# Competitive Distribution Estimation: Why is Good-Turing Good

**Alon Orlitsky**
UC San Diego
alon@ucsd.edu

**Ananda Theertha Suresh**
UC San Diego
asuresh@ucsd.edu

## Abstract

Estimating distributions over large alphabets is a fundamental machine-learning tenet. Yet no method is known to estimate all distributions well. For example, add-constant estimators are nearly min-max optimal but often perform poorly in practice, and practical estimators such as absolute discounting, Jelinek-Mercer, and Good-Turing are not known to be near optimal for essentially any distribution.

We describe the first universally near-optimal probability estimators. For every discrete distribution, they are provably nearly the best in the following two competitive ways. First they estimate every distribution nearly as well as the best estimator designed with prior knowledge of the distribution up to a permutation. Second, they estimate every distribution nearly as well as the best estimator designed with prior knowledge of the exact distribution, but as all natural estimators, restricted to assign the same probability to all symbols appearing the same number of times.

Specifically, for distributions over $k$ symbols and $n$ samples, we show that for both comparisons, a simple variant of Good-Turing estimator is always within KL divergence of $(3+o_n(1))/n^{1/3}$ from the best estimator, and that a more involved estimator is within $\tilde{\mathcal{O}}_n(\min(k/n, 1/\sqrt{n}))$. Conversely, we show that any estimator must have a KL divergence at least $\tilde{\Omega}_n(\min(k/n, 1/n^{2/3}))$ over the best estimator for the first comparison, and at least $\tilde{\Omega}_n(\min(k/n, 1/\sqrt{n}))$ for the second.

## 1 Introduction

### 1.1 Background

Many learning applications, ranging from language-processing staples such as speech recognition and machine translation to biological studies in virology and bioinformatics, call for estimating large discrete distributions from their samples. Probability estimation over large alphabets has therefore long been the subject of extensive research, both by practitioners deriving practical estimators [1, 2], and by theorists searching for optimal estimators [3].

Yet even after all this work, provably-optimal estimators remain elusive. The add-constant estimators frequently analyzed by theoreticians are nearly min-max optimal, yet perform poorly for many practical distributions, while common practical estimators, such as absolute discounting [4], Jelinek-Mercer [5], and Good-Turing [6], are not well understood and lack provable performance guarantees.

To understand the terminology and approach a solution we need a few definitions. The performance of an estimator $q$ for an underlying distribution $p$ is typically evaluated in terms of the *Kullback-*

*Leibler (KL) divergence* [7],

$$D(p||q) \stackrel{\text{def}}{=} \sum_x p_x \log \frac{p_x}{q_x},$$

reflecting the expected increase in the ambiguity about the outcome of $p$ when it is approximated by $q$. KL divergence is also the increase in the number of bits over the entropy that $q$ uses to compress the output of $p$, and is also the *log-loss* of estimating $p$ by $q$. It is therefore of interest to construct estimators that approximate a large class of distributions to within small KL divergence. We now describe one of the problem's simplest formulations.

### 1.2 Min-max loss

A distribution *estimator* over a support set $\mathcal{X}$ associates with any observed sample sequence $x^* \in \mathcal{X}^*$ a distribution $q(x^*)$ over $\mathcal{X}$. Given $n$ samples $X^n \stackrel{\text{def}}{=} X_1, X_2, \ldots, X_n$, generated independently according to a distribution $p$ over $\mathcal{X}$, the expected KL loss of $q$ is

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

Let $\mathcal{P}$ be a known collection of distributions over a discrete set $\mathcal{X}$. The worst-case loss of an estimator $q$ over all distributions in $\mathcal{P}$ is

$$r_n(q,\mathcal{P}) \stackrel{\text{def}}{=} \max_{p\in\mathcal{P}} r_n(q,p), \tag{1}$$

and the lowest worst-case loss for $\mathcal{P}$, achieved by the best estimator, is the min-max loss

$$r_n(\mathcal{P}) \stackrel{\text{def}}{=} \min_q r_n(q,\mathcal{P}) = \min_q \max_{p\in\mathcal{P}} r_n(q,p). \tag{2}$$

Min-max performance can be viewed as regret relative to an oracle that knows the underlying distribution. Hence from here on we refer to it as *regret*.

The most natural and important collection of distributions, and the one we study here, is the set of all discrete distributions over an alphabet of some size $k$, which without loss of generality we assume to be $[k] = \{1, 2, \ldots k\}$. Hence the set of all distributions is the *simplex* in $k$ dimensions, $\Delta_k \stackrel{\text{def}}{=} \{(p_1, \ldots, p_k) : p_i \geq 0 \text{ and } \sum p_i = 1\}$. Following [8], researchers have studied $r_n(\Delta_k)$ and related quantities, for example see [9]. We outline some of the results derived.

### 1.3 Add-constant estimators

The *add-$\beta$* estimator assigns to a symbol that appeared $t$ times a probability proportional to $t+\beta$. For example, if three coin tosses yield one heads and two tails, the add-$1/2$ estimator assigns probability $1.5/(1.5+2.5) = 3/8$ to heads, and $2.5/(1.5+2.5) = 5/8$ to tails. [10] showed that as for every $k$, as $n \to \infty$, an estimator related to add-3/4 is near optimal and achieves

$$r_n(\Delta_k) = \frac{k-1}{2n} \cdot (1 + o(1)). \tag{3}$$

The more challenging, and practical, regime is where the sample size $n$ is not overwhelmingly larger than the alphabet size $k$. For example in English text processing, we need to estimate the distribution of words following a context. But the number of times a context appears in a corpus may not be much larger than the vocabulary size. Several results are known for other regimes as well. When the sample size $n$ is linear in the alphabet size $k$, $r_n(\Delta_k)$ can be shown to be a constant, and [3] showed that as $k/n \to \infty$, add-constant estimators achieve the optimal

$$r_n(\Delta_k) = \log \frac{k}{n} \cdot (1 + o(1)), \tag{4}$$

While add-constant estimators are nearly min-max optimal, the distributions attaining the min-max regret are near uniform. In practice, large-alphabet distributions are rarely uniform, and instead, tend to follow a power-law. For these distributions, add-constant estimators under-perform the estimators described in the next subsection.

### 1.4 Practical estimators

For real applications, practitioners tend to use more sophisticated estimators, with better empirical performance. These include the Jelinek-Mercer estimator that cross-validates the sample to find the best fit for the observed data. Or the absolute-discounting estimators that rather than add a positive constant to each count, do the *opposite*, and subtract a positive constant.

Perhaps the most popular and enduring have been the *Good-Turing* estimator [6] and some of its variations. Let $n_x \stackrel{\text{def}}{=} n_x(x^n)$ be the number of times a symbol $x$ appears in $x^n$ and let $\varphi_t \stackrel{\text{def}}{=} \varphi_t(x^n)$ be the number of symbols appearing $t$ times in $x^n$. The basic Good-Turing estimator posits that if $n_x = t$,

$$q_x(x^n) = \frac{\varphi_{t+1}}{\varphi_t} \cdot \frac{t+1}{n},$$

surprisingly relating the probability of an element not just to the number of times it was observed, but also to the number other elements appearing as many, and one more, times. It is easy to see that this basic version of the estimator may not work well, as for example it assigns any element appearing $\geq n/2$ times 0 probability. Hence in practice the estimator is modified, for example, using empirical frequency to elements appearing many times.

The Good-Turing Estimator was published in 1953, and quickly adapted for language-modeling use, but for half a century no proofs of its performance were known. Following [11], several papers, e.g., [12, 13], showed that Good-Turing variants estimate the combined probability of symbols appearing any given number of times with accuracy that does not depend on the alphabet size, and [14] showed that a different variation of Good-Turing similarly estimates the probabilities of each previously-observed symbol, and all unseen symbols combined.

However, these results do not explain why Good-Turing estimators work well for the actual probability estimation problem, that of estimating the probability of each element, not of the combination of elements appearing a certain number of times. To define and derive uniformly-optimal estimators, we take a different, competitive, approach.

## 2 Competitive optimality

### 2.1 Overview

To evaluate an estimator, we compare its performance to the best possible performance of two estimators designed with some prior knowledge of the underlying distribution. The first estimator is designed with knowledge of the underlying distribution up to a permutation of the probabilities, namely knowledge of the probability multiset, e.g., $\{.5, .3, .2\}$, but not of the association between probabilities and symbols. The second estimator is designed with exact knowledge of the distribution, but like all *natural estimators*, forced to assign the same probabilities to symbols appearing the same number of times. For example, upon observing the sample $a, b, c, a, b, d, e$, the estimator must assign the same probability to $a$ and $b$, and the same probability to $c$, $d$, and $e$.

These estimators cannot be implemented in practice as in reality we do not have prior knowledge of the estimated distribution. But the prior information is chosen to allow us to determine the best performance of any estimator designed with that information, which in turn is better than the performance of any *data-driven* estimator designed without prior information. We then show that certain variations of the Good-Turing estimators, designed without any prior knowledge, approach the performance of both prior-knowledge estimators for every underlying distribution.

### 2.2 Competing with near full information

We first define the performance of an *oracle-aided* estimator, designed with some knowledge of the underlying distribution. Suppose that the estimator is designed with the aid of an oracle that knows the value of $f(p)$ for some given function $f$ over the class $\Delta_k$ of distributions.

The function $f$ partitions $\Delta_k$ into subsets, each corresponding to one possible value of $f$. We denote the subsets by $P$, and the partition by $\mathbb{P}$, and as before, denote the individual distributions by $p$. Then the oracle knows the unique partition part $P$ such that $p \in P \in \mathbb{P}$. For example, if $f(p)$ is

the multiset of $p$, then each subset $P$ corresponds to set of distributions with the same probability multiset, and the oracle knows the multiset of probabilities.

For every partition part $P \in \mathbb{P}$, an estimator $q$ incurs the worst-case regret in (1),

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

The oracle, knowing the unique partition part $P$, incurs the least worst-case regret (2),

$$r_n(P) = \min_q r_n(q, P).$$

The *competitive regret* of $q$ over the oracle, for all distributions in $P$ is

$$r_n(q, P) - r_n(P),$$

the competitive regret over all partition parts and all distributions in each is

$$r_n^{\mathbb{P}}(q, \Delta_k) \stackrel{\text{def}}{=} \max_{P \in \mathbb{P}} \left( r_n(q, P) - r_n(P) \right),$$

and the best possible competitive regret is

$$r_n^{\mathbb{P}}(\Delta_k) \stackrel{\text{def}}{=} \min_q r_n^{\mathbb{P}}(q, \Delta_k).$$

Consolidating the intermediate definitions,

$$r_n^{\mathbb{P}}(\Delta_k) = \min_q \max_{P \in \mathbb{P}} \left( \max_{p \in P} r_n(q, p) - r_n(P) \right).$$

Namely, an oracle-aided estimator who knows the partition part incurs a worst-case regret $r_n(P)$ over each part $P$, and the competitive regret $r_n^{\mathbb{P}}(\Delta_k)$ of data-driven estimators is the least overall increase in the part-wise regret due to not knowing $P$. In Appendix A.1, we give few examples of such partitions.

A partition $\mathbb{P}'$ *refines* a partition $\mathbb{P}$ if every part in $\mathbb{P}$ is partitioned by some parts in $\mathbb{P}'$. For example $\{\{a, b\}, \{c\}, \{d, e\}\}$ refines $\{\{a, b, c\}, \{d, e\}\}$. In Appendix A.2, we show that if $\mathbb{P}'$ refines $\mathbb{P}$ then for every $q$

$$r_n^{\mathbb{P}'}(q, \Delta_k) \geq r_n^{\mathbb{P}}(q, \Delta_k). \tag{5}$$

Considering the collection $\Delta_k$ of all distributions over $[k]$, it follows that as we start with single-part partition $\{\Delta_k\}$ and keep refining it till the oracle knows $p$, the competitive regret of estimators will increase from 0 to $r_n(q, \Delta_k)$. A natural question is therefore how much information can the oracle have and still keep the competitive regret low? We show that the oracle can know the distribution exactly up to permutation, and still the regret will be very small.

Two distributions $p$ and $p'$ *permutation equivalent* if for some permutation $\sigma$ of $[k]$,

$$p'_{\sigma(i)} = p_i,$$

for all $1 \leq i \leq k$. For example, $(0.5, 0.3, 0.2)$ and $(0.3, 0.5, 0.2)$ are permutation equivalent. Permutation equivalence is clearly an equivalence relation, and hence partitions the collection of distributions over $[k]$ into equivalence classes. Let $\mathbb{P}_\sigma$ be the corresponding partition. We construct estimators $q$ that uniformly bound $r_n^{\mathbb{P}_\sigma}(q, \Delta_k)$, thus the same estimator uniformly bounds $r_n^{\mathbb{P}}(q, \Delta_k)$ for any coarser partition of $\Delta_k$, such as partitions into classes of distributions with the same support size, or entropy. Note that the partition $\mathbb{P}_\sigma$ corresponds to knowing the underlying distribution up to permutation, hence $r_n^{\mathbb{P}_\sigma}(\Delta_k)$ is the additional KL loss compared to an estimator designed with knowledge of the underlying distribution up to permutation.

This notion of competitiveness has appeared in several contexts. In data compression it is called *twice-redundancy* [15, 16, 17, 18], while in statistics it is often called *adaptive* or *local min-max* [19, 20, 21, 22, 23], and recently in property testing it is referred as competitive [24, 25, 26] or *instance-by-instance* [27]. Subsequent to this work, [28] studied competitive estimation in $\ell_1$ distance, however their regret is poly$(1/\log n)$, compared to our $\tilde{\mathcal{O}}(1/\sqrt{n})$.

### 2.3 Competing with natural estimators

Our second comparison is with an estimator designed with exact knowledge of $p$, but forced to be *natural*, namely, to assign the same probability to all symbols appearing the same number of times in the sample. For example, for the observed sample $a, b, c, a, b, d, e$, the same probability must be assigned to $a$ and $b$, and the same probability to $c$, $d$, and $e$. Since data-driven estimators derive all their knowledge of the distribution from the data, we expect them to be natural.

We compare the regret of data-driven estimators to that of *natural oracle-aided* estimators. Let $\mathcal{Q}^{\text{nat}}$ be the set of all natural estimators. For a distribution $p$, the lowest regret of a natural estimator, designed with prior knowledge of $p$ is

$$r_n^{\text{nat}}(p) \stackrel{\text{def}}{=} \min_{q \in \mathcal{Q}^{\text{nat}}} r_n(q, p),$$

and the regret of an estimator $q$ relative to the least-regret natural-estimator is

$$r_n^{\text{nat}}(q, p) = r_n(q, p) - r_n^{\text{nat}}(p).$$

Thus the regret of an estimator $q$ over all distributions in $\Delta_k$ is

$$r_n^{\text{nat}}(q, \Delta_k) = \max_{p \in \Delta_k} r_n^{\text{nat}}(q, p),$$

and the best possible competitive regret is $r_n^{\text{nat}}(\Delta_k) = \min_q r_n^{\text{nat}}(q, \Delta_k)$.

In the next section we state the results, showing in particular that $r_n^{\text{nat}}(\Delta_k)$ is uniformly bounded. In Section 5, we outline the proofs, and in Section 4 we describe experiments comparing the performance of competitive estimators to that of min-max motivated estimators.

## 3 Results

Good-Turing estimators are often used in conjunction with empirical frequency, where Good-Turing estimates low probabilities and empirical frequency estimates large probabilities. We first show that even this simple Good-Turing version, defined in Appendix C and denoted $q'$, is uniformly optimal for all distributions. For simplicity we prove the result when the number of samples is $n' \sim \text{poi}(n)$, a Poisson random variable with mean $n$. Let $r^{\mathbb{P}_\sigma}_{\text{poi}(n)}(q', \Delta_k)$ and $r^{\text{nat}}_{\text{poi}(n)}(q', \Delta_k)$ be the regrets in this sampling process. A similar result holds with exactly $n$ samples, but the proof is more involved as the multiplicities are dependent.

**Theorem 1** (Appendix C)**.** *For any $k$ and $n$,*

$$r^{\mathbb{P}_\sigma}_{poi(n)}(q', \Delta_k) \le r^{nat}_{poi(n)}(q', \Delta_k) \le \frac{3 + o_n(1)}{n^{1/3}}.$$

Furthermore, a lower bound in [13] shows that this bound is optimal up to logarithmic factors.

A more complex variant of Good-Turing, denoted $q''$, was proposed in [13]. We show that its regret diminishes uniformly in both the partial-information and natural-estimator formulations.

**Theorem 2** (Section 5)**.** *For any $k$ and $n$,*

$$r^{\mathbb{P}_\sigma}_n(q'', \Delta_k) \le r^{nat}_n(q'', \Delta_k) \le \tilde{\mathcal{O}}_n\left(\min\left(\frac{1}{\sqrt{n}}, \frac{k}{n}\right)\right).$$

Where $\tilde{\mathcal{O}}_n$, and below also $\tilde{\Omega}_n$, hide multiplicative logarithmic factors in $n$. Lemma 6 in Section 5 and a lower bound in [13] can be combined to prove a matching lower bound on the competitive regret of any estimator for the second formulation,

$$r^{\text{nat}}_n(\Delta_k) \ge \tilde{\Omega}_n\left(\min\left(\frac{1}{\sqrt{n}}, \frac{k}{n}\right)\right).$$

Hence $q''$ has near-optimal competitive regret relative to natural estimators.

Fano's inequality usually yields lower bounds on KL loss, not regret. By carefully constructing distribution classes, we lower bound the competitive regret relative to the oracle-aided estimators.

**Theorem 3** (Appendix D)**.** *For any $k$ and $n$,*

$$r^{\mathbb{P}_\sigma}_n(\Delta_k) \ge \tilde{\Omega}_n\left(\min\left(\frac{1}{n^{2/3}}, \frac{k}{n}\right)\right).$$

### 3.1 Illustration and implications

Figure 1 demonstrates some of the results. The horizontal axis reflects the set $\Delta_k$ of distributions illustrated on one dimension. The vertical axis indicates the KL loss, or absolute regret, for clarity, shown for $k \gg n$. The blue line is the previously-known min-max upper bound on the regret, which by (4) is very high for this regime, $\log(k/n)$. The red line is the regret of the estimator designed with prior knowledge of the probability multiset. Observe that while for some probability multisets the regret approaches the $\log(k/n)$ min-max upper bound, for other probability multisets it is much lower, and for some, such as uniform over 1 or over k symbols, where the probability multiset determines the distribution it is even 0. For many practically relevant distributions, such as power-law distributions and sparse distributions, the regret is small compared to $\log(k/n)$. The green line is an upper bound on the absolute regret of the data-driven estimator $q''$. By Theorem 2, it is always at most $1/\sqrt{n}$ larger than the red line. It follows that for many distributions, possibly for distributions with more structure, such as those occurring in nature, the regret of $q''$ is significantly smaller than the pessimistic min-max bound implies.

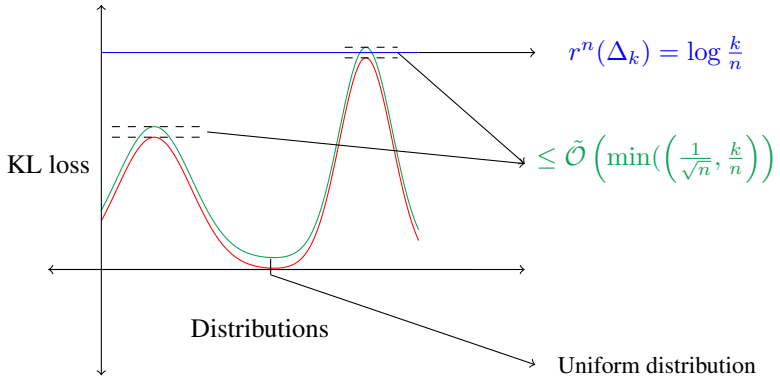



Figure 1: Qualitative behavior of the KL loss as a function of distributions in different formulations

We observe a few consequences of these results.

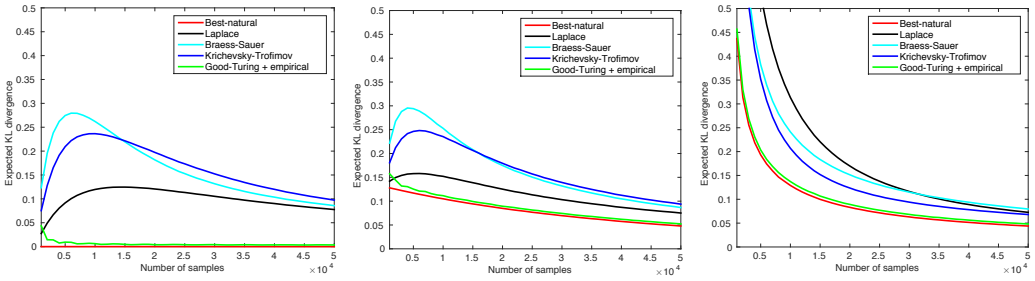
- Theorems 1 and 2 establish two uniformly-optimal estimators $q'$ and $q''$. Their relative regrets diminish to zero at least as fast as $1/n^{1/3}$, and $1/\sqrt{n}$ respectively, independent of how large the alphabet size $k$ is.
- Although the results are for relative regret, as shown in Figure 1, they lead to estimator with smaller absolute regret, namely, the expected KL divergence.
- The same regret upper bounds hold for all coarser partitions of $\Delta_k$ i.e., where instead of knowing the multiset, the oracle knows some property of multiset such as entropy.

## 4 Experiments

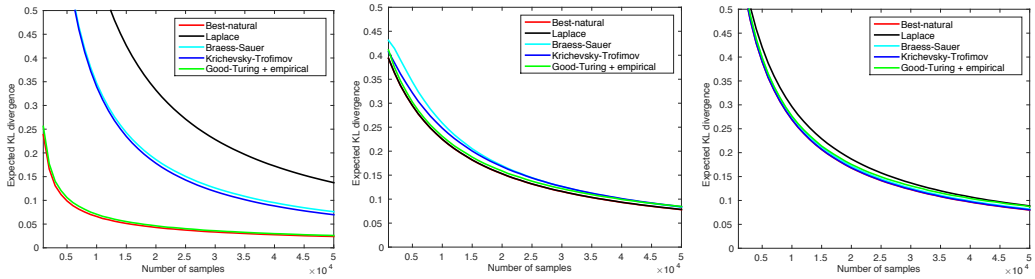
Recall that for a sequence $x^n$, $n_x$ denotes the number of times a symbol $x$ appears and $\varphi_t$ denotes the number of symbols appearing $t$ times. For small values of $n$ and $k$, the estimator proposed in [13] simplifies to a combination of Good-Turing and empirical estimators. By [13, Lemmas 10 and 11], for symbols appearing $t$ times, if $\varphi_{t+1} \geq \tilde{\Omega}(t)$, then the Good-Turing estimate is close to the underlying total probability mass, otherwise the empirical estimate is closer. Hence, for a symbol appearing $t$ times, if $\varphi_{t+1} \geq t$ we use the Good-Turing estimator, otherwise we use the empirical estimator. If $n_x = t$,

$$q_x = \begin{cases} \frac{t}{N} & \text{if } t > \varphi_{t+1}, \\ \frac{\varphi_{t+1}+1}{\varphi_t} \cdot \frac{t+1}{N} & \text{else,} \end{cases}$$

where $N$ is a normalization factor. Note that we have replaced $\varphi_{t+1}$ in the Good-Turing estimator by $\varphi_{t+1} + 1$ to ensure that every symbol is assigned a non-zero probability.




(a) Uniform (b) Step (c) Zipf with parameter 1




(d) Zipf with parameter 1.5 (e) Uniform prior (Dirichlet 1) (f) Dirichlet $1/2$ prior

Figure 2: Simulation results for support 10000, number of samples ranging from 1000 to 50000, averaged over 200 trials.

We compare the performance of this estimator to four estimators: three popular add-$\beta$ estimators and the optimal natural estimator. An add-beta estimator $\hat{S}$ has the form

$$q_x^{\hat{S}} = \frac{n_x + \beta_{n_x}^{\hat{S}}}{N(\hat{S})},$$

where $N(\hat{S})$ is a normalization factor to ensure that the probabilities add up to 1. The Laplace estimator, $\beta_t^L = 1 \,\forall\, t$, minimizes the expected loss when the underlying distribution is generated by a uniform prior over $\Delta_k$. The Krichevsky-Trofimov estimator, $\beta_t^{KT} = 1/2 \,\forall\, t$, is asymptotically min-max optimal for the cumulative regret, and minimizes the expected loss when the underlying distribution is generated according to a Dirichlet-$1/2$ prior. The Braess-Sauer estimator, $\beta_0^{BS} = 1/2, \beta_1^{BS} = 1, \beta_t^{BS} = 3/4 \;\forall\, t > 1$, is asymptotically min-max optimal for $r_n(\Delta_k)$. Finally, as shown in Lemma 10, the optimal estimator $q_x = \frac{S_{n_x}}{\varphi_{n_x}}$ achieves the lowest loss of any natural estimator designed with knowledge of the underlying distribution.

We compare the performance of the proposed estimator to that of the four estimators above. We consider six distributions: uniform distribution, step distribution with half the symbols having probability $1/2k$ and the other half have probability $3/2k$, Zipf distribution with parameter 1 ($p_i \propto i^{-1}$), Zipf distribution with parameter 1.5 ($p_i \propto i^{-1.5}$), a distribution generated by the uniform prior on $\Delta_k$, and a distribution generated from Dirichlet-$1/2$ prior. All distributions have support size $k = 10000$. $n$ ranges from 1000 to 50000 and the results are averaged over 200 trials.

Figure 2 shows the results. Observe that the proposed estimator performs similarly to the best natural estimator for all six distributions. It also significantly outperforms the other estimators for Zipf, uniform, and step distributions.

The performance of other estimators depends on the underlying distribution. For example, since Laplace is the optimal estimator when the underlying distribution is generated from the uniform prior, it performs well in Figure 2(e), however performs poorly on other distributions.

Furthermore, even though for distributions generated by Dirichlet priors, all the estimators have similar looking regrets (Figures 2(e), 2(f)), the proposed estimator performs better than estimators which are not designed specifically for that prior.

## 5 Proof sketch of Theorem 2

The proof consists of two parts. We first show that for every estimator $q$, $r_n^{\mathbb{P}_\sigma}(q, \Delta_k) \le r_n^{\text{nat}}(q, \Delta_k)$ and then upper bound $r_n^{\text{nat}}(q, \Delta_k)$ using results on combined probability mass.

**Lemma 4** (Appendix B.1)**.** *For every estimator $q$,*

$$r_n^{\mathbb{P}_\sigma}(q, \Delta_k) \le r_n^{nat}(q, \Delta_k).$$

The proof of the above lemma relies on showing that the optimal estimator for every class in $P \in \mathbb{P}_\sigma$ is natural.

### 5.1 Relation between $r_n^{\text{nat}}(q, \Delta_k)$ and combined probability estimation

We now relate the regret in estimating distribution to that of estimating the combined or total probability mass, defined as follows. Recall that $\varphi_t$ denotes the number of symbols appearing $t$ times. For a sequence $x^n$, let $S_t \stackrel{\text{def}}{=} S_t(x^n)$ denote the total probability of symbols appearing $t$ times. For notational convenience, we use $S_t$ to denote both $S_t(x^n)$ and $S_t(X^n)$ and the usage becomes clear in the context. Similar to KL divergence between distributions, we define KL divergence between $S$ and their estimates $\hat{S}$ as

$$D(S||\hat{S}) = \sum_{t=0}^{n} S_t \log \frac{S_t}{\hat{S}_t}.$$

Since the natural estimator assigns same probability to symbols that appear the same number of times, estimating probabilities is same as estimating the total probability of symbols appearing a given number of times. We formalize it in the next lemma.

**Lemma 5** (Appendix B.2)**.** *For a natural estimator $q$ let $\hat{S}_t(x^n) = \sum_{x:n_x=t} q_x(x^n)$, then*

$$r_n^{nat}(q, p) = \mathbb{E}[D(S||\hat{S})].$$

In Lemma 11(Appendix B.3), we show that there is a natural estimator that achieves $r_n^{\text{nat}}(\Delta_k)$. Taking maximum over all distributions $p$ and minimum over all estimators $q$ results in

**Lemma 6.** *For a natural estimator $q$ let $\hat{S}_t(x^n) = \sum_{x:n_x=t} q_x(x^n)$, then*

$$r_n^{nat}(q, \Delta_k) = \max_{p \in \Delta_k} \mathbb{E}[D(S||\hat{S})].$$

*Furthermore,*

$$r_n^{nat}(\Delta_k) = \min_{\hat{S}} \max_{p \in \Delta_k} \mathbb{E}[D(S||\hat{S})].$$

Thus finding the best competitive natural estimator is same as finding the best estimator for the combined probability mass $S$. [13] proposed an algorithm for estimating $S$ such that for all $k$ and for all $p \in \Delta_k$, with probability $\ge 1 - 1/n$ ,

$$D(S||\hat{S}) = \tilde{\mathcal{O}}_n\left(\frac{1}{\sqrt{n}}\right).$$

The result is stated in Theorem 2 of [13]. One can convert this result to a result on expectation easily using the property that their estimator is bounded below by $1/2n$ and show that

$$\max_{p \in \Delta_k} \mathbb{E}[D(S||\hat{S})] = \tilde{\mathcal{O}}_n\left(\frac{1}{\sqrt{n}}\right).$$

A slight modification of their proofs for Lemma 17 and Theorem 2 in their paper using $\sum_{t=1}^{n} \sqrt{\varphi_t} \le \sum_{t=1}^{n} \varphi_t \le k$ shows that their estimator $\hat{S}$ for the combined probability mass $S$ satisfies

$$\max_{p \in \Delta_k} \mathbb{E}[D(S||\hat{S})] = \tilde{\mathcal{O}}_n\left(\min\left(\frac{1}{\sqrt{n}}, \frac{k}{n}\right)\right).$$

The above equation together with Lemmas 4 and 6 results in Theorem 2.

## 6 Acknowledgements

We thank Jayadev Acharya, Moein Falahatgar, Paul Ginsparg, Ashkan Jafarpour, Mesrob Ohannessian, Venkatadheeraj Pichapati, Yihong Wu, and the anonymous reviewers for helpful comments.

## References


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

[2] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *ACL*, 1996.

[3] Liam Paninski. Variational minimax estimation of discrete distributions under KL loss. In *NIPS*, 2004.

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

[5] Fredrick Jelinek and Robert L. Mercer. Probability distribution estimation from sparse data. *IBM Tech. Disclosure Bull.*, 1984.

[6] Irving J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

[7] Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006.

[8] R. Krichevsky. *Universal Compression and Retrieval*. Dordrecht,The Netherlands: Kluwer, 1994.

[9] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *COLT*, 2015.

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

[11] David A. McAllester and Robert E. Schapire. On the convergence rate of Good-Turing estimators. In *COLT*, 2000.

[12] Evgeny Drukh and Yishay Mansour. Concentration bounds for unigrams language model. In *COLT*, 2004.

[13] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Optimal probability estimation with applications to prediction and classification. In *COLT*, 2013.

[14] Alon Orlitsky, Narayana P. Santhanam, and Junan Zhang. Always Good Turing: Asymptotically optimal probability estimation. In *FOCS*, 2003.

[15] Boris Yakovlevich Ryabko. Twice-universal coding. *Problemy Peredachi Informatsii*, 1984.

[16] Boris Yakovlevich Ryabko. Fast adaptive coding algorithm. *Problemy Peredachi Informatsii*, 26(4):24–37, 1990.

[17] Dominique Bontemps, Stéphane Boucheron, and Elisabeth Gassiat. About adaptive coding on countable alphabets. *IEEE Transactions on Information Theory*, 60(2):808–821, 2014.

[18] Stéphane Boucheron, Elisabeth Gassiat, and Mesrob I. Ohannessian. About adaptive coding on countable alphabets: Max-stable envelope classes. *CoRR*, abs/1402.6305, 2014.

[19] David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[20] Felix Abramovich, Yoav Benjamini, David L Donoho, and Iain M Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 2006.

[21] Peter J Bickel, Chris A Klaassen, YA'Acov Ritov, and Jon A Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

[22] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

[23] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2004.

[24] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Shengjun Pan. Competitive closeness testing. *COLT*, 2011.

[25] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Theertha Suresh. Competitive classification and closeness testing. In *COLT*, 2012.

[26] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. A competitive test for uniformity of monotone distributions. In *AISTATS*, 2013.

[27] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.

[28] Gregory Valiant and Paul Valiant. Instance optimal learning. *CoRR*, abs/1504.05321, 2015.

[29] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

# A Proofs for competitive formulation

## A.1 Examples of partitions

The following examples evaluate $r_n^{\mathbb{P}}(\Delta_k)$ for the two simplest partitions.

**Example 7.** *The* singleton partition *consists of $|\Delta_k|$ parts, each a single distribution in $\Delta_k$,*

$$\mathbb{P}_{|\Delta_k|} \stackrel{def}{=} \{\{p\} : p \in \Delta_k\}\,.$$

*An oracle-aided estimator that knows the part containing $p$ knows $p$. The competitive regret of data-driven estimators is therefore the min-max regret,*

$$\begin{aligned} r_n^{\mathbb{P}_{|\Delta_k|}}(\Delta_k) &= \min_q \max_{p\in\Delta_k} \left(r_n(q,\{p\}) - r_n(\{p\})\right) \\ &= \min_q \max_{p\in\Delta_k} r_n(q,p) \\ &= r_n(\Delta_k), \end{aligned}$$

*where the middle equality follows as $r_n(q,\{p\}) = r_n(q,p)$, and $r_n(\{p\}) = 0$.*

**Example 8.** *The* whole-collection *partition has only one part, the whole collection $\Delta_k$,*

$$\mathbb{P}_1 \stackrel{def}{=} \{\Delta_k\}\,.$$

*An estimator aided by an oracle that knows the part containing $p$ has no additional information, hence no advantage over a data-driven estimator, and the competitive regret is 0,*

$$\begin{aligned} r_n^{\mathbb{P}_1}(\Delta_k) &= \min_q \max_{P\in\{\Delta_k\}} \left(\max_{p\in P} r_n(q,p) - r_n(P)\right) \\ &= \min_q \left(\max_{p\in\Delta_k} r_n(q,p) - r_n(\Delta_k)\right) \\ &= \min_q \max_{p\in\Delta_k} \left(r_n(q,p)\right) - r_n(\Delta_k) \\ &= r_n(\Delta_k) - r_n(\Delta_k) \\ &= 0. \end{aligned}$$

The examples show that for the coarsest partition of $\Delta_k$, into a single part, the competitive regret is the lowest possible, 0, while for the finest partition, into singletons, the competitive regret is the highest possible, $r_n(\Delta_k)$.

## A.2 Proof of Equation (5)

The definition implies that if $P' \subseteq P$ then $r_n(P') \le r_n(P)$, for every distribution class $P$ and $P'$. Hence for every $q$,

$$\begin{aligned} r_n^{\mathbb{P}'}(q,\Delta_k) &= \max_{P'\in\mathbb{P}'} \left(r_n(q,P') - r_n(P')\right) \\ &= \max_{P\in\mathbb{P}} \max_{P\supseteq P'\in\mathbb{P}'} \left(r_n(q,P') - r_n(P')\right) \\ &\ge \max_{P\in\mathbb{P}} \max_{P\supseteq P'\in\mathbb{P}'} \left(r_n(q,P') - r_n(P)\right) \\ &= \max_{P\in\mathbb{P}} \big(\max_{P\supseteq P'\in\mathbb{P}'} r_n(q,P') - r_n(P)\big) \\ &= \max_{P\in\mathbb{P}} \left(r_n(q,P) - r_n(P)\right) \\ &= r_n^{\mathbb{P}}(q,\Delta_k). \end{aligned}$$

# B Upper bounds

For a distribution $p$ and sequence $x^n$, let $p(x^n)$ be the probability of observing $x^n$ under $p$. Recall that for a symbol $x$, we abbreviate $p(x)$ to be $p_x$.

### B.1 Proof of Lemma 4

The proof uses the following result.

**Lemma 9.** *For every class $P \in \mathbb{P}_\sigma$, $r_n(P) \geq \max_{p\in P} r_n^{nat}(p)$.*

*Proof.* We first show that there is an optimal estimator $q$ that is natural. In particular, let

$$q''_y(x^n) = \frac{\sum_{p\in P} p(x^n y)}{\sum_{p'\in P} p'(x^n)}.$$

We show that $q''_y(x^n)$ is an optimal estimator for $P$. Since $q''_y(x^n) = q''_{\sigma(y)}(\sigma(x^n))$ for any permutation $\sigma$, the estimator achieves the same loss for every $p \in P$,

$$\max_{p\in P} r_n(q'', p) = \frac{1}{k!}\sum_{p\in P} r_n(q'', p'). \tag{6}$$

For any estimator $q$,

$$\begin{aligned}
\max_{p\in P}\mathbb{E}[D(p||q)] &\overset{(a)}{\geq} \frac{1}{k!}\sum_{p\in P}\mathbb{E}_p[D(p||q)] \\
&\overset{(b)}{=} \frac{1}{k!}\sum_{p\in P}\sum_{x^n\in\mathcal{X}^n}\sum_{y\in\mathcal{X}} p(x^n y)\log\frac{1}{q_y(x^n)} - H(p) \\
&= \frac{1}{k!}\sum_{x^n\in\mathcal{X}^n}\sum_{y\in\mathcal{X}}\sum_{p\in P} p(x^n y)\log\frac{1}{q_y(x^n)} - H(p) \\
&\overset{(c)}{\geq} \frac{1}{k!}\sum_{x^n\in\mathcal{X}^n}\sum_{y\in\mathcal{X}}\sum_{p\in P} p(x^n y)\log\frac{\sum_{p'\in P} p'(x^n)}{\sum_{p''\in P} p''(x^n y)} - H(p) \\
&= \frac{1}{k!}\sum_{p\in P}\sum_{x^n\in\mathcal{X}^n}\sum_{y\in\mathcal{X}} p(x^n y)\log\frac{1}{q''_y(x^n)} - H(p) \\
&\overset{(d)}{=} \frac{1}{k!}\sum_{p\in P} r_n(q'', p).
\end{aligned}$$

$(a)$ follows from the fact that maximum is larger than the average. $(b)$ follows from the fact that every distribution in $P$ has the same entropy. Non-negativity of KL divergence implies $(c)$. All distributions in $P$ has the same entropy and hence $(d)$. Hence together with Equation (6)

$$\begin{aligned}
r_n(P) &= \min_q \max_{p\in P}\mathbb{E}[D(p||q)] \\
&\geq \frac{1}{k!}\sum_{p\in P} r_n(q'', p') \\
&= \max_{p\in P} r_n(q'', p).
\end{aligned}$$

Hence $q''$ is an optimal estimator. Recall that $n_y$ denote the number of times symbol $y$ appears in the sequence. $q''$ is natural as if $n_y = n_{y'}$, then $q''_y(x^n) = q''_{y'}(x^n)$. Since there is a natural estimator that achieves minimum in $r_n(P)$,

$$\begin{aligned}
r_n(P) &= \min_q \max_{p\in P}\mathbb{E}[D(p||q)] \\
&= \min_{q\in\mathcal{Q}^{\text{nat}}} \max_{p\in P}\mathbb{E}[D(p||q)] \\
&\geq \max_{p\in P} \min_{q\in\mathcal{Q}^{\text{nat}}}\mathbb{E}[D(p||q)] \\
&= \max_{p\in P} r_n^{\text{nat}}(p),
\end{aligned}$$

where the last inequality follows from the fact that min-max is bigger than max-min. □

We can now prove Lemma 4.

*Proof of Lemma 4.*

$$\begin{aligned} r_n^{\mathbb{P}_\sigma}(q,\Delta_k) &= \max_{P\in\mathbb{P}_\sigma}\left(\max_{p\in P}\mathbb{E}[D(p||q)] - r_n(P)\right) \\ &\overset{(a)}{\leq} \max_{P\in\mathbb{P}_\sigma}\left(\max_{p\in P}\mathbb{E}[D(p||q)] - \max_{p\in P} r_n^{\text{nat}}(p)\right) \\ &\overset{(b)}{\leq} \max_{P\in\mathbb{P}_\sigma}\max_{p\in P}\left(\mathbb{E}[D(p||q)] - r_n^{\text{nat}}(p)\right) \\ &= \max_{p\in\Delta_k}\left(\mathbb{E}[D(p||q)] - r_n^{\text{nat}}(p)\right) \\ &= r_n^{\text{nat}}(q,\Delta_k). \end{aligned}$$

Lemma 9 implies $(a)$. Difference of maximums is smaller than maximum of differences, hence $(b)$. □

### B.2 Proof of Lemma 5

The proof uses the following lemma which computes the best natural estimator. For a random sequence $X^n$, let $\Phi_t \overset{\text{def}}{=} \varphi_t(X^n)$. Recall that $S_t(x^n)$ is the sum of probabilities of symbols that appears $t$ times in $x^n$. For notational convenience we use $S_t$ to denote both $S_t(x^n)$ and $S_t(X^n)$.

**Lemma 10.** *Let* $q_x^*(x^n)=\frac{S_{n_x}}{\varphi_{n_x}}$*, then*

$$q^* = \arg\min_{q\in\mathcal{Q}^{nat}} r_n(q,p)$$

*and*

$$r_n^{nat}(p) = \mathbb{E}\left[\sum_{t=0}^{n} S_t \log\frac{\Phi_t}{S_t}\right] - H(p).$$

*Proof.* For a natural estimator $q$, if $n_y = n_{y'}$, then $q_y(x^n) = q_{y'}(x^n)$ . Hence, with a slight abuse of notation let $q_{n_y}(x^n) = q_y(x^n)$. For a sequence $x^n$ and estimator $q$,

$$\begin{aligned} \sum_{y\in\mathcal{X}} p_y \log\frac{1}{q_y(x^n)} - \sum_{t=0}^{n} S_t\log\frac{\varphi_t}{S_t} &= \sum_{t=0}^{n}\sum_{y:n_y=t} p_y\log\frac{1}{q_y(x^n)} - \sum_{t=0}^{n} S_t\log\frac{\varphi_t}{S_t} \\ &= \sum_{t=0}^{n} S_t\log\frac{1}{q_t(x^n)} - \sum_{t=0}^{n} S_t\log\frac{\varphi_t}{S_t} \\ &= \sum_{t=0}^{n} S_t\log\frac{S_t}{\varphi_t q_t(x^n)} \\ &\geq 0, \end{aligned}$$

where the last inequality follows from the fact that $\sum_{t=0}^{n} S_t = \sum_{t=0}^{n}\varphi_t q_t(x^n) = 1$ and KL divergence is non-negative. Furthermore, equality is achieved only by the estimator that assigns $q_x^* = \frac{S_{n_x}}{\varphi_{n_x}}$. Hence,

$$r_n^{\text{nat}}(p) = \min_{q\in\mathcal{Q}^{\text{nat}}}\mathbb{E}\left[\sum_{y\in\mathcal{X}} p_y\log\frac{p_y}{q_y(X^n)}\right] = -H(p) + \mathbb{E}\left[\sum_{t=0}^{n} S_t\log\frac{\Phi_t}{S_t}\right].$$

□

*Proof of Lemma 5.* As before, with a slight abuse of notation let $q_{n_y}(x^n) = q_y(x^n)$ for natural estimators $q$. For any natural estimator $q$ and sequence $x^n$,

$$\begin{aligned}\sum_{y\in\mathcal{X}} p_y \log\frac{1}{q_y(x^n)} &= \sum_{t=0}^{n}\sum_{y:n_y=t} p_y \log\frac{1}{q_y(x^n)}\\ &= \sum_{t=0}^{n} S_t \log\frac{S_t}{\varphi_t q_t(x^n)} + \sum_{t=0}^{n} S_t\log\frac{\varphi_t}{S_t}\\ &= \sum_{t=0}^{n} S_t \log\frac{S_t}{\hat{S}_t} + \sum_{t=0}^{n} S_t\log\frac{\varphi_t}{S_t}.\end{aligned}$$

Thus by Lemma 10,

$$\begin{aligned}r_n^{\text{nat}}(q,p) &= -H(p) + \mathbb{E}\left[\sum_{t=0}^{n} S_t\log\frac{S_t}{\hat{S}_t} + \sum_{t=0}^{n} S_t\log\frac{\Phi_t}{S_t}\right] + H(p) - \mathbb{E}\left[\sum_{t=0}^{n} S_t\log\frac{\Phi_t}{S_t}\right]\\ &= \mathbb{E}\left[\sum_{t=0}^{n} S_t\log\frac{S_t}{\hat{S}_t}\right]\\ &= \mathbb{E}[D(S||\hat{S})].\end{aligned}$$

□

### B.3 Optimality of natural estimators

We now show that exist natural estimators that achieve $r_n^{\text{nat}}(\Delta_k)$ and $r_n^{\mathbb{P}_\sigma}(\Delta_k)$.

**Lemma 11.** *The exists a natural estimator $q''$ such that*

$$r_n^{nat}(q'',\Delta_k) = r_n^{nat}(\Delta_k).$$

*Similar there exists a natural estimator $q'$ such that*

$$r_n^{\mathbb{P}_\sigma}(q',\Delta_k) = r_n^{\mathbb{P}_\sigma}(\Delta_k).$$

*Proof.* We prove the result for $r_n^{\text{nat}}(\Delta_k)$. The result for $r_n^{\mathbb{P}_\sigma}(\Delta_k)$ is similar and omitted. Let profile $\bar{\varphi}$ of a sequence $x^n$ be the vector of its prevalences i.e., $\bar{\varphi}(x^n) \stackrel{\text{def}}{=} (\varphi_0(x^n),\varphi_1(x^n),\varphi_2(x^n),\dots\varphi_n(x^n))$. For any optimal estimator $q$ and sequence $x^n y$ such that $\bar{\varphi}(x^n) = \bar{\varphi}_n$ and $n_y(x^n) = t$ , let

$$q''_y(x^n) = \frac{\sum_{w^n z:\bar{\varphi}(w^n)=\bar{\varphi}_n, n_z=t} q_z(w^n)}{\sum_{u^n v:\bar{\varphi}(u^n)=\bar{\varphi}_n, n_v=t} 1}.$$

$q''$ is a natural estimator as if for any sequence $x^n$, $n_y(x^n) = n_{y'}(x^n)$, then $q''_y(x^n) = q''_{y'}(x^n)$. We show that $q''$ is an optimal estimator. Observe that for any $P\in\mathbb{P}_\sigma$

$$r_n(q,P) \overset{(a)}{\geq} \frac{1}{k!}\sum_{p\in P} r_n(q,p) \overset{(b)}{\geq} \frac{1}{k!}\sum_{p\in P} r_n(q'',p) \overset{(c)}{=} r_n(q'',P). \tag{7}$$

Maximum is larger than average and hence $(a)$. Every distribution in $P$ has the same KL loss for $q''$ and hence $(c)$. To prove $(b)$, observe that

$$\begin{aligned}
\sum_{p\in P} r_n(q,p) &= \sum_{p\in P}\sum_{x^n\in\mathcal{X}^n}\sum_{y\in\mathcal{X}} p(x^n y)\log\frac{1}{q_y(x^n)} - H(p)\\
&= \sum_{x^n\in\mathcal{X}^n}\sum_{y\in\mathcal{X}}\sum_{p\in P} p(x^n y)\log\frac{1}{q_y(x^n)} - H(p)\\
&= \sum_{\bar\varphi_n,t}\sum_{x^n:\bar\varphi(x^n)=\bar\varphi_n}\sum_{y:n_y=t}\sum_{p\in P} p(x^n y)\log\frac{1}{q_y(x^n)} - H(p)\\
&\overset{(d)}{\geq} \sum_{\bar\varphi_n,t}\sum_{x^n:\bar\varphi(x^n)=\bar\varphi_n}\sum_{y:n_y=t}\sum_{p\in P} p(x^n y)\log\frac{\sum_{u^n,v:\bar\varphi(u^n)=\bar\varphi_n,n_v=t}1}{\sum_{w^n,z:\bar\varphi(w^n)=\bar\varphi_n,n_z=t} q_z(w^n)} - H(p)\\
&= \sum_{\bar\varphi_n,t}\sum_{x^n:\bar\varphi(x^n)=\bar\varphi_n}\sum_{y:n_y=t}\sum_{p\in P} p(x^n y)\log\frac{1}{q''_y(x^n)} - H(p)\\
&= \sum_{p\in P} r_n(q'',p),
\end{aligned}$$

For all sequences $x^n y$ with the same $\bar\varphi(x^n)$ and $n_y(x^n)$, $\sum_{p\in P} p(x^n y)$ is the same. Hence, applying log-sum inequality results in $(d)$. By Lemma 10, every $p\in P$ has the same $r_n^{\text{nat}}(p)$, hence subtracting $r_n^{\text{nat}}(p)$ from both sides of Equation (7) results in

$$\max_{p\in P}\left(r_n(q,p) - r_n^{\text{nat}}(p)\right) \geq \max_{p\in P}\left(r_n(q'',p) - r_n^{\text{nat}}(p)\right).$$

Hence for the optimal estimator $q$,

$$\begin{aligned}
r_n^{\text{nat}}(\Delta_k) &= \max_{p\in\Delta_k}\left(r_n(q,p) - r_n^{\text{nat}}(p)\right)\\
&= \max_{P\in\mathbb{P}_\sigma}\left(\max_{p\in P}\left(r_n(q,p) - r_n^{\text{nat}}(p)\right)\right)\\
&\geq \max_{P\in\mathbb{P}_\sigma}\left(\max_{p\in P}\left(r_n(q'',p) - r_n^{\text{nat}}(p)\right)\right)\\
&= \max_{p\in\Delta_k}\left(r_n(q'',p) - r_n^{\text{nat}}(p)\right)\\
&= r_n(q'',\Delta_k).
\end{aligned}$$

Thus $q''$ is an optimal estimator and furthermore it is natural, hence the lemma. □

# C Regret bounds on the Good-Turing estimator

## C.1 Preliminaries

In practice, often the Good-Turing estimator is used for small multiplicities and empirical estimators are used for large multiplicities. We analyze this estimator and bound its regret. For a symbol appearing $t$ times, we assign probability $q'_x = \hat{S}_t/\varphi_t$, where $\hat{S}_t = C_t/N$. $N$ is the normalization factor to ensure that $\sum_{t=0}^{\infty}\hat{S}_t = 1$ and

$$C_t = \begin{cases}\varphi_t\cdot\frac{t}{n} & \text{if } t\geq t_0,\\ (\varphi_{t+1}+1)\cdot\frac{t+1}{n} & \text{else.}\end{cases}$$

We set $t_0\propto n^{1/3}$ later. Similar to our experiments, we have modified the Good-Turing estimator to $(\varphi_{t+1}+1)\cdot\frac{t+1}{n}$, thus ensuring that we never assign a non-zero probability. However, unlike our experiments, where we decided between empirical and Good-Turing estimators depending on if $\varphi_{t+1}\geq t$, for our proofs we just decide it based on $t$ for convenience. We remark that in our experiments the estimator in Section 4 performed better than the one above.

Ideally we would like to analyze this estimator when the number of samples is $n$. However, such analysis is complicated as the number of times symbols appear are dependent, for example, they add to $n$. A standard approach to overcome the dependence, e.g., [29], samples the distribution a random number of times $\sim$ poi$(n)$, the Poisson distribution with parameter $n$. Some useful properties of Poisson sampling include: $(i)$ A symbol with probability $p$ appears poi$(np)$ times, $(ii)$ The numbers of times different symbols appear are independent of each other, $(iii)$ For any fixed $n_0$, conditioned on the length poi$(n) \geq n_0$, the distribution of the first $n_0$ elements is identical to sampling $p$ i.i.d. exactly $n_0$ times. Thus, to simplify the analysis of the estimator, we assume that the number of samples is a Poisson random variable with mean $n$. A similar result holds with $n$ samples.

We first relate the KL regret to a chi-squared like distance between $S$ and $C$.

**Lemma 12.** *For any distribution $p \in \Delta_k$,*

$$\mathbb{E}[D(S||\hat{S})] \leq \sum_{t=0}^{t_0-1} \mathbb{E}\left[\frac{(S_t - (t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right] + \sum_{t=t_0}^{\infty} \mathbb{E}\left[\frac{(S_t - t\Phi_t/n)^2}{\Phi_t t/n}\right].$$

*Proof.* Since $\log(1+y) \leq y$, $\sum_{t=0}^{\infty} S_t = 1$, and $\sum_{t=0}^{\infty} C_t = N$,

$$\begin{aligned}
D(S||\hat{S}) &= \sum_{t=0}^{\infty} S_t \log \frac{S_t}{\hat{S}_t} \\
&= \sum_{t=0}^{\infty} S_t \log \frac{N S_t}{C_t} \\
&= \sum_{t=0}^{\infty} S_t \log \frac{S_t}{C_t} + \sum_{t=0}^{\infty} S_t \log N \\
&= \sum_{t=0}^{\infty} S_t \log\left(1 + \frac{S_t - C_t}{C_t}\right) + \log N \\
&\leq \sum_{t=0}^{\infty} S_t \left(\frac{S_t - C_t}{C_t}\right) + \log N \\
&= \sum_{t=0}^{\infty} (S_t - C_t)\left(\frac{S_t - C_t}{C_t}\right) + \sum_{t=0}^{\infty} C_t \left(\frac{S_t - C_t}{C_t}\right) + \log N \\
&= \sum_{t=0}^{\infty} (S_t - C_t)\left(\frac{S_t - C_t}{C_t}\right) + \sum_{t=0}^{\infty} (S_t - C_t) + \log N \\
&= \sum_{t=0}^{\infty} \frac{(S_t - C_t)^2}{C_t} + 1 - N + \log N \\
&\leq \sum_{t=0}^{\infty} \frac{(S_t - C_t)^2}{C_t} \\
&= \sum_{t=0}^{t_0-1} \frac{(S_t - C_t)^2}{C_t} + \sum_{t=t_0}^{\infty} \frac{(S_t - C_t)^2}{C_t}.
\end{aligned}$$

Taking expectations on both sides and substituting $C_t$ results in the lemma. □

### C.2 Empirical estimators

All of our results including the next lemma hold for all distributions in $\Delta_k$ and hence stated without any condition on the underlying distribution. Let $N_x \stackrel{\text{def}}{=} n_x(X^n)$ for a random sequence $X^n$.

**Lemma 13.** *For any $n$ and $t_0$,*

$$\sum_{t=t_0}^{\infty} \mathbb{E}\left[\frac{(S_t - t\Phi_t/n)^2}{t\Phi_t/n}\right] \leq \frac{1}{t_0}.$$

*Proof.*

$$\begin{aligned}\sum_{t=t_0}^{\infty}\frac{(S_t-t\Phi_t/n)^2}{t\Phi_t/n} &\le \sum_{t=t_0}^{\infty}\frac{(S_t-t\Phi_t/n)^2}{\Phi_t t_0/n}\\ &\overset{(a)}{\le}\sum_{t=t_0}^{\infty}\sum_x 1_{N_x=t}\frac{(p_x-t/n)^2}{t_0/n}\\ &=\sum_x\sum_{t=t_0}^{\infty}1_{N_x=t}\frac{(p_x-t/n)^2}{t_0/n}\\ &\le\sum_x\sum_{t=0}^{\infty}1_{N_x=t}\frac{(p_x-t/n)^2}{t_0/n}.\end{aligned}$$

$(a)$ follows from the fact that $\frac{(\sum_{x=1}^m a_x)^2}{m}\le\sum_{i=1}^m a_x^2$ for $a_x=1_{N_x=t}(p_x-t/n)$ and $m=\Phi_t$. Taking expectations on both sides,

$$\begin{aligned}\sum_{t=t_0}^{\infty}\mathbb{E}\left[\frac{(S_t-t\Phi_t/n)^2]}{\Phi_t t/n}\right] &\le\sum_x\frac{\mathbb{E}[\sum_{t=0}^{\infty}1_{N_x=t}(p_x-t/n)^2]}{t_0/n}\\ &\le\sum_x\frac{p_x/n}{t_0/n}\\ &=\frac{1}{t_0},\end{aligned}$$

where the second inequality follows from observing that $\mathbb{E}[\sum_{t=0}^{\infty}1_{N_x=t}(p_x-t/n)^2]$ is the variance of a Poisson random variable with mean $np_x$. □

### C.3 Good-Turing estimators

To bound the regret corresponding to the Good-Turing estimator, we need few auxiliary results. The next set of equations follow from results in [13], For any $n$ and $t$,

$$\mathbb{E}[S_t]=\frac{t+1}{n}\cdot\mathbb{E}[\Phi_{t+1}]. \tag{8}$$

$$\text{Var}(S_t)\le\frac{(t+1)(t+2)}{n^2}\cdot\mathbb{E}[\Phi_{t+2}]. \tag{9}$$

$$\mathbb{E}\left[\left(S_t-\frac{(t+1)\Phi_{t+1}}{n}\right)^2\right]\le\frac{(t+1)(t+2)\mathbb{E}[\Phi_{t+2}]}{n^2}+\frac{(t+1)^2\mathbb{E}[\Phi_{t+1}]}{n^2}. \tag{10}$$

The next lemma relates $\mathbb{E}[\Phi_{t+1}]$ to $\mathbb{E}[\Phi_t]$.

**Lemma 14.** *For any $n$ and $t\ge 1$,*

$$\mathbb{E}[\Phi_{t+1}]\le\mathbb{E}[\Phi_t]\left(\frac{2}{t}\log n+\frac{t}{t+1}\right)+\frac{1}{t+1}.$$

*Proof.* Let $r \geq \frac{t}{t+1}$.

$$\begin{aligned}
\mathbb{E}[\Phi_{t+1}] &= \mathbb{E}\left[\sum_x 1_{N_x=t+1}\right] \\
&= \sum_x e^{-np_x}\frac{(np_x)^{t+1}}{(t+1)!} \\
&= \sum_x \frac{n}{t+1}\cdot e^{-np_x}\frac{(np_x)^t}{t!}p_x \\
&= \sum_{x:np_x\leq r(t+1)} \frac{n}{t+1}\cdot e^{-np_x}\frac{(np_x)^t}{t!}p_x + \sum_{x:np_x>r(t+1)} \frac{n}{t+1}\cdot e^{-np_x}\frac{(np_x)^t}{t!}p_x \\
&\overset{(a)}{\leq} r\sum_{x:np_x\leq r(t+1)} e^{-np_x}\frac{(np_x)^t}{t!} + \sum_{x:np_x>r(t+1)} \frac{n}{t+1}e^{-r(t+1)}\frac{(r(t+1))^t}{t!}p_x \\
&\leq r\sum_x e^{-np_x}\frac{(np_x)^t}{t!} + \sum_x \frac{n}{t+1}e^{-r(t+1)}\frac{(r(t+1))^t}{t!}p_x \\
&\overset{(b)}{\leq} r\sum_x e^{-np_x}\frac{(np_x)^t}{t!} + \sum_x \frac{n}{t+1}e^{-rt/2}p_x \\
&\leq r\mathbb{E}[\Phi_t] + \frac{n}{t+1}e^{-\frac{rt}{2}}.
\end{aligned}$$

$(a)$ follows from the fact that second term is a decreasing as a function of $np_x$ in the range $[r(t+1),\infty)$. $(b)$ follows from the fact that

$$e^{-r(t+1)}\frac{(r(t+1))^t}{t!} = e^{-rt}r^t\cdot e^{-t}\frac{(t+1)^t}{t!} \leq e^{-rt}r^t \leq e^{-rt/2}.$$

Choosing $r = \frac{2}{t}\log n + \frac{t}{t+1}$, yields

$$\mathbb{E}[\Phi_{t+1}] \leq \mathbb{E}[\Phi_t]\left(\frac{2}{t}\log n + \frac{t}{t+1}\right) + \frac{1}{t+1}.$$

□

The final auxiliary lemma bounds the inverse moment of Poisson binomial distributions.

**Lemma 15.** *Let $X_i$ for $1 \leq i \leq n$ be Bernoulli random variables, then*

$$\mathbb{E}\left[\frac{1}{\sum_{i=1}^n X_i + 1}\right] \leq \frac{1}{\sum_{i=1}^n \mathbb{E}[X_i]}.$$

*Proof.* Let $s_i = \mathbb{E}[X_i]$. We show that of all tuples $s_1, s_2, \ldots, s_n$ such that $\sum_{i=1}^n s_i = ns$, the one that maximizes the expectation is $s_i = s, \forall i$. Suppose for some $i, j$, $s_i > s_j$, we show that if we decrease $s_i$ and increase $s_j$ keeping the sum same, then the expectation increases. Let $Y = 1 + \sum_{k\notin\{i,j\}} X_k$. For any instance of $X^n$, taking expectation with respect to only $X_i$ and $X_j$.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{X_i + X_j + Y} \mid Y\right] &= \frac{(1-s_i)(1-s_j)}{Y} + \frac{s_i(1-s_j)+(1-s_i)s_j}{Y+1} + \frac{s_is_j}{Y+2} \\
&= \frac{1}{Y} + (s_i+s_j)\left(\frac{1}{Y+1} - \frac{1}{Y}\right) + s_is_j\frac{2}{Y(Y+1)(Y+2)}.
\end{aligned}$$

Thus if we decrease $s_i$ and increase $s_j$ (keeping $s_i + s_j$ fixed), then $s_is_j$ increases and hence the expectation increases. Hence the maximum occurs when $s_i = s_j$ for all $i, j$ and

$$\mathbb{E}\left[\frac{1}{\sum_{i=1}^n X_i + 1}\right] \leq \mathbb{E}\left[\frac{1}{Z+1}\right],$$

where $Z$ is a binomial random variable with parameters $n$ and $s = \sum_{i=1}^{n} \mathbb{E}[X_i]/n$.

The expectation can be bounded as

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{Z+1}\right] &= \sum_{j=0}^{n} \frac{1}{j+1}\binom{n}{j} s^j (1-s)^{n-j} \\
&= \frac{1}{(n+1)s} \sum_{j=0}^{n} \binom{n+1}{j+1} s^{j+1} (1-s)^{n+1-(j+1)} \\
&\leq \frac{1}{(n+1)s} \\
&\leq \frac{1}{ns} \\
&= \frac{1}{\sum_{i=1}^{n} \mathbb{E}[X_i]}.
\end{aligned}$$

□

Using the above lemma, we first bound the expectation of $S_t^2/(\Phi_{t+1}+1)$.

**Lemma 16.** *For any $n$ and $t$, if $\mathbb{E}[\Phi_{t+1}] > 2$, then*

$$\mathbb{E}\left[\frac{S_t^2}{\Phi_{t+1}+1}\right] \leq \frac{\mathbb{E}[S_t^2]}{\mathbb{E}[\Phi_{t+1}]-1}.$$

*Proof.* We first observe that for any $x$,

$$\mathbb{E}[1_{N_x=t+1}] = e^{-np_x}\frac{(np_x)^{t+1}}{(t+1)!} \leq e^{-t-1}\frac{(t+1)^{t+1}}{(t+1)!} \leq \frac{1}{e}. \tag{11}$$

Since $S_t = \sum_x p_x 1_{N_x=t}$ and $\Phi_{t+1} = \sum_x 1_{N_x=t+1}$,

$$\frac{S_t^2}{\Phi_{t+1}+1} = \frac{\sum_x \sum_y p_x p_y 1_{N_x=t} 1_{N_y=t}}{\sum_z 1_{N_z=t+1}+1} = \sum_x \sum_y \frac{p_x p_y 1_{N_x=t} 1_{N_y=t}}{\sum_{z:z\neq x, z\neq y} 1_{N_z=t+1}+1},$$

where the equality follows from the fact that symbol cannot appear both $t$ and $t+1$ times thus only one of $1_{N_x=t}$ and $1_{N_x=t+1}$ can be 1. The numerator and the denominator of the terms on RHS are independent of each other, hence

$$\begin{aligned}
\mathbb{E}\left[\frac{p_x p_y 1_{N_x=t} 1_{N_y=t}}{\sum_z 1_{N_z=t+1}+1}\right] &= \mathbb{E}\left[\frac{p_x p_y 1_{N_x=t} 1_{N_y=t}}{\sum_{z:z\neq x, z\neq y} 1_{N_z=t+1}+1}\right] \\
&= \mathbb{E}\left[p_x p_y 1_{N_x=t} 1_{N_y=t}\right] \mathbb{E}\left[\frac{1}{\sum_{z:z\neq x, z\neq y} 1_{N_z=t+1}+1}\right] \\
&\overset{(a)}{\leq} \frac{\mathbb{E}\left[p_x p_y 1_{N_x=t} 1_{N_y=t}\right]}{\sum_{z:z\neq x, z\neq y} \mathbb{E}[1_{N_z=t+1}]} \\
&\overset{(b)}{\leq} \frac{\mathbb{E}\left[p_x p_y 1_{N_x=t} 1_{N_y=t}\right]}{\mathbb{E}[\Phi_{t+1}-1]},
\end{aligned}$$

$(a)$ follows from Lemma 15 and $(b)$ follows from Equation (11) as

$$\sum_{z:z\neq x, z\neq y} \mathbb{E}[1_{N_z=t+1}] = \sum_z \mathbb{E}[1_{N_z=t+1}] - \mathbb{E}[1_{N_x=t+1}] - \mathbb{E}[1_{N_y=t+1}] \geq \mathbb{E}[\Phi_{t+1}] - 1.$$

Summing over $x$ and $y$ results in the lemma. □

We now have all the tools to bound the error of the Good-Turing estimator. We divide the set of values into two groups, depending on the value of $\mathbb{E}[\Phi_{t+1}]$.

**Lemma 17.** *For any $n$ and $t$ if $\mathbb{E}[\Phi_{t+1}] \leq 2$, then*

$$\mathbb{E}\left[\frac{(S_t - (t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right] \leq \frac{5t}{n} + \frac{4\log n}{n}\left(\frac{t+2}{t+1}\right) + \frac{6}{n}.$$

*Proof.* Let $Z = S_t - (t+1)\Phi_{t+1}/n$.

$$\begin{aligned}\mathbb{E}\left[\left(Z - \frac{t+1}{n}\right)^2\right] &\overset{(a)}{=} \mathbb{E}[Z^2] + \frac{(t+1)^2}{n^2} \\ &\overset{(b)}{\leq} \frac{(t+1)(t+2)\mathbb{E}[\Phi_{t+2}]}{n^2} + \frac{(t+1)^2\mathbb{E}[\Phi_{t+1}]}{n^2} + \frac{(t+1)^2}{n^2} \\ &\overset{(c)}{\leq} 2\frac{(t+1)(t+2)}{n^2}\cdot\left(\frac{2\log n}{t+1} + \frac{t+1}{t+2}\right) + \frac{(t+1)(t+2)}{n^2(t+2)} + \frac{3(t+1)^2}{n^2}.\end{aligned}$$

Equation (8) implies $Z$ is a zero mean random variable and hence $(a)$. Equation (10) implies $(b)$ and $(c)$ follows by Lemma 14 and the fact that $\mathbb{E}[\Phi_{t+1}] \leq 2$. Hence,

$$\begin{aligned}\mathbb{E}\left[\frac{(Z-(t+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right] &\leq \frac{\mathbb{E}[(Z-(t+1)/n)^2]}{(t+1)/n} \\ &\leq \frac{2(t+2)}{n}\cdot\left(\frac{2\log n}{t+1} + \frac{t+1}{t+2}\right) + \frac{1}{n} + \frac{3(t+1)}{n} \\ &= \frac{5t}{n} + \frac{4\log n(t+2)}{n(t+1)} + \frac{6}{n}.\end{aligned}$$

□

**Lemma 18.** *For any $n$ and $t$ if $\mathbb{E}[\Phi_{t+1}] > 2$, then*

$$\mathbb{E}\left[\frac{(S_t - (t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right] \leq \frac{5t}{n} + \frac{4\log n}{n}\left(\frac{t+2}{t+1}\right) + \frac{6}{n}.$$

*Proof.*

$$\frac{(S_t - (t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n} = \frac{S_t^2}{(\Phi_{t+1}+1)(t+1)/n} + \frac{(t+1)(\Phi_{t+1}+1)}{n} - 2S_t.$$

Thus by Equation (8),

$$\mathbb{E}\left[\frac{(S_t - (t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right] = \mathbb{E}\left[\frac{S_t^2}{(\Phi_{t+1}+1)(t+1)/n}\right] - \frac{(t+1)\mathbb{E}[\Phi_{t+1}]}{n} + \frac{t+1}{n}. \quad (12)$$

By Lemma 16 and Equations (8), (9),

$$\begin{aligned}\mathbb{E}\left[\frac{S_t^2}{(\Phi_{t+1}+1)(t+1)/n}\right] &\leq \frac{\mathbb{E}[S_t^2]}{\mathbb{E}[\Phi_{t+1}-1](t+1)/n} \\ &\leq \frac{t+1}{n}\frac{\mathbb{E}[\Phi_{t+1}]^2}{\mathbb{E}[\Phi_{t+1}-1]} + \frac{t+2}{n}\frac{\mathbb{E}[\Phi_{t+2}]}{\mathbb{E}[\Phi_{t+1}-1]}.\end{aligned}$$

Substituting the above equation in Equation (12) and simplifying,

$$\begin{aligned}\mathbb{E}\left[\frac{(S_t - (t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right] &\leq \frac{(t+1)\mathbb{E}[\Phi_{t+1}] + (t+2)\mathbb{E}[\Phi_{t+2}]}{n\mathbb{E}[\Phi_{t+1}-1]} + \frac{t+1}{n} \\ &\overset{(a)}{\leq} 2\frac{(t+1)\mathbb{E}[\Phi_{t+1}] + (t+2)\mathbb{E}[\Phi_{t+2}]}{n\mathbb{E}[\Phi_{t+1}]} + \frac{t+1}{n} \\ &\overset{(b)}{\leq} 2\left(\frac{t+1}{n} + \frac{t+2}{n}\left(\frac{2\log n}{t+1} + \frac{t+1}{t+2} + \frac{1}{2(t+2)}\right)\right) + \frac{t+1}{n} \\ &= \frac{5t}{n} + \frac{4\log n}{n}\left(\frac{t+2}{t+1}\right) + \frac{6}{n}.\end{aligned}$$

Since $\mathbb{E}[\Phi_{t+1}] \geq 2$, $\mathbb{E}[\Phi_{t+1}] - 1 \geq \mathbb{E}[\Phi_{t+1}]/2$ and hence $(a)$. Lemma 14 implies $(b)$. □

Combining the above two lemmas results in

**Lemma 19.** *For any $t_0 \geq 1$,*

$$\sum_{t=0}^{t_0-1} \mathbb{E}\left[\frac{(S_t - (t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right] \leq \frac{5t_0^2}{2n} + \frac{4\log n}{n}(t_0 + \log t_0 + 1) + \frac{7t_0}{2n}.$$

*Proof.* By Lemmas 17 and 18, regardless of the value of $\mathbb{E}[\Phi_{t+1}]$,

$$\mathbb{E}\left[\frac{(S_t - (t+1)(\Phi_{t+1}+1)/n)^2}{(\Phi_{t+1}+1)(t+1)/n}\right] \leq \frac{5t}{n} + \frac{4\log n}{n}\left(\frac{t+2}{t+1}\right) + \frac{6}{n}.$$

Summing the above expression for $0 \leq t \leq t_0 - 1$ results in the lemma. □

Substituting the results from Lemmas 13 and 19 in Lemma 12,

$$\mathbb{E}[D(S||\hat{S})] \leq \frac{1}{t_0} + \frac{5t_0^2}{2n} + \frac{4\log n}{n}(t_0 + \log t_0 + 1) + \frac{7t_0}{2n}.$$

Substituting $t_0 = n^{1/3}/5^{1/3}$ results in Theorem 1.

$$r^{\text{nat}}_{\text{poi}(n)}(q', \Delta_k) \leq \max_{p \in \Delta_k} \mathbb{E}[D(S||\hat{S})] \leq \frac{2.6}{n^{1/3}} + \frac{2.4\log n(n^{1/3} + \log n + 1)}{n} + \frac{2.1}{n^{2/3}} \leq \frac{3 + o_n(1)}{n^{1/3}}.$$

# D Proof of Theorem 3

To lower bound $r_n^{\mathbb{P}_\sigma}(\Delta_k)$ it is sufficient to lower bound $r_n^{\mathbb{P}_\sigma}(\mathcal{P})$ for any subset $\mathcal{P} \subseteq \Delta_k$. We construct a subset $\mathcal{P}$ by considering a set of distributions $\{p^{\bar{v}} : \bar{v} \in \{-1,1\}^{m-1}\}$ and all their possible permutations. The lower bound argument uses Fano's inequality and Gilbert Varshamov bounds.

We choose $\mathcal{P}$ to be the set of distributions whose probability multiset are close to that of a distribution $p^0$, where $p^0$ is defined as follows.

Let $c$ be a sufficiently large constant. Let $m$ be the largest odd number less than $\min(k, (n/(c^2\log^2 n))^{1/3})$. Let $p^0$ be the following distribution. For $1 \leq i \leq m-1$,

$$p_i^0 = \frac{\log n}{6n}\sqrt{\frac{c^2 n}{m}}\left(\sqrt{\frac{n}{c^2 m \log^2 n}} + i\right)$$

and $p_m^0 = 1 - \sum_{i=1}^{m-1} p_i^0$. Observe that for all $1 \leq i \leq m-1$, $1/(6m) \leq p_i^0 \leq 1/(3m)$ and $p_m^0 \geq 2/3$.

We choose the close-by distributions as follows. Let $\epsilon = \sqrt{\frac{c^*}{mn}}$, where $c^*$ is some sufficiently small constant. For a binary vector $\bar{v} \in \{-1,1\}^{m-1}$, let $p^{\bar{v}}$ be the distribution such that $p_i^{\bar{v}} = p_i^0 + \bar{v}_i\epsilon$ for $1 \leq i \leq m-1$ and $p^{\bar{v}}(m) = 1 - \sum_{i=1}^{m-1} p_i^{\bar{v}}$. Note that by the properties of $p^0$ and $\epsilon$, $p^{\bar{v}}$ is a valid distribution for every $\bar{v}$. Let $\mathcal{C}$ be the largest subset of $\{-1,1\}^{m-1}$ such that for every $\bar{v} \in \mathcal{C}$, $\sum_i \bar{v}_i = 0$ and for every pair $\bar{v}, \bar{v}' \in \mathcal{C}$, $\sum_i |\bar{v}_i - \bar{v}'_i| \geq c'(m-1)$ for some constant $c'$. The following variation of Gilbert Varshamov lemma lower bounds size of $\mathcal{C}$.

**Lemma 20.** *There exists a set of vectors $\mathcal{C}$ over $\{-1,1\}^{m-1}$ of size $2^{c''\cdot(m-1)}$ such that the minimum hamming distance between any two vectors is $\geq c'(m-1)$ for some universal constants $c' > 0, c'' > 0$ and $\sum_i \bar{v}_i = 0$ for all $\bar{v} \in \mathcal{C}$.*

Let $\mathcal{P}' = \{p^{\bar{v}} : \bar{v} \in \mathcal{C}\}$ and $P_{\bar{v}} = \{p^{\bar{v}}(\sigma(\cdot)) : \sigma \in \Sigma^{m-1}\}$ be the set of all permutations of a distribution $p^{\bar{v}}$, i.e., all distributions with the same multiset as $p^{\bar{v}}$. Let

$$\mathcal{P} = \cup_{\bar{v} \in \mathcal{C}} P_{\bar{v}}.$$

We first bound the regret of the induced permutation class $P_{\bar{v}}$ that contains all permutations of a distribution $p^{\bar{v}}$.

**Lemma 21.** *For every induced permutation class $P_{\bar{v}}$,*

$$r_n(P_{\bar{v}}) \leq \frac{1}{n}.$$

*Proof.* We prove the bound by constructing an estimator $q$. Consider the estimator $q$ which sorts the multiplicities and assigns the $i^{th}$-frequently occurred symbol probability $p_i^{\bar{v}}$. Since this is a natural estimator, it occurs the same loss for all distributions in $P_{\bar{v}}$ and hence,

$$\begin{aligned}
r_n(P_{\bar{v}}) &\leq \max_{p \in P_{\bar{v}}} \mathbb{E}[D(p||q)] \\
&= \mathbb{E}[D(p^{\bar{v}}||q)] \\
&\overset{(a)}{\leq} \Pr(\exists i, j : N_i > N_j,\, p_i^{\bar{v}} < p_j^{\bar{v}}) \log n \\
&\overset{(b)}{\leq} \binom{m}{2} e^{-2\log n} \log n \\
&\leq \frac{1}{n}.
\end{aligned}$$

$(a)$ follows from the fact that the estimator makes an error only if two multiplicities cross over and if it does make an error, the maximum KL divergence is at most $\log(p_{\max}/p_{\min}) \leq \log n$. Since probabilities for any two symbols $i$ and $j$ differ by at least $\frac{\log n}{6n} \cdot \sqrt{\frac{c^2 n}{m}}$ and the probabilities themselves lie between $1/(6m)$ and $1/(3m)$, by choosing a sufficiently large $c$, the cross over probability can be bounded by $e^{-2\log n}$ using the Chernoff bound and hence $(b)$. □

We now lower bound the KL divergence between $p^{\bar{v}}$ and $p^{\bar{v}'}$ for every pair of vectors $\bar{v}$ and $\bar{v}'$. Let the Hamming distance between two vectors $\bar{v}$ and $\bar{v}'$ be $||\bar{v} - \bar{v}'||_1 = \sum_{i=1}^{m-1} |\bar{v}_i - \bar{v}'_i|$.

**Lemma 22.** *For two distributions $p^{\bar{v}}$ and $p^{\bar{v}'}$ in $\mathcal{P}'$,*

$$\frac{1}{8}\left(c'\sqrt{\frac{mc^*}{n}}\right)^2 \leq \frac{1}{2}||p^{\bar{v}} - p^{\bar{v}'}||_1^2 \leq D(p^{\bar{v}}||p^{\bar{v}'}) \leq \frac{48mc^*}{n}.$$

*Proof.*

$$\begin{aligned}
D(p^{\bar{v}}||p^{\bar{v}\prime}) &\overset{(a)}{\leq} \sum_{i=1}^{m} \frac{(p_i^{\bar{v}} - p_i^{\bar{v}\prime})^2}{p_i^{\bar{v}\prime}} \\
&\overset{(b)}{\leq} 2\sum_{i=1}^{m} \frac{(p_i^{\bar{v}} - p_i^{\bar{v}\prime})^2}{p_i^0} \\
&\leq 2\sum_{i=1}^{m-1} \frac{(\bar{v}_i - \bar{v}'_i)^2(\sqrt{c^*/nm})^2}{1/(6m)} \\
&\leq 12\sum_{i=1}^{m-1} \frac{(\bar{v}_i - \bar{v}'_i)^2 c^*}{n} \\
&\leq 24\sum_{i=1}^{m-1} \frac{|\bar{v}_i - \bar{v}'_i| c^*}{n} \\
&= \frac{24||\bar{v} - \bar{v}'||_1 c^*}{n} \\
&\leq \frac{48mc^*}{n}.
\end{aligned}$$

$(a)$ follows from bounding the KL divergence by the Chi-squared distance and $(b)$ follows from the fact that $\epsilon \ll 1/m$. For the lower bound,

$$\begin{aligned} D(p^{\bar{v}}||p^{\bar{v}'}) &\overset{(a)}{\geq} \frac{1}{2}||p^{\bar{v}} - p^{\bar{v}'}||_1^2 \\ &= \frac{1}{2}\left(\frac{||\bar{v} - \bar{v}'||_1\sqrt{c^*}}{\sqrt{mn}}\right)^2 \\ &\overset{(b)}{\geq} \frac{1}{2}\left(\frac{c'(m-1)\sqrt{c^*}}{\sqrt{mn}}\right)^2 \\ &\overset{(c)}{\geq} \frac{1}{8}\left(c'\sqrt{\frac{mc^*}{n}}\right)^2, \end{aligned}$$

where $(a)$ follows from Pinsker's inequality, $(b)$ follows by construction, and $m - 1 \geq 2$ and hence $(c)$. □

We now state Fano's inequality for distribution estimation.

**Lemma 23.** *Let $p^1, p^2, \ldots p^{r+1}$ be distributions such that $D(p^i||p^j) \leq \beta$ and $||p^i - p^j||_1 \geq \alpha$, for all $i, j$. For any estimator $q$,*

$$\sup_i \mathbb{E}_i[||p^i - q||_1] \geq \frac{\alpha}{2}\left(1 - \frac{n\beta + \log 2}{\log r}\right).$$

We now have all the tools for the lower bound.

*Proof of Theorem 3.* For every permutation subclass $P_{\bar{v}}$ in $\mathcal{P}$, by Lemma 21

$$r_n(P_{\bar{v}}) \leq \frac{1}{n}.$$

Thus,

$$\begin{aligned} r_n^{\mathbb{P}_\sigma}(\mathcal{P}) &= \min_q \max_{\bar{v}} \Big(\max_{p \in P_{\bar{v}}} r_n(q,p) - r_n(P_{\bar{v}})\Big) \\ &\geq \min_q \max_{\bar{v}} \Big(\max_{p \in P_{\bar{v}}} r_n(q,p) - \frac{1}{n}\Big) \\ &= \min_q \max_{p \in \mathcal{P}} r_n(q,p) - \frac{1}{n} \\ &= \min_q \max_{p \in \mathcal{P}} \mathbb{E}[D(p||q)] - \frac{1}{n} \\ &\overset{(a)}{\geq} \min_q \max_{p \in \mathcal{P}'} \mathbb{E}[D(p||q)] - \frac{1}{n} \\ &\overset{(b)}{\geq} \min_q \max_{p \in \mathcal{P}'} \mathbb{E}\left[\frac{||p-q||_1^2}{2}\right] - \frac{1}{n} \\ &\overset{(c)}{\geq} \min_q \max_{p \in \mathcal{P}'} \frac{1}{2}\mathbb{E}\left[||p-q||_1\right]^2 - \frac{1}{n} \\ &\overset{(d)}{\geq} \Omega\left(\frac{m}{n}\right) - \frac{1}{n} \\ &\geq \Omega\left(\frac{m}{n}\right). \end{aligned}$$

$\mathcal{P}' \subset \mathcal{P}$, hence $(a)$. $(b)$ follows from Pinsker's inequality and $(c)$ follows from convexity. By construction, for every pair of distributions in $\mathcal{P}'$, $\beta = D(p||p') \leq 48c^*m/n$ and $\alpha = ||p - p'||_1 \geq \Omega(\sqrt{m/n})$ (Lemma 22). Furthermore by Lemma 20, $\mathcal{P}'$ has $r+1 = 2^{c''(m-1)}$ distributions. Setting $c^*$ to be a sufficiently small constant and applying Lemma 23 to $\mathcal{P}'$ with the above values of $\alpha, \beta$, and $r$ results in $(d)$. Substituting the value of $m$ in the above equation results in the Theorem. □