Data-driven weak universal redundancy

Narayana Santhanam
University of Hawaii, Manoa
nsanthan@hawaii.edu

Venkat Anantharam
UC Berkeley,
ananth@eecs.berkeley.edu

Aleksander Kavcic
University of Hawaii, Manoa,
kavcic@hawaii.edu

Wojciech Szpankowski
Purdue University,
spa@cs.purdue.edu

Abstract— **In applications involving estimation, the relevant model classes of probability distributions are often too complex to admit estimators that converge to the truth with convergence rates that can be uniformly bounded over the entire model class as the sample size increases (uniform consistency). While it is often possible to get pointwise guarantees, so that the convergence rate of the estimator can be bounded in a model-dependent way, such pointwise gaurantees are unsatisfactory — estimator performance is a function of the very unknown quantity that is being estimated. Therefore, even if an estimator is consistent, how well it is doing may not be clear no matter what the sample size.**

Departing from this traditional uniform/pointwise dichotomy, a new analysis framework is explored by characterizing model classes of probability distributions that may only admit pointwise guarantees, yet where all the information about the unknown model needed to gauge estimator accuracy can be inferred from the sample at hand. To provide a focus to this suggested broad new paradigm, we analyze the universal compression problem in this data-driven pointwise consistency framework.

I. Introduction

Nowadays, data accumulated in many biological, financial, and other statistical problems stands out not just because of its nature or size, but also because the questions we ask from the data are unlike anything we asked before. There is often a tension in these *big data* problems between the need for rich model classes to better represent the application and our ability to handle these classes at all from a mathematical point of view.

The model classes of probability distributions needed to adequately model such applications are often too complex to admit estimators which admit uniform convergence guarantees over the entire model class. While we can often come up with estimators with a guarantee of convergence irrespective of which element of the model class is in force, this may not be very useful — our gauge of how well the estimator is doing is dependent on the very quantity being estimated!

We are therefore led to challenge the dichotomy of *uniform* and *pointwise* consistency in the analysis of statistical estimators. Both uniform and pointwise guarantees have their own drawbacks. The former precludes the desired richness of model classes. While the latter allows for rich model classes, it does not provide practical guarantees that can be used in applications.

Instead, we consider a new paradigm positioned in between these two extremes. This framework modifies the world of pointwise consistent estimators—keeping as far as possible the richness of model classes possible but ensuring that all information needed about the unknown model to evaluate estimator accuracy can be gleaned from the data. We call this *data-driven pointwise consistency*.

To bring focus into the theoretical framework, we will formulate and characterize this approach in this document for weak compression over countably infinite alphabets. This approach generalizes a related prediction problem studied earlier by a subset of the current authors in [1].

II. Formulation of problem

Let $\mathcal{P}$ be a collection of probability distributions over the naturals $\mathbb{N} = \{1, 2, \ldots\}$. Let $\mathcal{P}^\infty$ be the measures induced over infinite sequences of numbers from $\mathbb{N}$ by *i.i.d.* sampling from distributions in $\mathcal{P}$. $\mathcal{P}^\infty$ is called *strongly compressible* if there is a measure q over infinite sequences of natural numbers satisfying

$$\limsup_{n\to\infty} \sup_{p\in\mathcal{P}^\infty} \frac{1}{n} E_p \log \frac{p(X^n)}{q(X^n)} = 0. \tag{1}$$

$\mathcal{P}^\infty$ is called *weakly compressible* if there exists a measure q over infinite sequences of natural numbers such that $\forall p \in \mathcal{P}^\infty$

$$\limsup_{n\to\infty} \frac{1}{n} E_p \log \frac{p(X^n)}{q(X^n)} = 0. \tag{2}$$

Remark Note that both (1) and (2) are usually phrased with encoders or distributions for length n sequences. However, since we will be concerned mainly with the limits, we can use the simpler formulations as above. See [2] for a formal explanation of why these formulations are completely equivalent. □

The issue that concerns us is the following problem with the weak compression formulation: While we know the measure q is a good *universal sequential encoding* of the unknown p for long enough sequences, how long is "long enough" depends on the unknown p since the convergence to limit may not be uniform in (2). Can we resolve this in (2) using the *data* generated?

Therefore, given any accuracy $\delta > 0$ we ask for an indicator function $\Phi : \mathbb{N}^* \to \{0, 1\}$ that will clarify this point. The function above observes a sequence in $\mathbb{N}^*$, and decides what sequence length is long enough that the normalized KL divergence in (2) above is below δ, and in addition will remain below δ for longer sequences.

From a notational point of view, we require $\Phi(x^i x_{i+1}) \geq \Phi(x^i)$—namely, once Φ indicates that the length is "long enough" that the normalized KL will remain below δ from that point on, it cannot renege later. When Φ turns 1, we say

978-1-4799-5186-4/14/$31.00 ©2014 IEEE

the scheme *enters the compression game*. Furthermore, we require that for all $p \in \mathcal{P}^\infty$,

$$p(\Phi \text{ enters}) = p(X^n : \lim_{n\to\infty} \Phi(X^n) = 1) = 1.$$

Fix a universal measure q, i.e. one satisfying (2) for the given model class $\mathcal{P}^\infty$. Given $\delta > 0$, the pair (Φ, q) is *δ-premature* for a source $p \in \mathcal{P}^\infty$ and string x_1^i if for some $j \leq i$,

$$\Phi(x_1^j) = 1 \text{ and } \frac{1}{j} E_p \log \frac{p(X^j)}{q(X^j)} > \delta.$$

Note that given $p \in \mathcal{P}^\infty$, the set of all strings on which (Φ, q) is $\delta-$premature can be identified with a prefix free set corresponding to the first times the accuracy condition was violated for the strings. The probability under p of (Φ, q) being $\delta-$premature is the probability of this prefix free set.

Definition 1. Given a weakly compressible class $\mathcal{P}^\infty$ we would like to find a universal measure q such that for any accuracy $\delta > 0$ and confidence $\eta > 0$, there is an indicator Φ such that no matter what $p \in \mathcal{P}^\infty$ is in force,

$$p((\Phi, q) \text{ is } \delta-\text{premature}) < \eta.$$

If this is possible, such a class is called weakly compressible in the *data-driven* sense (*d.w.c*). □

The operational justification for our formulation of *d.w.c* classes of *i.i.d.* sources can be articulated as follows. Given such a class, let q be any measure over infinite length sequences that verifies the definition, i.e. such that for every $\delta > 0$ and $\eta > 0$ there is some $\Phi_{\delta,\eta} : \mathbb{N}^* \mapsto \{0, 1\}$ for which the probability under every p in the model class that $(\Phi_{\delta,\eta}, q)$ is δ-premature is less than η.

As we observe the realization of the *i.i.d.* data samples from the (unknown) source p in the model class, we will eventually see a string of some (random) length $n = n(\delta, \eta, p)$ (say x_1^n) such that $\Phi_{\delta,\eta}(x_1^n) = 1$. Now, even though we do not know p, we get the guarantee (with confidence $\geq 1 - \eta$) that using q to compress any subsequent length-n or longer sequence of symbols in the usual way (*i.e.,* $-\log q(x^k)$ bits for a sequence x^k) incurs an expected per-symbol redundancy $\leq \delta$.

As an example, suppose $\mathcal{P}^\infty$ is strongly compressible in addition, namely there exists a measure q satisfying (1). For all $\delta > 0$, the sets

$$N_\delta = \{n : \sup_{p \in \mathcal{P}^\infty} \frac{1}{n} E_p \log \frac{p(X^n)}{q(X^n)} > \delta\}$$

are finite. Suppose we set for any δ, $\Phi(x^i) = 1$ if $i > \max N_\delta$ and 0 else, the we get for all $p \in \mathcal{P}^\infty$ that $p((\Phi, q) \text{ is } \delta-\text{premature}) = 0$.

In this paper, we obtain a condition that is both necessary and sufficient for an *i.i.d.* class $\mathcal{P}^\infty$ to be data-driven weakly compressible.

Remark The development in this paper generalizes the formulation of a prediction question that was studied in [1]. Suppose we have a collection $\mathcal{P}^\infty$ of *i.i.d.* measures, and samples $X_1, X_2, \ldots$ from an unknown $p \in \mathcal{P}^\infty$. Given a confidence $\eta > 0$, can we come up with a mapping $\Phi : \mathbb{N}^* \to \mathbb{R} \cup \infty$ such that for all p,

$$p(\Phi(X^i) < X_{i+1}) < \eta$$

and Φ is finite eventually almost surely? Please see the full version [3] for details on why this problem, which was formulated in [1], is a data-driven pointwise convergence formulation. □

III. Necessary and sufficient conditions for *d.w.c*

Since the probability distributions in $\mathcal{P}^\infty$ are *i.i.d.*, the sources therein can be identified without confusion using their single letter marginals as well. We will do this as needed, without comment.

For any two measures p and q, we let

$$D_n(p||q) \overset{\text{def}}{=} E_{p(X^n)} \log \frac{p(X^n)}{q(X^n)},$$

the KL divergence between the distributions induced over length n sequences by p and q respectively. Furthermore, for measures p and q,

$$\mathcal{J}(p, q) = D_1\left(p||\frac{p+q}{2}\right) + D_1\left(q||\frac{p+q}{2}\right),$$

where in the above, the KL divergences are taken between the single letter distributions corresponding to p and q.

An *$\epsilon-$neighborhood* of $p \in \mathcal{P}^\infty$ is the set $B(p, \epsilon)$ of all sources $p' \in \mathcal{P}^\infty$ such that $\mathcal{J}(p, p') < \epsilon$.

III-A. Deceptive measures

Roughly speaking, $p \in \mathcal{P}^\infty$ is *deceptive* if the strong redundancy of neighborhoods of p is bounded away from 0 in the limit as the neighborhood shrinks to 0. More precisely,

$$\lim_{\epsilon\to 0} \inf_q \limsup_{n\to\infty} \sup_{p' \in B(p,\epsilon)} \frac{1}{n} D_n(p'||q) > 0,$$

where the infimum is over measures on infinite sequences of naturals. For example, consider the class of all monotone sources over $\mathbb{N}$ (all sources such that the probability of $i \geq$ that of $i+1$). It is easy to see that the distribution that assigns probability 1 to 1 is deceptive, and in the same way that all sources in the collection are deceptive.

Lemma 1. If $p \in \mathcal{P}^\infty$ is not deceptive, then there is a measure q over infinite sequences of naturals such that

$$\lim_{\epsilon\to 0} \limsup_{n\to\infty} \sup_{p' \in B(p,\epsilon)} \frac{1}{n} D_n(p'||q^*) = 0. \qquad \Box$$

III-B. Main result

Our main result relates deceptive measures to *d.w.c*.

Theorem 2. $\mathcal{P}^\infty$ is *d.w.c* iff no $p \in \mathcal{P}^\infty$ is deceptive. □

IV. Examples

We provide a series of examples that highlight various aspects of our formulation as well as that of deceptive distributions. For proofs of the claims in the examples below, please see [3].

The first collection we consider is $\mathcal{U}$, the collection of all uniform distributions over finite supports of form $\{m, m+1, \ldots, M\}$ for all positive integers m and M with $m \leq M$. Let the sequence of losses be *i.i.d.* samples from distributions in $\mathcal{U}$—call the resulting model class over infinite loss sequences $\mathcal{U}^\infty$.

Example 1. $\mathcal{U}^\infty$ is not strongly compressible but is *d.w.c.* □

The second collection is the set $\mathcal{N}_1^\infty$ of all *i.i.d.* processes such that the one dimensional marginals have finite first moment. Namely, $\forall p \in \mathcal{N}_1^\infty$, $E_p X < \infty$ where $X \in \mathbb{N}$ is distributed according to the single letter marginal of p. Let $\mathcal{N}_1$ be the collection of single letter marginals from $\mathcal{N}_1^\infty$. It is easy to that $\mathcal{U} \subseteq \mathcal{N}$. It is easy to verify that every distribution in $\mathcal{N}_1$ is deceptive.

Example 2. $\mathcal{N}_1^\infty$ is weakly compressible but not *d.w.c.* □

A monotone probability distribution p on $\mathbb{N}$ is one that satisfies $p(y+1) \leq p(y)$ for all $y \in \mathbb{N}$. We will also consider $\mathcal{M}$, the collection of monotone distributions on $\mathbb{N}$ with finite entropy. Let $\mathcal{M}^\infty$ be the set of all *i.i.d.* processes, with one dimensional marginals from $\mathcal{M}$.

Example 3. $\mathcal{M}^\infty$ is weakly compressible but not *d.w.c.* □

Now for $h > 0$, we consider the set $\mathcal{M}_h \subset \mathcal{M}$ of all monotone distributions over $\mathbb{N}$ such that the second moment of the self information,

$$E_p\left(\log\frac{1}{p(X)}\right)^2$$

is bounded above by h. Let $\mathcal{M}_h^\infty$ be the set of all *i.i.d.* loss processes with one dimensional marginals from $\mathcal{M}_h$. Then

Example 4. $\mathcal{M}_h^\infty$ is strongly compressible, hence *d.w.c.* □

In the class $\mathcal{U}$ above, there was a neighborhood around each distribution $p \in \mathcal{U}$ with no other model from $\mathcal{U}$. Hence $\mathcal{U}$ trivially satisfied the local redundancy condition of Theorem 2. The $\mathcal{M}_h$ case falls into another extreme—the entire model class $\mathcal{M}_h$ is strongly compressible, and therefore the conditions of Theorem 2 was satisfied in a trivial way again. The following example illustrates a *d.w.c* class of models where neither extreme holds.

For a distribution p over $\mathbb{N}$ and a number $R \in \mathbb{N}$, let $p^{(R)}(i+R) = p(i)$ for all $i \in \mathbb{N}$. One can visualize $p^{(R)}$ as the distribution p shifted to the right by R. Furthermore let the *span* of any finite support probability distribution over naturals be the largest natural number which has non-zero probability. Then, let

$$\mathcal{F}_h = \big\{(1-\epsilon)p_1 + \epsilon p_2^{(\text{span}(p_1)+1)} : \\ \forall p_1 \in \mathcal{U}, p_2 \in \mathcal{M}_h \text{ and } 1 > \epsilon > 0\big\}.$$

As always $\mathcal{F}_h^\infty$ is the set of measures on infinite sequences formed by *i.i.d.* sampling from distributions in $\mathcal{F}_h$.

Example 5. $\mathcal{F}_h^\infty$ is *d.w.c.* □

V. Necessary part

This side of the characterization follows very naturally from the definition of deceptive measures.

Theorem 3. $\mathcal{P}^\infty$ is *d.w.c* only if no $p \in \mathcal{P}^\infty$ is deceptive.

Proof We prove the contrapositive of the statement above. Namely we show that if some $p \in \mathcal{P}^\infty$ is deceptive, then $\exists \eta > 0$ and $\delta > 0$ such that $\forall$ measures q on infinite sequences of naturals and $\forall$ indicator schemes Φ, there is some $p' \in \mathcal{P}^\infty$ such that $p'((\Phi, q)$ is $\delta-$premature$) > \eta$. To pick η, choose any $\alpha > 0$, and pick $\eta = 1 - \alpha$. Since p is deceptive, we can pick a δ that is > 0 and

$$< \lim_{\epsilon\to 0}\limsup_{n\to\infty}\sup_{p'\in B(p,\epsilon)}\frac{1}{n}D_n(p'||q).$$

The rest of the proof applies for all measures q and all indicator schemes Φ. For all $n \geq 1$, let

$$R_n \stackrel{\text{def}}{=} \{x^n : \Phi(x^n) = 1\}$$

be the set of sequences of length n on which Φ has entered and let $N \geq 4/\alpha$ be a number such that $p(R_N) > 1 - \alpha/2$. Set[1] $\epsilon = \frac{1}{16(\ln 2)N^8}$. Applying Lemma 6 to distributions over length-N sequences induced by p and any $\tilde{p} \in \mathcal{P}^\infty$ such that $\mathcal{J}(p, \tilde{p}) \leq \epsilon$, we have

$$\tilde{p}(R_N) \geq 1 - \alpha/2 - \frac{2}{N} \geq 1 - \alpha. \tag{3}$$

Note that

$$\inf_q \limsup_{n\to\infty}\sup_{p'\in B(p,\epsilon)}\frac{1}{n}D_n(p'||q)$$

is non-increasing with ϵ, and that the limit as $\epsilon \to 0$ is $> \delta$. Therefore, we can choose $n > N$ and $\tilde{p} \in B(p, \epsilon)$ such that

$$\tilde{p}(R_n) \geq 1 - \alpha \text{ and } \frac{1}{n}D_n(\tilde{p}||q) > \delta.$$

This in turn means for the choice of η and δ above, $\tilde{p}((\Phi, q)$ is $\delta-$premature $) > \eta$. Because Φ and q were arbitrary, the theorem follows. □

[1] Please note that in the interest of simplicity, we have not attempted to provide the best scaling for ϵ or the tightest possible bounds in arguments below

VI. Sufficient part

When no $p \in \mathcal{P}^\infty$ is deceptive, we construct a measure q such that given any confidence $\eta > 0$ and accuracy δ, there is a indicator scheme Φ such that (Φ, q) is $\delta-$premature with probability $\leq \eta$.

From Lemma 1, if no $p \in \mathcal{P}^\infty$ is deceptive, there is for each $p \in \mathcal{P}^\infty$ a neighborhood $B(p, \epsilon_p)$ such that for some measure q on infinite sequences of naturals, we have

$$\limsup_{n\to\infty} \sup_{p' \in B(p,\epsilon_p)} \frac{1}{n} D_n(p'||q) < \delta.$$

We pick such a neighborhood $B(p, \epsilon_p)$ for each $p \in \mathcal{P}$ and call it the *reach* of p. The reach of p will play the role of the set of measures in $\mathcal{P}^\infty$ for which it will be okay to eventually set indicators assuming p is in force.

VI-A. Topology of $\mathcal{P}$ with the ℓ_1 metric

To prove that $\mathcal{P}^\infty$ is *d.w.c* if no measure is deceptive, we will need to find a way to cover $\mathcal{P}$ with countably many sets of the form $B(p, \epsilon_p)$ above. Unfortunately, $\mathcal{J}(p, q)$ is not a metric, so it is not immediately clear how to go about doing this. On the other hand note that $\mathcal{J}(p', p) \leq |p - p'|_1/\ln 2$, where $|p - p'|_1$ denotes the ℓ_1 distance between the single letter marginals of p and p' (see Lemma 5 in the Appendix). Therefore, we can instead bootstrap off an understanding of the topology induced on $\mathcal{P}$ by the ℓ_1 metric.

The topology induced on $\mathcal{P}$ by the ℓ_1 metric is Lindelöf, i.e. any covering of $\mathcal{P}$ with open sets in the ℓ_1 topology has a countable subcover (see [4, Defn. 6.4] for definitions and properties of Lindelöf topological spaces). See [1] for the proof of why $\mathcal{P}$ is Lindelöf.

VI-B. Sufficient condition

We now have the machinery required to prove that if no $p \in \mathcal{P}^\infty$ is deceptive, then $\mathcal{P}^\infty$ is *d.w.c.*

Theorem 4. If no $p \in \mathcal{P}$ is deceptive, then $\mathcal{P}^\infty$ is *d.w.c.*
Proof The proof is constructive. For any confidence $0 < \eta < 1$ and accuracy δ, we find a measuer q^* on infinite sequences of naturals and an indicator scheme Φ such that for all $p \in \mathcal{P}^\infty$,

$$p((\Phi, q^*) \text{ is } \delta\text{−premature }) < \eta.$$

Wherever we use ℓ_1 distances $|p - \tau|_1$, it will be understood that we mean the one dimensional marginals of the measures p and τ respectively.

For $p \in \mathcal{P}$, define the following set

$$Q_p = \left\{ \tau : |p - \tau|_1 < \frac{{\epsilon_p}^2 (\ln 2)^2}{16} \right\},$$

where ϵ_p is the reach of p, and τ above is any distribution over $\mathbb{N}$ (not necessarily in $\mathcal{P}$). We will call Q_p as the *zone* of p. The set Q_p is non-empty when $\epsilon_p > 0$.

For large enough n, the set of sequences of length n with empirical distribution in Q_p will ensure that the indicator scheme Φ to be proposed enters with probability 1 when p is in force. Note that if $\epsilon_p > 0$ is small enough then $Q_p \cap \mathcal{P} \subset B(p, \epsilon_p)$—we will assume w.l.o.g. that $\epsilon_p > 0$ is always taken so that $Q_p \cap \mathcal{P} \subset B(p, \epsilon_p)$.

Since no $p \in \mathcal{P}$ is deceptive, none of the zones Q_p are empty and the space $\mathcal{P}$ of distributions can be covered by the sets $Q_p \cap \mathcal{P}$, namely

$$\mathcal{P} = \cup_{p \in \mathcal{P}} (Q_p \cap \mathcal{P}).$$

From Section VI-A, we know that $\mathcal{P}$ is Lindelöf under the ℓ_1 topology. Thus, there is a countable set $\tilde{\mathcal{P}} \subseteq \mathcal{P}$, such that $\mathcal{P}$ is covered by the collection of relatively open sets

$$\{Q_{\tilde{p}} \cap \mathcal{P} : \tilde{p} \in \tilde{\mathcal{P}}\}.$$

We let the above collection be denoted by $\mathcal{Q}_{\tilde{\mathcal{P}}}$. We will refer to $\tilde{\mathcal{P}}$ as the *quantization* of $\mathcal{P}$ and to elements of $\tilde{\mathcal{P}}$ as *centroids* of the quantization, borrowing from commonly used literature in classification.

We index the countable set of centroids, $\tilde{\mathcal{P}}$ (and reuse the index for the corresponding elements of $\mathcal{Q}_{\tilde{\mathcal{P}}}$) by $\iota : \tilde{\mathcal{P}} \to \mathbb{N}$.

a) Description of q^*: For each $\tilde{p} \in \tilde{\mathcal{P}}$, from Lemma 1 we have a measure $\tilde{q}$ on infinite sequences of naturals such that

$$\limsup_{n\to\infty} \sup_{p' \in B(\tilde{p},\epsilon_{\tilde{p}})} \frac{1}{n} D_n(p'||\tilde{q}) < \delta.$$

Let $\iota(\tilde{q})$ be the label assigned to the corresponding $\tilde{p}$ in the above enumeration of $\tilde{\mathcal{P}}$. Then for all sequences $\mathbf{x}$, let

$$q^*(\mathbf{x}) := \sum_{\tilde{q}} \frac{\tilde{q}(\mathbf{x})}{\iota(\tilde{q})(\iota(\tilde{q}) + 1)}$$

Observe again from Lemma 1 and the above quantization that for all $p \in \mathcal{P}^\infty$,

$$\limsup_{n\to\infty} \frac{1}{n} D_n(p||q^*) < \delta.$$

Moreover for all $\tilde{p} \in \tilde{\mathcal{P}}$,

$$\limsup_{n\to\infty} \sup_{p' \in B(\tilde{p},\epsilon_{\tilde{p}})} \frac{1}{n} D_n(p||q^*) < \delta. \qquad (4)$$

We now construct the indicator scheme Φ having the property that for all $p \in \mathcal{P}^\infty$,

$$p((\Phi, q^*) \text{ is } \delta\text{−premature }) < \eta.$$

b) Preliminaries: Consider a length-n sequence x^n on which Φ has not entered thus far. Let the empirical distribution of the sequence be q, and let

$$\mathcal{P}'_\tau := \{p' \in \tilde{\mathcal{P}} : \tau \in Q_{p'}\}$$

be the set of centroids in the quantization of $\mathcal{P}$ (elements of $\tilde{\mathcal{P}}$) which can potentially *capture* τ. Note that τ in general need not belong to $\tilde{\mathcal{P}}$ or $\mathcal{P}$.

If $\mathcal{P}'_\tau \neq \emptyset$, we will further refine the set of distributions that could capture τ further to $\mathcal{P}_\tau \subset \mathcal{P}'_\tau$ as described below. Refining $\mathcal{P}'_\tau$ to $\mathcal{P}_\tau$ ensures that models in $\mathcal{P}'_\tau$ do not $\delta-$prematurely capture sequences.

Let p be the model in force, which remains unknown. The idea is that we want sequences generated by (unknown) p to be captured by those centroids of the quantization $\tilde{\mathcal{P}}$ that have p in their reach. We will require (5) below to ensure that the probability (under the unknown p) of all sequences that may get captured by centroids $p' \in \mathcal{P}_\tau$ not having p in its reach remains small. In addition, we impose (6) as well to resolve a technical issue since τ need not, in general, belong to $\mathcal{P}$.

For $p' \in \mathcal{P}'_\tau$, let the reach of p' be $\epsilon_{_{p'}}$, and define

$$D_{_{p'}} := \frac{{\epsilon_{p'}}^4 (\ln 2)^4}{256} \ .$$

In case the underlying distribution p happens to be out of the reach of p' (wrong capture), the quantity $D_{_{p'}}$ will later lower bound the distance of the empirical τ in question from the underlying p.

Specifically, we place p' in $\mathcal{P}_\tau$ if n satisfies

$$\exp\left(-nD_{_{p'}}/18\right) \le \frac{\eta}{2C(p')\iota(p')^2 n(n+1)}, \qquad (5)$$

and

$$2F_\tau^{-1}(1 - \sqrt{D_{_{p'}}}/6) \le \log C(p'), \qquad (6)$$

where for any $0 < \gamma < 1$, $F_\tau^{-1}(1-\gamma)$ is the $1-\gamma$ percentile of τ as defined in [1]. where $C(p')$ is

$$C(p') := 2^{2\left(\sup_{r \in B(p', \epsilon_{_{p'}})} F_r^{-1}(1 - \sqrt{D_{_{p'}}}/6)\right)}.$$

Note that $C(p')$ is finite from Lemma 7 and because p' is not deceptive. See [1] for why the above equations look this way.

c) Description of Φ*:* For the sequence x^m with type τ, if $\mathcal{P}_\tau = \emptyset$ the scheme does not enter yet. If $\mathcal{P}_\tau \ne \emptyset$, let p_τ denote the distribution in $\mathcal{P}_\tau$ with the smallest index. All sequences with prefix x^m are then said to be *trapped* by p_τ.

From (4),

$$\limsup_{n \to \infty} \sup_{p' \in B(p_q, \epsilon_{p_q})} \frac{1}{n} D_n(p||q^*) < \delta,$$

therefore the set

$$N_{_{p_\tau}} = \{n : \sup_{p' \in B(p_q, \epsilon_{p_q})} \frac{1}{n} D_n(p||q^*) \ge \delta\}$$

is finite. If $m > \max N_{_{p_\tau}}$, we set $\Phi(x^m) = 1$, 0 else.

d) Φ *enters with probability 1:* Please see [3] for proof.

e) Probability Φ $\delta-$*premature* $\le \eta$*:* We now analyze the scheme. Consider any $p \in \mathcal{P}$. Among sequences on which Φ has entered, we will distinguish between those that are in *good* traps and those in *bad* traps. If a sequence x^n is trapped by p' such that $p \in B(p', \epsilon_{_{p'}})$, p' is a good trap. Conversely, if $p \notin B(p', \epsilon_{_{p'}})$, p' is a bad trap.

(Good traps) Suppose a length-n sequence x^n is in a good trap, namely, it is trapped by a distribution p' such that $p \in B(p', \epsilon_{_{p'}})$. In this case, we therefore have

$$p((\Phi, q^*) \text{ is } \delta\text{−premature}) = 0.$$

(Bad traps) We can show that the probability with which sequences generated by p fall into bad traps $\le \eta$ using an argument identical to [1]. Pessimistically, we assume that Φ is $\delta-$premature on every sequence that falls into a bad trap. The theorem follows. □

APPENDIX

The proofs of the following two Lemmas are in [1]

Lemma 5. Let p and q be probability distributions on $\mathbb{N}$. Then

$$\frac{1}{4\ln 2}|p-q|_1^2 \le \mathcal{J}(p,q) \le \frac{1}{\ln 2}|p-q|_1 \ .$$

If, in addition, r is a probability distribution on $\mathbb{N}$, then

$$\mathcal{J}(p,q) + \mathcal{J}(q,r) \ge \mathcal{J}^2(p,r)\frac{\ln 2}{8}. \qquad □$$

Lemma 6. Let p and q be probability distributions on a countable set $\mathcal{A}$ with $\mathcal{J}(p,q) \le \epsilon$. Let p^N and q^N be distributions over $\mathcal{A}^N$ obtained by *i.i.d.* sampling from p and q respectively (the distribution induced by the product measure). For any $R_N \subset \mathcal{A}^N$ and $\alpha > 0$, if $p^N(R_N) \ge 1 - \alpha$, then

$$q^N(R_N) \ge 1 - \alpha - 2N^3\sqrt{4\epsilon \ln 2} - \frac{1}{N}. \qquad □$$

The proof of this lemma is in [5]

Lemma 7. If a class $\mathcal{P}^\infty$ has bounded strong redundancy, then for any $\gamma > 0$

$$\sup_{p \in \mathcal{P}} F_p^{-1}(1-\gamma) < \infty. \qquad □$$

ACKNOWLEDGMENTS

N. Santhanam was supported by NSF Grants CCF-1065632 and CCF-1018984. V. Anantharam was supported by the ARO MURI grant W911NF- 08-1-0233, Tools for the Analysis and Design of Complex Multi-Scale Networks, the NSF grant CNS-0910702, the NSF Science & Technology Center grant CCF-0939370, Science of Information, Marvell Semiconductor Inc., and the U.C. Discovery program. A. Kavcic was supported by NSF Grant CCF-1018984. W. Szpankowski was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSA Grant H98230-11-1-0141, and NSF Grants DMS-0800568, and CCF-0830140.

REFERENCES

[1] N. Santhanam and V. Anantharam, "Agnostic insurance of model classes," *Submitted to Journal of Machine Learning Research*, 2012, full version available from arXiv doc id: 1212:3866.

[2] N. Santhanam, "Probability estimation and compression involving large alphabets," Ph.D. dissertation, University of California, San Diego, 2006.

[3] N. Santhanam, V. Anantharam, A. Kavcic, and W. Szpankowski, "Data driven weak universal redundancy," Submitted for publication, full version available form arXiv.

[4] J. Dugundji, *Topology*. Boston: Allyn and Bacon Inc., 1970.

[5] M. Hosseini and N. Santhanam, "On redundancy of memoryless sources over countable alphabets," Submitted for publication, full version available from arXiv.