

**The following resources related to this article are available online at [www.sciencemag.org](http://www.sciencemag.org) (this information is current as of October 29, 2009 ):**

**Updated information and services**, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/302/5644/427>

**Supporting Online Material** can be found at:

<http://www.sciencemag.org/cgi/content/full/302/5644/427/DC1>

This article **cites 13 articles**, 1 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/302/5644/427#otherarticles>

This article has been **cited by** 18 article(s) on the ISI Web of Science.

This article has been **cited by** 1 articles hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/302/5644/427#otherarticles>

This article appears in the following **subject collections**:

Computers, Mathematics

[http://www.sciencemag.org/cgi/collection/comp\\_math](http://www.sciencemag.org/cgi/collection/comp_math)

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:

<http://www.sciencemag.org/about/permissions.dtl>

better agreement, four broad diffraction peaks alone are insufficient for uniquely constraining the low-symmetry structure and bond lengths. Nevertheless, x-ray diffraction data provide additional information supporting the conclusion obtained from the IXS observation of  $\pi$ - $\sigma$  bonding change that the cold-compressed graphite is a new, distinct phase of carbon.

We observed exceptional hardness in the new phase, as indicated by the broadening of the ruby fluorescence lines. After releasing pressure from the high-pressure phase (without He pressure medium), the graphite sample left a ring crack indentation (9) on the diamond anvils following the original boundary of the sample in the gasket (Fig. 3). In normal DAC operation, ring cracks have only been observed when a diamond anvil is indented by another superhard material, such as an opposing beveled diamond anvil. The occurrence of ring cracks indicates that the

high-pressure form of graphite is harder than the strong materials commonly used in a DAC, e.g., rhenium gaskets, ruby crystals, and refractory oxides. The reversible, orders-of-magnitude change in strength from very soft graphite to superhard materials offers a possibility for intriguing applications as a pressure-dependent structural component (for instance, a composite gasket for high-pressure apparatus) (9).

#### References and Notes

1. E. D. Miller, D. C. Nesting, J. V. Badding, *Chem. Mater.* **9**, 18 (1997).
2. E. P. Bundy, J. S. Kasper, *J. Chem. Phys.* **46**, 3437 (1967).
3. M. Hanfland, K. Syassen, *Phys. Rev. B* **40**, 1951 (1989).
4. A. F. Goncharov, I. N. Makarenko, S. M. Stishov, *Sov. Phys. JETP* **69**, 380 (1989).
5. M. Hanfland, H. Beister, K. Syassen, *Phys. Rev. B* **39**, 12598 (1989).
6. W. Utsumi, T. Yagi, *Science* **252**, 1542 (1991).
7. T. Yagi, W. Utsumi, M. Yamakata, T. Kikegawa, O. Shimomura, *Phys. Rev. B* **46**, 6031 (1992).
8. Y. Zhao, I. L. Spain, *Phys. Rev. B* **40**, 993 (1989).

9. F. P. Bundy *et al.*, *Carbon* **34**, 141 (1996).
10. A. F. Goncharov, *High Pressure Res.* **8**, 607 (1992).
11. J. Xu, H. K. Mao, R. J. Hemley, *J. Phys. Condens. Matter* **14**, 11549 (2002).
12. C.-S. Zha, H. K. Mao, R. J. Hemley, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13494 (2000).
13. Materials and methods are available as supporting material on Science Online.
14. G. D. Cody, H. Ade, S. Wirick, G. D. Mitchell, A. Davis, *Org. Geochem.* **28**, 441 (1998).
15. J. Sung, *J. Mater. Sci.* **35**, 6041 (2000).
16. K. Takemura, *J. Appl. Phys.* **89**, 662 (2001).
17. We thank GSECARS, APS, and ANL for beam time and V. Prakapenka for help with x-ray diffraction. Use of the HPCAT facility was supported by U.S. Department of Energy (DOE)—Basic Energy Sciences, DOE—National Nuclear Security Administration, NSF, Department of Defense—Tank-Automotive and Armaments Command, and the W. M. Keck Foundation.

#### Supporting Online Material

[www.sciencemag.org/cgi/content/full/302/5644/425/DC1](http://www.sciencemag.org/cgi/content/full/302/5644/425/DC1)  
Materials and Methods  
Figs. S1 and S2

28 July 2003; accepted 9 September 2003

## Always Good Turing: Asymptotically Optimal Probability Estimation

Alon Orlitsky,<sup>1,2\*</sup> Narayana P. Santhanam,<sup>1</sup> Junan Zhang<sup>1</sup>

While deciphering the Enigma code, Good and Turing derived an unintuitive, yet effective, formula for estimating a probability distribution from a sample of data. We define the attenuation of a probability estimator as the largest possible ratio between the per-symbol probability assigned to an arbitrarily long sequence by any distribution, and the corresponding probability assigned by the estimator. We show that some common estimators have infinite attenuation and that the attenuation of the Good-Turing estimator is low, yet greater than 1. We then derive an estimator whose attenuation is 1; that is, asymptotically it does not underestimate the probability of any sequence.

In preparation for your next safari, you observe a random sample of African animals. You find three giraffes, one zebra, and two elephants. How would you estimate the probability of the various species you may encounter on your trip? A naïve, empirical-frequency, estimator may assign probability 1/2 to giraffes, 1/6 to zebras, and 1/3 to elephants. But the poor estimator will be completely unprepared for an encounter with an offended lion.

To address this unseen-elements problem, Laplace (1) proposed adding 1 to the count of each species, including to the collection of unseen ones, thereby assigning probability  $(3 + 1)/10 = 0.4$  to giraffes,  $(1 + 1)/10 = 0.2$  to

zebras,  $(2 + 1)/10 = 0.3$  to elephants, and  $(0 + 1)/10 = 0.1$  to unseen species. The Laplace and other add-constant estimators have since been applied and studied extensively. In particular, the add-half, or Krichevski-Trofimov (2), estimator was shown to possess certain optimality properties when the number of possible elements is fixed and the sample size increases to infinity (3, 4).

However, when the number of possible elements is large relative to the sample size, add-constant estimators are lacking (5). Suppose that during your safari trip you evaluate the distribution of animals' DNA sequences. You observe a large number  $n$  of animals and, predictably, find that each has a unique DNA sequence. You therefore have a sample of  $n$  sequences, each observed once, from which you would like to estimate the distribution of all sequences. An add- $c$  estimator would assign probability  $(1 + c)/(n + nc + c)$  to each observed sequence and probability  $c/(n + nc + c)$  to all unseen ones. It follows that the

probability  $(n + nc)/(n + nc + c)$  assigned to all observed sequences is close to 1, whereas that assigned to all unseen sequences is close to 0. Clearly, the opposite better represents the truth.

Good and Turing encountered this problem while trying to break the Enigma cipher during the Second World War (6). British intelligence was in possession of the *Kenngruppenbuch*, the German cipher book that contained all possible secret keys, and used previously decrypted messages to document the page numbers of keys used by various U-boat commanders. They wanted to use this information to estimate the distributions of pages that each U-boat commander picked secret keys from.

Good and Turing came up with a surprising estimator that bears little resemblance to either the empirical-frequency or the add-constant estimators above. After the war, Good published the estimator (7), mentioning that Turing had an "intuitive demonstration" for it but not describing what this intuition was.

The Good-Turing estimator has since been incorporated into a variety of applications such as information retrieval (8), spelling correction (9), and speech recognition [e.g., (10, 11)], where it is applied to estimate the probability distribution of words. Although the Good-Turing estimator performs well in general, it is suboptimal for elements that appear frequently, and hence it was modified in subsequent estimators [e.g., the Jelinek-Mercer, Katz, Witten-Bell, and Kneser-Ney estimators (11)].

On the theoretical side, interpretations of the Good-Turing estimator have been proposed (12–14), and its convergence rate was analyzed (15). Yet, lacking a measure for assessing the performance of an estimator, no objective evaluation or optimality

<sup>1</sup>Department of Electrical and Computer Engineering,

<sup>2</sup>Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093, USA.

\*To whom correspondence should be addressed. E-mail: alon@ucsd.edu

REPORTS

results for the Good-Turing estimator have been established.

To evaluate the performance of an estimator, we apply it not just once but repeatedly to a sequence of elements, all drawn according to the same underlying distribution. Before each element is revealed, we use the estimator to evaluate its conditional probability given the previous elements. Multiplying the conditional probability estimates together, we obtain the probability that the estimator assigns to the whole sequence.

The sequence attenuation of the estimator for a given sequence is then defined as the ratio between the highest probability assigned to the sequence by any distribution (including the one underlying the data) and the probability the estimator assigns to it. The symbol attenuation of the estimator for a length- $n$  sequence is the  $n$ th root of its sequence attenuation, namely, the ratio between the highest per-symbol probability assigned to the sequence by any distribution, and that assigned by the estimator. Finally, the (asymptotic, symbol) attenuation of the estimator is the highest symbol attenuation maximized over all sequences of increasing length.

This measure is similar to one used to evaluate estimators of distributions when the alphabet, the set of possible elements, is small and known (16, 17). Such a measure is used in a variety of fields, including universal compression [e.g., (18–20)], finance [e.g., (21)], online algorithms, and learning [e.g., (22–24)]. To extend it to unknown, potentially large, even infinite “alphabets,” we abstract the actual symbols that appear in the sequence and consider only their pattern, the order in which they appear. This allows us to enumerate sequences over infinite alphabets and to calculate their best probability assignment.

Every estimator corresponds to a probability distribution over sequences of any given length. Hence, the attenuation of any estimator is always at least 1. Because the number of distinct symbols in a sequence can be as large as its length, the sequence attenuation of an estimator on a length- $n$  sequence can be superexponential; hence, its attenuation can be infinite.

The attenuation of a constant  $c > 1$  implies that the estimator assigns to each  $n$ -symbol sequence a probability that is at most a factor of  $c^n$  lower than its best probability. An attenuation of 1, which we call diminishing attenuation, implies that the estimator assigns to each sequence a probability that is at most subexponentially smaller than the best possible, hence the per-symbol probability assigned by the estimator is asymptotically the best possible.

Our objective is to evaluate the performance of existing estimators and to construct diminishing-attenuation estimators. We show

that add-constant estimators have infinite attenuation and that the Good-Turing estimator performs well in the sense that its attenuation is low; however, for some sequences it assigns a probability that is exponentially smaller than the best possible, hence its attenuation is strictly  $> 1$ . We then derive two diminishing-attenuation estimators. The first is computationally more efficient and requires only a constant number of operations per symbol. The second is more complex, but its attenuation approaches 1 more quickly. To determine the estimators’ attenuations, we use potential functions and results of Hardy and Ramanujan (25) on the number of partitions of an integer.

To better understand diminishing-attenuation estimators, we study the probability that the low-complexity estimator assigns to some simple sequences. For some sequences (e.g., those where the symbols are all the same or all different), its estimate agrees with intuition. Yet for other sequences, such as where every symbol appears twice, its estimate differs from what we would intuitively expect.

To formally characterize the various concepts, we need some definitions. A sample is a sequence of elements. An estimator associates with every sample a probability distribution over the set of elements in the sample, plus “new.” For example, after observing the sample

giraffe, zebra, giraffe, elephant, elephant, giraffe,

an estimator postulates a distribution over the set {giraffe, zebra, elephant, “new”}, reflecting the probability that a randomly chosen element is any one of these animals, or “new,” which encompasses all unseen elements.

Because we assume no prior knowledge of the elements in the sample, a giraffe is no different to us from an elephant; hence, we replace the name of each animal by the order in which it appears. The sequence above can thus be expressed as 121331, which we call the pattern of the original sequence. The pattern of a sequence  $\bar{x} = x_1 x_2 \dots x_n$  is denoted by  $\Psi(\bar{x})$ .

A string of positive integers is the pattern of some sequence if and only if the first appearance of any  $i \geq 1$  precedes that of  $i + 1$ . For example, the empty string  $\Lambda$  and the strings 1, 12, and 121 are patterns (of the empty string and, say, “a,” “ad,” and “ada,” respectively), whereas 2, 21, and 132 are not.

Let  $\Psi^n$  denote the set of length- $n$  patterns. For example,  $\Psi^0 = \{\Lambda\}$ ,  $\Psi^1 = \{1\}$ ,  $\Psi^2 = \{11, 12\}$ , and  $\Psi^3 = \{111, 112, 121, 122, 123\}$ . It can be shown that every length- $n$  pattern corresponds to a partition of a set of cardinality  $n$ , hence  $|\Psi^n|$  is the  $n$ th Bell number.

Let  $\mathbb{A}^n$  denote the set of  $n$ -element sequences over an alphabet  $\mathbb{A}$ . For example,  $\{a,$

$b\}^2 = \{aa, ab, ba, bb\}$ . If  $p$  is a probability distribution over an alphabet  $\mathbb{A}$ , then for every  $n \in \mathbb{Z}^+ = \{1, 2, \dots\}$ ,  $p$  induces a probability distribution  $p^\Psi$  over  $\Psi^n$ , where

$$p^\Psi(\bar{\Psi}) \stackrel{\text{def}}{=} p\{\bar{x} \in \mathbb{A}^n : \Psi(\bar{x}) = \bar{\Psi}\} \quad (1)$$

denotes the probability that a sequence of elements, each selected according to  $p$ , will form the pattern  $\bar{\Psi} \in \Psi^n$ . For example, for any probability  $p$  over an alphabet  $\mathbb{A}$ ,  $p^\Psi(1) = p(\mathbb{A}) = 1$ , indicating that the first element of any pattern is 1 (“new”). If  $p$  is a distribution over  $\{a, b\}$  where  $p(a) = p$  and  $p(b) = 1 - p \stackrel{\text{def}}{=} \bar{p}$ , then  $p^\Psi(11) = p\{aa, bb\} = p^2 + \bar{p}^2$ , the probability that two elements will be identical, and  $p^\Psi(12) = p\{ab, ba\} = 2p\bar{p}$ , the probability that the two elements will be distinct.

Continuous (i.e., nonatomic) distributions induce probabilities over patterns as well. For example, if  $p$  is any continuous distribution, then for all  $n$ ,  $p^\Psi(1 \dots n) = p\{x_1 \dots x_n : x_i \neq x_j\} = 1$ , indicating that, with probability 1, a finite number of elements selected according to a continuous distribution are all distinct. It follows that for continuous distributions, every  $\bar{\Psi} \in \Psi^n - \{1 \dots n\}$ —that is, every pattern with repetitions—has  $p^\Psi(\bar{\Psi}) = 0$ .

Our goal is to derive an estimator that, though unaware of the underlying probability  $p$ , assigns to every pattern  $\bar{\Psi}$  a probability that is not much smaller than the induced probability  $p^\Psi(\bar{\Psi})$ . Because we do not know the underlying distribution, we consider the one that assigns to  $\bar{\Psi}$  the highest probability. The maximum probability of a pattern  $\bar{\Psi}$  is

$$\hat{p}^\Psi(\bar{\Psi}) \stackrel{\text{def}}{=} \max_p p^\Psi(\bar{\Psi}) \quad (2)$$

the highest probability assigned to the pattern by any distribution. For example, because any distribution  $p$  has  $p^\Psi(1) = 1$ , we have  $\hat{p}^\Psi(1) = 1$ . Because any distribution  $p$  concentrated on a single element has  $p^\Psi(1 \dots 1) = 1$  for any number of 1’s, we obtain  $\hat{p}^\Psi(1 \dots 1) = 1$ , and, because any continuous distribution  $p$  has  $p^\Psi(1 \dots n) = 1$ , we derive  $\hat{p}^\Psi(1 \dots n) = 1$ . In general, however, it is difficult to determine the maximum probability of a pattern. For example, some work (26) is needed to show that  $\hat{p}^\Psi(112) = 1/2$ .

Let  $m \stackrel{\text{def}}{=} m(\Psi^n) \stackrel{\text{def}}{=} |\{\psi_1 \dots \psi_n\}|$  be the number of distinct symbols appearing in a pattern  $\psi_1^n = \psi_1 \dots \psi_n \in \Psi^n$ . An estimator is a mapping  $q$  that associates with every pattern  $\psi_1^n$  a probability distribution  $q(\psi_{n+1} | \psi_1^n)$  over  $[m + 1] = \{1, \dots, m + 1\}$ , representing the probability that the estimator assigns to the possible values of  $\psi_{n+1}$ , after seeing  $\psi_1^n$ . For example,  $q(\psi_1) \stackrel{\text{def}}{=} q(\psi_1 | \Lambda)$  is a distribution over  $\{1\}$ , namely,  $q(1 | \Lambda) = 1$ , whereas  $q(\psi_3 | 12)$  and  $q(\psi_4 | 121)$  are distributions over  $\{1, 2, 3\}$ .

For a simple example, consider the add-one estimator mentioned earlier and henceforth denoted  $q_{+1}$ . After observing the pattern  $\psi_1^n$ , it assigns to any  $\psi_{n+1} \in [m + 1]$  a probability proportional to one more than the number of times  $\psi_{n+1}$  appeared in  $\psi_1^n$ . For instance, after observing the pattern 1, it estimates  $q_{+1}(1|1) = (1 + 1)/3 = 2/3$  and  $q_{+1}(2|1) = (0 + 1)/3 = 1/3$ .

For each  $n \in \mathbb{Z}^+$ , an estimator  $q$  induces a probability distribution over  $\Psi^n$  given by

$$q(\psi_1^n) = \prod_{i=0}^{n-1} q(\psi_{i+1} | \psi_i) \quad (3)$$

For example, the probability that the add-one estimator ascribes to the pattern 1213 is

$$\begin{aligned} q_{+1}(1213) &= q_{+1}(1|\Lambda) \cdot q_{+1}(2|1) \cdot q_{+1}(1|12) \cdot q_{+1}(3|121) \\ &= \frac{1}{1} \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{1}{6} = \frac{1}{45} \end{aligned} \quad (4)$$

Although for mathematical convenience we define the estimators and present the results in terms of patterns, they apply to the actual underlying sequences. For example, for the sequence  $g, z, g, e$ , the add-one estimator associates the probability

$$\begin{aligned} q_{+1}(\text{new}) \cdot q_{+1}(\text{new}|g) \cdot q_{+1}(g|g,z) \\ \cdot q_{+1}(\text{new}|g,z,g) \\ = \frac{1}{1} \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{1}{6} = \frac{1}{45} \end{aligned} \quad (5)$$

—the same probability it associates with its pattern 1213.

We would like to derive estimators that assign to every pattern a probability that is not much lower than the highest probability assigned to it by any distribution. We therefore define the sequence attenuation of an estimator  $q$  for a pattern  $\psi_1^n$  to be

$$R(q, \psi_1^n) \stackrel{\text{def}}{=} \frac{\hat{p}^\Psi(\psi_1^n)}{q(\psi_1^n)} \quad (6)$$

the ratio between the highest probability assigned to  $\psi_1^n$  by any distribution and the probability assigned to it by  $q$ . The worst-case sequence attenuation of  $q$  for length- $n$  patterns is

$$R^n(q) \stackrel{\text{def}}{=} \max_{\psi_1^n \in \Psi^n} R(q, \psi_1^n) \quad (7)$$

the largest sequence attenuation of  $q$  for any length- $n$  pattern. Note that  $[R^n(q)]^{1/n}$  is the worst-case symbol attenuation of  $q$  for length- $n$  patterns, namely, the largest possible ratio between the per-symbol probability assigned by any distribution to symbols of length- $n$  patterns and the corresponding probability assigned by  $q$ . Finally, the (asymptotic, worst-case, symbol) attenuation of  $q$  is

$$R^*(q) \stackrel{\text{def}}{=} \limsup_{n \rightarrow \infty} [R^n(q)]^{1/n} \quad (8)$$

the largest possible ratio between the per-symbol probability assigned to any asymptotically long pattern by any distribution, and the corresponding probability assigned by  $q$ .

As mentioned earlier,  $R^*(q) \geq 1$  for every estimator  $q$ . If  $R^*(q) > 1$ , then  $q$  assigns to some length- $n$  pattern a probability that is smaller than its highest possible probability by a factor of  $[R^*(q)]^n$ , and if  $R^*(q) = 1$ , then the probability that  $q$  assigns to every length- $n$  pattern is at most subexponentially smaller than the highest possible.

A few notes on the attenuation measure and the distribution assumptions are in order. Attenuation is a useful measure when underestimation of a probability entails a loss. Decisions based on low-attenuation estimations are then guaranteed to perform almost as well as those made with knowledge of the underlying probability.

Note, however, that because any estimator corresponds to a probability distribution, the overestimation of some event probabilities is inevitable. Low sequence attenuation is only weakly linked to low overestimation. Hence, if overestimation is a major issue, different estimators, optimizing another performance measure, should be sought.

Even when restricted to attenuation, different measures can be considered. The one addressed here requires the attenuation of every sequence to diminish. Alternatively, one could require the attenuation of all, or most, sequences, generated by any distribution to diminish at a rate that may depend on that distribution. Because the measure we consider is more stringent, the diminishing-attenuation estimators we present perform well according to these measures as well. However, estimators designed to optimize these measures may perform better on typical sequences or on the average.

Also, we assume that the observations are sampled (“with replacement”) from an underlying distribution. The observed elements are therefore independent and identically distributed. Extensions to sequences drawn from distributions with memory are clearly warranted.

We evaluate the attenuation of several common estimators. Add-constant estimators have unbounded attenuation. Consider, for example, the add-one estimator  $q_{+1}$ . To the pattern 1, 2, 3, . . . ,  $n$  it assigns probability  $q_{+1}(123 \dots n)$

$$\begin{aligned} &= \frac{1}{1} \cdot \frac{1}{3} \cdot \dots \cdot \frac{1}{2n+1} \\ &= \frac{2^n \cdot n!}{(2n+1)!} \end{aligned} \quad (9)$$

Because, as we saw,

$$\hat{p}^\Psi(12 \dots n) = 1 \quad (10)$$

we obtain that  $q_{+1}$  has symbol attenuation

$$[R^n(q_{+1})]^{1/n} \geq \frac{(2n+1)!}{2^n \cdot n!} \geq \frac{2n}{e} \quad (11)$$

hence

**Theorem 1:**  $R^*(q_{+1}) = \infty$ .

By applying the add-one estimator in two steps, we obtain a modified add-one estimator  $q_{+1}'$ , whose attenuation is between 1.69 and 2.85. The estimator uses the add-one rule to estimate the probability of the next symbol being new or repeated, and for repeated symbols it assigns a probability proportional to the number of occurrences of the symbol.

Recall that  $m$  is the number of distinct symbols appearing in the pattern  $\psi_1^n$ . The multiplicity  $\mu_\psi$  of  $\psi \in \mathbb{Z}^+$  in  $\psi_1^n$  is the number of times  $\psi$  appears in  $\psi_1^n$ . Then  $q_{+1}'$  assigns to each  $1 \leq \psi_{n+1} \leq m + 1$  the probability

$$q_{+1}'(\psi_{n+1} | \psi_1^n) \stackrel{\text{def}}{=} \begin{cases} \frac{m+1}{n+2} & \psi_{n+1} = m + 1 \\ \frac{n-m+1}{n+2} \cdot \frac{\mu_{\psi_{n+1}}}{n} & 1 \leq \psi_{n+1} \leq m \end{cases} \quad (12)$$

It can be shown that for sequences where the number of distinct symbols is a vanishing fraction of the sequence length, namely,  $m = o(n)$ , the modified add-one estimator has subexponential sequence attenuation, hence diminishing symbol attenuation. However, sequences with more symbols may have an exponential sequence attenuation. For example, to the pattern

$$\bar{\psi} \stackrel{\text{def}}{=} 12 \dots \frac{n}{2} 12 \dots \frac{n}{2} \quad (13)$$

$q_{+1}'$  assigns probability  $\approx 0.58^n n^{-n/2}$ , whereas the uniform distribution over an alphabet of size 0.628 $n$  assigns probability  $0.98^n n^{-n/2}$ . Therefore,  $R^*(q_{+1}') > 1.69$ . The attenuation of  $q_{+1}'$  is at most 2.85 (27), hence

**Theorem 2:**  $1.69 < R^*(q_{+1}') \leq 2.85$ .

We show that the attenuation of the Good-Turing estimator is a constant between 1.39 and 2. The prevalence  $\varphi_\mu \stackrel{\text{def}}{=} \varphi_\mu(\psi_1^n)$  of  $\mu \in \mathbb{Z}^+$  is the number of symbols appearing  $\mu$  times in  $\psi_1^n$ . Given  $\psi_1^{n+1}$ , let  $r \stackrel{\text{def}}{=} \mu_{\psi_{n+1}}(\psi_1^n)$  be the number of times  $\psi_{n+1}$  has appeared in  $\psi_1^n$ . The Good-Turing estimator (7) is then defined by

$$q(\psi_{n+1} | \psi_1^n) = \begin{cases} \frac{\varphi'_1}{n} & r = 0 \\ \frac{r+1}{n} \cdot \frac{\varphi'_{r+1}}{\varphi'_r} & r \geq 1 \end{cases} \quad (14)$$

where  $\varphi'_\mu$  is a smoothed value of  $\varphi_\mu$ . Smoothing is needed for a variety of reasons (7). One

REPORTS

of them is that if  $\varphi_{\mu+1} = 0$ , then without smoothing the estimator would assign  $q(\psi_{n+1} | \psi_1^n) = 0$  for the symbols appearing  $\mu$  times in  $\psi_1^n$ .

Many smoothing methods have been proposed; some seem too difficult to analyze. We considered three, all of which had attenuation of  $>1$ . Perhaps the simplest smoothed Good-Turing estimator,  $q_{GT}$ , is defined by

$$\varphi'_\mu = \max(\varphi_\mu, 1) \tag{15}$$

which ensures nonzero probabilities for all symbols in  $\{1, m(\psi_1^n) + 1\}$ . The sequence attenuation of  $q_{GT}$  is always at most  $2^n$ , hence its attenuation is at most 2 [see potential-function proof in (27)]. Yet to the pattern

$$\bar{\psi} \stackrel{\text{def}}{=} 12(132)^{n/3} \stackrel{\text{def}}{=} 12132132 \dots 132 \tag{16}$$

$q_{GT}$  assigns probability  $\Theta(72^{-n/3})$ , whereas the maximum probability of  $\bar{\psi}$  can be shown to be  $\Theta(3^{-n})$ . Hence,

**Theorem 3:**  $1.39 < R^*(q_{GT}) \leq 2$ .

The other two variants of Good-Turing estimators we considered were the “simple Good Turing” estimator developed by Gale (28), and an estimator that uses Good-Turing to predict the probability of infrequent symbols and empirical frequency for frequent symbols. These two estimators seem too complex to analyze mathematically, but simulation shows that although they perform well in general, for some sequences their attenuation does not approach 1 as sequence length increases.

We construct two diminishing-attenuation estimators. The first,  $q_{2/3}$ , has sequence attenuation of at most  $2^{O(n^{2/3})}$  (i.e., exponential in at most a constant multiple of  $n^{2/3}$ ), hence its symbol attenuation diminishes to 1 at a rate of at least  $2^{O(n^{-1/3})}$ . It uses a constant number of operations per symbol; hence, it has linear complexity for the whole sequence.

Recall that  $\mu_\psi$  is the multiplicity of  $\psi$  in  $\psi_1^n$ , that  $\varphi_\mu$  is the prevalence of  $\mu$  in  $\psi_1^n$ , and that  $r = \mu_{\psi_{n+1}}$ . For  $n \in \mathbb{Z}^+$ , let

$$f_n(\varphi) \stackrel{\text{def}}{=} \max(\varphi, \lceil n^{1/3} \rceil) \tag{17}$$

where the  $\lceil n^{1/3} \rceil$  term is chosen to balance two contributions to the attenuation appearing in the proof. The estimator  $q_{2/3}$  assigns

$$q_{2/3}(1) = 1 \tag{18}$$

and for all  $n \geq 1$  and  $\psi_1^n \in \Psi^n$  it assigns the conditional probability

$$q_{2/3}(\psi_{n+1} | \psi_1^n) = \frac{1}{S_{n+1}(\psi_1^n)} \cdot \begin{cases} f_{n+1}(\varphi_{r+1}) & r = 0 \\ (r+1) \frac{f_{n+1}(\varphi_{r+1}+1)}{f_{n+1}(\varphi_r)} & r > 0 \end{cases} \tag{19}$$

where  $S_{n+1}(\psi_1^n)$  is the normalization factor. Potential functions can be used to calculate

an upper bound on the estimator’s attenuation (27).

**Theorem 4:**  $R^n(q_{2/3}) = 2^{O(n^{2/3})}$ ,

where the implied constant is at most 10. It can be shown that the estimator requires only a constant number of calculations per symbol, hence has linear complexity for the whole sequence.

Building on an equivalence between set partitions and patterns, we derive another estimator,  $q_{1/2}$ , achieving a sequence attenuation of  $2^{O(n^{1/2})}$ , hence a symbol attenuation that diminishes to 1 at a rate of at least  $2^{O(n^{-1/2})}$ . However, the estimator has super-polynomial, albeit subexponential, complexity.

For a pattern  $\psi_1^n$ , let

$$z(\psi_1^n) \stackrel{\text{def}}{=} \frac{\prod_{\mu=1}^n \mu^{1^{\varphi_\mu}} \varphi_\mu!}{n!} \tag{20}$$

and define the distribution over  $\bar{\psi}$  over  $\Psi^n$  by

$$\bar{p}(\psi_1^n) = \frac{z(\psi_1^n)}{\sum_{\bar{\psi} \in \Psi^n} z(\bar{\psi})} \tag{21}$$

and let

$$\Psi^n(\psi_1^n) \stackrel{\text{def}}{=} \{\psi_1^n \in \Psi^n : \psi_1^n = \psi_1^n\} \tag{22}$$

be the set of patterns of length  $t_n$  with prefix  $\psi_1^n$ , where  $t_n \stackrel{\text{def}}{=} 2^{\lceil \log_2 n \rceil + 1}$  is the smallest power of 2 that is larger than  $n$ . Then  $q_{1/2}$  is defined by

$$q_{1/2}(1) = 1 \tag{23}$$

and for all  $n \geq 1$  and  $\psi_1^n \in \Psi^n$ ,

$$q_{1/2}(\psi_{n+1} | \psi_1^n) = \frac{\sum_{\bar{\psi} \in \Psi^n(\psi_1^n)} \bar{p}(\bar{\psi})}{\sum_{\bar{\psi} \in \Psi^n(\psi_1^n)} \bar{p}(\bar{\psi})} \tag{24}$$

To calculate an upper bound for the attenuation of  $q_{1/2}$ , we relate it to the number of partitions of an integer. A partition of an integer  $n$  is a sum  $a_1 + \dots + a_k = n$ , where  $a_1 \geq a_2 \geq \dots \geq a_k$  are positive integers. For example, 4, 3 + 1, 2 + 2, 2 + 1 + 1, and 1 + 1 + 1 + 1 are the five possible partitions of 4.

In what is considered by some to be “one of the jewels of 20th century mathematics” (29), Hardy and Ramanujan (25) gave an expression for the exact number of partitions of any positive integer  $n$ , and showed that it grows as  $\exp\{\pi(2/3)^{1/2}n^{1/2}[1 + o(1)]\}$ . We use this result (27) to show

**Theorem 5:** For all  $n$ ,

$$R^n(q_{1/2}) \leq \exp\left(\frac{4\pi}{\sqrt{3}(2 - \sqrt{2})} \sqrt{n}\right).$$

The theorem shows that the sequence attenuation of some estimators is always  $2^{O(n^{1/2})}$ . Yet sequence attenuation cannot be made arbitrarily small. The sequence attenuation of any estimator grows at least exponentially in the cube root of the sequence length [see (27) for a proof that

combines universal-coding techniques (30–32) with results from analysis (33)].

**Theorem 6:** For every estimator  $q$ ,

$$R^n(q) \geq \exp\left\{\frac{3}{2} n^{1/3} [1 - o(1)]\right\}.$$

To better understand the behavior of the diminishing-attenuation estimators, we consider the conditional probabilities that the low-complexity estimator  $q_{2/3}$  assigns to some simple sequences and compare it to what one would intuitively expect.

Upon observing the sequence  $aaa\dots$  where the same symbol always repeats, one would guess that the next symbol would be “a” as well. Indeed, after  $n$  elements, the estimator assigns probability  $1 - \Theta(1/n)$  for the next symbol being “a” and probability  $\Theta(1/n)$  to a new symbol. For the alternating sequence  $abab\dots$ , one would predict probability  $1/2$  for the next symbol being each of “a” and “b”. Correspondingly, the estimator assigns probability  $\Theta(1/n)$  to a new symbol and splits the remaining probability evenly between “a” and “b.”

Of course, we are more interested in the behavior of the estimator when the number of symbols appearing is large. In the extreme case where all symbols are different, for example, after observing the sequence  $abcde\dots$ , we would expect the next symbol to be new. Indeed, the estimator assigns probability  $1 - \Theta[1/(n^{2/3})]$  that the next symbol will be new.

But for large-alphabet sequences where the probability of new does not approach 1, intuition may not serve well. Consider the simplest such case, the sequence  $aabbcc\dots$ . After observing an even number  $n$  of symbols (e.g.,  $aabbcc$ ), the estimator assigns probability  $1/4$  to the next symbol being new and  $3/(2n)$  to each of the preceding symbols, and after observing an odd number  $n$  of symbols (e.g.,  $aabbc$ ), the estimator assigns probability approaching 1 to the next symbol being the same as the last one (e.g., “c” in this example).

These estimations may be at odds with the intuitive presumption that, because every other element so far was new, the next symbol will be new with probability  $1/2$ . One possible explanation for the lower probability of “new” assigned by the estimator is that it can be shown (26) that after seeing  $n$  symbols of the sequence, the most likely alphabet is of size  $0.62n$ ; hence, roughly speaking, the probability of seeing a new symbol is about  $(0.12n)/(0.62n) \approx 0.2$ .

References and Notes

1. P. Laplace, *Philosophical Essays on Probabilities* (Springer-Verlag, New York, 1995) (A. I. Dale, transl. from ed. 5, 1825).
2. R. Krichevsky, V. Trofimov, *IEEE Trans. Inform. Theory* **27**, 199 (1981).

3. I. Witten, T. Bell, *IEEE Trans. Inform. Theory* **37**, 1085 (1991).
4. B. Clarke, A. Barron, *J. Statist. Planning Inference* **41**, 37 (1994).
5. W. Gale, K. Church, in *Corpus Based Research into Language*, N. Oostdijk, P. de Haan, Eds. (Rodopi, Amsterdam, 1994), pp. 189–198.
6. F. Hinsley, A. Stripp, *Codebreakers: The Inside Story of Bletchley Park* (Oxford Univ. Press, Oxford, 1993).
7. I. Good, *Biometrika* **40**, 237 (1953).
8. F. Song, W. Croft, *Research and Development in Information Retrieval* (ACM Press, New York, 1999), pp. 279–280.
9. K. Church, W. Gale, *Statist. Comput.* **1**, 93 (1991).
10. F. Jelinek, *Statistical Methods for Speech Recognition* (MIT Press, Cambridge, MA, 1998).
11. S. Chen, J. Goodman, *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics* (Morgan Kaufmann, San Francisco, 1996), pp. 310–318.
12. A. Nadas, *IEEE Trans. Acoust. Speech Signal Proc.* **33**, 1414 (1985).
13. A. Nadas, *Am. J. Math. Manage. Sci.* **11**, 229 (1991).
14. I. Good, *J. Statist. Comput. Simul.* **66**, 101 (2000).
15. D. McAllester, R. Schapire, *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory* (2000).
16. L. Davison, *IEEE Trans. Inform. Theory* **19**, 783 (1973).
17. J. Shtarkov, in *Topics in Information Theory* (*Coll. Math. Soc. J. Bolyai, no. 16*), I. Csizsár, P. Elias, Eds. (North-Holland, Amsterdam, 1977), pp. 559–574.
18. J. Rissanen, *IEEE Trans. Inform. Theory* **42**, 40 (1996).
19. I. Csizsár, P. Shields, *IEEE Trans. Inform. Theory* **42**, 2065 (1996).
20. N. Merhav, M. Feder, *IEEE Trans. Inform. Theory* **44**, 2124 (1998).
21. T. Cover, *Math. Finance* **1**, 1 (1991).
22. N. Littlestone, M. Warmuth, *IEEE Symposium on Foundations of Computer Science* (1992).
23. V. Vovk, *J. Comput. Syst. Sci.* **56**, 153 (1998).
24. N. Cesa-Bianchi, G. Lugosi, *Proceedings of the Twelfth Annual Conference on Computational Learning Theory* (1999), pp. 12–18.
25. G. Hardy, S. Ramanujan, *Proc. London Math. Soc.* **17**, 75 (1918).
26. A. Orlitsky, N. Santhanam, K. Viswanathan, J. Zhang, in preparation.
27. See supporting material on Science Online.
28. W. Gale, *Good Turing Smoothing Without Tears* (AT&T Bell Laboratories, Murray Hill, NJ, 1994).
29. H. Wilf, *Generating Functionology* (Academic Press, San Diego, CA, 1990), p. 91.
30. J. Åberg, Y. Shtarkov, B. Smeets, *Proceedings of Compression and Complexity of Sequences* (1997).
31. N. Jevtić, A. Orlitsky, N. Santhanam, *Proceedings of IEEE Symposium on Information Theory* (2002).
32. A. Orlitsky, N. Santhanam, J. Zhang, in preparation.
33. W. Hayman, *J. Reine Angew. Math.* **196**, 67 (1956).
34. We thank N. Alon, Y. Freund, I. J. Good, N. Jevtić, Sajama, W. Szpankowski, and K. Viswanathan for helpful discussions. Supported by NSF grant CCR-0313367.

## Supporting Online Material

[www.sciencemag.org/cgi/content/full/302/5644/427](http://www.sciencemag.org/cgi/content/full/302/5644/427)

DC1

SOM Text

References

20 June 2003; accepted 3 September 2003

## Radar Evidence for Liquid Surfaces on Titan

Donald B. Campbell,<sup>1\*</sup> Gregory J. Black,<sup>2</sup> Lynn M. Carter,<sup>3</sup> Steven J. Ostro<sup>4</sup>

Arecibo radar observations of Titan at 13-centimeter wavelength indicate that most of the echo power is in a diffusely scattered component but that a small specular component is present for about 75% of the subearth locations observed. These specular echoes have properties consistent with those expected for areas of liquid hydrocarbons. Knowledge of the areal extent and depth of any deposits of liquid hydrocarbons could strongly constrain the history of Titan's atmosphere and surface.

As the largest satellite of Saturn and the only one in the solar system with a substantial atmosphere, Titan is of considerable interest, an interest heightened by the Cassini mission's rendezvous with the Saturn system in 2004. Photochemically produced haze layers in Titan's upper atmosphere have made it difficult to investigate its lower atmosphere and surface at optical wavelengths. However, it is possible to observe the surface with Earth- and spacecraft-based radar systems, because the atmosphere is transparent at radio wavelengths. The strength and variability of the radar backscatter cross sections measured at 3.5-cm wavelength (*I*) indicated that Titan's surface is not homogeneous and cast doubt on models

for Titan's atmosphere and surface that suggested the presence of a deep hydrocarbon ocean (2). Observations in the near infrared (IR) with the Hubble Space Telescope and ground-based telescopes using speckle imaging and adaptive optics techniques (3, 4, 5) have provided coarse surface maps of Titan. Especially notable is a bright, high-albedo region centered near 110° longitude and extending over ~90° in longitude. Near-IR spectroscopic observations (6, 7) of the bright region suggest that its composition is primarily that of water ice. Here we report on observations with the recently upgraded Arecibo 13-cm-wavelength radar system.

We observed Titan on 16 nights in November and December 2001 and on 9 nights in November and December 2002, transmitting at 13-cm wavelength with the 305-m Arecibo telescope and receiving the echo with Arecibo. Titan's rotational and orbital periods are 15.9 days, and our 2001 observations were obtained at a uniform 22.6° (~800 km) interval in longitude. The 9 observations in 2002 did not provide uniform coverage. The latitude of the subearth track was 25.9°S in 2001 and 26.2°S in 2002, its farthest southern excursion.

The round-trip light time to the Saturn system during the observations was 2 hours 15 min, and the limited tracking time of the Arecibo telescope meant that signal reception was restricted to ~30 min per day, corresponding to 0.5° of Titan rotation (20 km of motion of the subearth point). On one night in 2001 and for most of the 2002 observations (as well as others when we were attempting ranging measurements to Titan), the 100-m Green Bank Telescope (GBT) was also used to receive the echo for the full round-trip time. These data have lower signal-to-noise ratios than those obtained with Arecibo receiving the echo, but the longer receive time corresponding to 2.1° of Titan rotation allowed more subearth locations to be studied.

For each observation, a circularly polarized monochromatic signal was transmitted, and the echo power was measured as a function of Titan's rotational Doppler shift. Echo power spectra were obtained in both senses of received circular polarization: the opposite circular (OC) sense to that transmitted and the same circular (SC) sense. For a mirrorlike specular reflection, all of the echo power would be in the OC sense, whereas power in the SC sense arises from scattering from wavelength-size surface and/or subsurface structure. For surfaces other than ones composed of low-temperature water ice, the ratio of the SC to OC echo powers (the circular polarization ratio) is frequently used as an indicator of the degree of wavelength scale surface and/or near subsurface roughness. Radar observations of the icy Galilean satellites of Jupiter (8, 9) showed that the transparency of low-temperature water ice at radio wavelengths can result in high-backscatter cross sections and a polarization ratio greater than unity.

For the Titan spectra, 75 to 100% of the echo power is in a broad diffuse component. However, a small central specular

<sup>1</sup>National Astronomy and Ionosphere Center and Department of Astronomy, Space Sciences Building, Cornell University, Ithaca, NY 14853, USA. <sup>2</sup>Department of Astronomy, University of Virginia, Post Office Box 3818, Charlottesville, VA 22903, USA. <sup>3</sup>Department of Astronomy, Space Sciences Building, Cornell University, Ithaca, NY 14853, USA. <sup>4</sup>300-233 Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109–8099, USA.

\*To whom correspondence should be addressed. E-mail: [campbell@astro.cornell.edu](mailto:campbell@astro.cornell.edu)