Fiftieth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
October 1 - 5, 2012

Universal algorithms: building a case for pointwise convergence

Narayana Santhanam
Dept of ECE,
University of Hawaii
Honolulu, HI 96822.
nsanthan@hawaii.edu

Venkat Anantharam
Dept of EECS,
University of California, Berkeley
Berkeley, CA 94720.
ananth@eecs.berkeley.edu

***Abstract*— We consider algorithms for prediction, compression and entropy estimation in a universal setup. In each case, we estimate some function of an unknown distribution p over the set of natural numbers, using only n observations generated *i.i.d.* from p. While p is unknown, it belongs to a known collection $\mathcal{P}$ of possible models.**

When the supports of distributions in $\mathcal{P}$ are uniformly bounded, consistent algorithms exist for each of the problems. Namely, the convergence of the estimate to the true value can be bounded by a function depending only on the sample size, n, and not on the underlying distribution p. However, when the supports of distributions in $\mathcal{P}$ are not uniformly bounded, a more natural approach involves algorithms that are pointwise consistent, namely, the convergence to the true value is at a rate that depends on both n and the underlying (unknown) distribution p. The obvious practical difficulty with pointwise convergence is that the asymptotic consistency of the algorithm may indicate nothing about the performance of the algorithm for *any* fixed sample size, since the underlying distribution is unknown.

In this paper, we first note that for many complex model classes $\mathcal{P}$, we can still circumvent the above practical difficulty with pointwise convergence. Secondly, we take here a preliminary step towards characterizing a broad framework establishing the hierarchy of difficulty of problems involving pointwise convergence. We look for connections among the following problems which we define for a pointwise convergence scenario: (i) predicting good upper bounds on the next unseen sample, (ii) weak universal compression, and (iii) entropy estimation. We construct counter-examples to show that no two properties above imply the third.

Keywords: entropy estimation, insurance, non-parametric approaches, prediction, strong and weak universal compression.

Many applications require statistical inference when the amount of data available (*sample size*) is comparable to or even smaller than the *alphabet*, the set from which each symbol of the data comes from. Classical asymptotics do not apply here, and these regimes are better captured by considering asymptotics when the alphabet size is unbounded.

We take a non-parametric approach borrowing on the established universal compression framework. Specifically, we have a known class $\mathcal{P}$ of admissible models consisting of distributions over natural numbers $\mathbb{N}$. Put alternatively, given some problem, we make up a model class $\mathcal{P}$ such that it balances the necessary richness on the one hand with any property $\mathcal{P}$ may need for the problem to be solved feasibly on the other. We denote by $\mathcal{P}^\infty$ the class of *i.i.d.* measures on infinite length sequences of numbers whose single letter marginals belong to $\mathcal{P}$. We will abuse notation and use p to also denote the *i.i.d.* measure over infinite sequences whose single letter marginal is $p \in \mathcal{P}$,

Suppose we see a sample $X_1, \ldots, X_n$ generated from an unknown $p \in \mathcal{P}^\infty$. In this paper, as with a lot of universal compression literature, we focus on the following problem—what properties must $\mathcal{P}$ have if we are to solve a certain problem?

To bring focus on further development of ideas, we consider the following three problems using only knowledge of the data $X_1, \ldots, X_n$:

1) (insurability) predicting good upper bounds on X_{n+1} with a specified confidence η,
2) (compression) obtaining a measure q over infinite length sequences of numbers such that for all $p \in \mathcal{P}$, $\frac{1}{n} D(p(X^n)||q(X^n)) \to 0$,
3) (entropy estimation) estimating the entropy of p to a given accuracy ϵ with a specified confidence η.

In each case, we have a natural notion of *uniform* consistency. Suppose we have a sample of size n. For insurability, is it possible to find n independent of what p is, such that all our predicted upper bounds on $X_{n+1}, X_{n+2} \ldots$ hold with probability $\geq 1 - \eta$? Similarly, can the convergence of the KL divergence between p and q to 0 be independent of p and depend only on n? For the third problem, if we are given ϵ and η, can we find n independent of p so that our entropy estimates from time $n+1$ onwards are within the accuracy parameter with confidence $\geq 1 - \eta$? For each property above, if n cannot be independent of p no matter what algorithm is chosen, we say that only *pointwise* consistence is possible for the class $\mathcal{P}$.

When no upper bound is assumed on X_i, uniform

978-1-4673-4539-2/12/$31.00 ©2012 IEEE

consistency is often not possible. The performance of any algorithm may depend on the unknown distribution p. On the one hand, folk wisdom abhors such pointwise convergence. Since we do not know the model p and our algorithm's performance depends on p, it is possible we may never know if we are doing well no matter what the sample size is. Such an algorithm would clearly have no practical relevance. But at the same time, based on a host of applications—risk management (see *e.g.,* [1], [2], [3], [4]), probability estimation (see *e.g.,* [5], [6]), and entropy estimation (see *e.g.,* [7], [8], [9], [10]) being some examples—we want to consider richer classes of models than those where uniform consistency is possible. We consider two broad questions.

A. *Characterizations*

Suppose we have a model class $\mathcal{P}$ which admits pointwise convergence of the required properties. Even when we do not know the true model p, is it possible that we can tell if our algorithm is doing well or not? Such a situation would be a *desirable* pointwise estimation of the said properties. How do we characterize which model classes $\mathcal{P}$ lend themselves to desirable pointwise estimation?

At this point, we first formalize the above notions for the insurance and entropy estimation problems here. No characterization exists for entropy estimation yet, we have a condition on $\mathcal{P}$ that is both necessary and sufficient for it to be insurable. A notion of pointwise convergence (though not necessarily *desirable* in the sense below) for universal compression has been well developed in [11].

1) Insurability: Here we think of every sample point in the data as the aggregate *loss* incurred by the insured parties. We represent the loss at each time by numbers in $\mathbb{N} = \{0, 1, \ldots\}$, and denote the sequence of losses by $X_1, X_2 \ldots$ where $X_i \in \mathbb{N}$. A loss distribution is a distribution over $\mathbb{N}$, and let $\mathcal{P}$ be a set of loss distributions. As before $\mathcal{P}^\infty$ is the collection of *i.i.d.* measures over infinite sequences of symbols from $\mathbb{N}$ such that its single letter marginal belongs to $\mathcal{P}$.

An insurer's *scheme* Φ is a mapping from $\mathbb{N}^* \to \mathbb{R}^+ \cup \{\infty\}$, and is interpreted in terms of the premium demanded by the insurer from the insured after a loss sequence in $\mathbb{N}^*$ is observed. Note that Φ is supposed to work for *all* models in $\mathcal{P}$ and has no information on the underlying distribution other than through the samples from the distribution.

The insurer can observe the loss for any (finite) amount of time prior to entering the insurance game. However, we require the scheme enters the game with probability 1 no matter what loss model $p \in \mathcal{P}$ is in force. The insurer has to keep setting finite premiums from the point it enters. For convenience, we assume $\Phi(x^n) = \infty$ on every sequence x^n of losses on which Φ has not entered. Note that the point at which the insurer enters the game may depend on the loss model in general.

Definition 1. A class $\mathcal{P}^\infty$ of measures is insurable if $\forall\ \eta > 0$, there exists a premium scheme Φ such that $\forall$ $p \in \mathcal{P}^\infty$,

$$p(\Phi(X^n) < \infty \text{ and } \Phi(X^n) < X_{n+1}) < \eta$$

and if, in addition, for all $p \in \mathcal{P}^\infty$,

$$p(\lim_{n\to\infty} \min_{1\le j\le n} \Phi(X^j) < \infty) = 1. \qquad \square$$

Furthermore, if $\Phi(X^n) < \infty$, then $\Phi(X^m) < \infty$ $\forall$ sequences X^m that extend X^n (*i.e.,* contain X^n as a prefix).

2) Entropy estimation: Most approaches for entropy estimation do not consider pointwise convergence in the sense we want, with some exceptions (see, *e.g.,* [12]). As with insurance, we allow for observation of the sequence for any finite time. At some point, the estimator must decide that it can give an answer to accuracy ϵ with confidence $\geq 1 - \eta$ no matter what the underlying source is.

Formally, we enforce this situation by requiring that as long as the estimator is uncertain (given the confidence and accuracy parameters), it indicates so by assigning ∞ as the estimate. An estimate is output if it is determined to be within the confidence and accuracy parameters *no matter what the true distribution is.* In what follows, we will denote the entropy (equivalently entropy rate) of a source $p \in \mathcal{P}$ (equivalently $p \in \mathcal{P}^\infty$) by $H(p)$.

Definition 2. A class $\mathcal{P}^\infty$ of *i.i.d.* measures lends itself to entropy estimation if $\forall\ \eta > 0$ and $\epsilon > 0$, there exists an estimate $\hat{H}(X^n) : X^n \to \mathbb{R}^+ \cup \{\infty\}$ of entropy such that $\forall$ $p \in \mathcal{P}^\infty$

$$p(X^n : \hat{H}(X^n) < \infty \text{ and } |\hat{H}(X^n) - H(p)| > \epsilon) < \eta$$

while

$$p(\lim_{n\to\infty} \min_{1\le j\le n} X^j : \hat{H}(X^j) < \infty) = 1$$

Furthermore, if $\hat{H}(X^n) < \infty$, then $\hat{H}(X^m) < \infty$ $\forall$ sequences X^m that extend X^n. $\square$

Consider the following example where the number of samples needed to estimate the entropy depends on the unknown underlying distribution, yet we can always figure out by looking at the data if we have sufficient accuracy and confidence.

Example 1. Consider $\mathcal{U}$, the collection of all uniform distributions over finite subsets of naturals $\mathbb{N}$. We call X_i a repeated symbol if $X_i \in \{X_1, \ldots, X_{i-1}\}$. From [12], given any accuracy ϵ and confidence η, we can estimate the size of the support (and hence entropy) to within that accuracy and confidence by sampling till we obtain $r(\epsilon, \eta)$ repeated symbols, where $r(\epsilon, \eta) = \mathcal{O}(\log(1/\text{poly}(\epsilon\eta))/\text{poly}(\epsilon))$ is a function of ϵ and η. In a sense, this inverts the Birthday Paradox [13]. $\square$

B. Connections

While this notion of desirable pointwise consistency open up richer collections of models for use, it also poses the question of how the various properties are connected. Indeed, the three properties mentioned thus far—insurability, weak compression and desirable entropy estimation have non-trivial connections. We show, perhaps surprisingly, that no two properties imply the third. Specifically, we construct model classes where two of the three properties can be estimated in a desirable pointwise consistent manner, but not all three. Of significant interest will be a complete characterization of the hierarchy of difficulty of these problems, and of general functions of the unseen samples.

I. Insurability

A. Preliminaries

a) Distance between distributions: Insurability of $\mathcal{P}^\infty$ depends on the neighborhoods of the probability distributions among its single letter marginals $\mathcal{P}$. The relevant "distance" between distributions in $\mathcal{P}$ that decides the neighborhood is

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

b) Cumulative distribution functions: In this section, we phrase the notion of similarity in span in terms of the cumulative distribution function. Note that we are dealing with distributions over a discrete (countable) support, so a few non-standard definitions related to the cumulative distribution functions need to be clarified.

For our purposes the cumulative distribution function of any distribution p is a function from $\mathbb{R} \to [0,1]$, and will be denoted by F_p. We obtain F_p by first defining F_p on points in the support of p and the point at infinity. We define F_p for all other points by linearly interpolating between the values in the support of p.

Let $F_p^{-1}(1)$ be the smallest number y such that $F_p(y) = 1$, and let $F_p^{-1}(x) = 0$ for all $0 \le x < F_p(0)$. If p has infinite support then $F_p^{-1}(1) = \infty$. Note that for $0 \le x \le 1$, $F_p^{-1}(x)$ is now uniquely defined.

B. Characterization

Existence of close distributions with very different spans is what kills insurability. A scheme could be "deceived" by some process $p \in \mathcal{P}^\infty$ into setting low premiums, while a close enough distribution lurks with a high loss. The conditions for insurability of $\mathcal{P}^\infty$ are phrased in terms of its single letter marginals $\mathcal{P}$.

Formally, a distribution p in $\mathcal{P}$ is *deceptive* if $\forall$ neighborhoods $\epsilon_p > 0$, $\exists\ \delta > 0$ so that no matter what $f(\delta) \in \mathbb{R}$ is chosen, $\exists$ a (bad) distribution $q \in \mathcal{P}$ such that

$$\mathcal{J}(p,q) \le \epsilon_p$$

and

$$F_q^{-1}(1-\delta) > f(\delta).$$

In the above definition, $f(\delta)$ is simply an arbitrary number. However, it is useful to think of this number as the evaluation of a function $f : (0,1) \to \mathbb{R}$ at δ, particularly when thinking of the contrapositive of the definition as below. Equivalently, a distribution p in $\mathcal{P}$ is not *deceptive* if $\exists$ neighborhood $\epsilon_p > 0$, such that $\forall\ \delta > 0$, $\exists\ f(\delta) \in \mathbb{R}$, such that all distributions $q \in \mathcal{P}$ with

$$\mathcal{J}(p,q) \le \epsilon_p$$

satisfy

$$F_q^{-1}(1-\delta) \le f(\delta).$$

Theorem 1. $\mathcal{P}^\infty$ is insurable, iff no $p \in \mathcal{P}$ is deceptive.

Proof See [4]. □

II. Universal Compression

Recall that a class $\mathcal{P}^\infty$ of stationary ergodic measures on $\mathbb{N}^\infty$ is defined to be weakly compressible if there is a measure q on $\mathbb{N}^\infty$ that satisfies for all $p \in \mathcal{P}^\infty$,

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}_p \log \frac{1}{q(X^n)} = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_p \log \frac{1}{p(X^n)},$$

where X^n are sequences of natural numbers from p. The term on the right is the entropy rate of p. In particular, it can be shown that the above definition is equivalent to the more commonly used definition from [11], which uses a sequence $q_i : i \ge 1$ of distributions (q_i over length-i sequences) in the left limit. See, *e.g.,* [14], for the connection.

In other words, the expected codelength of length-n sequences using the distribution induced by q converges pointwise to the entropy rate over the class $\mathcal{P}^\infty$. Kieffer proved [11] that $\mathcal{P}^\infty$ is weakly compressible iff there exists a countable set $\mathcal{Q} = \{q_1, q_2, \ldots\}$ of (single letter) distributions over $\mathbb{N}$ such that for all $p \in \mathcal{P}^\infty$ with finite entropy rate, there exists some distribution $q_p \in \mathcal{Q}$ such that

$$\mathbb{E}_p \log \frac{1}{q_p(X_1)} < \infty,$$

where as before, X_1 is a number chosen from the distribution p. The following corollary of Kieffer's condition will be useful for our proofs.

Corollary 2. If class $\mathcal{P}^\infty$ of measures over $\mathbb{N}^\infty$ is weakly compressible, then there exists a distribution q over $\mathbb{N}$ such that for all $p \in \mathcal{P}^\infty$ with finite entropy rate,

$$\mathbb{E}_p \log \frac{1}{q(X_1)} < \infty.$$

Proof See [15]. □

III. Connections

A. Weakly compressible, Desirable Entropy estimation but not insurable

We first consider an example of a model class that is weakly compressible, lends itself to to desirable entropy estimation, but is not insurable.

Let

$$p_{L,\alpha}(X_i) = \begin{cases} 1-\alpha & \text{if } X_i = 0 \\ \alpha & \text{if } X_i = L. \end{cases}$$

For our first example, we consider the set $\mathcal{N}^\infty$ is the class of *i.i.d.* processes whose single letter marginals are in

$$\mathcal{N} = \{p_{L,\alpha} : L \in \mathbb{N}, \alpha \geq 0\}.$$

1) Weakly compressible: Let $q(n) = 1/2^n$. Now, q satisfies for all $p \in \mathcal{N}^\infty$,

$$\sum_{n\geq 1} p(n) \log \frac{1}{q(n)} = \sum_{n\geq 1} np(n) < \infty$$

where the last inequality follows since the summation has just two non-zero terms.

2) Desirable Entropy estimation: Let $p = p_{L,\alpha}$ be the unknown distribution. To estimate entropy, given ϵ and η, we estimate the unknown α by taking $n \geq \frac{2\log\frac{1}{\eta}}{(h^{-1}(\epsilon))^2}$ samples, where $h(\cdot)$ is the binary entropy function. Using the Chernoff bound, our estimate $\hat{\alpha}$ at this step satisfies

$$p\big(|\alpha - \hat{\alpha}| > h^{-1}(\epsilon)\big) < \eta.$$

Since for $1/2 \geq x \geq y \geq 0$, $h(x) - h(y) \leq h(x-y)$, we have with probability $\geq 1-\eta$ on length-n sequences,

$$|h(\alpha) - h(\hat{\alpha})| \leq h(|\alpha - \hat{\alpha}|) \leq \epsilon.$$

Note that the point at which we can estimate the entropy does not even depend on the unknown distribution for this case.

3) Not insurable: Note that the loss measure that puts probability 1 on the all-0 zero sequences exists in $\mathcal{N}^\infty$. Since we consider only schemes that enter with probability 1 no matter what $p \in \mathcal{N}^\infty$ is in force, every insurer must therefore enter after seeing a finite number of zeros. Consider any scheme, and denote the premiums charged at time i by $\Phi(X^i)$.

To show that $\mathcal{N}^\infty$ is not insurable, we show that $\exists \eta > 0$ such that no matter what the scheme Φ, $\exists p \in \mathcal{N}^\infty$ such that

$$p(\ \Phi \text{ goes bankrupt }) \geq \eta.$$

Suppose the scheme enters the game after seeing N losses of size 0. Fix $\delta = 1-\eta$. Let ϵ be small enough that

$$(1-\epsilon)^N > 1 - \delta/2,$$

and let M be a number large enough that

$$(1-\epsilon)^M < \delta/2.$$

Note that since $1-\delta/2 \geq \delta/2$, $N < M$. Let L be greater than any of premiums charged by Φ for the sequences $0^N, 0^{N+1}, \ldots 0^M$.

Note that if $p = p_{L,\epsilon} \in \mathcal{N}$ is in force, the insurer is bankrupted on all sequences that contain loss L in between the N'th and M'th step. The sequences in question have probabilities (under p)

$$(1-\epsilon)^N\epsilon, (1-\epsilon)^{N+1}\epsilon, \ldots, (1-\epsilon)^{N+M-1}$$

and they also form a prefix free set. Therefore, summing up the geometric series and using the assumptions on ϵ above,

$$p(\ \Phi \text{ is bankrupted }) \geq 1 - \delta/2 - \delta/2 = \eta. \qquad \square$$

B. Insurable, Desirable Entropy estimation but not Weakly Compressible

In order to find *i.i.d.* measures that are insurable but not weakly compressible, we construct a class $\mathcal{I}$ of distributions over $\mathbb{N}$. As with other classes, $\mathcal{I}^\infty$ is the set of *i.i.d.* measures formed whose single letter marginals are $\mathcal{I}$.

To do so, first partition the set of natural numbers into the sets T_k, where $T_1 = [2] = \{1,2\}$ and

$$T_k = [2^{k+1}-2] - \cup_{j=1}^{k-1} T_j \text{ for } k \geq 2.$$

Note that $|T_k| = 2^k$. Now, $\mathcal{I}$ is the collection of all possible distributions that can be formed as follows. For all $i \geq 1$, we pick exactly one element of T_i and assign it probability $6/(\pi^2 i^2)$. Note that $\mathcal{I}$ is not countable. Part of the rationale behind this construction is that for all $p \in \mathcal{I}$,

$$\sum_{n \geq 2^k - 2} p(n) = \frac{6}{\pi^2} \sum_{i \geq k} \frac{1}{i^2},$$

namely, all tails are uniformly bounded over the class $\mathcal{I}$ to ensure insurability.

1) Insurable: Put another way, for all $\delta > 0$ and all distributions $p \in \mathcal{I}$,

$$F_p^{-1}(1-\delta) \leq 2^{k(\delta)} - 2$$

where $k(\delta)$ is the smallest number such that

$$\delta > \frac{6}{\pi^2} \sum_{k(\delta)}^{\infty} \frac{1}{i^2}.$$

The set $\mathcal{I}^\infty$ of measures is insurable.

2) Desirable Entropy estimation: The set $\mathcal{I}^\infty$ of measures lends itself to desirable entropy estimation since every distribution in $\mathcal{I}$ has the same entropy.

3) Not weakly compressible: On the other hand, $\mathcal{I}^\infty$ is *not* weakly compressible.

Suppose q is any distribution over $\mathbb{N}$. We will show that $\exists p \in \mathcal{I}$ such that

$$\sum_{n \geq 1} p(n) \log \frac{1}{q(n)}$$

is not finite. Using the contrapositive of Corollary 2, we conclude that $\mathcal{I}^\infty$ is not weakly compressible.

Consider any distribution q over $\mathbb{N}$. Observe that for all i, $|T_i| = 2^i$. It follows that for all i there is $x_i \in T_i$ such that

$$q(x_i) \le \frac{1}{2^i}.$$

But by construction, $\mathcal{I}$ contains a distribution p that assigns to each x_i above the probability

$$p(x_i) = \frac{6}{\pi^2 i^2}.$$

Note that the KL divergence from p to q is not finite.

C. Insurable, Weakly Compressible, no Desirable Entropy Estimation

Let p_1 assign probability 1 to 0, and for $n \ge 2$, let p_n be a distribution over $\mathbb{N}$ that assigns probabilities

$$p_n(i) = \begin{cases} 1 - \frac{1}{n} & \text{if } i = 0 \\ \frac{1}{n2^{n^2}} & \text{if } 1 \le i \le 2^{n^2}. \end{cases}$$

Let $\mathcal{J} = \{p_n : n \ge 1\}$, and let $\mathcal{J}^\infty$ be the collection of *i.i.d.* measures on $\mathbb{N}^\infty$ whose single letter marginals are in $\mathcal{J}$.

1) Insurable: Like with the set $\mathcal{I}$ above, the $1 - \delta$ percentile of any distribution in $\mathcal{J}$ can be bounded easily. Note that if $X \sim p_n$ and $2^{n^2} \ge 2^{1/\delta^2}$, then

$$p_n(X > 2^{1/\delta^2}) < p_n(X > 0) = \frac{1}{n} \le \delta$$

If $X \sim p_n$ and $2^{n^2} \le 2^{1/\delta^2}$, then $p_n(X > 2^{1/\delta^2}) = 0$. Thus, no matter what $q \in \mathcal{J}$ we consider,

$$F_q^{-1}(1 - \delta) \le 2^{1/\delta^2}.$$

Using Theorem 1, we conclude $\mathcal{J}^\infty$ is insurable.

2) Weakly compressible: This is automatic from [11] since the class $\mathcal{J}$ itself is countable.

3) No desirable entropy estimation: Fix any ϵ and η, and consider any estimator $\hat{H}(X^n)$. Since p_1 is a measure that assigns probability 1 to the all zero sequence, clearly $|\hat{H}(x^n)|$ must be $\le \epsilon$ for some sequence of zeros with length $N(p_1)$.

It is easy to see that $H(p_m) \ge m$. Since we can find $m \ge 2$ so large that

$$p_m(0^{N(p_1)}) \ge 1 - \eta,$$

we have

$$p_m\Big(|H(p_m) - \hat{H}(X^n)| > \epsilon\Big) \ge \eta.$$

Acknowledgments

N. Santhanam was supported by NSF Grants CCF-1065632, CCF-1018984 and ECCS-1029081. 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.

References

[1] H. Cramer. Historical Review of Filip Lundberg's Work on Risk Theory. *Skandinavisk Aktuarietidskrift (Suppl.)*, 52:6–12, 1969. Reprinted in The Collected Works of Harald Cramér edited by Anders Martin-Löf, 2 volumes Springer 1994.

[2] S. Asmussen and H. Albrecher. *Ruin probabilities*. World Scientific Publishing Company, 2nd edition edition, 2010.

[3] N. Santhanam and V. Anantharam. What risks lead to ruin? In *Annual Allerton Conference on Communication, Control, and Computing*, 2010.

[4] N. Santhanam and V. Anantharam. Prediction over countable alphabets. In *Conference on Information Sciences and Systems*, 2012.

[5] I.J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237—264, December 1953.

[6] A. Orlitsky, N.P. Santhanam, and J. Zhang. Always Good Turing: Asymptotically optimal probability estimation. In *Proceedings of the 44th Annual Symposium on Foundations of Computer Sciece*, October 2003.

[7] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. In *Proceedings of the 34th Annual Symposium of the Theory of Computing*, Mar 2002.

[8] I. Nemenman, W. Bialek, and R. Steveninck. Entropy and information in neural spike trains: progress on the sampling problem. *Physical Review E*, 69, 2004.

[9] L. Paninski. Estimating entropy of m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50(9):2200—2203, September 2004.

[10] G. Valiant and P. Valiant. The power of linear estimators. In *Annual Symposium on Foundations of Computer Sciece*, 2011.

[11] J.C. Kieffer. A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, 24(6):674—682, November 1978.

[12] A. Orlitsky, N.P. Santhanam, and K. Viswanathan. Population estimation with performance guarantees. In *Proceedings of IEEE Symposium on Information Theory*, 2007.

[13] Birthday problem. `http://en.wikipedia.org/wiki/Birthday%5Fproblem`.

[14] Narayana Santhanam. *Probability estimation and compression involving large alphabets*. PhD thesis, University of California, San Diego, 2006.

[15] N. Santhanam and V. Anantharam. Agnostic insurance tasks and their relation to compression. In *International conference on signal processing and communications (SPCOM)*, 2012.