, introduced in the main body of the paper, that plays a role in the remainder of the proofs.

$$C^*_{s,\beta} = s^2 \begin{cases} 1-s & 0 < s < 1 \\ \frac{1}{\beta} & s = 1 \\ \frac{1}{R(s)} & s > 1 \end{cases} \tag{91}$$

Lemma B.6. *(The Good Event) Suppose $s > 0, \beta > 0$ and for $\ell \geq 2$, $\boldsymbol{p}^\ell \in \mathcal{P}_{f_s,\ell}$ (with permutation function denoted π_ℓ) and $k := \lfloor n^\beta \rfloor$ (for each n large enough so that $\lfloor n^\beta \rfloor \geq 2$). Also assume either $0 < \beta < \frac{1}{B(s)}$ or $s > 2$ and $0 < \epsilon < \frac{1}{\beta B(s)}$. And let*

$$Y_{1k}, Y_{2k}, \ldots, Y_{nk} \overset{iid}{\sim} \boldsymbol{p}^k$$

where for $j \in [k]$, $\boldsymbol{p}^k(j) = \mathbb{P}(Y_{1k} = j)$. Also, for $j \in [k]$, let $X_{jn} = \sum_{i=1}^n \mathbb{I}(Y_{ik} = j)$ and let

$$Z_{jn} = \frac{X_{jn}}{n\sqrt{\boldsymbol{p}^k(j)(1-\boldsymbol{p}^k(j))}}$$

and $\boldsymbol{Z}_n = (Z_{1n}, Z_{2n}, \ldots, Z_{kn})$. And define for $0 < \gamma \leq 1$

$$I(k,\gamma,\epsilon) := \begin{cases} \lfloor k^\gamma \rfloor, & k^\gamma \notin \mathbb{N}, 0 < \beta < \frac{1}{B(s)} \\ \lfloor k^\gamma \rfloor - 1 & k^\gamma \in \mathbb{N}, 0 < \beta < \frac{1}{B(s)} \\ \lfloor k^{\frac{1}{\beta(s+2)} - \epsilon} \rfloor & k^{\frac{1}{\beta(s+2)} - \epsilon} \notin \mathbb{N}, \beta \geq \frac{1}{B(s)}, s > 2 \\ \lfloor k^{\frac{1}{\beta(s+2)} - \epsilon} \rfloor - 1 & k^{\frac{1}{\beta(s+2)} - \epsilon} \in \mathbb{N}, \beta \geq \frac{1}{B(s)}, s > 2 \end{cases} \tag{92}$$

Then for any $0 < \omega < 1$ there is a N such that for every $\frac{1}{2} < \gamma \leq 1$ and $n \geq N$

$$\left[\|\boldsymbol{Z}_n - \mathbb{E}\boldsymbol{Z}_n\|_\infty < \frac{(1-\omega)\sqrt{C^*_{s,\beta}}}{2} n^{-\frac{1}{2}+e_{\gamma,2}} \log^{-e_{\gamma,2}}(n)\right] \subseteq$$
$$\left[X_{\pi_k^{-1}(1),n} > X_{\pi_k^{-1}(2),n} > \cdots > X_{\pi_k^{-1}(I(k,\gamma,\epsilon)-1),n} \text{ and } \forall j \in \{I(k,\gamma,\epsilon), \cdots, k\}, X_{\pi_k^{-1}(I(k,\gamma,\epsilon)-1),n} > X_{\pi_k^{-1}(j),n}\right] \tag{93}$$

And when $0 < \beta < \frac{1}{B(s)}$, for this same N and $n \geq N$

$$\left[\|\boldsymbol{Z}_n - \mathbb{E}\boldsymbol{Z}_n\|_\infty < \frac{(1-\omega)\sqrt{C^*_{s,\beta}}}{2} n^{-\frac{1}{2}+e_{1,1}} \log^{-e_{1,2}}(n)\right] \subseteq$$
$$\left[X_{\pi_k^{-1}(1),n} > X_{\pi_k^{-1}(2),n} > \cdots > X_{\pi_k^{-1}(k-1),n} > X_{\pi_k^{-1}(k),n}\right] \tag{94}$$

Proof. By lemma B.3 there exists a $J \in \{1, 2, 3, \ldots\}$ (depending on k) such that

$$\lfloor k^{\gamma\zeta_{J,\gamma}} \rfloor = I(k,\gamma,\epsilon)$$

Also by definition, $\zeta_{0,\gamma} = 0$ and by lemma B.3, $\zeta_{j,\gamma}$ is increasing in j. Therefore,

$$1 = \lfloor k^{\gamma\zeta_{0,\gamma}} \rfloor \leq \lfloor k^{\gamma\zeta_{1,\gamma}} \rfloor \leq \lfloor k^{\gamma\zeta_{2,\gamma}} \rfloor \leq \cdots \leq \lfloor k^{\gamma\zeta_{J-1,\gamma}} \rfloor \leq \lfloor k^{\gamma\zeta_{J,\gamma}} \rfloor = I(k,\gamma,\epsilon) \tag{95}$$

In particular, for each $i \in [I(k,\gamma,\epsilon)]$, there exists a $j_i \in [J]$ such that

$$\lfloor k^{\gamma\zeta_{j_i-1,\gamma}} \rfloor \leq i \leq \lfloor k^{\gamma\zeta_{j_i,\gamma}} \rfloor \tag{96}$$

Using this and the definition of $\boldsymbol{p}^k$ and lemma B.12, we have that

$$\sqrt{\boldsymbol{p}^k(\pi_k^{-1}(i))(1-\boldsymbol{p}^k(\pi_k^{-1}(i))} \leq \sqrt{\boldsymbol{p}^k(\pi_k^{-1}(\lfloor k^{\gamma\zeta_{j_i-1,\gamma}} \rfloor))(1-\boldsymbol{p}^k(\pi_k^{-1}(\lfloor k^{\gamma\zeta_{j_i-1,\gamma}} \rfloor)))} \tag{97}$$

Also recall that by lemma B.3, for each $\frac{1}{2} < \gamma \leq 1$ and $j \geq 1$

$$\gamma\zeta_{j,\gamma} \geq \frac{1}{2}\zeta_{j,\frac{1}{2}} \geq \frac{1}{2}\zeta_{1,\frac{1}{2}} > 0$$

In particular, for $0 < \omega_1 < 1$ and each n sufficiently large so that $k \geq (\frac{1}{\omega_1})^{\frac{2}{\zeta_{1,\frac{1}{2}}}}$ and $n \geq (\frac{1}{\omega_1})^{\frac{2}{\beta\zeta_{1,\frac{1}{2}}}}$ we have that for $\frac{1}{2} < \gamma \leq 1$ and $j \geq 1$

$$\lfloor k^{\gamma\zeta_{j,\gamma}} \rfloor \geq k^{\gamma\zeta_{j,\gamma}} - 1 \geq (1-\omega_1)k^{\gamma\zeta_{j,\gamma}} \geq (1-\omega_1)(n^{\beta\gamma\zeta_{j,\gamma}} - 1) \geq (1-\omega_1)^2 n^{\beta\gamma\zeta_{j,\gamma}} \tag{98}$$

By definition of $\boldsymbol{p}^k$ and $e_{\gamma,1}$ and the ζ terms and the lower bound on $H_{k,s}$ (lemma A.1) and equation 98, for $0 < \omega_1 < 1$, there exists an N such that for each $i \in [I(k,\gamma,\epsilon)]$ and $\frac{1}{2} < \gamma \leq 1$ and $n \geq N$

$$\begin{aligned}
n^{-\frac{1}{2}+e_{\gamma,1}} \log^{-e_{\gamma,2}}(n) \sqrt{\boldsymbol{p}^k(\pi_k^{-1}(\lfloor k^{\gamma\zeta_{j_i-1,\gamma}} \rfloor))(1-\boldsymbol{p}^k(\pi_k^{-1}(\lfloor k^{\gamma\zeta_{j_i-1,\gamma}} \rfloor)))} \leq \\
\log^{-e_{\gamma,2}}(n) H_{k,s}^{-\frac{1}{2}} n^{-\frac{1}{2}+e_{\gamma,1}} (\lfloor k^{\gamma\zeta_{j_i-1,\gamma}} \rfloor)^{-s/2} \leq \\
\log^{-e_{\gamma,2}}(n) \frac{(1-\omega_1)^{-s}}{\sqrt{H_{k,s}}} n^{-\frac{1}{2}+e_{\gamma,1}} n^{-\frac{\beta\gamma s\zeta_{j_i-1,\gamma}}{2}} \leq \\
\log^{-e_{\gamma,2}}(n) \frac{(1-\omega_1)^{-s}}{\sqrt{H_{k,s}}} n^{-\left(\frac{1+s\beta\gamma\zeta_{j_i-1,\gamma}}{2} - e_{\gamma,1}\right)} \leq \\
(1-\omega_1)^{-(s+\frac{1}{2})} n^{-\left(\frac{1+s\beta\gamma\zeta_{j_i-1,\gamma}}{2} - e_{\gamma,1}\right)} \times \begin{cases} (1-s)^{\frac{1}{2}} n^{-\frac{\beta(1-s)}{2}} & 0 < s < 1 \\ \beta^{-\frac{1}{2}} \log^{-1}(n) & s = 1 \\ R(s)^{-\frac{1}{2}} & s > 1 \end{cases}
\end{aligned} \tag{99}$$

where in the last line we have used the definition of $e_{\gamma,2}$ (see equation 67). Using equation 99 and lemma B.5 (and the definition of $C^*_{s,\beta}$ given in equation 6), for $0 < \omega_1 < 1$, we have the existence of a single N such that for each $\frac{1}{2} < \gamma \leq 1$ and $i \in [I(k,\gamma,\epsilon)]$ and $n \geq N$

$$\begin{aligned}
\frac{(1-\omega_1)^{s+3/2}\sqrt{C^*_{s,\beta}}}{2} n^{-\frac{1}{2}+e_{\gamma,1}} \log^{-e_{\gamma,2}}(n) \sqrt{\boldsymbol{p}^k(\pi_k^{-1}(\lfloor k^{\gamma\zeta_{j_i-1,\gamma}} \rfloor))(1-\boldsymbol{p}^k(\pi_k^{-1}(\lfloor k^{\gamma\zeta_{j_i-1,\gamma}} \rfloor)))} \leq \\
\frac{(1-\omega_1)\sqrt{C^*_{s,\beta}}}{2} n^{-\left(\frac{1+s\beta\gamma\zeta_{j_i-1,\gamma}}{2} - e_{\gamma,1}\right)} \times \begin{cases} (1-s)^{\frac{1}{2}} n^{-\frac{\beta(1-s)}{2}} & 0 < s < 1 \\ \beta^{-\frac{1}{2}} \log^{-1}(n) & s = 1 \\ R(s)^{-\frac{1}{2}} & s > 1 \end{cases} = \\
\frac{(1-\omega_1)}{2} n^{-\left(\frac{1+s\beta\gamma\zeta_{j_i-1,\gamma}}{2} - e_{\gamma,1}\right)} \times s \begin{cases} (1-s) n^{-\frac{\beta(1-s)}{2}} & 0 < s < 1 \\ \beta^{-1} \log^{-1}(n) & s = 1 \\ R(s)^{-1} & s > 1 \end{cases} \leq \\
\frac{1}{2} D_{\boldsymbol{p}^k, \pi_k^{-1}(\lfloor k^{\gamma\zeta_{j_i,\gamma}} \rfloor)}
\end{aligned} \tag{100}$$

Thus by lemma B.4 (2), and equations 96, 97 and 100, we have that for any $0 < \omega_1 < 1$ there is a single N such that for each $\frac{1}{2} < \gamma \leq 1$ and $i \in [I(k, \gamma, \epsilon)]$ and $n \geq N$

$$\frac{(1-\omega_1)^{s+\frac{3}{2}}\sqrt{C^*_{s,\beta}}}{2} n^{-1/2+e_{\gamma,1}} \log^{-e_{\gamma,2}}(n) \sqrt{\boldsymbol{p}^k(\pi_k^{-1}(i))(1-\boldsymbol{p}^k(\pi_k^{-1}(i)))} \leq \frac{1}{2} D_{\boldsymbol{p}^k, \pi_k^{-1}(i)} \tag{101}$$

Now let $\gamma \in (\frac{1}{2}, 1]$ and $n \geq N$ and suppose

$$\|\boldsymbol{Z}_n - \mathbb{E}\boldsymbol{Z}_n\|_\infty < \frac{(1-\omega_1)^{s+\frac{3}{2}}\sqrt{C^*_{s,\beta}}}{2} n^{-\frac{1}{2}+e_{\gamma,1}} \log^{-e_{\gamma,2}}(n) \tag{102}$$

and let $i \in [I(k, \gamma, \epsilon) - 1]$. Note that since $\gamma \leq 1$, by equation 92, $I(k, \gamma, \epsilon) < k$ and thus $i, i+1 < k$. So by equations 101 and 102 and lemma B.4 (1) and lemma B.11, we have that for $n \geq N$

$$\begin{aligned} \frac{X_{\pi_k^{-1}(i),n}}{n} > \frac{f_s(i)}{H_{k,s}} - \frac{1}{2} D_{\boldsymbol{p}^k, \pi_k^{-1}(i)} = \frac{f_s(i)}{H_{k,s}} - \frac{f_s(i) - f_s(i+1)}{2H_{k,s}} = \frac{f_s(i+1)}{H_{k,s}} + \frac{f_s(i) - f_s(i+1)}{2H_{k,s}} \geq \\ \frac{f_s(i+1)}{H_{k,s}} + \frac{f_s(i+1) - f_s(i+2)}{2H_{k,s}} = \\ \frac{f_s(i+1)}{H_{k,s}} + \frac{1}{2} D_{\boldsymbol{p}^k, \pi_k^{-1}(i+1)} > \\ \frac{X_{\pi_k^{-1}(i+1),n}}{n} \end{aligned} \tag{103}$$

Also, using that f_s is monotonically decreasing and lemma B.12 and using equations 101 (with $i = I(k, \gamma, \epsilon)$) and 102 and lemma B.4 (2) and using the first two inequalities of equation 103 with $i = I(k, \gamma, \epsilon) - 1$, we have that for $\ell \in \{I(k, \gamma, \epsilon), I(k, \gamma, \epsilon) + 1, \ldots, k\}$ and $n \geq N$

$$\begin{aligned} \frac{X_{\pi_k^{-1}(\ell),n}}{n} < \\ \frac{f_s(\ell)}{H_{k,s}} + \frac{(1-\omega_1)^{s+\frac{3}{2}}\sqrt{C^*_{s,\beta}}}{2} n^{-\frac{1}{2}+e_{\gamma,1}} \log^{-e_{\gamma,2}}(n) \sqrt{\boldsymbol{p}^k(\pi_k^{-1}(\ell))(1-\boldsymbol{p}^k(\pi_k^{-1}(\ell)))} \leq \\ \frac{f_s(I(k,\gamma,\epsilon))}{H_{k,s}} + \frac{(1-\omega_1)^{s+\frac{3}{2}}\sqrt{C^*_{s,\beta}}}{2} n^{-\frac{1}{2}+e_{\gamma,1}} \log^{-e_{\gamma,2}}(n) \sqrt{\boldsymbol{p}^k(\pi_k^{-1}(I(k,\gamma,\epsilon)))(1-\boldsymbol{p}^k(\pi_k^{-1}(I(k,\gamma,\epsilon))))} \leq \\ \frac{f_s(I(k,\gamma,\epsilon))}{H_{k,s}} + \frac{1}{2} D_{\boldsymbol{p}^k, \pi_k^{-1}(I(k,\gamma,\epsilon))} \leq \\ \frac{f_s(I(k,\gamma,\epsilon)-1)}{H_{ks}} - \frac{1}{2} D_{\boldsymbol{p}^k, \pi_k^{-1}(I(k,\gamma,\epsilon)-1)} < \\ \frac{X_{\pi_k^{-1}(I(k,\gamma,\epsilon)-1),n}}{n} \end{aligned} \tag{104}$$

Since equation 103 holds for every $i \in [I(k, \gamma, \epsilon) - 1]$ and equation 104 holds for every $\ell \in \{I(k, \gamma, \epsilon), I(k, \gamma, \epsilon) + 1, \ldots, k\}$, we conclude that when equation 102 holds and $n \geq N$

$$X_{\pi_k^{-1}(1),n} > X_{\pi_k^{-1}(2),n} > \cdots > X_{\pi_k^{-1}(I(k,\gamma,\epsilon)-1,n)} \text{ and } \forall \ell \in \{I(k,\gamma,\epsilon), I(k,\gamma,\epsilon)+1, \ldots, k\}, X_{\pi_k^{-1}(\ell),n} < X_{\pi_k^{-1}(I(k,\gamma,\epsilon)-1),n}$$

Equation 93 follows recalling that these arguments hold for each $\gamma \in (\frac{1}{2}, 1]$ whenever $n \geq N$ and that for every $0 < \omega_1 < 1$ there is such an N.

Finally, to conclude equation 94 when $0 < \beta < \frac{1}{B(s)}$, note that by equation 101 (with $i = I(k, 1, \epsilon) = k - 1$) and

equation 102 and lemma B.12 and lemma B.4 (1), we additionally have that when $n \geq N$

$$\begin{aligned}
\frac{X_{\pi_k^{-1}(k-1),n}}{n} &> \\
\frac{f_s(k-1)}{H_{k,s}} - \frac{1}{2}D_{\boldsymbol{p}^k,\pi_k^{-1}(k-1)} &= \\
\frac{f_s(k-1)}{H_{k,s}} - \frac{f_s(k-1)-f_s(k)}{2H_{k,s}} &= \\
\frac{f_s(k)}{H_{k,s}} + \frac{f_s(k-1)-f_s(k)}{2H_{k,s}} &= \\
\frac{f_s(k)}{H_{k,s}} + \frac{1}{2}D_{\boldsymbol{p}^k,\pi_k^{-1}(k-1)} &\geq \\
\frac{f_s(k)}{H_{k,s}} + \frac{(1-\omega_1)^{s+\frac{3}{2}}\sqrt{C^*_{s,\beta}}}{2}n^{-1/2+e_{\gamma,1}}\log^{-e_{\gamma,2}}(n)\sqrt{\boldsymbol{p}^k(\pi_k^{-1}(k-1))(1-\pi_k^{-1}(k-1))} &> \\
\frac{f_s(k)}{H_{k,s}} + \frac{(1-\omega_1)^{s+\frac{3}{2}}\sqrt{C^*_{s,\beta}}}{2}n^{-1/2+e_{\gamma,1}}\log^{-e_{\gamma,2}}(n)\sqrt{\boldsymbol{p}^k(\pi_k^{-1}(k))(1-\pi_k^{-1}(k))} &> \\
\frac{X_{\pi_k^{-1}(k),n}}{n} &
\end{aligned} \tag{105}$$

By equations 103 (which applies for each $i \in [k-2]$ when $\gamma = 1$ and $0 < \beta < \frac{1}{B(s)}$) and equation 104 (which when $\gamma = 1$ and $0 < \beta < \frac{1}{B(s)}$ ensures $X_{\pi_k^{-1}(k-1)}$ and $X_{\pi_k^{-1}(k)}$ are the two smallest counts) and equation 105, we have that under assumption 102 and for $n \geq N$ and provided $0 < \beta < \frac{1}{B(s)}$

$$X_{\pi_k^{-1}(1),n} > X_{\pi_k^{-1}(2),n} > \cdots > X_{\pi_k^{-1}(k-1),n} > X_{\pi_k^{-1}(k),n} \tag{106}$$

Thus equation 94 of the lemma statement holds. □

B.5 Upper and lower bounds when $\beta < \frac{1}{B(s)}$

Theorem B.7. *(Upper Bound For Sort and Snap when $0 < \beta < \frac{1}{B(s)}$) Suppose $s > 0$, $0 < \beta < \frac{1}{B(s)}$, and for $\ell \geq 2$, $\boldsymbol{p}^\ell \in \mathcal{P}_{f_s,\ell}$ (with permutation function denoted π_ℓ) and for each n large enough so that $\lfloor n^\beta \rfloor \geq 2$, $k := \lfloor n^\beta \rfloor$. Also,*

$$Y_{1k}, Y_{2k}, \ldots, Y_{nk} \overset{iid}{\sim} \boldsymbol{p}^k$$

Further, let $0 < \tau < 1 - \beta B(s)$. Then

$$\mathbb{E}\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1 \lesssim \begin{cases} n^{-\beta(B(s)-2)}\log^{-\mathbb{I}(s=1)}(n)\exp(-\frac{(1-\tau)C^*_{s,\beta}}{16}\frac{n^{1-\beta B(s)}}{\log^{\mathbb{I}(s=1)}(n)}) & 0 < \beta < \frac{1}{B(s)+1} \\ n^{-\beta(B(s)-2)}n^{2(\beta(B(s)+1)-(1-\tau))}\log^{-\mathbb{I}(s=1)}(n)\exp(-\frac{(1-\tau)C^*_{s,\beta}}{16}\frac{n^{1-\beta B(s)}}{\log^{\mathbb{I}(s=1)}(n)}) & \frac{1}{B(s)+1} \leq \beta < \frac{1}{B(s)} \end{cases} \tag{107}$$

where recall $C^*_{s,\beta}$ is defined in equation 6

Proof. Let

$$p := \begin{cases} 0 & 0 < \beta < \frac{1}{B(s)+1} \\ B(s) + 1 - \frac{1-\tau}{\beta} & \frac{1}{B(s)+1} \leq \beta < \frac{1}{B(s)} \end{cases} \tag{108}$$

and for $k \geq 2$

$$\gamma_k := 1 - \frac{1}{k^{1-p}\log(k)} \tag{109}$$

Now define two events for each k. Namely

$$A_{1,k} := \{\text{No two counts are equal and } \forall j \in [I(k,\gamma_k,1)-1], \hat{\pi}(j) = \pi_k^{-1}(j)\} \tag{110}$$

and

$$A_{2,k} := \{\text{No two counts are equal and } \forall j \in [k], \hat{\pi}(j) = \pi_k^{-1}(j)\} \tag{111}$$

where recall the I function is defined in equation 92 and $\hat{\pi}(j)$ is the index of the j^{th} largest count. $\hat{\pi}$ is unique under events $A_{1,k}$ and $A_{2,k}$. Note that

$$\begin{aligned}\mathbb{E}\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1 = \mathbb{E}\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1\mathbb{I}(A_{2k}) + \mathbb{E}\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1\mathbb{I}(A_{2k}^C \cap A_{1k}^C) + \mathbb{E}\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1\mathbb{I}(A_{2k}^C \cap A_{1k}) = \\ \mathbb{E}\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1\mathbb{I}(A_{2k}^C \cap A_{1k}^C) + \mathbb{E}\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1\mathbb{I}(A_{2k}^C \cap A_{1k})\end{aligned} \tag{112}$$

where the second equality is because event A_{2k} implies both that for every $j \in [k]$, the category with the j^{th} largest probability has the j^{th} largest count and there are no ties [3]. Thus Sort and Snap yields zero error.

Now note that by definition of $A_{1,k}$ and $A_{2,k}$

$$A_{1,k} \subseteq \left[X_{\pi_k^{-1}(1),n} > X_{\pi_k^{-1}(2),n} > \cdots > X_{\pi_k^{-1}(I(k,\gamma_k,1)-1),n} \text{ and } \forall j \in \{I(k,\gamma_k,1), \cdots, k\}, X_{\pi_k^{-1}(I(k,\gamma_k,1)-1)} > X_{\pi_k^{-1}(j),n}\right] \tag{113}$$

and

$$A_{2,k} \subseteq \left[X_{\pi_k^{-1}(1),n} > X_{\pi_k^{-1}(2),n} > \cdots > X_{\pi_k^{-1}(k-1),n} > X_{\pi_k^{-1}(k),n}\right] \tag{114}$$

And because $\tau < 1 - \beta B(s)$, we have that $p < 1$. Therefore $\gamma_k \to 1$ as $n \to \infty$. In particular, there exists an N sufficiently large, such that if $n \geq N$, then $\gamma_k > \frac{1}{2}$. Thus applying lemma B.6 and using equations 113 and 114 we have that for any $0 < \omega < 1$ the existence of some N such that for $n \geq N$, both

$$A_{1,k}^C \subseteq \left[\|\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}_n\|_\infty \geq \frac{(1-\omega)\sqrt{C^*_{s,\beta}}}{2} n^{-\frac{1}{2}+e_{\gamma_k,1}} \log^{-e_{\gamma_k,2}}(n)\right] \tag{115}$$

and

$$A_{2,k}^C \subseteq \left[\|\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}_n\|_\infty \geq \frac{(1-\omega)\sqrt{C^*_{s,\beta}}}{2} n^{-\frac{1}{2}+e_{1,1}} \log^{-e_{1,2}}(n)\right] \tag{116}$$

Now let $0 < \omega_1 < 1$. By equations 115, 116, and again using that $\gamma_k \to 1$ as $n \to \infty$ and applying lemma B.2 with $C = \frac{1}{1+\omega_1}\frac{(1-\omega)\sqrt{C^*_{s,\beta}}}{2}$ and ω_1 we have the existence of some N such that for $n \geq N$, both

$$\begin{aligned}A_{1,k}^C \subseteq \Bigg[\|\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}_n\|_\infty \geq \frac{(1-\omega)}{1+\omega_1}\frac{\sqrt{C^*_{s,\beta}}}{2}\left(n^{-\frac{1}{2}+e_{\gamma_k,1}} \log^{-e_{\gamma_k,2}}(n)\right) + \\ \left(n^{-\frac{1}{2}+e_{\gamma_k,1}} \log^{-e_{\gamma_k,2}}(n)\right)^2 \frac{\left(\frac{(1-\omega)}{1+\omega_1}\frac{\sqrt{C^*_{s,\beta}}}{2}\right)^2}{2} \frac{1}{\min_{j\in[k]}\sqrt{\boldsymbol{p}^k(j)(1-\boldsymbol{p}^k(j))}}\Bigg]\end{aligned} \tag{117}$$

and

$$\begin{aligned}A_{2,k}^C \subseteq \Bigg[\|\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}_n\|_\infty \geq \frac{(1-\omega)}{1+\omega_1}\frac{\sqrt{C^*_{s,\beta}}}{2} n^{-\frac{1}{2}+e_{1,1}} \log^{-e_{1,2}}(n) + \\ \left(n^{-\frac{1}{2}+e_{1,1}} \log^{-e_{1,2}}(n)\right)^2 \frac{\left(\frac{(1-\omega)}{1+\omega_1}\frac{\sqrt{C^*_{s,\beta}}}{2}\right)^2}{2} \frac{1}{\min_{j\in[k]}\sqrt{\boldsymbol{p}^k(j)(1-\boldsymbol{p}^k(j))}}\Bigg]\end{aligned} \tag{118}$$

[3] In defining Sort and Snap, we did not specify how to break ties. This is irrelevant in the analysis because the events $A_{1,k}$ and $A_{2,k}$ specify no ties

Now applying lemma B.1 with $u = n^{-\frac{1}{2}+e_{1,\gamma_k}} \log^{-e_{\gamma_k,2}}(n)$ and $C = \frac{(1-\omega)}{1+\omega_1}\frac{\sqrt{C^*_{s,\beta}}}{2}$ and using equation 117, we have an N such that for $n \geq N$

$$\begin{aligned}\mathbb{P}(A^C_{1,k}) \leq 2k\exp(-\frac{n^{1+2\left(-\frac{1}{2}+e_{\gamma_k,1}\right)}\log^{-2e_{\gamma_k,2}}(n)}{16}\left((\frac{(1-\omega)}{1+\omega_1})^2\right)C^*_{s,\beta}) = \\ 2k\exp(-\frac{C^*_{s,\beta}}{16}n^{1-\beta(\gamma_k(s+2)+\max(0,1-s))}\log^{-2e_{\gamma_k,2}}(n)(\frac{(1-\omega)}{1+\omega_1})^2)\end{aligned} \tag{119}$$

where we have used the definition of $e_{\gamma_k,1}$ (see equation 66). And similarly applying lemma B.1 with $u = n^{-\frac{1}{2}+e_{1,1}}\log^{-e_{1,2}}(n)$ and $C = \left(\frac{(1-\omega)}{1+\omega_1}\frac{\sqrt{C^*_{s,\beta}}}{2}\right)$ and using equation 118, we have that for $n \geq N$

$$\begin{aligned}\mathbb{P}(A^C_{2,k}) \leq 2k\exp(-\frac{n^{1+2\left(-\frac{1}{2}+e_{1,1}\right)}\log^{-2e_{1,2}}(n)}{16}C^*_{s,\beta}\left((\frac{(1-\omega)}{1+\omega_1})^2\right)) = \\ 2k\exp(-\frac{C^*_{s,\beta}}{16}n^{1-\beta(s+2+\max(0,1-s))}\log^{-2e_{1,2}}(n)(\frac{(1-\omega)}{1+\omega_1})^2)\end{aligned} \tag{120}$$

We must also produce an upper bound on the $\|\hat{\boldsymbol{p}}^k_n - \boldsymbol{p}^k\|_1$ under the event $A_{1,k} \cap A^C_{2,k}$. To do this, first note that since since $1 - x \leq \exp(-x)\ \forall x$, we have that for $q \in \mathbb{R}$ and $k \geq 1$

$$1 - \frac{1}{k^{1-q}} \leq \exp(-\frac{1}{k^{1-q}}) = k^{-\frac{1}{k^{1-q}\log(k)}} \tag{121}$$

In particular (after multiplying both sides of equation 121 by k, we have that for $q \in \mathbb{R}$ and $k \geq 1$

$$k - k^{1-\frac{1}{k^{1-q}\log(k)}} \leq \frac{1}{k^{-q}} = k^q \tag{122}$$

Using equation 122 (with $q = p$ where p is defined in equation 108) and the definition of γ_k (see equation 109) and the definition of the I function (see equation 92), we have that (for $k \geq 2$)

$$k - I(k,\gamma_k,1) \leq k - (\lfloor k^{\gamma_k}\rfloor - 1) \leq k - (k^{\gamma_k} - 1) + 1 = k - k^{\gamma_k} + 2 = k - k^{1-\frac{1}{k^{1-p}\log(k)}} + 2 \leq k^p + 2 \tag{123}$$

Now using the definition of $\hat{\boldsymbol{p}}^k_n$ (see equation 4) and the definition of $A_{1,k}$ (equation 110) and then using that f_s is monotonically decreasing, we have that

$$\begin{aligned}A_{1,k} \subseteq \\ \left[\|\hat{\boldsymbol{p}}^k_n - \boldsymbol{p}^k\|_1 = \sum_{j\in\{I(k,\gamma_k,1),I(k,\gamma_k,1)+1,\ldots,k\}} |\hat{\boldsymbol{p}}^k_n(\hat{\pi}(j)) - \boldsymbol{p}^k(\hat{\pi}(j))|\right] \cap \\ \left[\max_{j\in\{I(k,\gamma_k,1),I(k,\gamma_k,1)+1,\ldots,k\}} |\hat{\boldsymbol{p}}^k_n(\hat{\pi}(j)) - \boldsymbol{p}^k(\hat{\pi}(j))| \leq \frac{1}{H_{k,s}}|f_s(I(k,\gamma_k,1)) - f_s(k)|\right]\end{aligned} \tag{124}$$

Also, since $p < 1$ (which follows since $\tau < 1 - \beta B(s)$), there exists a K sufficiently large such that $k - k^p - 2 > 0$ for $k \geq K$. Using this and equations 123 and 124, and again using the monotonicity of f_s, we have that for $k \geq K$

$$A_{1,k} \subseteq \left[\|\hat{\boldsymbol{p}}^k_n - \boldsymbol{p}^k\|_1 \leq (k - I(k,\gamma_k,1) + 1)\frac{1}{H_{k,s}}|f_s(k - k^p - 2) - f_s(k)| \leq \frac{(k^p+3)|f_s(k-k^p-2) - f_s(k)|}{H_{k,s}}\right] \tag{125}$$

Now for $k \geq K$, using a two term taylor series expansion of the f_s function about $k - k^p - 2$,we have the existence of some $\phi \in [k - k^p - 2, k]$ such that for any $C > 0$ and N sufficiently large and $n \geq N$

$$f_s(k-k^p-2) - f_s(k) = sf_{s+1}(k-k^p-2)(k^p+2) + \frac{s(s+1)}{2}f_{s+2}(\phi)(k^p+2)^2 \leq s(1+C)f_{s+1}(k-k^p-2)(k^p+2) \tag{126}$$

where we used above that $f_{s+2}(k-k^p-2) = o(f_{s+1}(k-k^p-2)$. Now using that $(k^p+3)f_{s+1}(k-k^p-2)(k^p+2) \asymp k^{-(s+1)+2p}$ and equations 125 and 126 and lemma A.1 we have the existence of some constant $C_{s,\beta,1} > 0$ and a K_2 such that for $k \geq K_2$

$$A_{1,k} \subseteq \left[\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1 \leq C_{s,\beta,1} k^{-(s+1)+2p-max(0,1-s)} \log^{-\mathbb{I}(s=1)}(n) \right] \tag{127}$$

So by equations 112, 119, 120, and 127, and since the 1–norm is bounded by 2 for distributions on the simplex, we have that for an N and $n \geq N$

$$\begin{aligned} &\mathbb{E}\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1 \leq \\ &2\mathbb{P}(A_{1,k}^C) + C_{s,\beta,1} n^{-\beta(s+1)} n^{\beta(2p-\max(0,1-s))} \log^{-\mathbb{I}(s=1)}(n) \mathbb{P}(A_{2,k}^C) \leq \\ &2n^\beta \exp(-\frac{C_{s,\beta}^*}{16} n^{1-\beta(\gamma_k(s+2)+\max(0,1-s))} \log^{-2e_{\gamma_k,2}}(n)(\frac{(1-\omega)}{1+\omega_1})^2) + \\ &2n^\beta C_{s,\beta,1} n^{-\beta(s+1)} n^{\beta(2p-\max(0,1-s))} \log^{-\mathbb{I}(s=1)}(n) \exp(-\frac{C_{s,\beta}^*}{16} n^{1-\beta(s+2+\max(0,1-s))} \log^{-2e_{1,2}}(n)(\frac{(1-\omega)}{1+\omega_1})^2) \end{aligned} \tag{128}$$

Now by using the definition of γ_k (equation 109) and that $\exp(x) - 1 \geq +x$ for $x \in \mathbb{R}$, and the definition of p given in equation 108 and that $e_{\gamma,2} = \frac{1}{2}\mathbb{I}(s=1)$ for every $0 < \gamma \leq 1$ – (see equation 67) and the definition of $B(s)$ (see equation 65), and letting $C_{s,\beta,\omega,\omega_1} := \frac{C_{s,\beta}^*}{16}(\frac{(1-\omega)}{1+\omega_1})^2$, we have that

$$\begin{aligned} &\frac{\exp(-C_{s,\beta,\omega,\omega_1} n^{1-\beta(\gamma_k(s+2)+\max(0,1-s))} \log^{-2e_{\gamma_k,2}}(n))}{n^{-\beta(s+1)} n^{\beta(2p-\max(0,1-s))} \log^{-\mathbb{I}(s=1)}(n) \exp(-C_{s,\beta,\omega,\omega_1} n^{1-\beta(s+2+\max(0,1-s))} \log^{-(2e_{1,2})}(n))} = \\ &n^{\beta(s+1)} n^{-\beta(2p-\max(0,1-s))} \log^{\mathbb{I}(s=1)}(n) \exp\left(-\log^{-\mathbb{I}(s=1)}(n) C_{s,\beta,\omega,\omega_1} n^{1-\beta(s+2+\max(0,1-s))} (n^{\frac{\beta(s+2)}{k^{1-p}\log(k)}} - 1) \right) \leq \\ &n^{\beta(s+1)} n^{-\beta(2p-\max(0,1-s))} \log^{\mathbb{I}(s=1)}(n) \exp\left(-\log^{-\mathbb{I}(s=1)}(n) C_{s,\beta,\omega,\omega_1} n^{1-\beta(B(s))} (n^{\frac{\beta(s+2)}{\beta k^{1-p}\log(n)}} - 1) \right) = \\ &n^{\beta(s+1)} n^{-\beta(2p-\max(0,1-s))} \log^{\mathbb{I}(s=1)}(n) \exp\left(-\log^{-\mathbb{I}(s=1)}(n) C_{s,\beta,\omega,\omega_1} n^{1-\beta(B(s))} (e^{\frac{(s+2)}{k^{1-p}}} - 1) \right) \leq \\ &n^{\beta(s+1)} n^{-\beta(2p-\max(0,1-s))} \log^{\mathbb{I}(s=1)}(n) \exp\left(-\log^{-\mathbb{I}(s=1)}(n) C_{s,\beta,\omega,\omega_1} n^{1-\beta(B(s))} (\frac{(s+2)}{k^{1-p}}) \right) \leq \\ &n^{\beta(s+1)} n^{-\beta(2p-\max(0,1-s))} \log^{\mathbb{I}(s=1)}(n) \exp\left(-(s+2) \log^{-\mathbb{I}(s=1)}(n) C_{s,\beta,\omega,\omega_1} n^{1-\beta(B(s))-\beta(1-p)} \right) = \\ &n^{\beta(s+1)} n^{-\beta(2p-\max(0,1-s))} \log^{\mathbb{I}(s=1)}(n) \begin{cases} \exp\left(-(s+2) \log^{-\mathbb{I}(s=1)}(n) C_{s,\beta,\omega,\omega_1} n^{1-\beta(B(s)+1)} \right) & \beta < \frac{1}{B(s)+1} \\ \exp\left(-(s+2) \log^{-\mathbb{I}(s=1)}(n) C_{s,\beta,\omega,\omega_1} n^{\tau} \right) & \frac{1}{B(s)+1} \leq \beta < \frac{1}{B(s)} \end{cases} = \\ &\to 0 \text{ as } n \to \infty \end{aligned} \tag{129}$$

where the limit follows in the last step follows because for every $x, z \in \mathbb{R}$ and $c, y > 0$, $n^x \log^z(n) \exp(-c\frac{n^y}{\log^{z+1}(n)}) \to 0$ as $n \to \infty$.

By equations 129 and 128 and that $B(s) = s + 2 + \max(0, 1-s)$ (see equation 65), we have that

$$\begin{aligned} \mathbb{E}\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1 &\lesssim n^{-\beta(s+\max(0,1-s)-2p)} \log^{-\mathbb{I}(s=1)}(n) \exp\left(-\frac{C_{s,\beta}^*(\frac{(1-\omega)}{1+\omega_1})^2 n^{1-\beta(s+2+\max(0,1-s))}}{16 \log^{\mathbb{I}(s=1)}(n)} \right) = \\ &n^{-\beta(B(s)-2-2p)} \log^{-\mathbb{I}(s=1)}(n) \exp\left(-\frac{C_{s,\beta}^*(\frac{(1-\omega)}{1+\omega_1})^2 n^{1-\beta B(s)}}{16 \log^{\mathbb{I}(s=1)}(n)} \right) \end{aligned} \tag{130}$$

Since the argument holds for each $0 < \omega < 1$ and $0 < \omega_1 < 1$, and since the range of $(\frac{(1-\omega)}{1+\omega_1})^2$ for $0 < \omega, \omega_1 < 1$ is precisely $(0, 1)$ we conclude that for any $0 < \tau < 1$

$$\mathbb{E}\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}^k\|_1 \lesssim n^{-\beta(B(s)-2-2p)} \log^{-\mathbb{I}(s=1)}(n) \exp\left(-\frac{C_{s,\beta}^*(1-\tau) n^{1-\beta B(s)}}{16 \log^{\mathbb{I}(s=1)}(n)} \right) \tag{131}$$

Plugging in p (defined in equation 108) to the above equation yields the theorem statement.

□

Theorem B.8. *(Minimax Lower Bound For Zipfian Distribution Estimation when $\beta > 0$) Suppose $s > 0, 0 < \beta$, and for each n large enough so that $\lfloor n^\beta \rfloor \geq 2$, $k := \lfloor n^\beta \rfloor$. Also let $\tau > 0$. Then*

$$\inf_{\hat{\mu}_n} \sup_{\boldsymbol{p} \in \mathcal{P}_{f_s,k}} \mathbb{E}\|\hat{\mu}_n - \boldsymbol{p}\|_1 \gtrsim \begin{cases} n^{-\beta(B(s)-1)} \log^{-\mathbb{I}(s=1)}(n) \exp\left(-\frac{(1+\tau)C^*_{s,\beta} n^{1-\beta B(s)}}{\log^{\mathbb{I}(s=1)}(n)}\right) & 0 < \beta < \frac{1}{B(s)} \\ n^{-\left(\frac{s+1}{s+2} + \beta \max(0,1-s)(1-\frac{s+1}{s+2})\right)} \log^{-\mathbb{I}(s=1)(1-\frac{s+1}{s+2})}(n) & \beta \geq \frac{1}{B(s)} \end{cases} \tag{132}$$

where recall $\mathcal{P}_{f_s,k}$ (defined in equation 1) is the collection of all $k-$dimensional $s-$Zipfian probability distributions and $C^*_{s,\beta}$ is defined in equation 6. And the inf *is taken over all functions of the n samples $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \boldsymbol{p}$*

Proof. We will use the method of Le Cam (see Tibshirani and Wasserman (2017) for a good review) and first we will handle the case $\beta < \frac{1}{B(s)}$. Thus define $\boldsymbol{p}_{0k}, \boldsymbol{p}_{1k} \in \mathcal{P}_{f_s,k}$ such that for $j \in [k-2]$

$$\boldsymbol{p}_{0k}(j) = \boldsymbol{p}_{1k}(j) = \frac{j^{-s}}{H_{k,s}} \tag{133}$$

and

$$\boldsymbol{p}_{0k}(k-1) = \boldsymbol{p}_{1k}(k) = \frac{f_s(k-1)}{H_{k,s}} \tag{134}$$

and

$$\boldsymbol{p}_{0k}(k) = \boldsymbol{p}_{1k}(k-1) = \frac{f_s(k)}{H_{k,s}} \tag{135}$$

Thus

$$\|\boldsymbol{p}_{0k} - \boldsymbol{p}_{1k}\|_1 = \frac{2\left(f_s(k-1) - f_s(k)\right)}{H_{k,s}} \tag{136}$$

while

$$\begin{aligned} KL(\boldsymbol{p}_{0k}, \boldsymbol{p}_{1k}) = \\ \frac{f_s(k-1)}{H_{k,s}} \log(\frac{f_s(k-1)}{f_s(k)}) + \frac{f_s(k)}{H_{k,s}} \log(\frac{f_s(k)}{f_s(k-1)}) = \\ \log(\frac{f_s(k-1)}{f_s(k)}) \left(\frac{f_s(k-1) - f_s(k)}{H_{k,s}}\right) = \\ -\frac{s}{H_{k,s}} \log(1 - \frac{1}{k}) \left(f_s(k-1) - f_s(k)\right) \end{aligned} \tag{137}$$

Now by the Taylor Series expansion of the log function and the formula for a geoemtric series, we have that

$$-\log(1-\frac{1}{k}) = \sum_{j=1}^{\infty} \left(\frac{1}{k}\right)^j \frac{1}{j} \leq \frac{1}{k} + \frac{1}{2} \sum_{j=2}^{\infty} \left(\frac{1}{k}\right)^j = \frac{1}{k} + \frac{1}{2}(\frac{1}{1-\frac{1}{k}} - (1+\frac{1}{k})) = \frac{1}{k} + \frac{1}{2k(k-1)} = \frac{1}{k}\left(1 + \frac{1}{2(k-1)}\right) \tag{138}$$

Now using a Two term Taylor series expansion of f_s about $k-1$, there exists some $\phi \in [k-1, k]$ such that

$$f_s(k-1) - f_s(k) = s f_{s+1}(k-1) + \frac{s(s+1)}{2} f_{s+2}(\phi) \tag{139}$$

In particular, using that f_s functions are monotonically decreasing, we have that

$$s f_{s+1}(k-1) \leq f_s(k-1) - f_s(k) \leq s f_{s+1}(k-1) + \frac{s(s+1)}{2} f_{s+2}(k-1) \tag{140}$$

Using the above equation and that $f_{s+2}(k-1) = o(f_{s+1}(k-1))$ we have that for $\omega > 0$ some N sufficiently large such that for and $n \geq N$

$$s f_{s+1}(k-1) \leq f_s(k-1) - f_s(k) \leq s(1+\omega) f_{s+1}(k-1) \tag{141}$$

So by equations 136 and 141 (and using lemma A.1 and that $k = \lfloor n^\beta \rfloor$ and the definition of $B(s)$ – see equation 65), we have that for some N and constants $C_{s,\beta,2}, C_{s,\beta,3} > 0$ and for $n \geq N$

$$\begin{aligned} \|\boldsymbol{p}_{0k} - \boldsymbol{p}_{1k}\|_1 \geq \\ C_{s,\beta,2} \log^{-\mathbb{I}(s=1)}(n) k^{-\max(0,1-s)} f_{s+1}(k-1) \geq \\ C_{s,\beta,3} \log^{-\mathbb{I}(s=1)}(n) n^{-\beta(s+1+\max(0,1-s))} = \\ C_{s,\beta,3} \log^{-\mathbb{I}(s=1)}(n) n^{-\beta(B(s)-1)} \end{aligned} \tag{142}$$

and by equations 137 and 138 and 141 and lemma A.1 we have that for any $0 < \omega < 1$ there is an N such that for $n \geq N$

$$\begin{aligned} KL(\boldsymbol{p}_{0k}, \boldsymbol{p}_{1k}) \leq \\ \frac{s^2}{k H_{k,s}} (1+\omega) \left(1 + \frac{1}{2(k-1)}\right) f_{s+1}(k-1) \leq \\ (1+\omega)^2 n^{-\beta(s+2)} \frac{s^2}{H_{k,s}} \leq \\ s^2 \left(\frac{(1+\omega)^2}{1-\omega}\right) n^{-\beta((s+2)+\max(0,1-s))} \log^{-\mathbb{I}(s=1)}(n) \times \begin{cases} 1-s & 0 < s < 1 \\ \frac{1}{\beta} & s = 1 \\ \frac{1}{R(s)} & s > 1 \end{cases} = \\ \left(\frac{(1+\omega)^2}{1-\omega}\right) n^{-\beta B(s)} \log^{-\mathbb{I}(s=1)}(n) C^*_{s,\beta} \end{aligned} \tag{143}$$

So now applying theorem 4 of Tibshirani and Wasserman (2017) and using equations 142 and 143, we have that for $\beta < \frac{1}{B(s)}$ and any $\omega_1 > 1$

$$\inf_{\hat{\mu}_n} \sup_{\boldsymbol{p} \in \mathcal{P}_{f_s,k}} \mathbb{E}\|\hat{\mu}_n - \boldsymbol{p}\|_1 \gtrsim \log^{-\mathbb{I}(s=1)}(n) n^{-\beta(B(s)-1)} \exp\left(-\frac{n^{1-\beta B(s)} \omega_1 C^*_{s,\beta}}{\log^{\mathbb{I}(s=1)}(n)}\right) \tag{144}$$

Setting $\omega_1 = 1 + \tau$ concludes the $\beta < \frac{1}{B(s)}$ case

For the case $\beta \geq \frac{1}{B(s)}$, we redefine $\boldsymbol{p}_{0k}$ and $\boldsymbol{p}_{1k}$ so that for $j \in [k] - \{\lfloor \frac{n^{x_s}}{\log^{z_s}(n)} \rfloor, \lfloor \frac{n^{x_s}}{\log^{z_s}(n)} \rfloor - 1\}$,

$$\boldsymbol{p}_{0k}(j) = \boldsymbol{p}_{1k}(j) = \frac{j^{-s}}{H_{k,s}}$$

and

$$\boldsymbol{p}_{0k}(\lfloor \frac{n^{x_s}}{\log^{z_s}(n)} \rfloor - 1) = \boldsymbol{p}_{1k}(\lfloor \frac{n^{x_s}}{\log^{z_s}(n)} \rfloor) = \frac{f_s(\lfloor \frac{n^{x_s}}{\log^{z_s}(n)} \rfloor - 1)}{H_{k,s}}$$

and

$$\boldsymbol{p}_{0k}(\lfloor \frac{n^{x_s}}{\log^{z_s}(n)} \rfloor) = \boldsymbol{p}_{1k}(\lfloor \frac{n^{x_s}}{\log^{z_s}(n)} \rfloor - 1) = \frac{f_s(\lfloor \frac{n^{x_s}}{\log^{z_s}(n)} \rfloor)}{H_{k,s}}$$

where

$$x_s := \frac{1}{s+2}(1 - \beta \max(0, 1-s)) \tag{145}$$

and

$$z_s := \frac{1}{s+2} \mathbb{I}(s=1) \tag{146}$$

Now carrying out similar arguments to the $\beta < \frac{1}{B(s)}$ case we have that there exists a $C_{s,\beta,4} > 0$ and N such that for $n \geq N$

$$\|\boldsymbol{p}_{0k} - \boldsymbol{p}_{1k}\|_1 \geq C_{s,\beta,4} \log^{-\mathbb{I}(s=1)}(n) n^{-\beta(\max(0,1-s))} n^{-x_s(s+1)} \log^{z_s(s+1)}(n) \tag{147}$$

and also there is a $C_{s,\beta,5} > 0$ and N such that for $n \geq N$

$$\begin{aligned} KL(\boldsymbol{p}_{0k}, \boldsymbol{p}_{1k}) \leq \\ C_{s,\beta,5} \log^{-\mathbb{I}(s=1)}(n) n^{-\beta(\max(0,1-s))} n^{-x_s(s+1)} \log^{z_s(s+1)}(n) \left(n^{-x_s} \log^{z_s}(n)\right) = \\ C_{s,\beta,5} n^{-x_s(s+2)-\beta\max(0,1-s)} \log^{z_s(s+2)-\mathbb{I}(s=1)}(n) = \\ C_{s,\beta,5} n^{-1} \end{aligned} \tag{148}$$

where in the last line we have used the definitions of x_s and z_s (equations 145 and 146). By equations 147 and 148 and theorem 4 of Tibshirani and Wasserman (2017) we conclude that when $\beta \geq \frac{1}{B(s)}$

$$\begin{aligned} \inf_{\hat{\mu}_n} \sup_{\boldsymbol{p} \in \mathcal{P}_{f_s,k}} \mathbb{E}\|\hat{\mu}_n - \boldsymbol{p}\|_1 \gtrsim \log^{z_s(s+1)-\mathbb{I}(s=1)}(n) n^{-1(\beta\max(0,1-s)+x_s(s+1))} = \\ n^{-\left(\frac{s+1}{s+2}+\beta\max(0,1-s)\left(1-\frac{s+1}{s+2}\right)\right)} \log^{-\mathbb{I}(s=1)\left(1-\frac{s+1}{s+2}\right)}(n) \end{aligned} \tag{149}$$

where in the last equality we have again used the definitions of x_s and z_s □

Theorem 4.2 (SS Upper and Lower Bounds)**.** *Suppose $s > 0$, $0 < \beta < \frac{1}{B(s)}$ and for each n large enough so that $\lfloor n^\beta \rfloor \geq 2$, $k := \lfloor n^\beta \rfloor$ and $0 < \tau < 1 - \beta B(s)$. Then the Full Sort and Snap Estimator $\hat{\boldsymbol{p}}_n^k$ achieves that* $\sup_{\boldsymbol{p} \in \mathcal{P}_{f_s,k}} \mathbb{E}\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}\|_1$

$$\lesssim \frac{n^{-\beta(B(s)-2)}}{\log^{\mathbb{I}(s=1)}(n)} \exp\left(-\frac{C^*_{s,\beta}(1-\tau)}{16} \frac{n^{1-\beta B(s)}}{\log^{\mathbb{I}(s=1)}(n)}\right)$$

when $0 < \beta < \frac{1}{B(s)+1}$ and $\sup_{\boldsymbol{p} \in \mathcal{P}_{f_s,k}} \mathbb{E}\|\hat{\boldsymbol{p}}_n^k - \boldsymbol{p}\|_1$

$$\lesssim \frac{n^{-\beta(B(s)-2)}}{\log^{\mathbb{I}(s=1)}(n)} h_{n,s,\beta} \exp\left(-\frac{C^*_{s,\beta}(1-\tau)}{16} \frac{n^{1-\beta B(s)}}{\log^{\mathbb{I}(s=1)}(n)}\right)$$

when $\beta < \frac{1}{B(s)}$ and $h_{n,s,\beta} = n^{2(\beta(B(s)+1)-(1-\tau))}$. Also, $\inf_{\hat{\mu}_n} \sup_{\boldsymbol{p} \in \mathcal{P}_{f_s,k}} \mathbb{E}\|\hat{\mu}_n - \boldsymbol{p}\|_1$

$$\gtrsim \frac{n^{-\beta(B(s)-1)}}{\log^{\mathbb{I}(s=1)}(n)} \exp\left(-C^*_{s,\beta}(1+\tau) \frac{n^{1-\beta B(s)}}{\log^{\mathbb{I}(s=1)}(n)}\right)$$

where $\mathcal{P}_{f_s,k}$ is the collection of all k–dimensional s–Zipifan probability distributions. And the inf *is taken over all functions of the n samples $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \boldsymbol{p}$.*

Proof. Follows immediately from theorems B.8 and B.7. □

B.6 Upper bounds when $\frac{1}{B(s)} \leq \beta < \frac{1}{s}$ and $s > 2$

Theorem B.9. *(Upper Bound For Truncated Sort and Snap when $\frac{1}{s+2} \leq \beta < \frac{1}{s}$ and $s > 2$) Suppose $s > 2$, $\frac{1}{s+2} \leq \beta < \frac{1}{s}$ and for $\ell \geq 2$, $\boldsymbol{p}^\ell \in \mathcal{P}_{f_s,\ell}$ (with permutation function denoted π_ℓ) and $k = \lfloor n^\beta \rfloor$ for each n large enough so that $\lfloor n^\beta \rfloor \geq 2$. Also let $0 < \epsilon < \frac{1-s\beta}{\beta B(s)}$. Then letting $T_{n,\epsilon} := I(k, 1, \epsilon) - 1$ and the I function is defined in equation 92, there exists a $U_{s,\beta} > 0$ such that for any $\tau \in (0, U_{s,\beta})$*

$$\mathbb{E}\|\hat{\boldsymbol{p}}^k_{n,T_{n,\epsilon}} - \boldsymbol{p}^k\|_1 \lesssim n^{-\frac{s}{s+2}} n^\tau \tag{150}$$

Proof. Define the event

$$A_{3,k} := \{ \text{ No two counts are equal and } \forall j \in [T_{n,\epsilon}], \hat{\pi}(j) = \pi_k^{-1}(j) \} \tag{151}$$

Note that because event $A_{3,k}$ excludes ties, $A_{3,k}$ implies that any category that does not achieve one of the $[T_{n,\epsilon}]$ largest counts must also not be a category with one of the $[T_{n,\epsilon}]$ largest probabilities and vice versa. Using this

and the definition of $A_{3,k}$ (and the definition of Truncated Sort and Snap – see definition 5), we have that

$$\begin{aligned} A_{3,k} \subseteq \\ \left[\forall j \in [T_{n,\epsilon}], \hat{\boldsymbol{p}}^k_{n,T_{n,\epsilon}}(\pi_k^{-1}(j)) = \boldsymbol{p}^k(\pi_k^{-1}(j))\right] \cap \\ \left[\hat{\pi}\left(\{T_{n,\epsilon}+1, T_{n,\epsilon}+2, \ldots, k\}\right) = \pi_k^{-1}\left(\{T_{n,\epsilon}+1, T_{n,\epsilon}+2, \ldots, k\}\right)\right] \end{aligned} \tag{152}$$

Also, note that since $\hat{\boldsymbol{p}}^k_{n,T_{n,\epsilon}}$ is a concatenation of subsets of two probability distributions (the s–Zipfian law and the EPE), $\sum_{j=1}^k \hat{\boldsymbol{p}}^k_{n,T_{n,\epsilon}}(j) \leq 2$. Using this and that $\boldsymbol{p}^k \in \mathcal{S}_k$ and the definition of 1–norm, we have that

$$\|\hat{\boldsymbol{p}}^k_{n,T_{n,\epsilon}} - \boldsymbol{p}^k\|_1 \leq 3 \tag{153}$$

By the definition of Truncated Sort and Snap (see equation 5) and equation 152, and equation 153, and that for a sufficiently large $N_{s,\beta}$ and $n \geq N_{s,\beta},$, $T_{n,\epsilon} < k$, we have that for $n \geq N_{s,\beta}$

$$\begin{aligned} \mathbb{E}\|\hat{\boldsymbol{p}}^k_{n,T_{n,\epsilon}} - \boldsymbol{p}^k\| = \\ \mathbb{E}\|\hat{\boldsymbol{p}}^k_{n,T_{n,\epsilon}} - \boldsymbol{p}^k\|\mathbb{I}(A_{3,k}) + \mathbb{E}\|\hat{\boldsymbol{p}}^k_{n,T_{n,\epsilon}} - \boldsymbol{p}^k\|\mathbb{I}(A^C_{3,k}) \leq \\ \mathbb{E}\sum_{j=T_{n,\epsilon}+1}^{k} |\bar{\boldsymbol{p}}^k(\hat{\pi}(j)) - \boldsymbol{p}^k(\hat{\pi}(j))| + 3\mathbb{P}(A^C_{3,k}) = \\ \mathbb{E}\sum_{j=T_{n,\epsilon}+1}^{k} |\bar{\boldsymbol{p}}^k(\pi_k^{-1}(j)) - \boldsymbol{p}^k(\pi_k^{-1}(j))| + 3\mathbb{P}(A^C_{3,k}) \end{aligned} \tag{154}$$

Now note that by definition of $A_{3,k}$ and $T_{n,\epsilon}$, we have that

$$A_{3,k} \subseteq \left[X_{\pi_k^{-1}(1),n} > X_{\pi_k^{-1}(2),n} > \cdots > X_{\pi_k^{-1}(I(k,1,\epsilon)-1),n} \text{ and } \forall j \in \{I(k,1,\epsilon), \cdots, k\}, X_{\pi_k^{-1}(I(k,1,\epsilon)-1)} > X_{\pi_k^{-1}(j),n}\right] \tag{155}$$

By equation 155 and lemma B.6 (with $\omega = \frac{1}{2}$) and since $s > 2$ so $e_{\gamma,2} = 0$, there exists an N such that for $n \geq N$

$$A^C_{3,k} \subseteq \left[\|\boldsymbol{Z}_n - \mathbb{E}\boldsymbol{Z}_n\|_\infty \geq \frac{\sqrt{C^*_{s,\beta}}}{4} n^{-\frac{1}{2}+e_{\gamma,1}}\right] \tag{156}$$

Now note that by definition of $e_{\gamma,1}$ (equation 66) and since $0 < \epsilon < \frac{1-s\beta}{\beta B(s)}$ we have that

$$1 - \frac{s\beta}{2} - \epsilon\beta B(s) > \frac{1}{2} - \frac{\epsilon\beta B(s)}{2} = \frac{1}{2} - e_{\gamma,1}$$

Using this and that the definition of $e_{\gamma,1}$ (see equation 66), we have that

$$\begin{aligned} \left(\frac{\sqrt{C^*_{s,\beta}}}{4} n^{-\frac{1}{2}+e_{\gamma,1}}\right)^2 \frac{1}{\min_{j\in[k]}\sqrt{\boldsymbol{p}^k(j)(1-\boldsymbol{p}^k(j)}} \lesssim \\ n^{-(1-\frac{s\beta}{2}-\epsilon\beta B(s))} = \\ o\left(n^{-\left(\frac{1}{2}-e_{\gamma,1}\right)}\right) \end{aligned} \tag{157}$$

Thus using equation 156 and 157, we have that for some N and $n \geq N$

$$\begin{aligned} A^C_{3,k} \subseteq \\ \left[\|\boldsymbol{Z}_n - \mathbb{E}\boldsymbol{Z}_n\|_\infty \geq \frac{\sqrt{C^*_{s,\beta}}}{8} n^{-\frac{1}{2}+e_{\gamma,1}} + \right. \\ \left. \frac{1}{2}\left(\frac{\sqrt{C^*_{s,\beta}}}{8} n^{-\frac{1}{2}+e_{\gamma,1}}\right)^2 \frac{1}{\min_{j\in[k]}\sqrt{\boldsymbol{p}^k(j)(1-\boldsymbol{p}^k(j))}}\right] \end{aligned} \tag{158}$$

So applying lemma B.1 with $C = \frac{\sqrt{C^*_{s,\beta}}}{8}$ and $u = n^{-\frac{1}{2}+e_{\gamma,1}}$, we have that there is a constant $C_{s,\beta,6} > 0$ such that

$$\mathbb{P}(A^C_{3,k}) \lesssim k \exp\left(-C_{s,\beta,6} n^{2e_{\gamma,1}}\right) = k \exp\left(-C_{s,\beta,6} n^{\epsilon\beta B(s)}\right) \tag{159}$$

Next, noting that for $x \in \mathbb{R}$, $|x| = \sqrt{x^2}$ and using reverse Jensen's inequality for the $\sqrt{\cdot}$ function and then using the expression for the variance of an empirical proportion, we have that for an $N_{s,\beta}$ sufficiently large and $n \geq N_{s,\beta}$

$$\begin{gathered}\mathbb{E} \sum_{j=T_{n,\epsilon}+1}^{k} |\bar{\boldsymbol{p}}^k(\pi_k^{-1}(j)) - \boldsymbol{p}^k(\pi_k^{-1}(j))| \leq \\ \frac{1}{\sqrt{n}} \sum_{j=T_{n,\epsilon}+1}^{k} \sqrt{\boldsymbol{p}^k(\pi_k^{-1}(j))(1-\boldsymbol{p}^k(\pi_k^{-1}(j)))} \leq \\ \frac{1}{\sqrt{nH_{k,s}}} \sum_{j=T_{n,\epsilon}+1}^{k} j^{-s/2}\end{gathered} \tag{160}$$

Now because the summands $j^{-s/2}$ monotonically decrease, we use Riemann integration to bound the sum. Specifically,

$$\sum_{j=T_{n,\epsilon}+1}^{k} j^{-s/2} \leq \int_{T_{n,\epsilon}}^{k} x^{-s/2} dx = \frac{1}{\frac{s}{2}-1}\left(T_{n,\epsilon}^{1-\frac{s}{2}} - k^{1-\frac{s}{2}}\right) \tag{161}$$

Now recall that by definition $T_{n,\epsilon} = I(k,1,\epsilon) - 1$ and thus by the definition of $I(k,1,\epsilon)$ (equation 92), we have that

$$T_{n,\epsilon} \asymp n^{\frac{1}{s+2}-\beta\epsilon} \tag{162}$$

By equations 160 and 161 and 162 and since $H_{k,s}$ is convergent (since $s > 1$), we have that

$$\mathbb{E} \sum_{j=T_{n,\epsilon}+1}^{k} |\bar{\boldsymbol{p}}^k(\pi_k^{-1}(j)) - \boldsymbol{p}^k(\pi_k^{-1}(j))| \lesssim n^{-\frac{1}{2}+\frac{1-\frac{s}{2}}{s+2}-\beta\epsilon(1-\frac{s}{2})} \tag{163}$$

By the above equation and equations 154 and 159 and using that $-\frac{1}{2} + \frac{1-\frac{s}{2}}{s+2} = -\frac{s}{s+2}$, we conclude that

$$\mathbb{E}\|\hat{\boldsymbol{p}}^k_{n,T_{n,\epsilon}} - \boldsymbol{p}^k\|_1 \lesssim n^{-\frac{s}{s+2}+\beta\epsilon(\frac{s}{2}-1)} \tag{164}$$

Set $\tau = \beta\epsilon(\frac{s}{2}-1)$ and the theorem statement follows. □

B.7 Upper bounds when $\frac{1}{s} \leq \beta$ and $s > 2$

Theorem B.10. *(Upper Bound For Truncated Sort and Snap when $\beta \geq \frac{1}{s}$ and $s > 2$) Suppose $s > 2$, $\beta \geq \frac{1}{s}$ and for $\ell \geq 2$, $\boldsymbol{p}^\ell \in \mathcal{P}_{f_s,\ell}$ (with permutation function denoted π_ℓ) and $k = \lfloor n^\beta \rfloor$ for each n large enough so that $\lfloor n^\beta \rfloor \geq 2$. Also let $0 < \epsilon < \frac{1}{\beta(s+2)}$. Then letting $T_{n,\epsilon} = I(k,1,\epsilon) - 1$ (where $I(k,1,\epsilon)$ is defined in equation 92) there exists a $U_{s,\beta,2}$ such that for any $\tau \in (0, U_{s,\beta,2})$*

$$\mathbb{E}\|\hat{\boldsymbol{p}}^k_{n,T_{n,\epsilon}} - \boldsymbol{p}^k\|_1 \lesssim n^{-\frac{s}{s+2}} n^\tau \tag{165}$$

Proof. The proof strategy is almost exactly the same as in the $\frac{1}{s+2} \leq \beta < \frac{1}{s}$ case (see theorem B.9) with one major exception. This deviation of the proof strategy from that of theorem B.9 begins at equation 156. Instead of this equation, define a new, $\lfloor n^{1/s-\epsilon_2} \rfloor$ dimensional vector $\boldsymbol{Z}^*_n$ where

$$\frac{\epsilon\beta B(s)}{s} < \epsilon_2 < \frac{\left(\frac{s-1}{s} - \frac{s}{s+2}\right) + \beta\epsilon s}{s-1} \tag{166}$$

and for $j \in \{1, 2, \ldots, \lfloor n^{1/s-\epsilon_2} \rfloor - 1\}$,

$$\boldsymbol{Z}^*_{jn} := \frac{X_{\pi_k^{-1}(j),n}}{n\sqrt{\boldsymbol{p}^k(\pi_k^{-1}(j))(1-\boldsymbol{p}^k(\pi_k^{-1}(j)))}} \tag{167}$$

and

$$\boldsymbol{p}_{E,n} := \sum_{j=\lfloor n^{1/s-\epsilon_2} \rfloor}^{k} \boldsymbol{p}^k(\pi_k^{-1}(j)) \tag{168}$$

and

$$X_{E,n} := \sum_{j=\lfloor n^{1/s-\epsilon_2} \rfloor}^{k} X_{\pi_k^{-1}(j),n} \tag{169}$$

and the final entry of the vector is defined as

$$\boldsymbol{Z}^*_{\lfloor n^{1/s-\epsilon_2} \rfloor n} := \frac{X_{E,n}}{n\sqrt{\boldsymbol{p}_{E,n}(1-\boldsymbol{p}_{E,n})}} \tag{170}$$

Note that the upper bound on ϵ_2 is positive because $s > 2$ and that the lower bound on ϵ_2 is indeed below the upper bound on ϵ_2 because of the assumed upper bound on ϵ.

Now we need a slight modification of lemma B.6 to be able to argue that

$$A^C_{3,k} \subseteq \left[\|\boldsymbol{Z}^*_n - \mathbb{E}\boldsymbol{Z}^*_n\|_\infty \geq \frac{\sqrt{C^*_{s,\beta}}}{4} n^{-\frac{1}{2}+e_{\gamma,1}} \right] \tag{171}$$

where $A_{3,k}$ is as defined in theorem B.9. The modification is required because unlike for the $\boldsymbol{Z}_n$ used in lemma B.6, the last probability of $\boldsymbol{Z}^*_n$ is not an s–Zipfian probability. So we must still check carefully that $\boldsymbol{p}_{E,n} \leq \boldsymbol{p}^k(\pi_k^{-1}(I(k,1,\epsilon))$; this ensures (using arguments identical to lemma B.6) that $X_{E,n} \leq X_{\pi_k^{-1}(I(k,1,\epsilon)-1),n}$ under event $A_{3,k}$.

First note that by definition of $I(k,1,\epsilon)$ (see equation 92)

$$\boldsymbol{p}^k(\pi_k^{-1}(I(k,1,\epsilon)) \asymp n^{-\left(\frac{s}{s+2}-\beta\epsilon s\right)} \tag{172}$$

And using Riemann integration,

$$\boldsymbol{p}_{E,n} \lesssim \int_{\lfloor n^{1/s-\epsilon_2} \rfloor - 1}^{k} x^{-s} dx \lesssim n^{\frac{1}{s}(1-s)-\epsilon_2(1-s)} = n^{-\left(\frac{s-1}{s}+\epsilon_2(1-s)\right)} = o\left(n^{-\left(\frac{s}{s+2}-\beta\epsilon s\right)}\right) \tag{173}$$

where the last equality in the above line is because $\epsilon_2 < \frac{\left(\frac{s-1}{s}-\frac{s}{s+2}\right)+\beta\epsilon s}{s-1}$ implies that $\frac{s-1}{s}+\epsilon_2(1-s) > \frac{s}{s+2}-\beta\epsilon s$. Using equations 172 and 173 we conclude that

$$\boldsymbol{p}_{E,n} = o\left(\boldsymbol{p}^k(\pi_k^{-1}(I(k,1,\epsilon))\right)$$

. Using the above equation and arguments that are identical to those of lemma B.6, equation 171 follows (when $s > 2$).

The next adjustment to the argument of theorem B.9 we need is regarding the dominating factor in the use of Bernstein's inequality. Analogous to equation 157 in theorem B.9 (and using lemma B.12), we have that

$$\begin{aligned} \left(n^{-\frac{1}{2}+e_{\gamma,1}}\right)^2 \frac{1}{\min\left(\left(\min_{j\in\lfloor n^{1/s-\epsilon_2}\rfloor-1}\sqrt{\boldsymbol{p}^k(\pi_k^{-1}(j))(1-\boldsymbol{p}^k(\pi_k^{-1}(j)))}\right), \sqrt{\boldsymbol{p}_{E,n}(1-\boldsymbol{p}_{E,n})}\right)} &\lesssim \\ n^{-1+\epsilon\beta B(s)} n^{\frac{1}{2}-\frac{\epsilon_2 s}{2}} &\lesssim \\ n^{-1/2} n^{\epsilon\beta B(s)-\frac{\epsilon_2 s}{2}} &= \\ o\left(n^{-1/2+e_{\gamma,1}}\right) \end{aligned} \tag{174}$$

where the last equality in the above equation follows because of the lower bound on ϵ_2 (given in equation 166) and the definition of $e_{\gamma,1}$ (see equation 66).

Using equations 171 and 174, we have that for sufficiently large N and $n \geq N$

$$A_{3,k}^C \subseteq \left[\|\boldsymbol{Z}_n^* - \mathbb{E}\boldsymbol{Z}_n^*\|_\infty \geq \frac{\sqrt{C_{s,\beta}^*}}{8} n^{-\frac{1}{2}+e_{\gamma,1}} + \frac{1}{2}\left(\frac{\sqrt{C_{s,\beta}^*}}{8} n^{-\frac{1}{2}+e_{\gamma,1}}\right)^2 \frac{1}{\min\left(\left(\min_{j\in\lfloor n^{1/s-\epsilon_2}\rfloor-1} \sqrt{\boldsymbol{p}^k(\pi_k^{-1}(j))(1-\boldsymbol{p}^k(\pi_k^{-1}(j)))}\right), \sqrt{\boldsymbol{p}_{E,n}(1-\boldsymbol{p}_{E,n})}\right)}\right] \tag{175}$$

Finally, applying lemma B.1 for the Multinomial distribution indicated by $\boldsymbol{Z}_n^*$ and with $C = \frac{\sqrt{C_{s,\beta}^*}}{8}$ and

$$u = n^{-\frac{1}{2}+e_{\gamma,1}}$$

we conclude that for some $C_{s,\beta,7} > 0$

$$\mathbb{P}(A_{3,k}^C) \lesssim k \exp(-C_{s,\beta,7} n^{\epsilon\beta B(s)}) \tag{176}$$

The remainder of the proof proceeds identically to theorem B.9 □

B.8 Miscellaneous facts about the Zipfian function

Recall the notation $f_s(x) := x^{-s}$.

Lemma B.11. *If $f : (0,\infty) \to (0,\infty)$ is convex, then for any $x_1, x_2 \in \{1,2,3,\dots\}$ such that $x_1 < x_2$*

$$f(x_1) - f(x_1+1) \geq f(x_2) - f(x_2+1) \tag{177}$$

In particular this holds for f_s

Proof. We will prove the lemma by induction for $x_2 \in \{x_1+1, x_1+2, \dots\}$. For the base case, suppose $x_2 = x_1+1$. Then since f is convex,

$$f(\frac{x_1}{2} + \frac{x_1+2}{2}) \leq \frac{f(x_1) + f(x_1+2)}{2}$$

This implies that

$$2f(x_1+1) \leq f(x_1) + f(x_1+2)$$

Which implies that

$$f(x_2) - f(x_2+1) = f(x_1+1) - f(x_1+2) \leq f(x_1) - f(x_1+1)$$

completing the base case. Now suppose that for some $x_2 \in \{x_1+1, x_1+2, \dots\}$ that

$$f(x_2) - f(x_2+1) \leq f(x_1) - f(x_1+1) \tag{178}$$

Since f is convex,

$$f(\frac{x_2}{2} + \frac{x_2+2}{2}) \leq \frac{f(x_2) + f(x_2+2)}{2}$$

Applying the same algebraic simplifications as in the base case to the above equation yields

$$f(x_2+1) - f(x_2+2) \leq f(x_2) - f(x_2+1)$$

Combining this with the inductive assumption yields

$$f(x_2+1) - f((x_2+1)+1) \leq f(x_1) - f(x_1+1)$$

In particular equation 177 holds for x_2+1. By induction equation 177 follows for every $x_2 \in \{x_1+1, x_1+2, \dots\}$. Finally note that $x_1 \in \{1,2,\dots\}$ was arbitrary equation 177 follows for any $x_1, x_2 \in \{1,2,\dots\}$ such that $x_1 < x_2$.

Also, note that for $x > 0$, $f_s''(x) = s(s+1)f_{s+2}(x) > 0$. Thus f_s is convex and so property 177 holds for f_s. □

Lemma B.12. *For $k \geq 2$, if $\boldsymbol{p} \in \mathcal{S}_k$ and π^{-1} indexes the sorted order of $\boldsymbol{p}$, that is,*

$$\boldsymbol{p}(\pi^{-1}(1)) \geq \boldsymbol{p}(\pi^{-1}(2)) \geq \cdots \geq \boldsymbol{p}(\pi^{-1}(k))$$

then for any $1 \leq i_1 < i_2 \leq k$

$$\boldsymbol{p}(\pi^{-1}(i_1))(1 - \boldsymbol{p}(\pi^{-1}(i_1))) \geq \boldsymbol{p}(\pi^{-1}(i_2))(1 - \boldsymbol{p}(\pi^{-1}(i_2)))$$

In particular

$$\frac{f_s(i_1)}{H_{k,s}}(1 - \frac{f_s(i_1)}{H_{k,s}}) \geq \frac{f_s(i_2)}{H_{k,s}}(1 - \frac{f_s(i_2)}{H_{k,s}}) \tag{179}$$

Proof. Since $\boldsymbol{p}(\pi^{-1}(1))$ is the largest probability and $\boldsymbol{p} \in \mathcal{S}_k$ we have that

$$\forall j \in \{2, 3, \ldots, k\}, \boldsymbol{p}(\pi^{-1}(j)) \leq 1/2 \tag{180}$$

If $\max(\boldsymbol{p}(\pi^{-1}(i_1)), \boldsymbol{p}(\pi^{-1}(i_2))) \leq \frac{1}{2}$, then since $g(x) = x(1-x)$ is monotonically increasing on $[0, 1/2]$ and $i_1 > i_2$

$$g(\boldsymbol{p}(\pi^{-1}(i_1))) \geq g(\boldsymbol{p}(\pi^{-1}(i_2)))$$

If $\max(\boldsymbol{p}(\pi^{-1}(i_1)), \boldsymbol{p}(\pi^{-1}(i_2))) > \frac{1}{2}$, then since $i_1 < i_2$, $i_1 = 1$ and in particular, $\boldsymbol{p}(\pi^{-1}(1)) > 1/2$. Therefore

$$\boldsymbol{p}(\pi^{-1}(i_2)) \leq 1 - \boldsymbol{p}(\pi^{-1}(i_1)) \leq \frac{1}{2} \tag{181}$$

So again since g is monotonically increasing on $[0, 1/2]$, using the above equation we have that

$$g(\boldsymbol{p}(\pi^{-1}(i_2))) \leq g(1 - \boldsymbol{p}(\pi^{-1}(i_1))) = g(\boldsymbol{p}(\pi^{-1}(i_1)))$$

where the equality in the above line is because $g(x) = g(1-x)$ for $x \in [0, 1]$.

Finally, note that equation 179 follows because f_s is monotonically decreasing and $H_{k,s}$ normalizes f_s into a probability measure. □