Poissonization and universal compression of envelope classes

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh

ECE UCSD, {jacharya, ashkan, alon, asuresh}@ucsd.edu

***Abstract*—Poisson sampling is a method for eliminating dependence among symbols in a random sequence. It helps improve algorithm design, strengthen bounds, and simplify proofs. We relate the redundancy of fixed-length and Poisson-sampled sequences, use this result to derive a simple formula for the redundancy of general envelope classes, and apply this formula to obtain simple and tight bounds on the redundancy of power-law and exponential envelope classes, in particular answering a question posed in [1].**

I. Introduction

Universal compression is the design of a single scheme to encode an unknown source from a known class of distributions. Let P be a distribution over $\mathcal{A}$. From the source coding theorem we know that $H(P)$ bits are necessary and sufficient to encode P. Any distribution Q over $\mathcal{A}$ implies a code that uses $-\log Q(a)$ bits to encode $a \in \mathcal{A}$. The code implied by the underlying distribution uses $-\log P(a)$ bits and expected code length is the entropy $H(P)$. The (worst-case) redundancy of a distribution Q with respect to P is

$$\sup_{a\in\mathcal{A}} \log \frac{P(a)}{Q(a)},$$

the largest difference between the number of bits used and the optimal. We assume the distributions belong to a known class $\mathcal{P}$. Then

$$\hat{R}(\mathcal{P}) = \inf_{Q} \sup_{P\in\mathcal{P}} \sup_{a\in\mathcal{A}} \log \frac{P(a)}{Q(a)},$$

is the worst case redundancy, or minimax regret of $\mathcal{P}$.

Let $\hat{P}(a) \stackrel{\text{def}}{=} \sup_{P\in\mathcal{P}} P(a)$ be the largest probability assigned to symbol a by a distribution in $\mathcal{P}$. Then,

$$S(\mathcal{P}) \stackrel{\text{def}}{=} \sum_{a\in\mathcal{A}} \hat{P}(a)$$

is the Shtarkov sum of $\mathcal{P}$. Shtarkov [2] showed that

$$\hat{R}(\mathcal{P}) = \log\left(S(\mathcal{P})\right),$$

and the optimal distribution assigns probability $\hat{P}(a)/S(\mathcal{P})$ to $a \in \mathcal{A}$.

In this work we study coding of *i.i.d.* samples from unknown distributions. For a distribution P on an underlying alphabet $\mathcal{X}$ let P^n be the product distribution $P \times P \ldots \times P$ over $\mathcal{X}^n$. Let P belong to a known class $\mathcal{P}$. Let $\mathcal{P}^n$ be the class of all distributions P^n with $P \in \mathcal{P}$. Coding length-n sequences over $\mathcal{X}$ equivalently implies encoding symbols from $\mathcal{A} = \mathcal{X}^n$. In such block encoding we treat each $x_1^n \in \mathcal{X}^n$ as a symbol a. As before we consider the class of all distributions (not only *i.i.d.*) Q_n over $\mathcal{X}^n$ to define

$$\hat{R}(\mathcal{P}^n) \stackrel{\text{def}}{=} \inf_{Q_n} \sup_{P^n} \sup_{x_1^n\in\mathcal{X}^n} \log \frac{P^n(x_1^n)}{Q_n(x_1^n)},$$

as the worst case redundancy, or minimax regret of $\mathcal{P}^n$. The class $\mathcal{P}$ is said to be universal if $\hat{R}(\mathcal{P}) = o(n)$, *i.e.,* redundancy is sublinear in the block-length.

The most extensively studied [3–12] class of distributions is $\mathcal{I}_k^n$, the collection of all *i.i.d.* distributions over length-n sequences from an underlying alphabet of size k, *e.g.,* over $[k]$. It is now well established that

1) For $k = o(n)$

$$\hat{R}(\mathcal{I}_k^n) = \frac{k-1}{2}\log\frac{n}{k} + \frac{k}{2} + o(k),$$

2) For $n = o(k)$

$$\hat{R}(\mathcal{I}_k^n) = n\log\frac{k}{n} + o(n).$$

In particular [3, 4] showed that class of all distributions over $\mathbb{N}$ has infinite redundancy.

While the problem was traditionally studied in the setting of finite underlying alphabet size and large block lengths, however in numerous applications the underlying natural alphabet best describing the data might be large. For example, a text document only contains a small set of the entire dictionary, a natural image has many possible pixel values.

The class of all *i.i.d.* distributions over $\mathbb{N}$ is therefore *too large* to provide meaningful compression schemes for all distributions. These impossibility results led researchers to consider natural subclasses of all *i.i.d.* distributions and then compress sequences generated from distributions in these or to consider all *i.i.d.* distributions but decompose sequences into natural components and compress each part separately.

We now describe three such approaches that have been proposed in recent years to characterize large alphabet compression. [13] propose to encode the order of symbols (pattern) in the sequences and the dictionary separately. [14, 15] studied the compression of patterns and in particular show that patterns can be compressed with sublinear worst case redundancy *i.e.,* they are universally compressible.

[16] consider the class $\mathcal{M}$ of all monotone distributions over $\mathbb{N}$. They first show that even this class is not universally compressible. Despite this, they designed universal codes for

monotone distributions with finite entropy. [17] study the redundancy of $\mathcal{M}_k$, monotone distributions over alphabet size k, as a function of k and the block length n.

In this paper, we consider a third, natural and elegant approach, proposed by [1]. They study a fairly general class of distributions, called *envelope classes*, which as the name suggests are distributions that are bounded by an envelope.

They provide general bounds on the redundancy of envelope classes. The upper bounds on the worst-case redundancy are obtained by bounding the Shtarkov sum. They provide bounds on the more stringent (for lower bounds) average case redundancy, where instead of the supremum over all possible sequences, they consider the expected log ratio, *i.e.,* the KL divergence. For this quantity, they provide a more complex employing the redundancy-capacity theorem.

We now introduce the Poisson sampling model, and later use it to provide redundancy bounds on envelope classes.

II. Poisson Model

Poisson sampling provides simplified analysis in a number of problems in machine learning and statistics [18]. To the best of our knowledge, the first application of Poisson sampling in universal coding was in [19], and [20] to prove near-optimal bounds on the pattern redundancy. More recently, [21] also apply Poisson sampling to provide simpler coding schemes for compressing *i.i.d.* distributions. In this work we are interested in coding/prediction over countably infinite alphabets described bounded by an envelope.

Let $\mathrm{poi}(\lambda)$ be a Poisson distribution with parameter λ and

$$\mathrm{poi}(\lambda, \mu) \stackrel{\text{def}}{=} e^{-\lambda} \frac{\lambda^\mu}{\mu!}$$

is the probability of μ under a $\mathrm{poi}(\lambda)$ distribution. For a distribution P over $\mathcal{X}$, recall that P^n is a distribution over $\mathcal{X}^n$. Similarly, we define a distribution $P^{\mathrm{poi}(n)}$ over $\mathcal{X}^* \stackrel{\text{def}}{=} \mathcal{X} \cup \mathcal{X}^2 \cup \mathcal{X}^3 \ldots$ as follows. First generate $n' \sim \mathrm{poi}(n)$, then sample the distribution *i.i.d.* n' times. Therefore the probability of any sequence $x_1^{n'} \in \mathcal{X}^*$ is

$$P^{\mathrm{poi}(n)}(x_1^{n'}) = \mathrm{poi}(n, n') P^{n'}(x_1^{n'}) = \mathrm{poi}(n, n') \prod_{i=1}^{n'} P(x_i).$$

Similar to $\mathcal{P}^n$, let

$$\mathcal{P}^{\mathrm{poi}(n)} \stackrel{\text{def}}{=} \{P^{\mathrm{poi}(n)} : P \in \mathcal{P}\}$$

be the class of distributions over $\mathcal{X}^*$ via sampling a distribution *i.i.d.* $\mathrm{poi}(n)$ times. Under this model, the advantage is that the number of appearances of symbols (multiplicities) are independent of each other.

Note that even though the length of the sequence is random, it is concentrated around n as a Poisson random variable.

For large n, the length n' of a Poisson sampled sequence is concentrated around n. Using this, we show in Theorem 2 that it is sufficient to consider $\mathcal{P}^{\mathrm{poi}(n)}$ instead of $\mathcal{P}^n$.

We first describe some properties of Poisson sampled distributions.

A. Properties of Poisson sampling

The *multiplicity* of a symbol in a sequence is the number of times it appears in it. The *type* of a sequence x_1^n over $\mathcal{X} = \{a_1, \ldots, \}$ is

$$\tau(x_1^n) \stackrel{\text{def}}{=} (\mu(a_1), \mu(a_2), \ldots),$$

the tuple of multiplicities of the symbols in the sequence x_1^n [22, 23]. For example, if $a_i = i$, for $i = 1, \ldots, 6$ denotes the possible outcomes of a die. Then the sequence of outcomes $2, 3, 1, 6, 1, 3, 3, 4, 6$ has type $\tau = (2, 1, 3, 1, 0, 2)$. Any product distribution assigns the same probability to sequences with the same type. Therefore, for most statistical problems, under independent sampling, it suffices to consider only the type as the random variable.

When $|\mathcal{X}| = k$, this corresponds to the balls and bins model with n balls and k bins. The main difficulty in analyzing such models is the dependencies that arise among the multiplicities, *e.g.,* they add to n [18]. However, if we consider $\mathrm{poi}(n)$ sampling, the distribution $P^{\mathrm{poi}(n)}$ has the following properties.

Lemma 1 ([18]): For any distribution $P^{\mathrm{poi}(n)}$,

1) conditioned on n', the distribution on $\mathcal{X}^{n'}$ is $P^{n'}$ for discrete distributions.
2) A symbol $a \in \mathcal{X}$ with $P(a) = p$ appears $\mathrm{poi}(np)$ times *independently* of all other symbols.

This shows that the multiplicities are independent under Poisson sampling.

In the next result we relate the redundancy of Poisson sampling to fixed length sampling. In particular, we show that it suffices to consider $P^{\mathrm{poi}(n)}$.

Theorem 2: For any $\epsilon > 0$, for $n > n_0(\epsilon, \mathcal{P})$

$$\hat{R}(\mathcal{P}^{\mathrm{poi}(n(1-\epsilon))}) - 1 \le \hat{R}(\mathcal{P}^n) \le \hat{R}(\mathcal{P}^{\mathrm{poi}(n)}) + 1.$$

III. Related work and new results

Definition 3: The envelope class of a function f is the class

$$\mathcal{P}_f \stackrel{\text{def}}{=} \{(p_1, p_2, \ldots) : p_i \le f_i, \text{ and } p_1 + p_2 + \ldots = 1\}$$

of all distributions such that the symbol probability is bounded by the value of the function at that point.

Analogous to $\mathcal{P}^n$, we can define $\mathcal{P}_f^n$ and $\mathcal{P}_f^{\mathrm{poi}(n)}$.

The study of universal coding for envelope classes was initiated in [1], where they provide upper and lower bounds on the redundancy. They show that if $\sum f_i = \infty$ the redundancy is infinite for any n, and therefore consider only absolutely integrable envelopes. Their upper bound states that

$$\hat{R}(\mathcal{P}_f^n) \le \min_{u \le n} \left[n \sum_{i \ge u} f(i) + \frac{u-1}{2} \log n \right] + O(1).$$

They prove a more complex lower bound on the average-redundancy. More recently, [21] provide bounds of similar order for envelope classes using Poisson sampling.

We provide simple bounds on $\hat{R}(\mathcal{P}_f^{\mathrm{poi}(n)})$ in terms of the redundancy of a simple one dimensional class of distributions characterized by a single parameter $\lambda^{\max}$, as

$$\mathrm{POI}(\lambda^{\max}) \stackrel{\text{def}}{=} \{\mathrm{poi}(\lambda) : \lambda < \lambda^{\max}\}.$$

This is the class of all Poisson distributions with parameter bounded above. This class is simple enough, so that we can bound its redundancy tightly.

For an envelope characterized by f, let $\lambda_i^{\max} \stackrel{\text{def}}{=} nf_i$. Let l_f be the smallest integer such that

$$\sum_{j \geq l_f} f_j \leq 1 - \epsilon.$$

This also implies that $\sum_{j \geq l_f} \lambda_j^{\max} \leq n(1-\epsilon)$.

We now state our main result on envelope classes.

Theorem 4: For the envelope class $\mathcal{P}_f$,

$$\sum_{i=l_f}^{\infty} \hat{R}\left(\mathrm{POI}(\lambda_i^{\max})\right) \leq \hat{R}\left(\mathcal{P}_f^{\mathrm{poi}(n)}\right) \leq \sum_{i=1}^{\infty} \hat{R}\left(\mathrm{POI}(\lambda_i^{\max})\right),$$

where $\lambda_i^{\max} = nf_i$.

Since, we use $\mathrm{POI}(\lambda^{\max})$ as a primitive, we now show simple bounds that will be used to compute redundancies of specific classes later.

Lemma 5: For $\lambda \leq 1$, $\hat{R}\Big(\mathrm{POI}(\lambda)\Big) = \log\left(2 - \exp(-\lambda)\right)$, and for $\lambda \geq 1$,

$$\sqrt{\frac{2(\lambda+1)}{\pi}} \leq \hat{R}\Big(\mathrm{POI}(\lambda)\Big) \leq 2 + \sqrt{\frac{2\lambda}{\pi}}.$$

We now consider power-law and exponential envelopes and apply Theorem 4 to obtain sharp bounds on redundancies of these classes improving the previous results.

To prove the efficacy of these bounds we apply them on the Power-law and exponential envelopes.

Definition 6: The *power-law* envelope class with parameters $\alpha > 1$ and c is the class of distributions $\Lambda_{c \cdot^{-\alpha}}$ over $\mathbb{N}$ such that $\forall i \in \mathbb{N}$, $p_i \leq \frac{c}{i^\alpha}$.

Definition 7: The *exponential-law* envelope class with parameters α and c is the class of distributions $\Lambda_{ce^{-\alpha \cdot}}$ over $\mathbb{N}$ such that $\forall i \in \mathbb{N}$, $p_i \leq ce^{-\alpha i}$.

[1] obtain bounds on the redundancy of these classes. For power-law envelopes they show that for large n,

$$C_0 n^{\frac{1}{\alpha}} \leq \hat{R}(\Lambda^n_{c \cdot^{-\alpha}}) \leq \left(\frac{2cn}{\alpha - 1}\right)^{\frac{1}{\alpha}} (\log n)^{1-\frac{1}{\alpha}} + O(1), \quad (1)$$

where C_0 is a constant (function of α and c). For exponential envelopes they prove that

$$\frac{\log^2 n}{8\alpha}(1 + o(1)) \leq \hat{R}(\Lambda^n_{ce^{-\alpha \cdot}}) \leq \frac{\log^2 n}{2\alpha} + \mathcal{O}(1).$$

[24] improve the bounds for $\Lambda^n_{ce^{-\alpha \cdot}}$ and show that

$$\hat{R}(\Lambda^n_{ce^{-\alpha \cdot}}) = \frac{\log^2 n}{4\alpha} + O(\log n \log\log n).$$

More recently, [25] extend the arguments of [24] to find tight universal codes for the larger class of sub-exponential distributions, which have strictly faster decay than power-law classes, but slower than exponential.

However these results do not find the optimal redundancy of power-law envelopes. Their techniques seem to rely on the fact that exponential envelopes restrict the support sizes to be poly-logarithmic and break down for more heavy-tailed distributions, such as the power-law envelopes.

We show that simply applying Theorem 4 to these classes and bounding the resulting expressions gives tight redundancy bounds. In particular, for $\Lambda^n_{c \cdot^{-\alpha}}$, we show that

Theorem 8: For large n

$$\frac{(cn)^{1/\alpha}}{2}\Big[\alpha + \frac{1}{\alpha - 1} - \log 3\Big] - 1 \leq \hat{R}(\Lambda^n_{c \cdot^{-\alpha}}) \leq (cn)^{1/\alpha}\Big[\frac{\alpha}{2} + \frac{1}{\alpha - 1} + \log 3\Big] + 1.$$

For $\Lambda^n_{ce^{-\alpha \cdot}}$, we prove that

Theorem 9: For large n

$$\hat{R}(\Lambda^n_{ce^{-\alpha \cdot}}) = \frac{\log^2 n}{4\alpha} + O(\log c \log n).$$

Remark This result can also be generalized to provide tight bounds for sub-exponential envelope class considered in [25].

In the next three sections we prove these results.

IV. Proofs

A. Proof of Theorem 2

For $x_1^n \in \mathcal{X}^n$, let $\hat{P}^n(x_1^n) = \sup_{P^n \in \mathcal{P}^n} P^n(x_1^n)$ be the maximum likelihood (ML) probability of x_1^n. Then,

$$S(\mathcal{P}^n) = \sum_{x_1^n \in \mathcal{X}^n} \hat{P}^n(x_1^n).$$

By the first part of Lemma 1, it follows that

$$S(\mathcal{P}^{\mathrm{poi}(n)}) = \sum_{n'} \mathrm{poi}(n, n') S(\mathcal{P}^{n'}). \quad (2)$$

By conditioning on x_1^n in the expression for $S(\mathcal{P}^{n+1})$ it is easy to see that $S(\mathcal{P}^n) \leq S(\mathcal{P}^{n+1})$.

Therefore,

$$S(\mathcal{P}^{\mathrm{poi}(n)}) \stackrel{(a)}{=} \sum_{n' \geq 0} \mathrm{poi}(n, n') S(\mathcal{P}^{n'}) \stackrel{(b)}{\geq} S(\mathcal{P}^n) \sum_{n' \geq n} \mathrm{poi}(n, n') \stackrel{(c)}{\geq} \frac{1}{2} S(\mathcal{P}^n),$$

where (a) follows from Equation (2), (b) from monotonicity of $S(\mathcal{P}^n)$ and (c) from the fact that median of a Poisson distribution close to its mean. Taking logarithms gives the second part of the theorem.

The first part of the theorem uses (2) again along with tail bounds on Poisson random variables and is omitted here. It essentially uses the fact that a $\mathrm{poi}(n(1-\epsilon))$ random variable exceeds n with exponentially small probability. We omit the proof due to lack of space.

B. Proof of Theorem 4

For *i.i.d.* sampling types are a sufficient statistic of the sequence, namely all sequences with the same type have the same probability. Let $\tau(P^n)$ be the distribution induced by P^n over types. Let

$$\tau(\mathcal{P}^n) = \{\tau(P^n) : P \in \mathcal{P}\}$$

be all distributions over types from distributions of the form P^n. By the second item in Lemma 1, for a distribution $P = (p_1, p_2, , \ldots)$ over $\mathcal{X} = \{1, 2, \ldots\}$

$$\tau(P^{\mathrm{poi}(n)}) = (\mathrm{poi}(np_1), \mathrm{poi}(np_2), \ldots),$$

where each coordinate is an independent Poisson distribution.

We first show that the redundancy of sequences is the same as the redundancy of types.

Lemma 10: $\hat{R}(\tau(\mathcal{P}^n)) = \hat{R}(\mathcal{P}^n)$ and $\hat{R}(\tau(\mathcal{P}^{\mathrm{poi}(n)})) = \hat{R}(\mathcal{P}^{\mathrm{poi}(n)})$.

Proof: Any *i.i.d.* distribution assigns the same probability to all sequences with the same type. Therefore, such sequences have the same maximum likelihood probability. We show that the Shtarkov sums of the two classes are the same and hence they have same redundancy.

$$\begin{aligned} S(\mathcal{P}^n) = \sum_{x_1^n \in \mathcal{X}^n} \hat{P}^n(x_1^n) &= \sum_{\tau} \sum_{x_1^n : \tau(x_1^n) = \tau} \hat{P}^n(x_1^n) \\ &= \sum_{\tau} \hat{P}^n(\tau) = S(\tau(\mathcal{P}^n)). \end{aligned}$$

The proof also applies to Poisson sampling. ∎

Under Poisson sampling, the type of a distribution is a tuple of independent random variables, namely the multiplicities. In general suppose $\mathcal{P}$ be a collection of product (independent) distributions over $\mathcal{A} \times \mathcal{B}$, *i.e.,* each element in $\mathcal{P}$ is a distribution of the form $P_1 \times P_2$, where P_1 and P_2 are distributions over $\mathcal{A}$ and $\mathcal{B}$ respectively. Let $\mathcal{P}_{\mathcal{A}}$ and $\mathcal{P}_{\mathcal{B}}$ be the class of marginals over $\mathcal{A}$ and $\mathcal{B}$ respectively. It can be shown, *e.g.,* [20], that the redundancy of $\mathcal{P}$ is at most the sum of the marginal redundancies.

Lemma 11 (Redundancy of products): For a collection $\mathcal{P}$ of product distributions over $\mathcal{A} \times \mathcal{B}$,

$$\hat{R}(\mathcal{P}) \le \hat{R}(\mathcal{P}_{\mathcal{A}}) + \hat{R}(\mathcal{P}_{\mathcal{B}}).$$

Furthermore, if $\mathcal{P} = \mathcal{P}_{\mathcal{A}} \times \mathcal{P}_{\mathcal{B}}$ then equality holds.

Proof: For any $(a, b) \in \mathcal{A} \times \mathcal{B}$,

$$\sup_{(P_1, P_2) \in \mathcal{P}} P_1(a) P_1(b) \le \sup_{P_1 \in \mathcal{P}_{\mathcal{A}}} P_1(a) \sup_{P_2 \in \mathcal{P}_{\mathcal{B}}} P_2(b).$$

Now,

$$\begin{aligned} S(\mathcal{P}) &= \sum_{(a,b) \in \mathcal{A} \times \mathcal{B}} \sup_{(P_1, P_2) \in \mathcal{P}} P_1(a) P_2(b) \\ &\le \sum_{a \in \mathcal{A}} \sup_{P_1 \in \mathcal{P}_{\mathcal{A}}} P(a) \sum_{b \in \mathcal{B}} \sup_{P_2 \in \mathcal{P}_{\mathcal{B}}} Q(b) = S(\mathcal{P}_{\mathcal{A}}) S(\mathcal{P}_{\mathcal{B}}), \end{aligned}$$

where the inequality follows from the equation above, and the lemma follows by taking logarithms. When all possible combinations of marginals is possible, then the inequality becomes an equality. ∎

The lemma generalizes to more than two product classes. Similar to the characterization of $\tau(\mathcal{P}^{\mathrm{poi}(n)})$ it follows that

$$\tau\left(\mathcal{P}_f^{\mathrm{poi}(n)}\right) = \left\{(\mathrm{poi}(\lambda_1), \mathrm{poi}(\lambda_2) \ldots) : \lambda_i \le n f_i, \sum \lambda_i = n\right\}.$$

Notice that each $\lambda_i \le \lambda_i^{\max} \stackrel{\mathrm{def}}{=} n \cdot f_i$, and therefore the class marginal distributions of element i is a Poisson random variable with parameter $\le \lambda_i^{\max}$ and hence a subset of $\mathrm{POI}(\lambda_i^{\max})$. Therefore, Lemma 11 yields

$$\hat{R}\left(\tau(\mathcal{P}_f^{\mathrm{poi}(n)})\right) \le \hat{R}\left(\mathrm{POI}(\lambda_1^{\max})\right) + \hat{R}\left(\mathrm{POI}(\lambda_2^{\max})\right) + \ldots$$

For the lower bound note for any choice of $\lambda_i < \lambda_i^{\max}$ for $i \ge l_f$ corresponds to a distribution in $\mathcal{P}_f^{\mathrm{poi}(n)}$. In other words, all product distributions in

$$\mathrm{POI}(\lambda_{l_f}^{\max}) \times \mathrm{POI}(\lambda_{l_f+1}^{\max}) \times \ldots$$

are valid projections of a distribution in $\mathcal{P}_f^{\mathrm{poi}(n)}$ along the coordinates $i \ge l_f$. Therefore, for the elements along these coordinates the redundancy of types under Poisson sampling is determined precisely by Lemma 11.

$$\hat{R}\left(\tau(\mathcal{P}_f^{\mathrm{poi}(n)})\right) \ge \hat{R}\left(\mathrm{POI}(\lambda_{l_f}^{\max})\right) + \hat{R}\left(\mathrm{POI}(\lambda_{l_f+1}^{\max})\right) + \ldots.$$

Combining these with Lemma 10 proves the theorem.

C. Proof of Lemma 5

The Poisson distribution assigning highest probability to $i \in \mathbb{N}$ is $\mathrm{poi}(i)$. Therefore the ML distribution of j in $\mathrm{POI}(\lambda)$ is

$$\arg\max_{\mathrm{POI}(\lambda)} P(i) = \begin{cases} \mathrm{poi}(i) & \text{if } i \le \lfloor \lambda \rfloor \\ \mathrm{poi}(\lambda) & \text{otherwise.} \end{cases}$$

Using this, the Shtarkov sum of the class is

$$\begin{aligned} S\Big(\mathrm{POI}(\lambda)\Big) &= \sum_{i=0}^{\lfloor \lambda \rfloor} e^{-i} \frac{i^i}{i!} + \sum_{\lfloor \lambda \rfloor + 1}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} \\ &\stackrel{(a)}{=} 1 + \sum_{i=0}^{\lfloor \lambda \rfloor} \left(e^{-i} \frac{i^i}{i!} - e^{-\lambda} \frac{\lambda^i}{i!}\right), \end{aligned}$$

where (a) uses that $\sum_{\mu} \mathrm{poi}(\lambda, \mu) = 1$. Using the second expression shows the case $\lambda \le 1$. Using the first along with the following Stirling's approximation proves the case of $\lambda > 1$.

Lemma 12 (Stirling's Approximation): For any $n \ge 1$, there is a $\theta_n \in (\frac{1}{12n+1}, \frac{1}{12n})$ such that

$$n! = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\theta_n}.$$

D. Power-law: Proof of Theorem 8

Proof: By Definition 6, for power-law class $\Lambda_{c\cdot^{-\alpha}}$, $\lambda_i^{\max} = \frac{cn}{i^\alpha}$. Let $b \stackrel{\text{def}}{=} (cn)^{1/\alpha}$, then $\lambda_i^{\max} \geq 1$ for $i \leq b$ and $\lambda_i^{\max} < 1$ otherwise.

Then,

$$\hat{R}(\Lambda^n_{c\cdot^{-\alpha}}) \stackrel{(a)}{\leq} \sum_{i\leq b} \hat{R}(\mathrm{POI}(\lambda_i^{\max})) + \sum_{i>b} \hat{R}(\mathrm{POI}(\lambda_i^{\max})) + 1$$
$$\stackrel{(b)}{\leq} \sum_{i\leq b} \log\left(2 + \sqrt{\frac{2\lambda_i^{\max}}{\pi}}\right) + \sum_{i>b}^{\infty} \log(2 - e^{-\lambda_i^{\max}}) + 1,$$

where (a) follows from Theorem 4 and (b) from Lemma 5.

We consider the two summations separately.

For the first term, we note that for $\lambda \geq 1$, $2 + \sqrt{2\lambda/\pi} < 3\sqrt{\lambda}$ and use it with the following simplification.

$$\sum_{i=1}^{B} \log \frac{B}{i} = \log \frac{B^B}{B!} \leq B,$$

which follows from $B! > (B/e)^B$. Therefore,

$$\sum_{i=1}^{b} \log\left(2 + \sqrt{\frac{2\lambda_i^{\max}}{\pi}}\right) < \sum_{i=1}^{b} \log\left(3\sqrt{\frac{cn}{i^\alpha}}\right)$$
$$= b\log(3) + \frac{\alpha}{2}\sum_{i=1}^{b} \log\left(\frac{(cn)^{\frac{1}{\alpha}}}{i}\right)$$
$$\stackrel{(a)}{<} (cn)^{1/\alpha}\Big(\log(3) + \frac{\alpha}{2}\Big),$$

where (a) follows since $b = (cn)^{1/\alpha}$.

Taking the second term,

$$\sum_{i=b+1}^{\infty} \log(2 - e^{-\lambda_i^{\max}}) \stackrel{(a)}{\leq} \sum_{i=b+1}^{\infty} \lambda_i^{\max} = cn \sum_{i=b+1}^{\infty} \frac{1}{i^\alpha}$$
$$= \frac{c^{1/\alpha}}{\alpha - 1} n^{1/\alpha},$$

where (a) uses $e^x \geq 2 - e^{-x}$, and (b) follows by using $s = b + 1 = (cn)^{1/\alpha} + 1$ in

$$\sum_{i=s}^{\infty} \frac{1}{i^r} \leq \int_s^{\infty} \frac{1}{(x-1)^r} \leq \frac{(s-1)^{1-r}}{(r-1)}.$$

Combining these results proves the theorem.

For the lower bound we only use the terms $\lambda_i^{\max}$ for $j \geq l_f$. Using the lower bound on $S(\mathrm{POI}(\lambda))$ from Lemma 5 and again applying the Stirling's approximation (lower bound instead of upper) proves the result. Since the ideas are the same, we omit the proof. ∎

E. Exponential class: Proof of Theorem 9

Proof: By Definition 7 for $\Lambda^n_{ce^{-\alpha\cdot}}$, $i \leq \frac{\log(cn)}{\alpha}$ if and only if $\lambda_i^{\max} \geq 1$. Let $b \stackrel{\text{def}}{=} \frac{\log(cn)}{\alpha}$. Then similar to the proof of power-law class we derive the two summations and bound them individually.

Once again for $\lambda > 1$, $2 + \sqrt{2\lambda/\pi} < 3\sqrt{\lambda}$. Therefore

$$\sum_{i=1}^{b} \log\left(2 + \sqrt{\frac{2\lambda_i^{\max}}{\pi}}\right) = b\log 3 + \frac{1}{2}\sum_{i=1}^{b} \log[cne^{-\alpha i}]$$
$$= b\log 3 + \frac{1}{2}\sum_{i=1}^{b} \alpha(b-i)$$
$$< b\log(3) + \frac{b^2\alpha}{4}.$$

For the second term, since $e^{\alpha b} = cn$,

$$\sum_{i=b}^{\infty} \log(2 - e^{-\lambda_i^{\max}}) \leq \sum_{i=b}^{\infty} \lambda_i^{\max} = \sum_{i=b}^{\infty} cne^{-\alpha i} \frac{1}{1 - e^{-\alpha}}.$$

Substituting $b = \log(cn)/\alpha$ proves the upper bound. Once again the lower bound is very similar and is omitted. ∎

REFERENCES

[1] S. Boucheron, A. Garivier, and E. Gassiat, "Coding on countably infinite alphabets," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 358–373, 2009.

[2] Y. M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, no. 3, pp. 3–17, 1987.

[3] J. Kieffer, "A unified approach to weak universal source coding," *IEEE Transactions on Information Theory*, vol. 24, no. 6, pp. 674–682, Nov. 1978.

[4] L. Davisson, "Universal noiseless coding," *IEEE Transactions on Information Theory*, vol. 19, no. 6, pp. 783–795, Nov. 1973.

[5] L. D. Davisson, R. J. McEliece, M. B. Pursley, and M. S. Wallace, "Efficient universal noiseless source codes," *IEEE Transactions on Information Theory*, vol. 27, no. 3, pp. 269–279, 1981.

[6] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: basic properties." *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[7] T. Cover, "Universal portfolios," *Mathematical Finance*, vol. 1, no. 1, pp. 1–29, January 1991.

[8] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, January 1996.

[9] W. Szpankowski, "On asymptotics of certain recurrences arising in universal coding," *Problems of Information Transmission*, vol. 34, no. 2, pp. 142–146, 1998.

[10] Q. Xie and A. R. Barron, "Asymptotic minimax regret for data compression, gambling and prediction," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 431–445, 2000.

[11] A. Orlitsky and N. Santhanam, "Speaking of infinity [i.i.d. strings]," *itt*, vol. 50, no. 10, pp. 2215 – 2230, oct 2004.

[12] W. Szpankowski and M. J. Weinberger, "Minimax redundancy for large alphabets," in *ISIT*, 2010, pp. 1488–1492.

[13] J. Åberg, Y. M. Shtarkov, and B. J. M. Smeets, "Multialphabet coding with separate alphabet description," in *Proceedings of Compression and Complexity of Sequences*, 1997.

[14] A. Orlitsky, N. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1469– 1481, July 2004.

[15] ——, "Always Good Turing: Asymptotically optimal probability estimation," *Science*, vol. 302, no. 5644, pp. 427–431, October 17 2003, see also Proceedings of the 44th Annual Symposium on Foundations of Computer Science, October 2003.

[16] D. Foster, R. Stine, and A. Wyner, "Universal codes for finite sequences of integers drawn from a monotone distribution," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1713–1720, June 2002.

[17] G. I. Shamir, "Universal source coding for monotonic and fast decaying monotonic distributions," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7194–7211, 2013.

[18] M. Mitzenmacher and E. Upfal, *Probability and computing - randomized algorithms and probabilistic analysis*. Cambridge Univ. Press, 2005.

[19] J. Acharya, H. Das, and A. Orlitsky, "Tight bounds on profile redundancy and distinguishability," in *NIPS*, 2012.

[20] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and A. Suresh, "Tight bounds for universal compression of large alphabets," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, 2013, pp. 2875–2879.

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

[22] H. Das, "Competitive tests and estimators for properties of distributions," Ph.D. dissertation, UCSD, 2012.

[23] T. Cover and J. Thomas, *Elements of Information Theory, 2nd Ed.* Wiley Interscience, 2006.

[24] D. Bontemps, "Universal coding on infinite alphabets: Exponentially decreasing envelopes," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1466–1478, 2011.

[25] D. Bontemps, S. Boucheron, and E. Gassiat, "About adaptive coding on countable alphabets," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 808–821, 2012.