The power of absolute discounting: all-dimensional distribution estimation

Moein Falahatgar
UCSD
moein@ucsd.edu

Mesrob Ohannessian
TTIC
mesrob@gmail.com

Alon Orlitsky
UCSD
alon@ucsd.edu

Venkatadheeraj Pichapati
UCSD
dheerajpv7@ucsd.edu

Abstract

Categorical models are a natural fit for many problems. When learning the distribution of categories from samples, high-dimensionality may dilute the data. Minimax optimality is too pessimistic to remedy this issue. A serendipitously discovered estimator, absolute discounting, corrects empirical frequencies by subtracting a constant from observed categories, which it then redistributes among the unobserved. It outperforms classical estimators empirically, and has been used extensively in natural language modeling. In this paper, we rigorously explain the prowess of this estimator using less pessimistic notions. We show that (1) absolute discounting recovers classical minimax KL-risk rates, (2) it is *adaptive* to an effective dimension rather than the true dimension, (3) it is strongly related to the Good–Turing estimator and inherits its *competitive* properties. We use power-law distributions as the cornerstone of these results. We validate the theory via synthetic data and an application to the Global Terrorism Database.

1 Introduction

Many natural problems involve uncertainties about categorical objects. When modeling language, we reason about words, meanings, and queries. When inferring about mutations, we manipulate genes, SNPs, and phenotypes. It is sometimes possible to embed these discrete objects into continuous spaces, which allows us to use the arsenal of the latest machine learning tools that often (though admittedly not always) need numerically meaningful data. But why not operate in the discrete space directly? One of the main obstacles to this is the dilution of data due to the high-dimensional aspect of the problem, where dimension in this case refers to the number k of categories.

The classical framework of categorical distribution estimation, studied at length by the information theory community, involves a fixed small k, [BS04]. Add-contant estimators are sufficient for this purpose. Some of the impetus to understanding the large k regime came from the neuroscience world, [Pan04]. But this extended the pessimistic worst-case perspective of the earlier framework, resulting in guarantees that left a lot to be desired. This is because high-dimension often also comes with additional structure. In particular, if a distribution produces only roughly d distinct categories in a sample of size n, then we ought to think of d (and not k) as the *effective* dimension of the problem. There are also some ubiquitous structures, like power-law distributions. Natural language is a flagship example of this, which was observed as early as by Zipf in [Zip35]. Species and genera, rainfall, terror incidents, to mention just a few all obey power-laws [SLE$^+$03, CSN09, ADW13].

Are there estimators that mold to both dimension *and* structure? It turns out we don't need to search far. In natural language processing (NLP) it was first discovered that an estimator proposed by Good and Turing worked very well [Goo53]. Only recently did we start understanding why and how [OSZ03, OD12, AJOS13, OS15]. And the best explanation thus far is that it implicitly *competes* with the best estimator in a very small neighborhood of the true distribution. But NLP researchers [NEK94, KN95, CG96] have long realized that another simpler estimator, *absolute discounting*, is equally good. But why and how this is the case was never properly determined, save some mention in [OD12] and in [FNT16], where the focus is primarily on form.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

In this paper, we first show that absolute discounting, defined in Section 3, recovers pessimistic minimax optimality in both the low- and high-dimensional regimes. This is an immediate consequence of an upper bound that we provide in Section 5. We then study lower bounds with classes defined by the number of distinct categories d and also power-law structure in Section 6. This reveals that absolute discounting in fact *adapts* to the family of these classes. We further unravel the relationship of absolute discounting with the Good–Turing estimator, for power-law distributions. Interestingly, this leads to a further refinement of this estimator's performance in terms of *competitivity*. Lastly, we give some synthetic experiments in Section 8 and then explore forecasting global terror incidents on real data [LDMN16], which showcases very well the "all-dimensional" learning power of absolute discounting. These contributions are summarized in more detail in Section 4. We start out in Section 2 with laying out what we mean by these notions of optimality.

2 Optimal distribution learning

In this section we concretely formulate the optimal distribution learning framework and take the opportunity to point out related work.

Problem setting Let $p = (p_1, p_2, \dots, p_k)$ be a distribution over $[k] := \{1, 2, \dots, k\}$ categories. Let $[k]^*$ be the set of finite sequences over $[k]$. An estimator q is a mapping that assigns to every sequence $x^n \in [k]^*$ a distribution $q(x^n)$ over $[k]$. We model p as being the underlying distribution over the categories. We have access to data consisting of n samples $X^n = X_1, X_2, ..., X_n$ generated *i.i.d.* from p. Intuitively, our goal is to find a choice of q that is guaranteed to be as close as any other estimator can be to p, in average. We first need to quantify how performance is measured.

General notation: Let $(\mu_j \ : \ j = 1, \cdots, k)$ denote the empirical counts, i.e. the number of times symbol j appears in X^n and let D be the number of *distinct* categories appearing in X^n, i.e. $D = \sum_j \mathbb{1}\{\mu_j > 0\}$. We denote by $d := \mathbb{E}[D]$ its expectation. Let $(\Phi_\mu \ : \ \mu = 0, \cdots, n)$, be the total number of categories appearing exactly μ times, $\Phi_\mu := \sum_j \mathbb{1}\{\mu_j = \mu\}$. Note that $D = \sum_{\mu > 0} \Phi_\mu$. Also let $(S_\mu \ : \ \mu = 0, \cdots, n)$, be the total probability within each such group, $S_\mu := \sum_j p_j \mathbb{1}\{\mu_j = \mu\}$. Lastly, denote the empirical distribution by $q_j^{+0} := \mu_j / n$.

KL-Risk We adopt the Kullback-Leibler (KL) divergence as a measure of loss between two distributions. When a distribution p is approximated by another q, the KL divergence is given by $\mathsf{KL}(p||q) := \sum_{j=1}^k p_j \log \frac{p_j}{q_j}$. We can then measure the performance of an estimator q that depends on data in terms of the *KL-risk*, the expectation of the divergence with respect to the samples. We use the following notation to express the KL-risk of q after observing n samples X^n:

$$r_n(p, q) := \mathop{\mathbb{E}}_{X^n \sim p^n} [\mathsf{KL}(p||q(X^n))].$$

An estimator that is identical to p regardless of the data is unbeatable, since $r_n(p, q) = 0$. Therefore it is important to model our ignorance of p and gauge the optimality of an estimator q accordingly. This can be done in various ways. We elaborate the three most relevant such perspectives: *minimax*, *adaptive*, and *competitive* distribution learning.

Minimax In the *minimax* setting, p is only known to belong to some class of distributions $\mathcal{P}$, but we don't know which one. We would like to perform well, no matter which distribution it is. To each q corresponds a distribution $p \in \mathcal{P}$ (assuming the class is finite or closed) on which q has its *worst* performance:

$$r_n(\mathcal{P}, q) := \max_{p \in \mathcal{P}} r_n(p, q).$$

The minimax risk is the *least* worst-case KL-risk achieved by *any* estimator q,

$$r_n(\mathcal{P}) := \min_q r_n(\mathcal{P}, q).$$

The minimax risk depends only on the class $\mathcal{P}$. It is a *lower bound*: no estimator can beat it *for all* p, i.e. it's not possible that $r_n(p, q) < r_n(\mathcal{P})$ for all $p \in \mathcal{P}$. An estimator q that satisfies an *upper bound* of the form $r_n(\mathcal{P}, q) = (1 + o(1)) r_n(\mathcal{P})$ is said to be minimax *optimal* "even to the constant" (an informal but informative expression that we adopt in this paper). If instead $r_n(\mathcal{P}, q) = \mathcal{O}(1) r_n(\mathcal{P})$, we say that q is *rate optimal*. Near-optimality notions are also possible, but we don't dwell on them. As an aside, note that *universal compression* is minimax optimality using *cumulative* risk. See [FJO$^+$15] for such related work on universal compression for power laws.

Adaptive The minimax perspective captures our ignorance of p in a pessimistic fashion. This is because $r_n(\mathcal{P})$ may be large, but for a specific $p \in \mathcal{P}$ we may have a much smaller $r_n(p, q)$. How can we go beyond this pessimism? Observe that when a class is smaller, then $r_n(\mathcal{P})$ is smaller. This is because we'd be maximizing on a smaller set. In the extreme case noted earlier, when $\mathcal{P}$ contains only a single distribution, we have $r_n(\mathcal{P}) = 0$. The *adaptive* learning setting finds an intermediate ground where we have a *family* of distribution classes $\mathcal{F} = \{\mathcal{P}_s : s \in \mathcal{S}\}$ indexed by a (not necessarily countable) index set $\mathcal{S}$. For each s, we have a corresponding $r_n(\mathcal{P}_s)$ which is often much smaller than $r_n\left(\bigcup_{s\in\mathcal{S}} \mathcal{P}_s\right)$, and we would like the estimator to achieve the risk bound corresponding to the smaller class. We say that an estimator q is *adaptive* to the family $\mathcal{F}$ if for all $s \in \mathcal{S}$:

$$r_n(p, q) \leq\ O_s(1)\, r_n(\mathcal{P}_s) \quad \forall p \in \mathcal{P}_s \quad \iff \quad r_n(\mathcal{P}_s, q) \leq\ O_s(1)\, r_n(\mathcal{P}_s)$$

There often is a price to adaptivity, which is a function of the granularity of $\mathcal{F}$ and is paid in the form of varying/large leading constants per class. This framework has been particularly successful in density estimation with smoothness classes [Tsy09] and has been recently used in the discrete setting for universal compression [BGO15].

Competitive The adaptive perspective can be tightened by demanding that, rather than a multiplicative constant, the KL-risk tracks the risk up to a vanishingly small *additive* term:

$$r_n(p, q) = r_n(\mathcal{P}_s) + \epsilon_n(\mathcal{P}_s, q) \quad \forall p \in \mathcal{P}_s.$$

Ideally, we would like the *competitive loss* $\epsilon_n(\mathcal{P}_s, q)$ to be negligible compared to the risk of each class $r_n(\mathcal{P}_s)$. If $\epsilon_n(\mathcal{P}_s, q) = O_s(1) r_n(\mathcal{P}_s)$ for all s, then we recover adaptivity. And when $\epsilon_n(\mathcal{P}_s, q) = o_s(1) r_n(\mathcal{P}_s)$ for all $s \in \mathcal{S}$, we have minimax optimality even to the constant within each class, which is a much stronger form of adaptivity. We then say that the estimator is *competitive* with respect to the family $\mathcal{F}$. We may also evaluate the *worst-case* competitive loss, over $\mathcal{S}$.

This formulation was recently introduced in [OS15] in the context of distribution learning. This work shows that the celebrated Good–Turing estimator [Goo53], combined with the empirical estimator, has small worst-case competitive loss over the family of classes defined by any given distribution and all its permutations. Most importantly, this loss was shown to stay bounded, even as the dimension increases. This provided a rigorous theoretical explanation for the performance of the Good–Turing estimator in high-dimensions. A similar framework is also studied for ℓ_1-loss in [VV15].

3 Absolute discounting

One of the first things to observe is that the empirical distribution is particularly ill-suited to handle KL-risk. This is most easily seen by the fact that we'd have infinite blow-up when any $\mu_j = 0$, which *will* happen with positive probability. Instead, one could resort to an add-constant estimator, which for a positive β is of the form $q_j^{+\beta} := (\mu_j + \beta)/(n + k\beta)$.

The most widely-studied class of distributions is the one that includes all of them: the $k-$dimensional simplex, $\Delta_k := \{(p_1, p_2, \ldots, p_k), : \sum\limits_i p_i = 1,\ p_i \geq 0\ \forall i \in [k]\}$. In the low-dimensional scaling, when $n/k \to \infty$ (the "dimension" here being the support size k), the minimax risk is

$$r_n(\Delta_k) = (1 + o(1))\, \frac{k-1}{2n},$$

In [BS04], a variant of the add-constant estimator is shown to achieve this risk even to the constant. Furthermore, any add-constant estimator is rate optimal when k is fixed. But in the very high-dimensional setting, when $k/n \to \infty$, [Pan04] showed that the minimax risk behaves as

$$r_n(\Delta_k) = (1 + o(1))\ \log \frac{k}{n},$$

achieved by an add-constant estimator, but with a constant that depends on the ratio of k and n.

Despite these classical results on minimax optimal estimators, in practice people often use other estimators that have better empirical performance. This was a long-running mystery in the language modeling community [CG96], where variants of the Good–Turing estimator were shown to perform the best [JM85, GS95]. The gap in performance was only understood recently, using the notion of competitivity [OS15]. In essence, the Good–Turing estimator works well in *both* low- and high-dimensional regimes, and in-between. Another estimator, *absolute discounting*, unlike add-constant estimators, simply *subtracts* a positive constant from the empirical counts and redistributes the subtracted amount to unseen categories. For a discount parameter $\delta \in [0, 1)$, it is defined as:

$$q_j^{-\delta} := \begin{cases} \frac{\mu_j - \delta}{n} & \text{if } \mu_j > 0, \\ \frac{D\delta}{n(k-D)} & \text{if } \mu_j = 0. \end{cases} \tag{1}$$

Starting with the work of [NEK94], absolute discounting soon supplanted the Good–Turing estimator, due to both its simplicity and comparable performance. Kneser-Ney smoothing [KN95], which uses absolute discounting at its core was long held as the preferred way to train N-gram models. Even to this day, the state-of-the-art language models are combined systems where one usually interpolates between recurrent neural networks and Kneser-Ney smoothing [JVS$^+$16]. Can this success be explained?

Kneser-Ney is for the most part a principled implementation of the notion of back-off, which we only touch upon in the conclusion. The use of absolute discounting is critical however, as performance deteriorates if we back-off with care but use a more naïve add-constant or even Katz-style smoothing [Kat87], which switches from the Good–Turing to the empirical distribution at a fixed frequency point. It is also important to mention the Bayesian approach of [Teh06] that performs similarly to Kneser-Ney, called the Hierarchical Pitman-Yor language model. The hierarchies in this model reprise the role of back-off, while the two-parameter Poisson-Dirichlet prior proposed by Pitman and Yor [PY97] results in estimators that are very similar to absolute discounting. The latter is not a surprise because this prior almost surely generates a power law distribution, which is intimately related to absolute discounting as we study in this paper. Though our theory applies more generally, it can in fact be straightforwardly adapted to give guarantees to estimators built upon this prior.

4 Contributions

We investigate the reason behind the auspicious behavior of the absolute discounting estimator. We achieve this by demonstrating the adaptivity and competitivity of this estimator for many relevant families of distribution classes. In summary:

- We analyze the performance of the absolute discounting estimator by upper bounding the KL-risk for each class in a family of distribution classes defined by the expected number of distinct categories. [Section 5, Theorem 1] This result implies that absolute discounting achieves classical minimax rate-optimality in *both* the low- and high-dimensional regimes over the whole simplex Δ_k, as outlined in Section 2.
- We provide a generic lower bound to the minimax risk of classes defined by a single distribution and all of its permutations. We then show that if the defining distribution is a truncated (possibly perturbed) power-law, then this lower bound matches the upper bound of absolute discounting, up to a constant factor. [Section 6, Corollaries 3 and 4]
- This implies that absolute discounting is adaptive to the family of classes defined by a truncated power-law distribution and its permutations. Also, since classes defined by the expected number of distinct categories necessarily includes a power-law, absolute discounting is also adaptive to this family. This is a strict refinement of classical minimax rate-optimality.
- We give an equivalence between the absolute discounting and Good–Turing estimators in the high-dimensional setting, whenever the distribution is a truncated power-law. This is a finite-sample guarantee, as compared to the asymptotic version of [OD12]. As a consequence, absolute-discounting becomes competitive with respect to the family of classes defined by permutations of power-laws, inheriting Good–Turing's behavior [OS15]. [Section 7, Lemma 5 and Theorem 6]

We corroborate the theoretical results with synthetic experiments that reproduce the theoretical minimax risk bounds. We also show that the prowess of absolute discounting on real data is not restricted only to language modeling. In particular, we explore a striking application to forecasting global terror incidents and show that, unlike naive estimators, absolute discounting gives accurate predictions simultaneously in all of low-, medium-, and high-activity zones. [Section 8]

5 Upper bound and classical minimax optimality

We now give an upper bound for the risk of the absolute discounting estimator and show that it recovers classical minimax rates in the low- and high-dimensional regimes. Recall that $d := \mathbb{E}[D]$ is the expected number of distinct categories in the samples. The upper bound that we derive can be written as function of only d, k, and n, and is non-decreasing in d. For a given n and k, let $\mathcal{P}_d$ be the set of all distributions for which $\mathbb{E}[D] \leq d$. The upper bound is thus also a worst-case bound over $\mathcal{P}_d$.

Theorem 1 (Upper bound)**.** *Consider the absolute discounting estimator $q = q^{-\delta}$, defined in (1). Let p be such that $\mathbb{E}[D] = d$. Given a discount $0 < \delta < 1$, there exists a constant c that may depend on δ and only δ, such that*

$$r_n(p, q) \leq \begin{cases} \dfrac{d}{n} \log \dfrac{k - \frac{d}{2}}{\frac{d}{2}} + c\dfrac{d}{n} & \text{if} \quad d \geq 10 \log\log k, \\ \dfrac{d}{n} \log k + c\dfrac{d}{n} & \text{if} \quad d < 10 \log\log k. \end{cases} \tag{2}$$

The same bound holds for $r_n(\mathcal{P}_d, q)$.

We defer the proof the theorem to the supplementary material. Here are the immediate implications. For the low-dimensional regime $\frac{n}{k} \to \infty$ and the class Δ_k, the largest d can be once $n > k$ is k. The risk of absolute discounting is thus bounded by $c(1 + o(1))\frac{k}{n} = \mathcal{O}(1)\frac{k}{n}$. This is minimax rate-optimal [BS04]. For the high-dimensional regime $\frac{k}{n} \to \infty$ and the class Δ_k, the largest d can be when $k > n$ is n. The risk of absolute discounting is thus dominated by the first term, which reduces to $(1 + o(1)) \log \frac{k}{n}$. This is the optimal risk for the class Δ_k [Pan04], even to the constant.

Therefore on the two extreme ranges of k and n absolute discounting recovers the best performance, either as rate-optimal or optimal even to the constant. These results are for the entire $k-$dimensional simplex Δ_k. Furthermore, for smaller classes, it characterizes the worst-case risk of the class by the d, the expected number of distinct categories. Is this characterization tight?

6 Lower bounds and adaptivity

In order to lower bound the minimax risk of a given class $\mathcal{P}$, we use a finer granularity than the $\mathcal{P}_d$ classes described in Section 5. In particular, let $\mathcal{P}_p$ be the *permutation class* of distributions consisting of a single distribution p and all of its permutations. Note that the multiset of probabilities is the same for all distributions in $\mathcal{P}_p$, and since the expected number of distinct categories only depends on the multiset ($d = \sum_j [1 - (1 - p_j)^n]$) it follows that $\mathcal{P}_p \subset \mathcal{P}_d$[1]. To find a good lower bound for $\mathcal{P}_d$, we need a p that is "worst case". We first give the following generic lower bound.

Theorem 2 (Generic lower bound)**.** *Let $\mathcal{P}_p$ be a permutation class defined by a distribution p and let $\gamma > 1$. Then for $k > \gamma d$, the minimax risk is bounded by:*

$$r_n(\mathcal{P}_p) \geq \left(1 - \frac{1}{\gamma}\right) \left(\sum_{j=\gamma d}^{k} p_j\right) \log \frac{k - \gamma d}{\sum_{j=\gamma d}^{k} p_j} + \sum_{i=\gamma d} p_j \log p_j \tag{3}$$

Equation (3) can be used as a starting point for more concrete lower bounds on various distribution classes. We illustrate this for two cases. First, let us choose p to be a truncated power-law distribution with power α: $p_j \propto j^{-\alpha}, \quad$ for $j = 1, \cdots, k$. We always assume $\alpha \geq \alpha_0 > 1$. This leads to the following lower bound.

Corollary 3. *Let $\mathcal{P}$ be all permutations of a single power-law distribution with power α truncated over k categories. Then there exists a constant $c > 0$ and large enough n_0 such that when $n > n_0$ and $k > \max\{n, 1.2^{\frac{1}{\alpha-1}} n^{\frac{1}{\alpha}}\}$,*

$$r_n(\mathcal{P}) \geq c\frac{d}{n} \log \frac{k - 2d}{2d}.$$

Next, we use a different choice of p for $\mathcal{P}_p$ to provide a lower bound whenever d grows linearly with n. This essentially closes the gap of the previous corollary when α approaches 1.

[1] We abuse notation by distinguishing the classes by the letter used, while at the same time using the letters to denote actual quantities. From the context we understand that d is the expected number of distinct categories for p, at the given n.

Corollary 4. *Let $\rho \in (1, 1.75)$ and let $\mathcal{P}$ be all permutations of a single uniform distribution over a subset $k' = \frac{n}{\rho}$ out of k categories. Then $d \sim (1 - e^{-\rho})n/\rho$ and there exists a constant $c > 0$ and large enough n_0 such that when $n > n_0$ and $k > n^5$,*

$$r_n(\mathcal{P}) \geq c\frac{d}{n} \log \frac{k - 1.2d}{d} \ .$$

We defer the proofs of the theorem and its corollaries to the supplementary material. The upper bound of Theorem 1 and the lower bounds of Corollaries 3 and 4 are within constant factors of each other. The immediate consequence is that absolute discounting is adaptive with respect to the families of classes of the Corollaries. Furthermore, over the family of classes $\mathcal{P}_d$ where we can write d as $n^{\frac{1}{\alpha}}$ for some $\alpha > 1$ or $d \propto n$, we can select a distribution from the Corollaries among each class and use the corresponding lower bound to match the upper bound of Theorem 1 up to a constant factor. Therefore absolute discounting is adaptive to this family of classes. Intuitively, adaptivity to these classes establishes optimality in the intermediate range between low- and high-dimensional settings in a distribution-dependent fashion and governed by the expected number of distinct categories d, which we may regard as the *effective* dimension of the problem.

7 Relationship to Good–Turing and competitivity

We now establish a relationship between the absolute discounting and Good–Turing estimators and refine the adaptivity results of the previous section into competitivity results. When [OS15] introduced the notion of competitive optimality, they showed that a variation of the Good–Turing estimator is worst-case competitive with respect to the family of distribution classes defined by any given probability distribution and its permutations. In light of the results of Sections 5 and 6, it is natural to ask whether absolute discounting enjoys the same kind of competitive properties. Not only that, but it was observed empirically by [NEK94] and shown theoretically in [OD12] that *asymptotically* Good–Turing behaves exactly like absolute discounting, when the underlying distribution is a (possibly perturbed) power-law. We therefore choose this family of classes to prove competitivity for. We first make the aforementioned equivalence concrete by establishing a *finite sample* version. We use the following *idealized* version of the Good–Turing estimator [Goo53]:

$$q_j^{\mathsf{GT}} := \begin{cases} \frac{\mu_j+1}{n}\frac{\mathbb{E}[\Phi_{\mu_j+1}]}{\mathbb{E}[\Phi_{\mu_j}]} & \text{if } \mu_j > 0, \\ \frac{\mathbb{E}[\Phi_1]}{n(k-D)} & \text{if } \mu_j = 0. \end{cases} \tag{4}$$

Lemma 5. *Let p be a power law with power α truncated over k categories. Then for $k > \max\{n, n^{\frac{1}{\alpha-1}}\}$, we have the equivalence:*

$$q_j^{\mathsf{GT}} = \frac{\mu_j - \frac{1}{\alpha}}{n}\left(1 + \mathcal{O}\left(n^{-\frac{1}{2}\frac{3}{2\alpha+1}}\right)\right) \sim \frac{\mu_j - \frac{1}{\alpha}}{n} \qquad \forall\, \mu_j \in \left\{1, \cdots, n^{\frac{1}{2\alpha+1}}\right\}.$$

An interesting outcome of the equivalence of Lemma 5 is that it suggests a choice of the discount δ in terms of the power, $1/\alpha$. To give a data-driven version of $1/\alpha$, we will use a robust version of the ratio Φ_1/D proposed in [OD12, BBO17], which is a strongly consistent estimator when $k = \infty$.

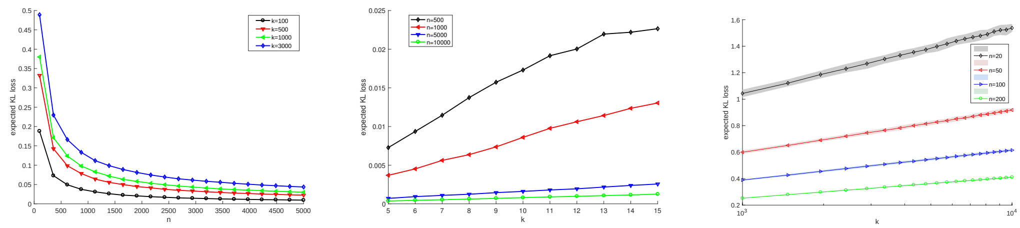
Theorem 6. *Let $\mathcal{P}$ be all permutations of a truncated power law p with power α. Let q be the absolute discounting estimator with $\delta = \min\left\{\frac{\max\{\Phi_1, 1\}}{D}, \delta_{\max}\right\}$, for a suitable choice of $\delta_{\max}$. Then for $k > \max\{n, n^{\frac{1}{\alpha-1}}\}$, the competitive loss is*

$$\epsilon_n(\mathcal{P}_p, q) = \mathcal{O}\left(n^{-\frac{2\alpha-1}{2\alpha+1}}\right) \ .$$

The implications are as follows. For the union of all such classes above a given α, we find that we beat the $n^{-1/3}$ rate of the worst-case competitive loss obtained for the estimator in [OS15]. Theorem 6 and the bounds of Sections 5 and 6, together imply that absolute discounting is not only worst-case competitive, but also *class-by-class* competitive with respect to the power-law permutation family. In other words, it in fact achieves minimax optimality even to the constant. One of the advantages of absolute discounting is that it gradually transitions between values that are close to the empirical distribution for abundant categories (since μ then dominates the discount δ), to a behavior that imitates the Good–Turing estimator for rare categories (as established by Lemma 5). In contrast, the estimator proposed in [OS15], and its antecedents starting from [Kat87], have to carefully choose a threshold where they switch abruptly from one estimator to the other.

8 Experiments

We now illustrate the theory with some experimental results. Our purpose is to (1) validate the functional form of the risk as given by our lower and upper bounds and (2) compare absolute discounting on both synthetic and real data to estimators that have various optimality guarantees. In all synthetic experiments, we use 500 Monte Carlo iterations. Also, we set the discount value based on data, $\delta = \min\{\frac{\max(\Phi_1,1)}{D}, 0.9\}$. This is as suggested in Section 7, assuming $\delta_{\max} = 0.9$ is sufficient.



(a) k fixed (b) n fixed, $k << n$ (c) n fixed, $k >> n$

Figure 1: Risk of absolute discounting in different ranges of k and n for a power-law with $\alpha = 2$

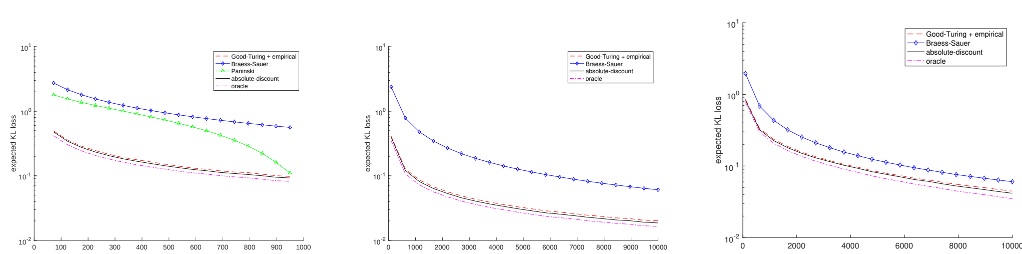
Validation For our first goal, we consider absolute discounting in isolation. Figure 1(a) shows the decay of KL-risk with the number of samples n for a power-law distribution. The dependence of the risk on the number of categories k is captured in Figures 1(b) (linear x-axis) and 1(c) (logarithmic x-axis). Note the linear growth when k is small and the logarithmic growth when k is large. For the last plot we give 95% confidence intervals for the simulations, by performing 100 restarts.

Synthetic data For our second goal, we start with synthetic data. In Figure 2, we pit absolute discounting against a number of distributions related to power-laws. The estimators used for our comparisons are: empirical $q^{+0}(x) = \frac{\mu_x}{n}$, add-beta $q^{+\beta}(x) = \frac{\mu_x + \beta_{\mu_x}}{N}$, and its two variants:

- Braess and Sauer, q^{BS} [BS04] $q^{+\beta}$ with $\beta_0 = 0.5$, $\beta_1 = 1$, and $\beta_i = 0.75\ \forall i \geq 2$
- Paninski, q^{Pan} [Pan04] $q^{+\beta}$ with $\beta_i = \frac{n}{k} \log \frac{k}{n}\ \forall i$,

absolute discounting, $q^{-\delta}$, described in 1, Good–Turing + empirical q^{GT} in [OS15], and an oracle-aided estimator where S_μ is known.

In Figures 2(a) and 2(b), samples are generated according to a power-law distribution with power $\alpha = 2$ over $k = 1{,}000$ categories. However, the underlying distribution in Figure 2(c) is a piece-wise power-law. It consists of three equal-length pieces, with powers 1.3, 2, and 1.5. Paninski's estimator is not shown in Figures 2(b) and 2(c) since it is not well-defined in this range (it is designed for the case $k > n$ only). Unsurprisingly, absolute discounting dominates these experiments. What is more interesting is that it does not seem to need a pure power-law (similar results hold for other kinds of perturbations, such as mixtures and noise). Also Good–Turing is a tight second.



(a) pure power-law (b) pure power-law (c) piece-wise power-law

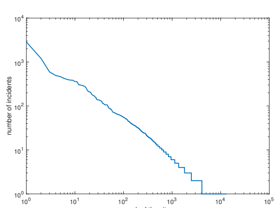
Figure 2: Comparing estimators for power-law variants with power $\alpha = 2$ and $k = 1000$.

Real data One of the chief motivations to investigate absolute discounting is natural language modeling. But there have been such extensive empirical studies that have verified over and over the power of absolute discounting (see the classical survey of [CG96]) that we chose to use this space for something new. We use the START *Global terrorism database* from the University of Maryland [LDMN16] and explore how well we can forecast the number of terrorist incidents in different cities. The data contains the record of more than $50,000$ terror incidents between the years 1992 and 2010, in more than $12,000$ different cities around the world. First, we display in Figure 3(a) the frequency of incidents across the entire dataset versus the activity rank of the city in $\log$-$\log$ scale, showing a striking adherence to a power-law (see [CSN09] for more on this).

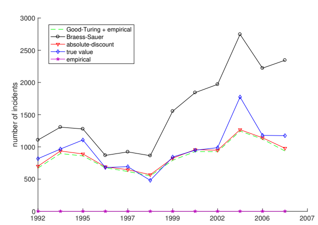
The forecasting problem that we solve is to estimate the number of total incidents in a subset of the cities over the coming year, using the current year's data from all cities. In order to emulate the various dimension regimes, we look at three subsets: (1) low-activity cities with *no* incidents in the current year and less than 20 incidents in the whole data, (2) medium-activity cities, with *some* incidents in the current year and less than 20 incidents in the whole data, and (3) high-activity individual cities with a large number of overall incidents.

The results for (1) are in Figure 3(b). The frequency estimator trivially estimates zero. Braess-Sauer does something meaningful. But absolute discounting and Good–Turing estimators, indistinguishable from each other, are remarkably on spot. And this, without having observed any of the cities! This nicely captures the importance of using structure when dimensionality is so high and data is so scarce. The results for (2) are in Figure 3(c). The frequency estimator markedly overestimates. But now absolute discounting, Good–Turing, and Braess-Sauer, perform similarly. This is a lower dimensional regime than in (1), but still not adequate for simply using frequencies. This changes in case (3), illustrated in Figure 4. To take advantage of the abundance of data, in this case at each time point we used the previous $2,000$ incidents for learning, and predicted the share of each city for the next $2,000$ incidents. In fact, incidents are so abundant that we can simply rely on the previous window's count. Note how Braess-Sauer over-penalizes such abundant categories and suffers, whereas absolute discounting and Good–Turing continue to hold their own, mimicking the performance of the empirical counts. This is a very low-dimensional regime.

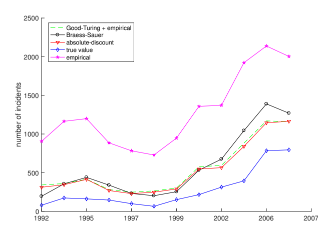
The closeness of the Good–Turing estimator to absolute discounting in all of our experiments validates the equivalence result of Lemma 5. The robustness in various regimes and the improvement in performance over such minimax optimal estimators as Braess-Sauer's and Paninski's are evidence that absolute discounting truly molds to both the raw dimension and effective dimension / structure.



(a) frequency vs rank



(b) unobserved cities

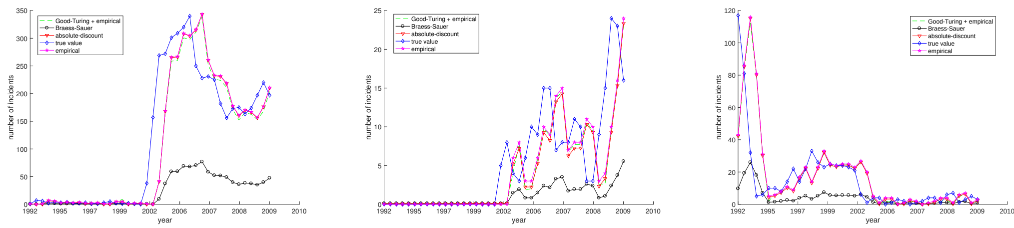


(c) observed cities

Figure 3: (a) power-law behavior of frequency vs rank in terror incidents, (b), and (c) comparing forecasts of the number of incidents in unobserved cities and observed ones, respectively.

9 Conclusion

In this paper, we offered a rigorous analysis of the absolute discounting estimator for categorical distributions. We showed that it recovers classical minimax optimality. The true reason for its success, however, is in adapting to distributions much more intimately, by recovering the right dependence on the distinct observed categories d, which can be regarded as an effective dimension, and optimally tracking structure such as power-laws. We also tightened its relationship with the celebrated Good–Turing estimator.



(a) Baghdad (b) Fallujah (c) Belfast

Figure 4: Estimating the number of incidents based on previous data for different cities

Some of our analysis could possibly be tightened, in particular in terms of the range of applicability over n, k, and d. Also, the limiting case of $\alpha = 1$ (very heavy tails, known as "fast variation" [BBO17]) to which our results don't directly apply, merits investigation. But more importantly, absolute discounting is often a module. For example, we already note how it is widely used in N-gram back-off models [KN95]. Also, recently, it has been successfully applied to smoothing low-rank probability matrices [FOO16]. Perhaps to further understand its power, it is worthwhile to study how it interacts with such larger systems.

Acknowledgements We thank Vaishakh Ravindrakumar for very helpful suggestions, and NSF for supporting this work through grants CIF-1564355 and CIF-1619448.

References

[ADW13] Armen E. Allahverdyan, Weibing Deng, and Q. A. Wang. Explaining Zipf's law via a mental lexicon. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 88(6), 2013. 1

[AJOS13] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Optimal Probability Estimation with Applications to Prediction and Classification. In *COLT*, pages 764–796, 2013. 1

[BBO17] Anna Ben Hamou, Stéphane Boucheron, and Mesrob I Ohannessian. Concentration Inequalities in the Infinite Urn Scheme for Occupancy Counts and the Missing Mass, with Applications. *Bernoulli*, 2017. 7, 9

[BGO15] Stéphane Boucheron, Elisabeth Gassiat, and Mesrob I Ohannessian. About adaptive coding on countable alphabets: Max-stable envelope classes. *IEEE Transactions on Information Theory*, 61(9), 2015. 2

[BS04] Dietrich Braess and Thomas Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004. 1, 3, 5, 8

[CG96] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics, 1996. 1, 3, 8

[CSN09] Aaron Clauset, Cosma Rohilla Shalizi, and Mark E J Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009. 1, 8

[FJO$^+$15] Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Universal compression of power-law distributions. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 2001–2005. IEEE, 2015. 2

[FNT16] Stefano Favaro, Bernardo Nipoti, and Yee Whye Teh. Rediscovery of {Good–Turing} estimators via {B}ayesian nonparametrics. *Biometrics*, 72(1):136–145, 2016. 1

[FOO16] Moein Falahatgar, Mesrob I Ohannessian, and Alon Orlitsky. Near-Optimal Smoothing of Structured Conditional Probability Matrices. In *NIPS*, pages 4860–4868, 2016. 9

[Goo53] Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, pages 237–264, 1953. 1, 2, 7

[GS95] William A Gale and Geoffrey Sampson. {Good–Turing} frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237, 1995. 3

[JM85] Frederick Jelinek and Robert Mercer. Probability distribution estimation from sparse data. *IBM technical disclosure bulletin*, 28:2591–2594, 1985. 3

[JVS+16] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016. 3

[Kat87] Slava M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer, 1987. 3, 7

[KN95] Reinhard Kneser and Hermann Ney. Improved Backing-Off for {M}-Gram Language Modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–184, Detroit, MI, may 1995. 1, 3, 9

[LDMN16] Gary LaFree, Laura Dugan, Erin Miller, and National Consortium for the Study of Terrorism and Responses to Terrorism. Global Terrorism Database, 2016. 1, 8

[NEK94] Hermann Ney, Ute Essen, and Reinhard Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language*, 8(1):1–38, 1994. 1, 3, 7

[Neu13] Edward Neuman. Inequalities and bounds for the incomplete gamma function. *Results in Mathematics*, pages 1–6, 2013. 20

[NP00] Pierpaolo Natalini and Biagio Palumbo. Inequalities for the incomplete gamma function. *Mathematical Inequalities & Applications*, 3(1):69–77, 2000. 19

[OD12] Mesrob I Ohannessian and Munther A Dahleh. Rare Probability Estimation under Regularly Varying Heavy Tails. In *COLT*, page 21, 2012. 1, 4, 7, 7, 17

[OS15] Alon Orlitsky and Ananda Theertha Suresh. Competitive distribution estimation: Why is {Good–Turing} good. In *NIPS*, pages 2143–2151, 2015. 1, 2, 3, 4, 7, 7, 8, C.2

[OSZ03] Alon Orlitsky, Narayana P Santhanam, and Junan Zhang. Always {Good–Turing}: Asymptotically optimal probability estimation. *Science*, 302(5644):427–431, 2003. 1

[Pan04] Liam Paninski. Variational Minimax Estimation of Discrete Distributions under KL Loss. In *NIPS*, pages 1033–1040, 2004. 1, 3, 5, 8

[PY97] Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997. 3

[SLE+03] Felisa A Smith, S Kathleen Lyons, S K Ernest, Kate E Jones, Dawn M Kaufman, Tamar Dayan, Pablo A Marquet, James H Brown, and John P Haskell. Body mass of late Quaternary mammals. *Ecology*, 84(12):3403, 2003. 1

[Teh06] Yee-Whye Teh. A Hierarchical Bayesian Language Model Based on Pitman-Yor processe. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, (July):985–992, 2006. 3

[Tsy09] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation.* Springer series in statistics. Springer, 2009. 2

[VV15] Gregory Valiant and Paul Valiant. Instance optimal learning. *arXiv preprint arXiv:1504.05321*, 2015. 2

[Zip35] George Kingsley Zipf. *The psycho-biology of language.* Houghton, Mifflin, 1935. 1

A Proof of Theorem 1, upper bound

We start with a technical note. Though we presented the framework for a fixed sample size n, the entirety of the paper analyzes the "Poissonized" version. In the Poisson sampling model, the number of samples is in fact $N \sim \text{POI}(n)$, a Poisson random variable with mean n. This is often the more natural model when data is collected within a fixed time window, in contrast to until a certain number of samples are collected. Or we can think of Poisson sampling as a convenience because it makes all counts independent and distributed according to $\mu_j \sim \text{POI}(np_j)$. It is possible to "de-Poissonize" the results, but we omit this for brevity.

In proof of the theorem, we show a more general upper bound. We upper bound the instantaneous risk of a class of distributions based on d, $\mathbb{E}[\Phi_1]$, and $\mathbb{E}[\Phi_2]$, the expected number of distinct categories, categories that appeared once, and twice respectively. Namely, we show for some constant c,

$$\max_{p \in \mathcal{P}_d} \mathbb{E}_{x^n}\left[\mathsf{KL}(p||q(x^n))\right] \leq \frac{\mathbb{E}[\Phi_1]}{n} \log \frac{2k-d}{d\delta} + \frac{\mathbb{E}[\Phi_1]}{n} + \frac{2\mathbb{E}[\Phi_2]}{n} \log \frac{1}{1-\delta} + \frac{c \cdot d}{n}.$$

Proof.

$$\begin{aligned}
&\mathbb{E}_{X^n \sim p^n}\left[\mathsf{KL}(p||q(X^n))\right] \\
&= \mathbb{E}_{X^n \sim p^n}\left[\sum_{j=1}^{k} p_j \log \frac{p_j}{q_j(X^n)}\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{k} \mathbb{1}_j^0 p_j \log \frac{np_j(k-D)}{D\delta} + \sum_{j=1}^{k}\sum_{i=1}^{\infty} \mathbb{1}_j^i p_j \log \frac{np_j}{i-\delta}\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{k} \mathbb{1}_j^0 p_j \log np_j + \mathbb{1}_j^0 p_j \log \frac{(k-D)}{D\delta} + \sum_{j=1}^{k}\sum_{i=1}^{\infty} \mathbb{1}_j^i p_j \log \frac{np_j}{i-\delta}\right] \\
&\overset{(a)}{=} \frac{1}{n}\sum_{j=1}^{k} e^{-\lambda_j}\lambda_j \log \lambda_j + \frac{1}{n}\sum_{j=1}^{k} \lambda_j \mathbb{E}\left[\mathbb{1}_j^0 \log \frac{k-D}{D\delta}\right] + \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{\infty} \lambda_j \log \frac{\lambda_j}{i-\delta}\text{poi}(np_j, i) \\
&\overset{(b)}{=} \frac{1}{n}\sum_{j=1}^{k} \lambda_j \mathbb{E}\left[\mathbb{1}_j^0 \log \frac{k-D}{D\delta}\right] + \frac{1}{n}\sum_{j=1}^{k}\left(\lambda_j \log \lambda_j + \sum_{i=1}^{\infty} \lambda_j \log \frac{1}{i-\delta}\text{poi}(\lambda_j, i)\right) \qquad (5)
\end{aligned}$$

where (a) is by Poisson sampling and replacing $\lambda_j := np_j$, and (b) is by combining the first and last expressions. Now we state two lemmas that are helpful in bounding each of the two terms in (5).

Lemma 7. *For all $p \in \mathcal{P}_d$ and with the assumption of $D > 2$, for $d > 10 \log\log k$,*

$$\underset{X^n \sim p^n}{\mathbb{E}}\left[\mathbb{1}_j^0 \log \frac{k-D}{D}\right] \leq e^{-\lambda_j}\left(1 + \log \frac{k - \frac{d}{2}}{\frac{d}{2}}\right)$$

and for $d < 10 \log\log k$,

$$\underset{X^n \sim p^n}{\mathbb{E}}\left[\mathbb{1}_j^0 \log \frac{k-D}{D}\right] \leq e^{-\lambda_j} \log k.$$

Lemma 8. *For $x > 0$ and for $0 < \delta < 1$, $x \log x + \sum_{i=2}^{\infty} x \log \frac{1}{i-\delta}\text{poi}(x, i) < c'$ for some constant c'.*

We can write the second part in (5) as

$$\frac{1}{n}\sum_{j=1}^{k} \lambda_j \log \frac{1}{1-\delta}\text{poi}(\lambda_j, 1) + \frac{1}{n}\sum_{j=1}^{k}\left(\lambda_j \log \lambda_j + \sum_{i=2}^{\infty} \lambda_j \log \frac{1}{i-\delta}\text{poi}(\lambda_j, i)\right), \qquad (6)$$

and since the second term in (6) is negative for all $\lambda_j < 1$, (6) is upper bounded by

$$\frac{1}{n}\sum_{j=1}^{k} \lambda_j^2 e^{-\lambda_j} \log \frac{1}{1-\delta} + \frac{1}{n}\sum_{\lambda_j \geq 1} \left(\lambda_j \log \lambda_j + \sum_{i=2}^{\infty} \lambda_j \log \frac{1}{i-\delta} \mathrm{poi}(\lambda_j, i) \right).$$

Continuing from (5), using Lemmas 7, 8 and the definitions of $\mathbb{E}[\Phi_1] = \sum_{j=1}^{k} e^{-\lambda_j}\lambda_j$ and $\mathbb{E}[\Phi_2] = \sum_{j=1}^{k} e^{-\lambda_j}\frac{\lambda_j^2}{2}$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{X^n \sim p^n}[\mathsf{KL}(p||q(X^n))] &\leq \frac{\mathbb{E}[\Phi_1]}{n}\left(\log\frac{2k-d}{d\delta} + 1\right) + \frac{2\mathbb{E}[\Phi_2]}{n}\log\frac{1}{1-\delta} + \frac{1}{n}\sum_{j:\lambda_j \geq 1} c' \\
&\leq \frac{\mathbb{E}[\Phi_1]}{n}\left(\log\frac{2k-d}{d\delta} + 1\right) + \frac{2\mathbb{E}[\Phi_2]}{n}\log\frac{1}{1-\delta} + \frac{c' \cdot d}{n(1-e^{-1})},
\end{aligned}$$

where the last line is because $d = \sum_j (1-e^{\lambda_j}) \geq \sum_{\lambda_j \geq 1}(1-e^{\lambda_j}) \geq |\{j : \lambda_j \geq 1\}|(1-e^{-0.7})$ ☐

A.1 Proof of Lemma 7

Proof. Using Lemma 17,

$$\begin{aligned}
\mathop{\mathbb{E}}_{X^n \sim p^n}[\mathbb{1}_j^0 \log\frac{k-D}{D}] &= e^{-\lambda_j}\mathbb{E}\left[\log\frac{k-D}{D}\middle| D < d - \sqrt{2vs}, \mu_j = 0\right]\Pr(D < d - \sqrt{2vs}) \\
&+ e^{-\lambda_j}\mathbb{E}\left[\log\frac{k-D}{D}\middle| D > d - \sqrt{2vs}, \mu_j = 0\right]\Pr(D > d - \sqrt{2vs}) \\
&\leq e^{-\lambda_j}\left(e^{-s}\log(k-1) + \log\frac{k-d+\sqrt{2vs}}{d-\sqrt{2vs}}\right).
\end{aligned}$$

Choosing $s = \log\log k$ and assuming $D \geq 2$ and $d > 10\log\log k$ yield the results. Note that if $\mu_j = 0$, it can change D by at most one and its effect can be ignored. Also when $d < 10\log\log k$ we can use the naive bound of $\log k$, since $\log\frac{k-D}{D} < \log k$ for $D > 1$. ☐

A.2 Proof of Lemma 8

Proof. We first assume $x > 100$ and prove the lemma.

$$\begin{aligned}
&\sum_{i=2}^{\infty}\mathrm{poi}(x,i)\log(i-\delta) &(7)\\
&\geq \sum_{i=x-x_0}^{x+x_0}\mathrm{poi}(x,i)\log(i-\delta) \\
&= \mathrm{poi}(x,x)\log(x-\delta) + \sum_{a=1}^{x_0}\mathrm{poi}(x,x-a)\log(x-a-\delta) + \mathrm{poi}(x,x+a)\log(x+a-\delta) \\
&\geq \mathrm{poi}(x,x)\log(x-\delta) + \sum_{a=1}^{x_0}\mathrm{poi}(x,x-a)\Big[\log(x-a-\delta) + \log(x+a-\delta)\Big] \\
&= \sum_{a=0}^{x_0}\frac{\mathrm{poi}(x,x-a)+\mathrm{poi}(x,x+a)}{2}\Big[\log(x-a-\delta)+\log(x+a-\delta)\Big] \\
&+ \sum_{a=1}^{x_0}\frac{\mathrm{poi}(x,x-a)-\mathrm{poi}(x,x+a)}{2}\Big[\log(x-a-\delta)+\log(x+a-\delta)\Big] &(8)
\end{aligned}$$

By Lemma 18,

$$\begin{aligned}
\sum_{a=0}^{x_0}\mathrm{poi}(x,x-a)+\mathrm{poi}(x,x+a) &= \mathrm{poi}(x,x) + 1 - \Pr(\mathrm{POI}(x) > x+x_0) - \Pr(\mathrm{POI}(x) < x-x_0) \\
&\geq \frac{1}{e\sqrt{x}} + 1 - 2\cdot e^{x_0-(x+x_0)\ln(1+\frac{x_0}{x})},
\end{aligned}$$

Also we can lower bound the bracket in (8) as

$$\begin{aligned}
\log(x+a-\delta)+\log(x-a+\delta) &= \log\left((x-\delta)^2-a^2\right) \\
&= \log(x^2-2x\delta+\delta^2-a^2) \\
&= \log\left(x^2(1-\frac{2\delta}{x}+\frac{\delta^2-a^2}{x^2})\right) \\
&= 2\log x+\log(1-\frac{2\delta}{x}+\frac{\delta^2-a^2}{x^2}) \\
&\geq 2\log x-\frac{4\delta}{x}-\frac{2(a^2-\delta^2)}{x^2}.
\end{aligned}$$

Thus for some constant c_1 and $x_0=x^{0.8}$,

$$\begin{aligned}
&\sum_{a=0}^{x_0}\frac{\mathrm{poi}(x,x-a)+\mathrm{poi}(x,x+a)}{2}\Big[\log(x-a-\delta)+\log(x+a-\delta)\Big] \\
&\geq \left(1+\frac{1}{e\sqrt{x}}-2e^{x_0-(x+x_0)\ln(1+\frac{x_0}{x})}\right)\left(\log x-\frac{2\delta}{x}\right) \\
&-\sum_{a=0}^{x_0}\Big(\mathrm{poi}(x,x-a)+\mathrm{poi}(x,x+a)\Big)\left(\frac{a^2}{x^2}\right) \\
&= \log x-\frac{2\delta}{x}-2e^{x_0-(x+x_0)\ln(1+\frac{x_0}{x})}(\log x-\frac{2\delta}{x}) \\
&-\sum_{a=0}^{x_0}\Big(\mathrm{poi}(x,x-a)+\mathrm{poi}(x,x+a)\Big)\left(\frac{a^2}{x^2}\right) \\
&\geq \log x-\frac{c_1}{x}. \qquad (9)
\end{aligned}$$

where the last line is due to the following lemma.

Lemma 9. *For $x_0=x^{0.8}$ there exists a constant c_1 such that*

$$\sum_{a=0}^{x_0}\Big[\mathrm{poi}(x,x-a)+\mathrm{poi}(x,x+a)\Big](\frac{a^2}{x^2})\leq\frac{c_1}{x}.$$

The difference in probabilities of two equidistant points from the mean of a Poisson distribution is bounded by

$$\begin{aligned}
\mathrm{poi}(x,x+a)-\mathrm{poi}(x,x-a) &= \frac{e^{-x}x^{x-a}}{(x-a)!}\Big[\frac{1}{(1+\frac{a}{x})(1+\frac{a-1}{x})\dots(1+\frac{1-a}{x})}-1\Big] \\
&= \frac{e^{-x}x^{x-a}e^{x-a}}{(x-a)^{x-a}\sqrt{2\pi(x-a)}}\Big[\frac{1}{(1+\frac{a}{x})(1+\frac{a-1}{x})\dots(1+\frac{1-a}{x})}-1\Big] \\
&= \frac{e^{-a}}{\sqrt{2\pi(x-a)}}\Big[\frac{1}{(1+\frac{a}{x})(1+\frac{a-1}{x})\dots(1+\frac{1-a}{x})}-1\Big] \\
&\approx \frac{e^{-a}}{\sqrt{2\pi(x-a)}}\frac{4}{x},
\end{aligned}$$

and therefore for $x_0 = x^{0.8}$ and some constant c_5,

$$
\begin{aligned}
&\sum_{a=1}^{x_0} \frac{\mathrm{poi}(x, x-a) - \mathrm{poi}(x, x+a)}{2} \Big[\log(x-a-\delta) + \log(x+a-\delta) \Big] \\
&\geq -\sum_{a=1}^{x_0} \frac{e^{-a}}{\sqrt{2\pi(x-a)}} \frac{2}{x} \log\left((x-\delta)^2 - a^2\right) \\
&\geq -\sum_{a=1}^{x_0} \frac{e^{-a}}{\sqrt{2\pi(x-a)}} \frac{2}{x} \log x^2 \\
&\geq -\frac{\sum_{a=1}^{\infty} e^{-a}}{\sqrt{2\pi(x-x_0)}} \frac{4}{x} \log x \\
&\geq -\frac{4 \log x}{x\sqrt{2\pi(x-x_0)}} \\
&\geq -\frac{c_5}{x}.
\end{aligned} \tag{10}
$$

Selecting $c > c_1 + c_5$ leads to the Lemma. It can be shown that the lemma is valid for $x < 100$ by plotting the function. ☐

A.3 Proof of Lemma 9

Proof.

$$
\begin{aligned}
&\sum_{a=0}^{x_0} \Big[\mathrm{poi}(x, x-a) + \mathrm{poi}(x, x+a) \Big] \left(\frac{a^2}{x^2}\right) \\
&\leq \sum_{a=0}^{x_0} \frac{2a^2}{x^2} \mathrm{poi}(x, x-a) \\
&= \sum_{a=0}^{x_0} \frac{2a^2}{x^2} \frac{e^{-x} x^{x-a}}{(x-a)!} \\
&\overset{(a)}{\leq} \sum_{a=0}^{x_0} \frac{2a^2}{x^2} \Big[\frac{e^{-x+x-a} x^{x-a}}{(x-a)^{x-a}\sqrt{2\pi(x-a)}} \Big] \\
&= \sum_{a=0}^{x_0} \frac{2a^2}{x^2} \Big[\frac{e^{-a}}{\sqrt{2\pi(x-a)}} \left(1 + \frac{a}{x-a}\right)^{x-a} \Big] \\
&= \sum_{a=0}^{x_0} \frac{2a^2}{x^2} \Big[\frac{e^{-a}}{\sqrt{2\pi(x-a)}} e^{(x-a)\ln(1+\frac{a}{x-a})} \Big] \\
&\overset{(b)}{\leq} \sum_{a=0}^{x_0} \frac{2a^2}{x^2} \Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}} \Big],
\end{aligned}
$$

where (a) is by Stirling's approximation and (b) is because $\ln(1+x) < x - \frac{x^2}{4}$ for $x < 1$. We can decompose the last summation to three different summations as

$$
\begin{aligned}
&\sum_{a=0}^{x_0} \frac{2a^2}{x^2} \Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}} \Big] \\
&= \sum_{a=0}^{\sqrt{x}} \frac{2a^2}{x^2} \Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}} \Big] + \sum_{a=\sqrt{x}+1}^{\sqrt{x}\ln x} \frac{2a^2}{x^2} \Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}} \Big] + \sum_{a=\sqrt{x}\ln x}^{x_0} \frac{2a^2}{x^2} \Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}} \Big]
\end{aligned} \tag{11}
$$

Now we bound each term in (11). For the first term and for some constant c_2:

$$\begin{aligned}\sum_{a=0}^{\sqrt{x}} \frac{2a^2}{x^2}\Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}}\Big] &\le 2\sqrt{x}\frac{x}{x^2}\frac{1}{\sqrt{2\pi(x-\sqrt{x})}} \\ &\le \sqrt{\frac{2}{\pi}}\frac{1}{x}\frac{1}{\sqrt{1-\frac{\sqrt{x}}{x}}} \\ &\le \sqrt{\frac{2}{\pi}}\frac{1}{x}(1+\frac{\sqrt{x}}{2x}) \le \frac{c_2}{x}.\end{aligned}$$

Also for the middle term in (11) and some constant c_4:

$$\begin{aligned}&\sum_{a=\sqrt{x}+1}^{\sqrt{x}\ln x} \frac{2a^2}{x^2}\Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}}\Big] \\ &\le \sqrt{\frac{2}{\pi}}\frac{1}{x^2}\frac{1}{\sqrt{x-\sqrt{x}\ln x}}\sum_{a=\sqrt{x}+1}^{\sqrt{x}\ln x} a^2 e^{-\frac{a^2}{4x}} \\ &\le \sqrt{\frac{2}{\pi}}\frac{1}{x^2}\frac{1}{\sqrt{x-\sqrt{x}\ln x}}\int_{\sqrt{x}}^{\sqrt{x}\ln x} a^2 e^{-\frac{a^2}{4x}}\,da \\ &= \sqrt{\frac{2}{\pi}}\frac{1}{x^2}\frac{1}{\sqrt{x-\sqrt{x}\ln x}} 2\Big[x\sqrt{x}e^{-\frac{1}{4}} - x\sqrt{x}e^{-\frac{x\ln^2 x}{4x}}\ln x + 2\sqrt{\pi}\left(\mathrm{Erf}(\frac{\ln x}{2}) - \mathrm{Erf}(\frac{1}{2})\right)\Big] \\ &\le \sqrt{\frac{2}{\pi}}\frac{1}{x^2\sqrt{x}}\frac{1}{\sqrt{1-\frac{\sqrt{x}\ln x}{x}}}(4x^{\frac{3}{2}}e^{-\frac{1}{4}}) \\ &\le \frac{c_4}{x}.\end{aligned}$$

Similarly for the third term in (11) and for some constant c_3, we can write

$$\begin{aligned}\sum_{a=\sqrt{x}\ln x}^{x_0} \frac{2a^2}{x^2}\Big[\frac{e^{-\frac{a^2}{4(x-a)}}}{\sqrt{2\pi(x-a)}}\Big] &\le (x_0-\sqrt{x}\ln x)\frac{2x_0^2}{x^2}\Big[\frac{e^{-\frac{(\sqrt{x}\ln x)^2}{4x}}}{\sqrt{2\pi(x-x_0)}}\Big] \\ &\le \sqrt{\frac{2}{\pi}}\frac{1}{x^2}\frac{x_0^3}{\sqrt{x-x_0}}e^{-\frac{\ln^2 x}{4}} \\ &= \sqrt{\frac{2}{\pi}}\frac{1}{x^2}\frac{x_0^3}{\sqrt{x-x_0}}\frac{1}{x^{\frac{\ln x}{4}}} \le \frac{c_3}{x}.\end{aligned}$$

Choosing $c_1 \ge c_2 + c_3 + c_4$ leads to the lemma. □

B Proofs of lower bound

In this part we provide the proofs of Theorem 2 as well as Corollaries 3 and 4. In order to lower bound the minimax risk of a given class $\mathcal{P}$, we can resort to two simplifications. First, we consider classes at a much a finer granularity than the $\mathcal{P}_d$ classes described in Section 5. In particular, let $\mathcal{P}_p$ be the *permutation class* of distributions consisting of a single distribution p and all of its permutations. Note that the multiset of probabilities is the same for all distributions in $\mathcal{P}_p$, and since the expected number of distinct categories only depends on the multiset ($d = \sum_j [1-(1-p_j)^n]$) it follows that $\mathcal{P}_p \subset \mathcal{P}_d$. [2]. To find a good lower bound for $\mathcal{P}_d$, we need a p that is "worst case" among all those who have the same value of d and then use the corresponding lower bound for $\mathcal{P}_p$. In what follows, we start by giving a lower bound for $\mathcal{P}_p$, and then specialize it for $\mathcal{P}_d$.

[2]We abuse notation by distinguishing the classes by the letter used, while at the same time using the letters to denote actual quantities. From the context we understand that d is the expected number of distinct categories for p, at the given n.

We also assume that an oracle specifies the *true* probability of all observed categories. With this side-information, the best estimator *has* to use the true probabilities for the observed categories. For the unobserved categories, it needs to redistribute all the missing mass (the total probability of unobserved categories). Since the multiset of probabilities is fixed and any permutation of the remaining categories is equally probable, by symmetry there is no advantage in favoring one over the other. Therefore the best oracle-aided estimator is uniquely specified: exact probabilities for seen categories and uniform redistribution of the missing mass (S_0) over the unobserved categories. This argument can be proven formally via the maximin trick: substitute the maximum with a mean against an arbitrary prior over p, at which point the optimal q is the posterior, and then optimize over priors. It then suffices to use the convexity of $p \log \frac{p}{q}$ with respect to q.

B.1 Proof of Theorem 2

Proof. Without loss of generality assume that $p_1 \geq p_2 \geq p_3 \geq \ldots \geq p_k$. Let $\gamma > 1$, we have:

$$
\begin{aligned}
r_n(\mathcal{P}_p) = \min_q \max_{p \in \mathcal{P}_p} \mathbb{E}\left[\sum_{j=1}^{k} p_j \log \frac{p_j}{q_j}\right] &\geq \mathbb{E}\left[\sum_{j=D+1}^{k} p_j \log \frac{p_j}{\frac{\sum_{j=D+1}^{k} p_j}{k-D}}\right] \\
&= \mathbb{E}\left[\sum_{j=D+1}^{k} p_j \log \frac{p_j(k-D)}{n\sum_{j=D+1}^{k} p_j} \;\middle|\; D < \gamma d\right] \Pr\left(D < \gamma d\right) \\
&+ \mathbb{E}\left[\sum_{j=D+1}^{k} p_j \log \frac{p_j(k-D)}{\sum_{j=D+1}^{k} p_j} \;\middle|\; D \geq \gamma d\right] \Pr\left(D \geq \gamma d\right) \\
&\overset{(a)}{\geq} \left(1 - \frac{1}{\gamma}\right) \sum_{j=\gamma d}^{k} p_j \log \frac{p_j(k-\gamma d)}{\sum_{j=\gamma d}^{k} p_j}
\end{aligned}
$$

where (a) is by the following arguments: By Markov's inequality we have $\Pr\left(D \geq \gamma d\right) \leq \frac{1}{\gamma}$. Also, $\sum_{j=D+1}^{k} p_j \log \frac{np_j(k-D)}{n\sum_{j=D+1}^{k} p_j}$ is positive and decreasing in D (in the extreme case, when $D = k$ is zero). Therefore,

$$
r_n(\mathcal{P}_p) \geq \left(1 - \frac{1}{\gamma}\right)\left(\sum_{j=\gamma d}^{k} p_j\right) \log \frac{k-\gamma d}{\sum_{j=\gamma d}^{k} p_j} + \sum_{i=\gamma d} p_j \log p_j
$$

This completes the proof. For any specific classes of distributions, we can find a lower bound by calculating d, $\sum_{j=\gamma d}^{k} p_j$, and $\sum_{j=\gamma d}^{k} p_j \log p_j$ for some $\gamma > 1$. □

B.2 Proof of Corollary 3

Proof. To use Theorem 2, we first calculate d, $\sum_{j>L} p_j$, and $\sum_{j>L} p_j \log p_j$ and then let $L = \gamma d$ for $\gamma = 2$.

$$
\begin{aligned}
\sum_{j=L+1}^{k} p_j &= \sum_{j=L+1}^{k} \frac{c}{j^\alpha} \\
&\overset{(a)}{\geq} \int_{L+1}^{k+1} \frac{c}{x^\alpha} dx \\
&= \frac{c}{\alpha - 1}\left[(L+1)^{1-\alpha} - (k+1)^{1-\alpha}\right],
\end{aligned}
$$

where (a) is by integration bound for monotone series. Similarly, we can show:

$$
\sum_{j=L+1}^{k} p_j \leq \int_{L}^{k} \frac{c}{x^\alpha} dx = \frac{c}{\alpha - 1}\left[L^{1-\alpha} - k^{1-\alpha}\right].
$$

For the last summation in the lower bound of Theorem 2 we have:

$$
\begin{aligned}
&\sum_{j=L+1}^{k} p_j \log p_j \\
&= \sum_{j=L+1}^{k} \frac{c}{j^\alpha} \log \frac{c}{j^\alpha} \\
&= c \sum_{j=L+1}^{k} \frac{1}{j^\alpha} \log \frac{1}{j^\alpha} + \log c \sum_{j=L+1}^{k} \frac{c}{j^\alpha} \\
&\overset{(a)}{\geq} c \int_{L+1}^{k+1} \frac{1}{j^\alpha} \log \frac{1}{j^\alpha} dj + \log c \sum_{j=L+1}^{k} p_j \\
&\overset{(b)}{=} \frac{c}{\alpha} \int_{(k+1)^{-\alpha}}^{(L+1)^{-\alpha}} x^{-\frac{1}{\alpha}} \log x \, dx + \log c \sum_{j=L+1}^{k} p_j \\
&\geq \frac{c}{\alpha-1} \left[x^{1-\frac{1}{\alpha}} \log x \right]_{(k+1)^{-\alpha}}^{(L+1)^{-\alpha}} - \frac{c}{\alpha-1} \int_{(k+1)^{-\alpha}}^{(L+1)^{-\alpha}} x^{-\frac{1}{\alpha}} + \frac{c \log c}{\alpha-1} \left[(L+1)^{1-\alpha} - (k+1)^{1-\alpha} \right]
\end{aligned}
$$

Using Theorem 2, if $k > \max\{n, \left(\frac{10}{9}^{\frac{1}{\alpha-1}}\right) n^{\frac{1}{\alpha}}\}$ we have,

$$
\begin{aligned}
&r_n(\mathcal{P}) \\
&\geq \frac{c}{\alpha-1}(L+1)^{1-\alpha} \log \frac{k-2d}{\frac{c}{\alpha-1}(L+1)^{1-\alpha}} + (L+1)^{1-\alpha} \left[\frac{c}{\alpha-1} \log (L+1)^{-\alpha} - \frac{c\alpha}{(\alpha-1)^2} + \frac{c \log c}{\alpha-1} \right] \\
&\geq \frac{c}{10(\alpha-1)}(L+1)^{1-\alpha} \log \frac{k-2d}{(L+1)} + \frac{c}{\alpha-1}(L+1)^{1-\alpha} \left[\frac{1}{10} \log(\alpha-1) - \frac{\alpha}{\alpha-1} \right]
\end{aligned}
$$

where (a) is by integration bound for monotone series, and (b) is by change of variable $x = \frac{1}{j^\alpha}$. Using Equation (3) , choosing $L = 2d$,

$$
r_n(\mathcal{P}) \geq \frac{2^{1-\alpha}c}{\alpha-1} d^{1-\alpha} \log \frac{k-2d}{2d} + \frac{2^{1-\alpha}c}{\alpha-1} d^{1-\alpha} \left[\log(\alpha-1) - \frac{\alpha}{\alpha-1} \right],
$$

and since for power-law distributions, d grows proportionally to $n^{\frac{1}{\alpha}}$, we can write

$$
r_n(\mathcal{P}) \geq c_1 \frac{d}{n} \log \frac{k-2d}{2d} (1 - o(1)),
$$

for some constants c_1 and c_2. To compare this with the upper bound in the proof of Theorem 1, note that we always have $\mathbb{E}[\Phi_1] \leq d$, but for power law distributions both expressions grow proportionally to $n^{\frac{1}{\alpha}}$ and furthermore $\mathbb{E}[\Phi_1]/d$ converges to a constant, $\frac{1}{\alpha}$. This shows that the upper and lower bounds for power-law distributions are tight in the first order term, when k is large. □

B.3 proof of Corollary 4

Proof. To use Theorem 2, we first calculate d, $\sum_{j>\gamma d} p_j$, and $\sum_{j>\gamma d} p_j \log p_j$. For the expected number of distinct categories,

$$
d = \sum_{j=1}^{k'} 1 - e^{-np_j} = k'(1 - e^{-\frac{n}{k'}}) = \frac{1-e^{-\rho}}{\rho} n.
$$

For the sum of probabilities of unobserved categories,

$$
\sum_{j>\gamma d} p_j = \frac{k' - \gamma d}{k'} = 1 - \frac{\gamma n (1-e^{-\rho})}{\rho k'} = 1 - \gamma(1 - e^{-\rho}),
$$

and for the last summation in (3),

$$\sum_{j=\gamma d+1}^{k} p_j \log p_j = \frac{k'-\gamma d}{k'} \log(\frac{1}{k'}) = \left(1-\gamma(1-e^{-\rho})\right) \log\left(\frac{\rho}{n}\right).$$

Therefore, by (3) we have $r_n(\mathcal{P}) \geq \left(1-\frac{1}{\gamma}\right)(1-\gamma(1-e^{-\rho})) \log \frac{k-\gamma d}{1-\gamma(1-e^{-\rho})} + (1-\gamma(1-e^{-\rho})) \log\left(\frac{\rho}{n}\right)$, which can also be written as $r_n(\mathcal{P}) \geq \left(1-\frac{1}{\gamma}\right) \frac{\rho\left(1-\gamma(1-e^{-\rho})\right)}{1-e^{-\rho}} \frac{d}{n} \log \frac{k-\gamma d}{d} + (1-\gamma(1-e^{-\rho})) \log \frac{1-e^{-\rho}}{(1-\gamma(1-e^{-\rho}))} + \frac{1}{\gamma}(1-\gamma(1-e^{-\rho})) \log \frac{1-e^{-\rho}}{d}$. Choosing $\gamma = 1.2$ and having $k > n^5$, the corollary follows for $\rho \leq 1.75$. □

C Proofs of Good–Turing and absolute-discount relationship

C.1 Proof of Lemma 5

Proof. For notational simplicity we define $C(\mu) := \frac{c^{\frac{1}{\alpha}} \Gamma(\mu-\frac{1}{\alpha})}{\alpha \mu!} n^{\frac{1}{\alpha}}$. Using Lemma 15,

$$\begin{aligned}
\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]} &\overset{(a)}{\leq} \frac{C(\mu+1)+\mathcal{O}(\mu^{-\frac{1}{2}})}{C(\mu)\left(1-\mathcal{O}(\mu^{-1}n^{-\frac{1}{\alpha}})\right)-\mathcal{O}(\mu^{-\frac{1}{2}})} \\
&\leq \frac{C(\mu+1)}{C(\mu)}\left(\frac{1+\mathcal{O}(\mu^{-\frac{1}{2}}C^{-1}(\mu+1))}{1-\mathcal{O}(\mu^{-1}n^{-\frac{1}{\alpha}})-\mathcal{O}(\mu^{-\frac{1}{2}}C^{-1}(\mu))}\right) \\
&\leq \frac{C(\mu+1)}{C(\mu)}\left(1+\mathcal{O}(\mu^{-\frac{1}{2}}C^{-1}(\mu+1))+\mathcal{O}(\mu^{-1}n^{-\frac{1}{\alpha}})\right) \\
&\overset{(b)}{\leq} \frac{C(\mu+1)}{C(\mu)}\left(1+\mathcal{O}(\mu^{-\frac{1}{2}+1+\frac{1}{\alpha}}n^{\frac{-1}{\alpha}})\right) \\
&\overset{(c)}{=} \frac{\mu-\frac{1}{\alpha}}{\mu+1}\left(1+\mathcal{O}(n^{\frac{-3}{2(2\alpha+1)}})\right)
\end{aligned}$$

The inequality in (a) and (b) are by Lemma 15 and (c) is by the fact that $\mu < n^{\frac{1}{2\alpha+1}}$. □

C.2 Proof of Theorem 6

Recall that S_μ denotes the total probability of symbols appearing μ times, and let $\hat{S}_\mu$ be the probability assigned to those symbols by an estimator. Note that, given the samples, we may think of S and $\hat{S}$ as legitimate probability distributions on the set $\mu = 0, 1, \cdots, n$. In [OS15], it was shown that the competitive loss of an estimator over a class defined by a single distribution p and its permutations can be bounded by:

$$\epsilon_n(\mathcal{P}_p, q) = r_n(p,q) - r_n(\mathcal{P}_p) \leq \mathbb{E}[\mathsf{KL}(S||\hat{S})].$$

This is well defined, since S and $\hat{S}$ only refer to the multiset probabilities, which stays invariant over all distributions in the class. Using this bound and the equivalence of Lemma 5, we can proceed with the proof. In the proof, we analyze the absolute-discount estimator with discount $\delta = \min\{\frac{\max\{\Phi_1,1\}}{D}, \delta_{\max}\}$.

Proof. We have:

$$\begin{aligned}
&\mathsf{KL}(S||\hat{S}) \\
&= \sum_{\mu=0}^{\infty} S_\mu \log \frac{S_\mu}{\hat{S}_\mu} \\
&\overset{(a)}{\leq} \sum_{\mu=0}^{\infty} \frac{(S_\mu - \hat{S}_\mu)^2}{\hat{S}_\mu} \\
&= \frac{(S_0 - \hat{S}_0)^2}{\hat{S}_0} + \sum_{\mu=1}^{\mu_0} \frac{(S_\mu - \hat{S}_\mu)^2}{\hat{S}_\mu} + \sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \hat{S}_\mu)^2}{\hat{S}_\mu} \\
&= \frac{(S_0 - \frac{D\delta}{n})^2}{\frac{D\delta}{n}} + \sum_{\mu=1}^{\mu_0} \frac{(S_\mu - \frac{\mu-\frac{1}{\alpha}}{n}\Phi_\mu + \frac{\mu-\frac{1}{\alpha}}{n}\Phi_\mu - \frac{\mu-\delta}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu} + \sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \frac{\mu-\delta}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu} \\
&\overset{(b)}{\leq} \frac{(S_0 - \frac{D\delta}{n})^2}{\frac{D\delta}{n}} + 2\sum_{\mu=1}^{\mu_0} \frac{(S_\mu - \frac{\mu-\frac{1}{\alpha}}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu} + 2\sum_{\mu=1}^{\mu_0} \frac{(\frac{\mu-\frac{1}{\alpha}}{n} - \frac{\mu-\delta}{n})^2\Phi_\mu^2}{\frac{\mu-\delta}{n}\Phi_\mu} + \sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \frac{\mu-\delta}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}
\end{aligned} \tag{12}$$

where (a) is by Lemma 14 and (b) is by $(a+b)^2 \leq 2a^2 + 2b^2$. We choose $\mu_0 = n^{\frac{1}{2\alpha+1}}$ and show the proof for the case when $n^{\frac{1}{2\alpha+1}} \geq 20 \log n$, namely, $\alpha \leq \frac{\log n}{2(\log\log n + \log 20)} - \frac{1}{2}$. For $\alpha > \frac{\log n}{2(\log\log n+\log 20)} - \frac{1}{2}$, the proof follows the same lines, but by a different choice of μ_0. Lemmas 10, 11, 12, and 13 bound each term in Equation (12) separately, and hence

$$\mathbb{E}[\mathsf{KL}(S||\hat{S})] = \mathcal{O}\left(\frac{1}{n^{\frac{2\alpha-1}{2\alpha+1}}}\right). \qquad \square$$

Lemma 10. *For a power-law distribution with exponent $\alpha > \alpha_0 > 1$, and the choice of $\delta = \min\{\frac{\max\{\Phi_1,1\}}{D}, \delta_{\max}\}$,*

$$\mathbb{E}\Big[\frac{(S_0 - \frac{D\delta}{n})^2}{\frac{D\delta}{n}}\Big] = \mathcal{O}\left(\frac{1}{n}\right).$$

Proof. To upper bound the first term of the KL loss in Equation (12), namely the loss of proposed estimator for the missing mass, let A be the event $(1-t)\mathbb{E}[\Phi_1] \leq \Phi_1 \leq (1+t)\mathbb{E}[\Phi_1]$ and $\frac{1-t}{1+t}\frac{\mathbb{E}[\Phi_1]}{d} \leq$

$\frac{\Phi_1}{D} \leq \frac{1+t}{1-t}\frac{\mathbb{E}[\Phi_1]}{d}$ for some $0 < t < 1$,

$$\begin{aligned}
\mathbb{E}\Big[\frac{(S_0-\frac{D\delta}{n})^2}{\frac{D\delta}{n}}\Big] &= \mathbb{E}\Big[\frac{(S_0-\frac{D\delta}{n})^2}{\frac{D\delta}{n}}\ \Big|\ A\Big]\Pr(A) + \mathbb{E}\Big[\frac{(S_0-\frac{D\delta}{n})^2}{\frac{D\delta}{n}}\ \Big|\ A^c\Big]\Pr(A^c) \\
&\overset{(a)}{\leq} \mathbb{E}\Big[\frac{(S_0-\frac{D\delta}{n})^2}{\frac{D\delta}{n}}\ \Big|\ A\Big]\Pr(A) + 4\exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right)n^2 \\
&\overset{(b)}{\leq} \mathbb{E}\Big[\frac{2(S_0-\frac{\mathbb{E}[\Phi_1]}{n})^2 + 2(\frac{\mathbb{E}[\Phi_1]}{n}-\frac{\Phi_1}{n})^2}{\frac{\Phi_1}{n}}\ \Big|\ A\Big]\Pr(A) + 4n^2\exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right) \\
&\overset{(c)}{\leq} \frac{\mathbb{E}\Big[2(S_0-\frac{\mathbb{E}[\Phi_1]}{n})^2 + 2(\frac{\mathbb{E}[\Phi_1]}{n}-\frac{\Phi_1}{n})^2\ \Big|\ A\Big]\Pr(A)}{\frac{(1-t)\mathbb{E}[\Phi_1]}{n}} + 4n^2\exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right) \\
&\overset{(d)}{\leq} \frac{2\mathrm{Var}(S_0)+\frac{2}{n^2}\mathrm{Var}(\Phi_1)}{\frac{\mathbb{E}[\Phi_1]}{2n}} + o\left(\frac{1}{n}\right) \\
&\overset{(e)}{\leq} \frac{\frac{4}{n^2}\mathbb{E}[\Phi_2]+\frac{2}{n^2}\mathbb{E}[\Phi_1]}{\frac{\mathbb{E}[\Phi_1]}{2n}} + o\left(\frac{1}{n}\right) \\
&= \mathcal{O}\left(\frac{1}{n}\right),
\end{aligned}$$

where (b) is by choosing t such that $\frac{1+t}{1-t}\frac{1}{\alpha_0} < \delta_{\max}$ and therefore conditioned on A, $\delta = \frac{\Phi_1}{D}$. Also, (c) is by concentration of Φ_1 (see Lemma 16), (d) is by choosing $t = n^{-\frac{1}{4\alpha}}$, and (e) is because $\mathrm{Var}(\Phi_1) \leq \mathbb{E}[\Phi_1]$ and $\mathrm{Var}(S_0) \leq \frac{2}{n^2}\mathbb{E}[\Phi_2]$ (see Lemma 21). ☐

Lemma 11. *For a power-law distribution with exponent α and choice of $\mu_0 = \mathcal{O}(n^{\frac{1}{2\alpha+1}})$,*

$$\mathbb{E}\Big[\sum_{\mu=1}^{\mu_0}\frac{(S_\mu-\frac{\mu-\frac{1}{\alpha}}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\Big] = \mathcal{O}\left(n^{\frac{1-2\alpha}{2\alpha+1}}\right)$$

Proof. Using Lemma 5 and $(a+b)^2 \leq 2a^2+2b^2$, we bound the second term in (12):

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{\mu=1}^{\mu_0}\frac{(S_\mu-\frac{\mu-\frac{1}{\alpha}}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\Big] \\
&\leq 2\mathbb{E}\left[\sum_{\mu=1}^{\mu_0}\frac{(\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu\mathcal{O}(n^{\frac{-3}{2(2\alpha+1)}}))^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] + 2\mathbb{E}\left[\sum_{\mu=1}^{\mu_0}\frac{(S_\mu-\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right]
\end{aligned} \tag{13}$$

For the first term in right hand side of Equation (13),

$$\begin{aligned}
\mathbb{E}\left[\sum_{\mu=1}^{\mu_0}\frac{(\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu\mathcal{O}(n^{\frac{-3}{2(2\alpha+1)}}))^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] &\leq n^{-1-\frac{3}{2\alpha+1}}\sum_{\mu=1}^{\mu_0}\frac{(\mu+1)^2\left(\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\right)^2\mathbb{E}[\Phi_\mu]}{\mu-\delta} \\
&\leq \frac{4}{1-\delta_{\max}}n^{\frac{1}{\alpha}-1-\frac{3}{2\alpha+1}}\sum_{\mu=1}^{\mu_0}\mu^{-\frac{1}{\alpha}} \\
&\leq \frac{4}{1-\delta_{\max}}n^{\frac{1}{\alpha}-1-\frac{3}{2\alpha+1}}\mu_0^{1-\frac{1}{\alpha}} = \mathcal{O}\left(\frac{1}{n}\right),
\end{aligned}$$

where the last line is by choosing $\mu_0 = n^{\frac{1}{2\alpha+1}}$. For the second term in Equation 13, using $(a+b)^2 \leq 2a^2 + 2b^2$ we have,

$$\left(S_\mu - \frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu\right)^2 = \left[S_\mu - \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}] + \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}] - \frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu\right]^2$$
$$\leq 2\left(S_\mu - \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}]\right)^2 + 2\left(\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu - \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}]\right)^2,$$

and therefore:

$$\mathbb{E}\left[\sum_{\mu=1}^{\mu_0}\frac{(S_\mu - \frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right]$$
$$= \mathbb{E}\left[\sum_{\mu=1}^{\mu_0}\frac{(S_\mu - \frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\;\Big|\;\Phi_\mu \geq \frac{\mathbb{E}[\Phi_\mu]}{2}\right]\Pr\left(\Phi_\mu \geq \frac{\mathbb{E}[\Phi_\mu]}{2}\right) +$$
$$\mathbb{E}\left[\sum_{\mu=1}^{\mu_0}\frac{(S_\mu - \frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\;\Big|\;\Phi_\mu < \frac{\mathbb{E}[\Phi_\mu]}{2}\right]\Pr\left(\Phi_\mu < \frac{\mathbb{E}[\Phi_\mu]}{2}\right)$$
$$\overset{(a)}{\leq} \mathbb{E}\left[\sum_{\mu=1}^{\mu_0}\frac{(S_\mu - \frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\;\Big|\;\Phi_\mu \geq \frac{\mathbb{E}[\Phi_\mu]}{2}\right]\Pr\left(\Phi_\mu \geq \frac{\mathbb{E}[\Phi_\mu]}{2}\right) + n^2\exp\left(-\frac{1}{6\mu_0}\left(\frac{n}{\mu_0}\right)^{\frac{1}{\alpha}}\right)$$
$$\leq \left(\sum_{\mu=1}^{\mu_0}\frac{\mathbb{E}[(S_\mu - \frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu)^2]}{\frac{\mu-\delta}{2n}\mathbb{E}[\Phi_\mu]}\;\Big|\;\Phi_\mu \geq \frac{\mathbb{E}[\Phi_\mu]}{2}\right)\Pr\left(\Phi_\mu \geq \frac{\mathbb{E}[\Phi_\mu]}{2}\right) + n^2\exp\left(-\frac{1}{6\mu_0}\left(\frac{n}{\mu_0}\right)^{\frac{1}{\alpha}}\right)$$
$$\overset{(b)}{\leq} \sum_{\mu=1}^{\mu_0}\frac{\mathbb{E}\left[2\left(S_\mu - \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}]\right)^2 + 2\left(\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\Phi_\mu - \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}]\right)^2\right]}{\frac{\mu-\delta}{2n}\mathbb{E}[\Phi_\mu]} + n^2\exp\left(-\frac{1}{6\mu_0}\left(\frac{n}{\mu_0}\right)^{\frac{1}{\alpha}}\right)$$
$$\leq \sum_{\mu=1}^{\mu_0}\frac{2\mathrm{Var}(S_\mu) + 2\left(\frac{\mu+1}{n}\frac{\mathbb{E}[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}\right)^2\mathrm{Var}(\Phi_\mu)}{\frac{\mu-\delta_{\max}}{2n}\mathbb{E}[\Phi_\mu]} + n^2\exp\left(-\frac{1}{6\mu_0}\left(\frac{n}{\mu_0}\right)^{\frac{1}{\alpha}}\right)$$
$$\overset{(d)}{\leq} \sum_{\mu=1}^{\mu_0}\frac{2\frac{(\mu+2)^2}{n^2}\mathbb{E}[\Phi_{\mu+2}] + 2\frac{(\mu+1)^2}{n^2}\frac{\mathbb{E}^2[\Phi_{\mu+1}]}{\mathbb{E}[\Phi_\mu]}}{\frac{\mu-\delta_{\max}}{2n}\mathbb{E}[\Phi_\mu]} + n^2\exp\left(-\frac{1}{6\mu_0}\left(\frac{n}{\mu_0}\right)^{\frac{1}{\alpha}}\right)$$
$$\overset{(e)}{\leq} \sum_{\mu=1}^{\mu_0}\frac{3}{1-\delta_{\max}}\frac{\mu}{n} + o\left(\frac{1}{n}\right)$$
$$\overset{(f)}{\leq} \frac{3}{n(1-\delta_{\max})}\left(\frac{\mu_0^2}{2} + 2\mu_0\right) + o\left(\frac{1}{n}\right) = \mathcal{O}\left(n^{\frac{1-2\alpha}{2\alpha+1}}\right).$$

Note that (a) follows from Lemma 16, (b) from $(x+y)^2 \leq 2x^2 + 2y^2$, and (c) from $\mathbb{E}[S_\mu] = \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}]$ and $\delta < \delta_{\max}$. Also, (d) results from $\mathrm{Var}[\Phi_\mu] \leq \mathbb{E}[\Phi_\mu]$ and $\mathrm{Var}[S_\mu] \leq$

$\frac{(\mu+2)^2}{n^2}\mathbb{E}[\Phi_{\mu+2}]$ (see Lemma 21), (e) is by Lemma 15, and (f) results from the choice of $\mu_0 = n^{\frac{1}{2\alpha+1}}$. $\square$

Lemma 12. *For a power-law distribution with exponent α and the choice of $\mu_0 = n^{\frac{1}{2\alpha+1}}$,*

$$\mathbb{E}\left[\sum_{\mu=1}^{\mu_0} \frac{\left(\frac{\mu-\frac{1}{\alpha}}{n} - \frac{\mu-\delta}{n}\right)^2 \Phi_\mu^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] = \mathcal{O}\left(\frac{n^{\frac{1}{2\alpha}}}{n}\right).$$

Proof.

$$\mathbb{E}\left[\sum_{\mu=1}^{\mu_0} \frac{\left(\frac{\mu-\frac{1}{\alpha}}{n} - \frac{\mu-\delta}{n}\right)^2 \Phi_\mu^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] \leq \frac{1}{n}\sum_{\mu=1}^{\mu_0} \frac{\mathbb{E}\left[(\frac{1}{\alpha}-\delta)^2\Phi_\mu\right]}{\mu-\delta_{\max}}.$$

Similar to the proof of Lemma 10, let A be the event $(1-t)\mathbb{E}[\Phi_1] \leq \Phi_1 \leq (1+t)\mathbb{E}[\Phi_1]$ and $\frac{1-t}{1+t}\frac{\mathbb{E}[\Phi_1]}{d} \leq \frac{\Phi_1}{D} \leq \frac{1+t}{1-t}\frac{\mathbb{E}[\Phi_1]}{d}$ for some $0<t<1$. Thus,

$$\begin{aligned}
\mathbb{E}\left[(\frac{1}{\alpha}-\delta)^2\Phi_\mu\right] &= \mathbb{E}\left[(\frac{1}{\alpha}-\delta)^2\Phi_\mu \,\middle|\, A\right]\Pr(A) + \mathbb{E}\left[(\frac{1}{\alpha}-\delta)^2\Phi_\mu \,\middle|\, A^c\right]\Pr(A^c) \\
&\leq t^2 \Pr(A)\,\mathbb{E}\left[\Phi_\mu \,\middle|\, A\right] + \Pr(A^c)\,\mathbb{E}\left[\Phi_\mu \,\middle|\, A^c\right] \\
&\leq n^{-\frac{1}{2\alpha}}\mathbb{E}\left[\Phi_\mu\right]
\end{aligned}$$

where the last line is by choosing $t = n^{-\frac{1}{4\alpha}}$ and using Lemma 22. Hence, we have

$$\mathbb{E}\left[\sum_{\mu=1}^{\mu_0} \frac{\left(\frac{\mu-\frac{1}{\alpha}}{n} - \frac{\mu-\delta}{n}\right)^2 \Phi_\mu^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] \leq \frac{n^{-\frac{1}{2\alpha}}}{n}\sum_{\mu=1}^{\mu_0}\frac{\mathbb{E}\left[\Phi_\mu\right]}{\mu-\delta_{\max}} = \mathcal{O}\left(\frac{n^{\frac{1}{2\alpha}}}{n}\right),$$

where the constant depends on $\delta_{\max}$ and therefore on α_0. $\square$

Lemma 13. *For a power-law distribution with exponent α, and $\mu_0 = n^{\frac{1-2\alpha}{2\alpha+1}}$,*

$$\mathbb{E}\left[\sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \frac{\mu-\delta}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] = \mathcal{O}\left(n^{\frac{1-2\alpha}{2\alpha+1}}\right)$$

Proof. For the last part in Equation 12:

$$\mathbb{E}\left[\sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \frac{\mu-\delta}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] \leq \mathbb{E}\left[\sum_{\mu=\mu_0+1}^{\infty} \frac{2(S_\mu - \frac{\mu}{n}\Phi_\mu)^2 + 2(\frac{\delta\Phi_\mu}{n})^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right].$$

We bound both terms in the above expression separately. For the second term, we have:

$$\mathbb{E}\left[\sum_{\mu=\mu_0+1}^{\infty} \frac{(\frac{\delta\Phi_\mu}{n})^2}{\frac{\mu-\delta}{n}\Phi_\mu}\right] \leq \frac{1}{n}\sum_{\mu=\mu_0+1}^{\infty}\frac{\mathbb{E}[\Phi_\mu]}{\mu-\delta_{\max}} \leq \frac{2c^{\frac{1}{\alpha}}\Gamma\left(1-\frac{1}{\alpha}\right)}{\alpha n}\frac{1}{\mu_0}\left(\frac{n}{\mu_0}\right)^{\frac{1}{\alpha}} + \frac{2}{n\sqrt{\mu_0}} = \mathcal{O}\left(n^{-\frac{2\alpha}{2\alpha+1}}\right),$$

and for the first part, we have:

$$
\begin{aligned}
&\sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \frac{\mu}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu} \\
&\overset{(a)}{\leq} \sum_{\mu=\mu_0+1}^{\infty} \sum_{x} \mathbb{1}_x^\mu \frac{(p_x - \frac{\mu}{n})^2}{\frac{\mu-1}{n}} \\
&\leq 2 \sum_{x\colon np_x \geq \frac{\mu_0}{2}} \sum_{\mu=\mu_0+1}^{\infty} \mathbb{1}_x^\mu \frac{(p_x - \frac{\mu}{n})^2}{\frac{\mu}{n}} + 2 \sum_{x\colon np_x < \frac{\mu_0}{2}} \sum_{\mu=\mu_0+1}^{\infty} \mathbb{1}_x^\mu \frac{(p_x - \frac{\mu}{n})^2}{\frac{\mu}{n}} \\
&\leq 2 \sum_{x\colon np_x \geq \frac{\mu_0}{2}} \sum_{\mu=1}^{\infty} \mathbb{1}_x^\mu \frac{(p_x - \frac{\mu}{n})^2}{\frac{\mu}{n}} + 2 \sum_{x\colon np_x < \frac{\mu_0}{2}} \frac{n}{\mu_0} \mathbb{1}_x^{>\mu_0},
\end{aligned}
$$

where (a) follows from $(\sum_{i=1}^n a_i)^2 \leq n(\sum_{i=1}^n a_i^2)$. Taking expectations of both sides:

$$
\begin{aligned}
\mathbb{E}\Big[\sum_{\mu=\mu_0+1}^{\infty} \frac{(S_\mu - \frac{\mu}{n}\Phi_\mu)^2}{\frac{\mu-\delta}{n}\Phi_\mu}\Big] &\overset{(a)}{\leq} 2\left(\frac{2nc}{\mu_0}\right)^{\frac{1}{\alpha}} \frac{1}{n}\mathbb{E}\Big[\frac{(np_x)^2 - 2\mu np_x + \mu^2}{\mu}\Big] + 2 \sum_{x\colon np_x < \frac{\mu_0}{2}} \frac{n}{\mu_0}\mathbb{E}[\mathbb{1}_x^{>\mu_0}] \\
&\overset{(b)}{\leq} 2\left(\frac{2nc}{\mu_0}\right)^{\frac{1}{\alpha}} \frac{3}{n} + 2 \sum_{x\colon np_x < \frac{\mu_0}{2}} \frac{n}{\mu_0}\mathbb{E}[\mathbb{1}_x^{>\mu_0}] \\
&\overset{(c)}{\leq} \left(\frac{2nc}{\mu_0}\right)^{\frac{1}{\alpha}} \frac{6}{n} + 2 \sum_{x\colon np_x < \frac{\mu_0}{2}} \frac{n}{\mu_0} \exp\left(\mu_0 - np_x - \mu_0 \ln\left(\frac{\mu_0}{np_x}\right)\right) \\
&\overset{(d)}{\leq} \left(\frac{2nc}{\mu_0}\right)^{\frac{1}{\alpha}} \frac{6}{n} + 2 \sum_{x\colon np_x < \frac{\mu_0}{2}} \frac{n}{\mu_0} \exp\left(\frac{np_x - \mu_0}{3}\right) \\
&\overset{(e)}{\leq} \left(\frac{2nc}{\mu_0}\right)^{\frac{1}{\alpha}} \frac{4}{n} + 2e^{-\frac{\mu_0}{6}} \left(\frac{n}{\mu_0}\right)^2 \\
&= \mathcal{O}(n^{\frac{1-2\alpha}{2\alpha+1}}),
\end{aligned}
$$

where (a) is by bounding the number of elements with probability greater than $\mu_0/2n$, (b) follows from the fact that $\mathbb{E}[\frac{1}{\mu}]$ when μ is a Poisson distribution with mean λ, is bounded by $\frac{1}{\lambda} + \frac{3}{\lambda^2}$(note that $\mu = 0$ is excluded). Also, (c) follows from Lemma 18, (d) follows from $3(x - 1 - x\ln x) \leq 1 - x$ for $x > 2$, and (e) is by convexity of the exponential term in p_x and the fact that a convex function is maximized at the boundaries. □

D Tools

This section provides a summary of tools used in the proofs throughout the paper.

Lemma 14. *For two distributions p and q,*

$$\mathsf{KL}(p||q) := \sum_i p_i \log \frac{p_i}{q_i} \leq \sum_i \frac{(p_i - q_i)^2}{q_i}$$

Lemma 15. *For a power-law distribution with power $\alpha > \alpha_0 > 1$ and normalization factor c, for $\mu \geq 1$*

$$\mathbb{E}[\Phi_\mu] \leq \frac{c^{\frac{1}{\alpha}} \Gamma\left(\mu - \frac{1}{\alpha}\right)}{\alpha \mu!} n^{\frac{1}{\alpha}} + \frac{1}{\sqrt{2\pi\mu}} \leq \frac{c^{\frac{1}{\alpha}} \Gamma(1 - \frac{1}{\alpha})}{\mu\alpha} \left(\frac{n}{\mu}\right)^{\frac{1}{\alpha}} + \frac{1}{\sqrt{2\pi\mu}}.$$

Also, for $1 \leq \mu < n^{\frac{1}{\alpha+1}}$, and $k > n^{\frac{1}{\alpha-1}}$,

$$\mathbb{E}[\Phi_\mu] \geq \frac{c^{\frac{1}{\alpha}} \Gamma\left(\mu - \frac{1}{\alpha}\right)}{\alpha \mu!} n^{\frac{1}{\alpha}} - \frac{1}{\sqrt{2\pi\mu}}.$$

Proof. For the upper bound on the expected number of elements that appeared μ times:

$$\begin{aligned}
\mathbb{E}[\Phi_\mu] &= \mathbb{E}\left[\sum_{x=1}^{k} \mathbb{1}_x^\mu\right] \\
&= \sum_{x=1}^{k} e^{-np_x} \frac{(np_x)^\mu}{\mu!} \\
&= \sum_{x=1}^{k} e^{-\frac{nc}{x^\alpha}} \frac{(\frac{nc}{x^\alpha})^\mu}{\mu!} \\
&\overset{(a)}{\leq} \int_1^k e^{-\frac{nc}{x^\alpha}} \frac{(\frac{nc}{x^\alpha})^\mu}{\mu!} dx + \max_x \{e^{-\frac{nc}{x^\alpha}} \frac{(\frac{nc}{x^\alpha})^\mu}{\mu!}\} \\
&\overset{(b)}{=} \frac{(nc)^{\frac{1}{\alpha}}}{\alpha \mu!} \int_{\frac{nc}{k^\alpha}}^{nc} e^{-y} y^{\mu - 1 - \frac{1}{\alpha}} dy + \max_x \{e^{-\frac{nc}{x^\alpha}} \frac{(\frac{nc}{x^\alpha})^\mu}{\mu!}\} \\
&\overset{(c)}{\leq} \frac{(nc)^{\frac{1}{\alpha}}}{\alpha \mu!} \left[\Gamma\left(\mu - \frac{1}{\alpha}, \frac{nc}{k^\alpha}\right) - \Gamma\left(\mu - \frac{1}{\alpha}, nc\right)\right] + \frac{1}{\sqrt{2\pi\mu}} \\
&= \frac{c^{\frac{1}{\alpha}} \Gamma\left(\mu - \frac{1}{\alpha}\right)}{\alpha \mu!} n^{\frac{1}{\alpha}} + \frac{1}{\sqrt{2\pi\mu}},
\end{aligned} \tag{14}$$

where (a) is followed by the integration bound for a uni-modal series, (b) is by changing of variables $\frac{nc}{x^\alpha} = y$. Also (c) is by the definition of Gamma function and the fact that $e^{-t}t^\mu$ is maximized at $t = \mu$ followed by Stirling's approximation. By further simplifying the Gamma function term:

$$\begin{aligned}
\frac{\Gamma(\mu - \frac{1}{\alpha})}{\mu!} &= \frac{(\mu - 1 - \frac{1}{\alpha})(\mu - 2 - \frac{1}{\alpha}) \ldots (1 - \frac{1}{\alpha}) \Gamma(1 - \frac{1}{\alpha})}{\mu!} \\
&= \frac{1}{\mu} \prod_{j=1}^{\mu-1} \left(1 - \frac{1}{j\alpha}\right) \Gamma\left(1 - \frac{1}{\alpha}\right) \\
&= \frac{1}{\mu} \exp\left(\sum_{j=1}^{\mu-1} \log\left(1 - \frac{1}{j\alpha}\right)\right) \Gamma\left(1 - \frac{1}{\alpha}\right) \\
&\overset{(a)}{\leq} \frac{1}{\mu} \exp\left(-\frac{1}{\alpha} \sum_{j=1}^{\mu-1} \frac{1}{j}\right) \Gamma\left(1 - \frac{1}{\alpha}\right) \\
&\overset{(b)}{\leq} \frac{1}{\mu} \mu^{-\frac{1}{\alpha}} \Gamma\left(1 - \frac{1}{\alpha}\right).
\end{aligned}$$

where (a) is by $\log(1-x) \leq -x$ for $0 < x < 1$, and (b) is because $\sum_{j=1}^{t} \frac{1}{j} \geq \log(t+1)$. Similarly for the lower bound we have:

$$\begin{aligned}
\mathbb{E}[\Phi_\mu] &= \mathbb{E}\left[\sum_{x=1}^{k} \mathbb{1}_x^\mu\right] \\
&= \sum_{x=1}^{k} e^{-np_x} \frac{(np_x)^\mu}{\mu!} \\
&= \sum_{x=1}^{k} e^{-\frac{nc}{x^\alpha}} \frac{(\frac{nc}{x^\alpha})^\mu}{\mu!} \\
&\overset{(a)}{\geq} \int_1^k e^{-\frac{nc}{x^\alpha}} \frac{(\frac{nc}{x^\alpha})^\mu}{\mu!} dx - \max_x \{e^{-\frac{nc}{x^\alpha}} \frac{(\frac{nc}{x^\alpha})^\mu}{\mu!}\} \\
&\overset{(b)}{=} \frac{(nc)^{\frac{1}{\alpha}}}{\alpha\mu!} \int_{\frac{nc}{k^\alpha}}^{nc} e^{-y} y^{\mu-1-\frac{1}{\alpha}} dy - \frac{1}{\sqrt{2\pi\mu}} \\
&\overset{(c)}{=} \frac{(nc)^{\frac{1}{\alpha}}}{\alpha\mu!} \left[\Gamma\left(\mu - \frac{1}{\alpha}\right) - \gamma\left(\mu - \frac{1}{\alpha}, \frac{nc}{k^\alpha}\right) - \Gamma\left(\mu - \frac{1}{\alpha}, nc\right)\right] - \frac{1}{\sqrt{2\pi\mu}} \\
&\overset{(d)}{\geq} \frac{c^{\frac{1}{\alpha}} \Gamma\left(\mu - \frac{1}{\alpha}\right)}{\alpha\mu!} n^{\frac{1}{\alpha}} \left(1 - \mathcal{O}\left(\mu^{-1} n^{-\frac{1}{\alpha}}\right)\right) - \frac{1}{\sqrt{2\pi\mu}},
\end{aligned} \tag{15}$$

By Lemma 20 we have $\gamma\left(\mu - \frac{1}{\alpha}, \frac{nc}{k^{\alpha}}\right) \leq \frac{1}{\mu+1-\frac{1}{\alpha}} \left(1 + (\mu - \frac{1}{\alpha}) e^{-\frac{nc}{k^\alpha}}\right) \frac{\left(\frac{nc}{k^\alpha}\right)^{\mu-\frac{1}{\alpha}}}{\mu - \frac{1}{\alpha}}$ which leads to $\gamma\left(\mu - \frac{1}{\alpha}, \frac{nc}{k^{\alpha}}\right) = \mathcal{O}(\mu^{-1} n^{-\frac{1}{\alpha}})$ when $k > n^{\frac{1}{\alpha-1}}$. Lemma 19 implies that $\Gamma\left(\mu - \frac{1}{\alpha}, nc\right) \leq B(nc)^{\mu-\frac{1}{\alpha}} e^{-nc}$ for some constant B and for every $1 < \mu < n^{\frac{1}{\alpha+1}}$. This and the recursion $\Gamma(s+1, t) = s\Gamma(s, x) + x^s e^{-x}$ lead to $\Gamma\left(\mu - \frac{1}{\alpha}, nc\right) = \mathcal{O}(\frac{1}{n})$ for $1 \leq \mu < n^{\frac{1}{\alpha+1}}$ and therefore (d). Also, (a) is followed by the integration bound for a uni-modal series, (b) is by changing of variables $\frac{nc}{x^\alpha} = y$, and (c) is by the definition of Gamma function and the fact that $e^{-t}t^\mu$ is maximized at $t = \mu$ followed by Stirling's approximation. $\square$

Lemma 16. *For a power-law distribution with power α and $\mu < n^{\frac{1}{\alpha+1}}$,*

$$\Pr\left[\Phi_\mu < \frac{\mathbb{E}[\Phi_\mu]}{2}\right] \leq \exp\left(-\frac{1}{6\mu}\left(\frac{n}{\mu}\right)^{\frac{1}{\alpha}}\right)$$

Proof. $\Phi_\mu = \sum_x \mathbb{1}_x^\mu$, and therefore is a sum of independent random variables $\mathbb{1}_x^\mu$. By Bernstein's inequality

$$\Pr\left[\left|\Phi_\mu - \mathbb{E}[\Phi_\mu]\right| > t\right] \leq 2\exp\left(-\frac{t^2/2}{\mathrm{Var}(\Phi_\mu) + t/3}\right)$$

Substituting $\mathbb{E}[\Phi_\mu]$ from Lemma 15 and using $\mathrm{Var}(\Phi_\mu) \leq \mathbb{E}[\Phi_\mu]$, for $t = \frac{\mathbb{E}[\Phi_\mu]}{2}$ we have the lemma. $\square$

Lemma 17 ([OD12]). *Let D be the number of distinct categories and $d = \mathbb{E}[D]$. Also let $v = \mathbb{E}[\Phi_1]$ be the expected number of categories that appeared once. Then,*

$$\Pr[D < d - \sqrt{2vs}] \leq e^{-s}$$

Lemma 18. *Let $X \sim POI(x)$, then for $x_0 > 0$, $\Pr(X \geq x + x_0) \leq e^{x_0 - (x+x_0)\ln(1+\frac{x_0}{x})}$, and Also* $\Pr(X \leq x - x_0) \leq e^{x_0 - (x+x_0)\ln(1+\frac{x_0}{x})}$.

Proof. Chernoff bound suggests that for every $t > 0$

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[e^{tX}]}{e^{t \cdot a}},$$

and similarly for every $t < 0$,

$$\Pr(X \leq a) \leq \frac{\mathbb{E}[e^{tX}]}{e^{t \cdot a}}.$$

Moment generating function, $\mathbb{E}[e^{tX}]$ for X distributed according to POI(x) is $e^{x(e^t-1)}$. Therefore,

$$\begin{aligned}\Pr(X \geq a) &\leq \inf_{t>0} \frac{e^{x(e^t-1)}}{e^{t \cdot a}} \\ &= \inf_{t>0} e^{x(e^t-1)-t \cdot a} \\ &= e^{a-x-a \ln \frac{a}{x}}.\end{aligned}$$

Substituting a by $x + x_0$ leads to the lemma. □

Lemma 19 ([NP00]). *For $a > 1$, $B > 1$, and $x > \frac{B}{B-1}(a-1)$, we have*

$$x^{a-1}e^{-x} < |\Gamma(a, x)| < Bx^{a-1}e^{-x}.$$

Lemma 20 (Theorem 4.1 in [Neu13]). *For $a > 0$ and $x > 0$, we have*

$$\exp\left(-\frac{ax}{a+1}\right) \leq \frac{a}{x^a}\gamma(a, x) \leq \frac{1}{a+1}(1 + ae^{-x}).$$

Lemma 21. *For every distribution, $\mu \geq 1$, and in the presence of Poisson sampling,*

$$\mathrm{Var}(\Phi_\mu) \leq \mathbb{E}[\Phi_\mu], \quad \mathrm{Var}(S_\mu) \leq \frac{(\mu+1)(\mu+2)}{n^2}\mathbb{E}[\Phi_{\mu+2}], \quad \mathbb{E}[S_\mu] = \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}]$$

Proof. We use the property of the Poisson sampling that the counts are independent. For the variance of Φ_μ we can write:

$$\begin{aligned}\mathrm{Var}(\Phi_\mu) &= \mathrm{Var}\left(\sum_j \mathbb{1}_j^\mu\right) \\ &= \sum_j \mathrm{Var}(\mathbb{1}_j^\mu) \\ &\leq \sum_j \mathbb{E}[\mathbb{1}_j^\mu] \\ &= \mathbb{E}[\Phi_\mu].\end{aligned}$$

Also, for the expected value of the sum of probabilities that appeared μ times, we have:

$$\begin{aligned}\mathbb{E}[S_\mu] &= \mathbb{E}\Big[\sum_j p_j \mathbb{1}_j^\mu\Big] \\ &= \sum_j p_j e^{-np_j}\frac{(np_j)^\mu}{\mu!} \\ &= \frac{\mu+1}{n}\sum_j e^{-np_j}\frac{(np_j)^{\mu+1}}{(\mu+1)!} \\ &= \frac{\mu+1}{n}\sum_j \mathbb{E}[\mathbb{1}_j^{\mu+1}] \\ &= \frac{\mu+1}{n}\mathbb{E}[\Phi_{\mu+1}],\end{aligned}$$

and for their variance, we can write:

$$\begin{aligned}\mathrm{Var}[S_\mu] &= \mathrm{Var}\Big[\sum_j p_j \mathbb{1}_j^\mu\Big] \\ &= \sum_j p_j^2 \mathrm{Var}(\mathbb{1}_j^\mu) \\ &= \sum_j p_j^2 \mathbb{E}[\mathbb{1}_j^\mu] \\ &= \sum_j p_j^2 e^{-np_j}\frac{(np_j)^\mu}{\mu!} \\ &= \frac{(\mu+1)(\mu+2)}{n^2}\sum_j e^{-np_j}\frac{(np_j)^{\mu+2}}{(\mu+2)!} \\ &= \frac{(\mu+1)(\mu+2)}{n^2}\sum_j \mathbb{E}[\mathbb{1}_j^{\mu+2}] \\ &= \frac{(\mu+1)(\mu+2)}{n}\mathbb{E}[\Phi_{\mu+2}].\end{aligned}$$

□

Lemma 22. *Let Φ_1 be number of categories appeared once. Also let D be the number of distinct categories observed and $d = \mathbb{E}[D]$, then for $0 < t < 1$,*

$$\Pr\left(\Big|\frac{\Phi_1}{D} - \frac{\mathbb{E}[\Phi_1]}{d}\Big| > \frac{2t}{1-t}\right) \le 4\exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right)$$

Proof. Using Lemma 16 we have

$$\begin{aligned}\Pr\left(\Big|\frac{\Phi_1}{\mathbb{E}[\Phi_1]} - 1\Big| > t\right) &= \Pr\left(\Big|\Phi_1 - \mathbb{E}[\Phi_1]\Big| > t\mathbb{E}[\Phi_1]\right) \\ &\le \exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right)\end{aligned}$$

Similarly for number of distinct elements we have

$$\begin{aligned}\Pr\left(\Big|\frac{D}{d} - 1\Big| > t\right) &= \Pr\left(|D-d| > td\right) \\ &\overset{(a)}{\le} \Pr\left(|D-d| > t\mathbb{E}[\Phi_1]\right) \\ &\overset{(b)}{\le} \exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right)\end{aligned}$$

where (a) is because $\mathbb{E}[\Phi_1] \le d$ and (b) is because $\mathrm{Var}(D) = \mathbb{E}[\Phi_1]$. Hence, with probability at least $1-4\exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right)$ we have $1-t \le \frac{\Phi_1}{\mathbb{E}[\Phi_1]} \le 1+t$ and $1-t \le \frac{D}{d} \le 1+t$. With probability $\ge 1-4\exp\left(-\frac{t^2\mathbb{E}[\Phi_1]}{2(1+t/3)}\right)$, $\frac{1-t}{1+t} \le \frac{\Phi_1}{D}\frac{d}{\mathbb{E}[\Phi_1]} \le \frac{1+t}{1-t}$, namely $\left|\frac{\Phi_1}{D} - \frac{\mathbb{E}[\Phi_1]}{d}\right| \le \max\left(\frac{1+t}{1-t} - 1, 1 - \frac{1-t}{1+t}\right) = \frac{2t}{1-t}$. □