An Optimal Anytime Estimation Algorithm
Gavaldà, R.
Research Report LSI-04-56-R

Departament de Llenguatges i Sistemes Informàtics

UNIVERSITAT POLITÈCNICA DE CATALUNYA

An Optimal Anytime Estimation Algorithm

Ricard Gavaldà*
Univ. Politècnica de Catalunya
gavalda@lsi.upc.es

December 7th, 2004

Abstract

In many applications a key step is estimating some unknown quantity μ from a sequence of trials, each having expected value μ. Optimal algorithms are known when the task is to estimate μ within a multiplicative factor of $1 \pm \epsilon$, for an ϵ given in advance. In this paper we consider *anytime* approximation algorithms, i.e., algorithms that must give a reliable approximation after each trial, and whose approximations have to be increasingly accurate as the number of trials grows. We give an anytime algorithm for this task when the only a-priori known property of μ is its range, and show that it is asymptotically optimal in some cases, in the sense that no correct anytime algorithm can give asymptotically better approximations. The key ingredient is a new large deviation bound for the supremum of the deviations in an infinite sequence of trials, which can be seen as a non-limit analog of the classical Law of the Iterated Logarithm.

1 Introduction

In many applications, an algorithm has access to a sequence of random trials X_1, X_2, ..., X_i, ..., and is required to approximate a quantity μ underlying the generation of the trials. In this paper, we consider the case in which the X_i are independent and identically distributed and $\mu = E[X_i]$. Nothing is known initially on the X_i except that they take values in a known interval $[a, b]$.

One way of formalizing this problem is as follows: the algorithm receives two inputs ϵ, δ and, with probability at least $1 - \delta$, after reading a finite number of trials must stop and output an number $\hat{\mu}$ such that

$$(1 - \epsilon) \cdot \mu \leq \hat{\mu} \leq (1 + \epsilon) \cdot \mu.$$

*Department of Software (LSI), Universitat Politècnica de Catalunya. Jordi Girona Salgado 1–3, E-08034 Barcelona, Spain. Work supported in part by the 6th Framework Program of EU through the integrated project DELIS (#001907) and by MCYT TIC 2002-04019-C03-1 (MOISES). Partly done while visiting McGill University.

We call this type of algorithm a *one-shot* estimator. Hoeffding's inequality lets us compute a number n from ϵ and δ (hence, before reading any trials) such that the average of the first n trials will have the properties required of $\hat{\mu}$. However, Hoeffding's inequality is far from optimal for small μ or for trials with small variance, hence this approach leads to unnecessarily large n.

Sequential sampling strategies try to be more efficient by detecting, from the trials themselves, that the distribution is not the worst case, hence probably stopping much earlier than predicted by the worst-case bounds. See e.g. [2, 17, 18] for some examples where this approach is likely to be much more efficient than the previous one. The efficiency of such an algorithm can be defined as the expected running time until it stops – a random variable that presumably depends on μ itself.

In particular, Dagum *et al* [2] gave a sequential algorithm for this task that is provably optimal. Here "optimal" means that no other algorithm can be more efficient with respect to the stopping time criterion, up to multiplicative constants. For the case of Bernoulli trials, essentially optimal algorithms are also given in [17, 18, 19].

In many applications, however, one desires *anytime* algorithms, that is, algorithms that can provide some reliable approximation after seeing each trial (or, say, after every block of k consecutive trials). Additionally, one expects that approximations get more and more accurate as more and more trials have been seen. Two reasons why this anytime behavior is needed may be: 1) the required accuracy ϵ is not known from the start, as it may itself depend on other quantities that are being estimated, or another series of trials that is read sequentially; hence, it is a waste of time to use a very small ϵ if it turns out that a coarse approximation suffices; 2) some other tasks may depend on this estimate, can be started as soon as a reasonable estimate is allowed, but can later incorporate of a better estimate.

We expect anytime estimators to come up in the processing and analysis of Data Streams [1] and on-line Machine Learning and Data Mining. In particular, several papers have used sequential sampling to speed up learning or mining algorithms; thus, [5, 6, 8, 14, 15, 16, 22, 23, 24, 25] use distribution independent bounds (e.g., Hoeffding), while [3, 4, 8, 9] use techniques that are sensitive to the actual distribution of the data, in the spirit of [2, 17, 18, 19].

More formally, an estimator algorithm performs as follows: after reading the nth trial, it outputs a pair $(\hat{\mu}_n, \epsilon_n)$ where $\hat{\mu}_n$ is its current approximation and ϵ_n an uncertainty margin. We require that, with probability $1 - \delta$, *all* approximations are correct, i.e.,

$$\Pr[\,\forall n:\ (1-\epsilon_n)\cdot\mu \le \hat{\mu}_n \le (1+\epsilon_n)\cdot\mu\,] \ge 1-\delta$$

and, furthermore, that ϵ_n tends to 0 with probability 1. We want this to hold even if the trial sequence is infinite, i.e., the estimator never stops. Thus, it makes no sense to use "stopping time" as an efficiency measure. Rather, we take as a measure of efficiency the rate at which ϵ_n tends to 0.

The requirement to be correct at each step may seem a strong one, but it may be necessary in many cases. Indeed, allowing error probability δ at each step

allows for substantial probability of infinitely many errors occurring, and even for very long runs of consecutive errors. This may be unacceptable in situations where the sequence of trials is monitored until a decision is made based on the current (or latest) estimations. In this case, a single largely incorrect estimation may lead to an awfully wrong decision.

The main result of this paper is an anytime algorithm that satisfies these strong requirements and that is optimal up to multiplicative constants for the case of Bernoulli trials; we conjecture it is in fact optimal for essentially all trials with a bounded range. The two ingredients are the one-shot estimator algorithm of Dagum *et al* [2] and a (to our knowledge, new) bound on the probability that a large enough deviation occurs in an infinite sequence of trials. This bound can be seen as a non-limit version of the classical Law of the Iterated Logarithm [11, 12, 7] in the same sense that the Chernoff-Hoeffding-Bernstein inequalities can be seen as non-limit versions of the Central Limit Theorem.

It is easy to see that an anytime estimator A can be used to build a one-shot algorithm B: on some input ϵ, run A algorithm until it outputs a pair (μ_n, ϵ_n) with $\epsilon_n < \epsilon$, then output μ_n. However, B may not be an optimal one-shot algorithm even if A is an optimal anytime algorithm: intuitively, the values of ϵ_n output by A must be somewhat conservative in order to be correct for every n. As a consequence of our results, it turns out that this is in fact the case: the optimal anytime algorithm is less efficient than the optimal one-shot algorithm by a multiplicative factor of about $\log\log$.

The paper is organized as follows: In Section 2 we define more formally the estimation problems we consider, and state known results for this task. In Section 3.1 we sketch the kind of large deviation bound required for anytime estimators. In Section 3.2 we prove such a bound together with an essentially matching lower bound; the combination of both bounds is the non-limit version of the Law of the Iterated Logarithm. In Section 3.3 we specialize these bounds to variables to which Bernstein's inequality applies. In Section 4 we present the optimal anytime estimator algorithm, which is shown to be optimal in Section 5 for Bernoulli variables. Finally, in Section 6 we indicate some possible improvements and future work.

2 Estimators and Previous Results

We start giving a formal definition of two kinds of estimators, *one-shot* and *anytime* estimator algorithms.

Let X_1, X_2, ..., X_t, ... be an infinite sequence of outcomes drawn from infinitely many independent copies of a single random variable X. If $\mathcal{X}$ is the range of X, let F map $\mathcal{X}^\star$ to real numbers such that for every t it holds that $E[F(X_1, \ldots, X_t)] = \mu$, for some μ. We can think of μ as some value associated to μ that we want to estimate from the trials.

The following two definitions are essentially those in [2] and [8], respectively.

Definition 1 *A one-shot estimator algorithm A performs as follows: At the start, it reads parameters ϵ and δ. Then, at each time-step t, it reads the value*

of X_t, performs some internal computation, and decides either to continue or to output a number and stop. Let $\hat{\mu}$ be the number (a random variable) output by A when it stops; it is undefined in case A does not stop. The estimator is correct if

(1) A stops with probability 1, *and*

(2) $\Pr[\,(1-\epsilon)\mu \leq \hat{\mu} \leq (1+\epsilon)\mu\,] \geq 1-\delta$,

where the probabilities are taken over all infinite sequences $X_1, X_2, \ldots$

Definition 2 *An anytime estimator algorithm A performs as follows: At the start, it reads* one *parameter δ. Then, at each time-step t, it reads the value of X_t, outputs a pair of numbers $(\hat{\mu}_t, \epsilon_t)$, and proceeds to the next step. The estimator is correct if*

(1) With probability 1, $\lim_{t\to\infty} \epsilon_t = 0$, *and*

(2) $\Pr[\,\forall t \;:\; (1-\epsilon_t)\mu \leq \hat{\mu}_t \leq (1+\epsilon_t)\mu\,] \geq 1-\delta$,

where the probabilities are taken over all infinite sequences $X_1, X_2, \ldots$

Remarks:

- Strictly speaking an anytime estimator is not an algorithm since it never stops.
- All that the algorithms know about X initially are two numbers a and b such that $a \leq X \leq b$; that is, they are required to satisfy the definition for any random variable X such that $a \leq X \leq b$.
- We can identify an anytime estimator algorithm with a pair of real-valued functions $(\hat{\mu}, \epsilon)$, each of which takes as parameters a sequence $(X_1, \ldots, X_t)$ and a number δ, and which satisfy conditions (1) and (2) above.
- We take as a measure of efficiency of a one-shot algorithm its expected running time, whichis a function of ϵ, δ, μ, and possibly other characteristics of X (such as its variance).
- We take as a measure of efficiency of an anytime algorithm the rate of convergence of ϵ_t to 0. This is usually the main criterion for choosing an estimator, although there may be others (computational cost, bias of $\hat{\mu}$ w.r.t. μ, etc.).
- Although we took different efficiency measures of one-shot and anytime algorithms, it still makes sense to compare their relative efficiencies. We can, for example, measure the time that an anytime algorithm reads until it outputs some ϵ_t less than given ϵ.

In this paper we consider only the case where X takes real values and F is the average function so $\mu = E[X]$. This is also the case considered in [2]. In [8], the framework allows in principle for any F, although some F will not allow the design of efficient estimator algorithms (e.g., if F is not sufficiently smooth). The necessity for estimating other F appears naturally in some Machine Learning or Data Mining taks; see e.g. [4, 5, 6, 14].

When X is Bernoulli, F the average counts the number of successes, and we denote it as freq (for frequency). We call estimators that work for Bernoulli variables *frequency estimators*.

Some known results about the efficiency of frequency estimators are:

1. [2, 17, 18, 19] There are one-shot frequency estimators $A = (\text{freq}, \epsilon)$ whose expected stopping time is $O\left(\frac{1}{\epsilon^2 \mu} \ln \frac{1}{\delta}\right)$. No one-shot estimator can have asymptotically smaller stopping time.

2. [4, 8] There is a function ϵ such that the pair $A = (\text{freq}, \epsilon)$ is a correct anytime estimator algorithm and the value ϵ_t is (in probability)

$$\epsilon_t = O\left(\sqrt{\frac{1}{\mu\, t} \cdot (\ln \frac{1}{\delta} + \ln \frac{1}{\mu\, t})}\right).$$

3. ([8], also follows from [2]) If $A = (\text{freq}, \epsilon)$ is a correct anytime estimator, then with it must hold (in probability) that

$$\epsilon_t = \Omega\left(\sqrt{\frac{1}{\mu\, t} \cdot \ln \frac{1}{\delta}}\right)$$

That is, (1) determines the the efficiency of one-shot frequency estimators, and (2) and (3) provide upper and lower bounds on the efficiency of anytime frequency estimators. These results are not completely satisfactory for two reasons:

First, the lower bound (3) applies *only* to estimators that use $\hat{\mu} = \text{freq}$, that is, whose estimate for μ is always the observed frequency of 1s. Although this is a very natural choice, it requires proof that there is no other choice for $\hat{\mu}$ that allows for larger efficiency (i.e., that allows for a faster-decreasing function ϵ.)

Second, while the upper and lower bounds for one-shot estimators match up to multiplicative constants, the bounds for anytime estimators differ by a gap of $\sqrt{\ln t}$.

We solve both problems in Sections 4 and 5, by showing: for any constant λ, let $\epsilon_\lambda(t)$ be

$$\epsilon_\lambda(t) = \lambda \cdot \sqrt{\frac{2}{\mu\, t} \cdot (\ln \frac{1}{\delta} + \ln\ln \frac{1}{\mu_t\, t})} \cdot (1 + o(1)).$$

If $\lambda > 1$, then $(\text{freq}, \epsilon_\lambda)$ is a correct anytime frequency estimator, and, if $\lambda < 1$, no pair $(\hat{\mu}, \epsilon_\lambda)$ is a correct anytime estimator frequency for any $\hat{\mu}$ whatsoever. Hence, our frequency estimator is asymptotically optimal up to small-order terms.

Note that this bound is a $\sqrt{\ln\ln}$-factor larger than the sufficient and necessary one for one-shot algorithms. Since $\ln\ln x \leq 4$ for all quantities x likely to appear in (today's) application, this is a moderate price to pay for being anytime.

It is worth noting that an estimator for Bernoulli variables which is only $O(\sqrt{\ln\ln t})$ worse than the best one-shot estimator was given in [4]. However, that efficiency was achieved at the price of refraining from producing an estimation at most steps. More precisely, that algorithm outputs estimation only at time steps t which are powers of some fixed constant γ (thus, increasingly spaced), so it is not an anytime algorithm according to our definition.

For variables that are not Bernoulli, but Bernstein, Dagum *et al.* [2] gave a one-shot estimator whose asymptotic running time is

$$O\left(\frac{\sigma^2}{\epsilon^2\mu}\ln\frac{1}{\delta}\right)$$

and showed that this is again optimal. We give an anytime estimator algorithm that is only a log log factor less efficient than this one. We argue that if Bernstein's inequality is tight for X, then this anytime algorithm is asymptotically optimal, up to multiplicative constants.

Finally, estimators are very close in spirit of the strategies studied in the classical area of statistics known as sequential analysis initiated by Wald in the 1940's [10, 26]. We are not aware, however, of any work in that area dealing with anytime estimators or yielding efficiency analysis similar to ours.

3 Large Deviation Bounds for Suprema of Partial Sums

3.1 Preliminaries

At the core of our estimation task is the following question in probability: For some given random variable X generating independent trials X_1, X_2, ..., what bound $\epsilon(n, \delta\{, \mu, \ldots\})$ can we give so that

$$\Pr\left[\,\exists n : |S_n - \mu n| \leq \epsilon(n,\delta)\cdot\mu n\,\right] \geq 1-\delta$$

holds? (Recall: $S_n = X_1 + \ldots + X_n$ and $\mu = E[X_1]$). Note that we allow ϵ to depend on the *unknown* quantity μ and even other parameters such as $Var(X_1)$; we can later deal with this problem by bootstrapping argument [4, 8]. We write $\epsilon(n,\delta)$ for conciseness.

If we have any bound θ for the deviations for fixed n, one can find such an ϵ as follows: Assume that

$$\forall n : \Pr\left[\,|S_n - \mu\, n| \geq \theta(n,\delta)\cdot\mu\, n\,\right] \leq \delta.$$

Then let $\epsilon(n,\delta) = \theta(n, \delta/(n(n+1))$. Using the union bound,

$$\Pr[\,\exists n \quad : \quad |S_n - \mu\, n| \leq \epsilon(n,\delta)\cdot\mu\, n\,]$$

$$\leq \sum_n \Pr[\,|S_n - \mu\, n| \leq \theta(n, \delta/(n(n+1))) \cdot \mu\, n\,]$$
$$\leq \sum_n \frac{\delta}{n(n+1)} = \delta.$$

This is the trick used in [4] to derive an anytime algorithm which is a log factor less efficient than a one-shot algorithm.

However, the union bound ignores the fact that the events being summed are strongly dependent: if there is a large deviation at n, it is much more likely that there is a large deviation at $n+1$. We need a finer bound that takes this dependence into account.

The form of the bound that we can expect in the limit is given by the so-called Laws of the Iterated Logarithm. For Bernoulli variables, for example, such a law is stated as follows:

Theorem 3 (Law of the Iterated Logarithm) *[11, 12, 7, 21] Let X_1, X_2, $\ldots X_n$, $\ldots$ be i.i.d. Bernoulli random variables with $E[X_i] = p = 1 - q$, and let S_n be $X_1 + \ldots + X_n$. Then with probability 1:*

$$\limsup_{n\to\infty} \frac{S_n - p\,n}{\sqrt{2\,p\,q\,n\,\ln\ln n}} = 1.$$

In other words: for $\lambda < 1$, with probability one only finitely many of the events

$$S_n \geq p\,n + \sqrt{2\,\lambda\,p\,q\,n\,\ln\ln n} \tag{1}$$

occur, and for $\lambda > 1$ with probability one infinitely many events (1) occur.

This statement is taken from Feller's book [7], where it is attributed to Khintchine [11], as well as to Kolmogorov [12] for more general random variables. Alternative proofs are given by Minozzo [20] and (in one direction only) by Rényi [21].

Unfortunately, this is a statement about limit behaviors. It does not quantify neither the extreme deviations we can expect in a finite number of variables n, nor how quickly the probability converges to 1. For example, it is conceivable that convergence was exponentially slow (say, in $1/p$), in which case we would never perceive it in many applications. We would like a statement that provides a specific bound that can be applied at a given n and for given ϵ. The bound we look is to this Law of the Iterated Logarithm like Chernoff-Bernstein's inequalities are to the Central Limit Theorem.

In the next two Sections 3.2 and 3.3 we state a quantitative version of this law that tells what precise deviations we can expect for each partial sum for a given probability level. The bound determines the form of dependence on all parameters and is strong enough so that Khintchine's Law of the Iterated Logarithm can be deduced from it.

3.2 General Bounds

The two theorems in this section show how to transfer any tight large deviation bound for the sum of n variables to a tight large deviation bound for the supremum of the infinite sequence of partial sums.

Theorem 4 *Let $X_1, X_2, \ldots X_n, \ldots$ be i.i.d. random variables with 0 mean and variance σ^2, and let S_n be $X_1 + \ldots + X_n$. Let $\epsilon(n,\delta)$ be a nonnegative function satisfying*

$$\Pr[\,|S_n| > \epsilon(n,\delta)\cdot n\,] \le \delta.$$

For every $\lambda > 1$ there is a constant $c = c(\lambda)$ such that if function $\theta(n,\delta)$ is increasing in n and satisfies

$$\theta(n,\delta) \ge \lambda\cdot\epsilon(\lambda n, c\,\delta/(\ln n)^{\lambda}) + \sqrt{2\sigma^2/n}$$

then for every $\delta \in (0,1)$ it holds

$$\Pr\,[\exists n \ge 2 : |S_n| > \theta(n,\delta)\cdot n] \le \delta.$$

The converse states that for any slightly smaller function $\theta(n,\delta)$, not only infinitely many large deviations occur, but in fact occur polynomially often.

Theorem 5 *Let $X_1, X_2, \ldots X_n, \ldots$ be i.i.d. random variables with 0 mean, and let S_n be $X_1 + \ldots + X_n$. Let $\epsilon(n,\delta)$ be a nonnegative function satisfying*

$$\Pr[\,|S_n| > \epsilon(n,\delta)\cdot n\,] \ge \delta.$$

For every $\lambda < 1$, $\lambda > 0$, and every c there is a constant $d = d(\lambda,c) > 1$ such that if function $\theta(n,\delta)$ satisfies

$$\theta(n,\delta) + (1-\lambda)\,\theta((1-\lambda)n,\delta) \le \lambda\,\epsilon(\lambda n, c\,\delta/\ln n)$$

then for every $\delta \in (0,3/4)$ and every $N > N(\lambda)$ it holds

$$\Pr\left[\,\exists n \in [N, N^d] : |S_n| > \theta(n,\delta)\cdot n\right] \ge \delta.$$

To prove Theorems 4 and 5 we use the following two lemmas, whose proofs are given in the Appendix. The following is an easy variation of Lemma E in [21] (Ch. VII, §4), and is proved for self-containment.

Lemma 6 *[21] Let $X_1, X_2, \ldots, X_n$ be independent random variables with variance σ^2. Let S_i ($i \le n$) be $X_1 + X_2 + \ldots + X_i$. Then for any x it holds*

1) $\Pr\,[\exists i \le n : S_i \ge x\,] \le 2\Pr\left[\,S_n \ge x - \sqrt{2\sigma^2 n}\right]$.

2) $\Pr\,[\exists i \le n : S_i \le x\,] \le 2\Pr\left[\,S_n \le x + \sqrt{2\sigma^2 n}\right]$.

Lemma 7 *Let $B_1, \ldots, B_k, \ldots$ be independent events, and assume that $\Pr[B_k] \ge \delta/(2k\cdot(\alpha-1))$, with $\delta < 3/4$ and some α. Then $\Pr[\exists k : k_0 \le k \le \alpha\cdot k_0 : B_k] \ge \delta$.*

Proof of Theorem 4. Fix $\lambda > 1$ and for each $k \geq 1$, define the integer $n_k = \lceil \lambda^k \rceil$. To simplify writing, let θ_n denote $\theta(n, \delta)$ in the following.

Let B_k be the event "$\exists n \in (n_k \ldots n_{k+1}] : |S_n| > \theta_n n$ ". Then clearly

$$\Pr[\ \exists n \geq 2 : |S_n| > \theta_n n\ ] = \Pr[\ \exists k \geq 1 : B_k] \leq \sum_{k \geq 1} \Pr[\ B_k\ ].$$

We bound $\Pr[B_k]$ as follows:

$$\begin{aligned}
\Pr[\ B_k\ ] &= \Pr[\ \exists n \in (n_k \ldots n_{k+1}] : |S_n| > \theta_n n\ ] \\
&\leq \Pr[\ \exists n \in (n_k \ldots n_{k+1}] : |S_n| > \theta_{n_k} n_k\ ] \\
&\leq 4\ \Pr[\ |S_{n_{k+1}}| > \theta_{n_k} n_k - \sqrt{2\sigma^2 n_{k+1}}\ ] \\
&\leq 4\ \Pr[\ |S_{n_{k+1}}| > \lambda \cdot \epsilon(\lambda\, n_k, c\delta/(\ln n_k)^\lambda) \cdot n_k + \sqrt{2\lambda\sigma^2 n_k} - \sqrt{2\lambda\sigma^2 n_k}\ ] \\
&\leq 4\ \Pr[\ |S_{n_{k+1}}| > \epsilon(n_{k+1}, c\delta/(\ln n_k)^\lambda) \cdot n_{k+1}\ ],
\end{aligned}$$

where the first inequality holds because $\theta_n n$ is increasing, the second by Lemma 6, and the third and fourth by definition of θ_n and $n_{k+1} \cong \lambda n_k$. By assumption on ϵ we have

$$\Pr[\ B_k\ ] \leq 4\, \frac{c\delta}{(\ln n_k)^\lambda} \leq 4\, \frac{c\delta}{(\ln \lambda^k)^\lambda} = 4c\delta\, (\ln \lambda)^{-\lambda} k^{-\lambda}.$$

To conclude, note that

$$\sum_{k \geq 1} \Pr[\ B_k\ ] \leq 4c\delta\, (\ln \lambda)^{-\lambda} \sum_{k \geq 1} k^{-\lambda}.$$

The sum converges for $\lambda > 1$, so the theorem holds if $c^{-1} = 4(\ln \lambda)^{-\lambda} \cdot \sum_{k \geq 1} k^{-\lambda}$. ∎ (Theorem 4)

Proof of Theorem 5. Fix $\lambda < 1$ and c and define $\gamma = 1/(1 - \lambda)$, so that $\gamma > 1$ and $\lambda = 1 - 1/\gamma$. Define for each $k \geq 1$ the integer $n_k = \lceil \gamma^k \rceil$. To simplify writing, let θ_n denote $\theta(n, \delta)$ in the following. Also, let

$$\varphi(n, \delta) = \lambda\, \epsilon(\lambda n, c\delta / \ln n);$$

so that the condition on $\theta(n, \delta)$ in the statement of the theorem reads

$$\varphi(n, \delta) \geq \theta(n, \delta) + \theta(n/\gamma, \delta)/\gamma \tag{2}$$

Fix $N > \gamma^4$; for the constant $d = d(\lambda, c) > 2$ to be defined later, let (k_0, k_1) be a maximal interval such that $[n_{k_0}, n_{k_1}] \subseteq [N, N^d]$. Note that $4 \leq k_0 < k_1$. Clearly,

$$\Pr[\ \exists n \in [N, N^d] : |S_n| \geq \theta_n \cdot n\ ] \geq \Pr[\ \exists k \in [k_0, k_1] : |S_{n_k}| \geq \theta_{n_k} \cdot n_k\ ].$$

We want to show that this quantity is at least δ. To do this, define $D_k = S_{n_k} - S_{n_{k-1}}$ and let B_k be the event "$|D_k| \geq \varphi(n_k, \delta)\, n_k$". We will show that

$$\Pr[\,\exists k \in (k_0, k_1] \,:\, B_k\,] \geq \delta \tag{3}$$

This suffices because for every k,

$$\begin{aligned}
B_k &\iff |D_k| \geq \varphi(n_k, \delta) \\
&\implies |S_{n_k} - S_{n_{k-1}}| \geq (\theta(n_k, \delta) + \theta(n_k/\gamma, \delta)/\gamma) \cdot n_k \\
&\qquad\qquad \cong \theta(n_k, \delta) \cdot n_k + \theta(n_{k-1}, \delta) \cdot n_{k-1} \\
&\implies \text{either } |S_{n_k}| \geq \theta(n_k, \delta) \cdot n_k \text{ or } |S_{n_{k-1}}| \geq \theta(n_{k-1}, \delta) \cdot n_{k-1} \\
&\implies |S_{n_{k'}}| \geq \theta(n_{k'}, \delta) \cdot n_{k'} \text{ for } k' \in \{k, k-1\}.
\end{aligned}$$

and therefore

$$\Pr[\,\exists k \in [k_0, k_1] \,:\, |S_{n_k}| \geq \theta_{n_k} \cdot n_k\,] \geq \Pr[\,\exists k \in (k_0, k_1] \,:\, B_k\,].$$

To prove (3), let $m_k = n_k - n_{k-1} = (1 - 1/\gamma)n_k \cong \lambda n_k$ and observe that the random variable $|D_k|$ has the same distribution as $|S_{m_k}|$ (i.e., both are sums of m_k of the X_i variables, which are all i.i.d.). Then

$$\begin{aligned}
\Pr[\,B_k\,] &= \Pr[\,|S_{m_k}| \geq \varphi(n_k, \delta) n_k\,] \\
&= \Pr[\,|S_{m_k}| \geq \lambda\, \epsilon(\lambda n_k, c\delta/\ln n_k) n_k\,] \\
&= \Pr[\,|S_{m_k}| \geq \epsilon(m_k, c\delta/\ln n_k) m_k\,] \\
&\geq \frac{c\,\delta}{\ln n_k} \cong \frac{c\,\delta}{\ln \gamma^k} = \frac{c\,\delta}{k\,\ln\gamma} \geq \frac{2\delta}{(d/2-1)k}
\end{aligned}$$

where the first equality is by definition of φ and γ, the second by $m_k \cong (1 - 1/\gamma) \cdot n_k$, and the last inequality holds for an appropriate constant $d = d(\lambda, c)$.

Now observe that D_k depends only on $X_{n_k+1}, \ldots, X_{n_{k+1}}$ so D_k and D'_k are independent for $k \neq k'$, and so are the events B_k and $B_{k'}$. Note also that from the definition of k_0 and k_1 it follows $\gamma^{k_1+1} \geq N^{d/2} \geq (\gamma^{k_0-1})^{d/2}$, hence $k_1 \geq d(k_0 - 1) - 1 \geq (d/2)(k_0 + 1)$. By Lemma 7,

$$\Pr[\,\exists k \in (k_0, k_1] \,:\, B_k\,] \geq \Pr[\,\exists k \in [k_0 + 1, (d/2)(k_0 + 1)] \,:\, B_k] \geq \delta,$$

which proves inequality (3) and, so, the theorem. ∎ (Theorem 5)

3.3 Special cases: Bernstein and Bernoulli variables

In this section we apply the two previous theorems to specific types of random variables and specific bounds. First, we derive an upper bound based on Theorem 4 and Bernstein's inequality. Then, we derive a lower bound for Bernoulli variables, based on Theorem 5, which matches the upper bound up to smaller-order terms. In particular, we show that from these upper and lower bound we

can recover the Law of the Iterated Logarithm for Bernoulli variables, as given by Theorem 3.

Bernstein's inequality is about the most general large deviation bound that can be proved for sums of bounded random variables. We use the following form of Bernstein's inequality ([13], see also [21]).

Lemma 8 *(Bernstein's inequality) Let X_1, X_2, ..., X_n be independent real-valued random variables with $a \leq X_i \leq b$, $E[X_i] = 0$, $\sigma^2 = (\sum_{i=1}^n Var(X_i))/n$, and $S_n = \sum_{i=0}^n X_i$. Then for any $\epsilon > 0$*

$$\Pr[\,|S_n| > \epsilon\, n\,] \leq 2\, \exp\left(-\frac{1}{2}\,\frac{\epsilon^2 n}{\sigma^2 + \epsilon(b-a)/3}\right)\,.$$

(The inequality is proved in [13] with the condition $|X_i| \leq c$ instead of $a \leq X_i \leq b$; it is easy to extend the proof there to this slightly stronger statement).

To have a closed formula for ϵ, it is convenient for us to rephrase it in the following form (proved in the Appendix):

Lemma 9 *Under the conditions of Lemma 8, for any $\gamma > 1$ and β, if*

$$\epsilon \geq \max\left\{\sqrt{\gamma \cdot \frac{2\beta\sigma^2}{n}}, \frac{\gamma}{\gamma - 1} \cdot \frac{2\beta(b-a)}{3n}\right\}$$

then

$$\Pr[\,|S_n| > \epsilon\, n\,] \leq 2\exp(-\beta).$$

From Theorem 4 and Lemma 9 is is easy to prove the following:

Theorem 10 *Let X_1, X_2, ... X_n, ... be infinitely many i.i.d. random variables with $E[X_i] = \mu$, $Var(X_i) = \sigma^2$, and let S_n be $X_1 + \ldots + X_n$.*

For any $\lambda > 1$ there is a constant c such that if $\epsilon(n,\delta)$ is defined by

$$\begin{aligned}\epsilon(n,\delta) \quad = \quad & \max\left\{\sqrt{\frac{2\lambda\sigma^2\,\ell(n,\delta,c)}{\mu^2 n}}\,,\ \frac{\lambda}{\lambda-1}\cdot\frac{2(b-a)\ell(n,\delta,c)}{3\mu n}\right\} \\ & +\sqrt{2\sigma^2/n}\end{aligned}$$

and $\ell(n,\delta,c) = (\ln(1/\delta) + \ln\ln n + c)$ then

$$\Pr\,[\,\exists n \geq 2 : |S_n - \mu\, n| > \epsilon(n,\delta)\,\mu\, n\,] \leq \delta.$$

Let us now consider the special case of Bernoulli variables. When $X_i = 1$ with probability p and 0 with probability $q = 1 - p$, the definition of $\epsilon(n,\delta)$ in Theorem 10 becomes:

$$\epsilon(n,\delta) \quad = \quad \max\left\{\sqrt{\frac{2\lambda q\,\ell(n,\delta,c)}{pn}}\,,\ \frac{\lambda}{\lambda-1}\cdot\frac{2\ell(n,\delta,c)}{3pn}\right\} \quad + \sqrt{2pq/n}. \quad (4)$$

To show that this bound is essentially tight for Bernoulli variables, we use the following converse to Chernoff bound, together with Theorem 5.

Lemma 11 *For all $\lambda < 1$, all p, all n sufficiently large, and all ϵ sufficiently small,*

$$\Pr[\, S_n \geq (1+\epsilon)pn \,] \quad \geq \quad \exp\left(-\frac{1}{2\lambda} \left(\epsilon + \sqrt{2\pi q/pn} \right)^2 \frac{1}{q} pn \right) .$$

Plugging this bound in Theorem 5, one obtains (details are omitted):

Theorem 12 *Let $X_1, X_2, \ldots X_n, \ldots$ be infinitely many i.i.d. Bernoulli variables with $E[X_i] = p = q - 1$, and let S_n be $X_1 + \ldots + X_n$.*

For any $\lambda < 1$ and any c there is a d such that if $\epsilon(n, \delta)$ is defined by

$$\epsilon(n,\delta) \quad = \quad \sqrt{\frac{2\lambda q\, \ell(n,\delta,c)}{pn}}$$

then for any N sufficiently large

$$\Pr\left[\, \exists n \in [N, N^d] : |S_n - pn| > \epsilon(n,\delta)\, p\, n \,\right] \geq \delta.$$

Note that the Law of the Iterated Logarithm as given in Theorem 3 follows from Theorems 10 and 12. Indeed, consider Bernoulli variables with $E[X_i] = p = 1 - q$. For fixed $\lambda > 1$ and δ, the value of $\epsilon(n, \delta)$ given by (4) is asymptotically dominated by the first argument in the $\max\{\ldots\}$ in its definition. This implies that for any λ' slightly larger than λ, the probability that there are infinitely many n such that

$$|S_n - pn| > \sqrt{2\lambda' pqn\, (\ln(1/\delta) + \ln\ln n + c)}$$

is less than δ for an appropriate $c = c(\lambda)$. Since this is true for all δ, the event occurs with probability 0. Conversely, fix $\lambda < 1$, c, and δ. For any N, let B_N be the event "$\exists n : N \leq n < N^d : |S_n - pn| > \sqrt{2\lambda' pqn\, (\ln(1/\delta) + \ln\ln n + c)}$", where λ' is any value between λ and 1. Theorem 12 states that for appropriate $d = d(\lambda', c)$ and sufficiently large N, $\Pr[B_N] \geq \delta$. On the other hand, events B_N and $B_{N'}$ are essentially independent for $N' \gg N$. Therefore, the probability that infinitely many events B_N occur is 1 (details are omitted).

4 An Anytime Estimator Algorithm

In this section we describe an anytime estimation algorithm for Bernoulli variables which, by the results in the next section, has essentially optimal efficiency. We describe in less detail a more general estimation algorithm that works for all variables to which Bernstein's inequality can be applied; the algorithm will be optimal to the extent that Bernstein's inequality is tight for the random variable in question.

The pseudocode for the anytime frequency estimator is given in Figure 1. Observe that its structure is quite straightforward and that it performs a constant amount of work per iteration. The only difficulty is to compute an appropriate ϵ_n, which we do as follows.

procedure *FrequencyEstimator*
  input δ;
  $t := 0$; $\vec{X}_0 :=$ the empty sequence;
  while true **do**
    $t := t + 1$;
    compute $\mu_t := \mathrm{freq}(X_1, \ldots, X_t)$;
    compute ϵ_t (by some formula given in the analysis);
    output (μ_t, ϵ_t);
  end while

Figure 1: The Optimal Anytime Frequency Estimator Algorithm

Fix some $\lambda > 1$ and define $\theta_t(p)$ as given by the formula for $\epsilon_t(p)$ in Equation (4) (θ_t depends on other quantities besides t and p, but this is irrelevant now). By Theorem 10, with probability at least $1 - \delta$ we have

$$\forall t \,:\, (1 - \theta_t(p)) \cdot p \leq \mu_t \leq (1 + \theta_t(p)) \cdot p\,. \tag{5}$$

We want to define a formula for $\epsilon_t(\mu_t)$, which may depend on other parameters but not on p, such that condition (5) implies

$$\forall t \,:\, (1 - \epsilon_t(\mu_t)) \cdot p \leq \mu_t \leq (1 + \epsilon_t(\mu_t)) \cdot p\,. \tag{6}$$

That is, we want

$$|\mu_t - p| \leq \theta_t(p)\, p \Longrightarrow |\mu_t - p| \leq \epsilon_t(\mu_t)\, \mu_t$$

Observe that (ignoring a small term), θ_t is defined as $\max\{\sqrt{\alpha/p},\, \beta/p\}$, for quantities α and β that do not depend on p. Considering the two cases separately, one can show that condition (7)

$$\epsilon_t(\mu_t) = \max\left\{ \frac{1}{2\mu_t}\left(\alpha + \sqrt{\alpha^2 + 4\mu_t\alpha}\right)\,,\, \beta/\mu_t \right\} \tag{7}$$

implies condition (6). This proves:

Theorem 13 *Procedure* FrequencyEstimator *given in Figure 1 is a correct anytime frequency estimator algorithm if $\epsilon_t(\mu_t)$ is defined as in condition (7).*

Observe that α tends to 0 as t increases, and that β does so even faster. Therefore, for fixed μ_t, $\epsilon_t(\mu_t)$ is dominated by the term

$$\frac{1}{2\mu_t} \cdot \sqrt{4\mu_t\alpha}$$

and so (expanding α)

$$\epsilon_t(\mu_t) = \sqrt{\frac{2\lambda\, q\, (\ln(1/\delta) + \ln\ln n)}{\mu_t\, t}} \cdot (1 + o(1)).$$

Assume now that the random variable X generating trials X_1, X_2, ... is not Bernoulli. Instead of Equation (4), we would like to use the more general bound given by Theorem 10. But that bound involves, besides $\mu = E[X]$, the variance σ^2 and the range $[a, b]$ of X, which may be unknown in general. We do assume that the range $[a, b]$ is known.

If any upper bound $\hat{\sigma}^2$ is known, then using it in the formula for $\epsilon(n, \delta)$ of Theorem 10 still yields a valid statement. Then, for any fixed $\lambda > 1$, an anytime frequency estimator for $\mu = E[X]$ can be designed exactly as in the Bernoulli case.

If no upper bound on σ^2 is known, the the technique in [2] can be used: they design a simple variant of their one-shot algorithm that suffices for obtaining approximation within a constant factor and that requires no variance information. The the anytime algorithm has three phases: First, use the simple estimator to obtain an estimator $\hat{\mu}$ of μ within some small constant factor. Second, use $\hat{\mu}$ and the simple estimator to compute an estimation $\hat{\sigma}^2$ of $\sigma^2 = Var(X)$ within a factor λ', where where λ' is an appropriate constant between 1 and λ. In particular, $\hat{\sigma}^2$ is the desired upper bound on σ. Until this point, the anytime outputs a trivially correct ϵ_t (e.g., making the confidence interval to be $[a, b]$). Then the third phase starts: using $\hat{\sigma^2}$ in place of σ^2, the anytime estimator uses the bound given by Theorem 10 to compute a decreasing sequence of ϵ_t.

With this idea one can prove:

Theorem 14 *For every $\lambda > 1$ and every $[a, b]$ there is an anytime estimator $A = (\text{freq}, \epsilon)$ for $\mu = E[X]$ that is correct for all X provided $a \leq X \leq b$ and $\mu > 0$, and such that*

$$\epsilon(t, \delta, \mu_t) \leq \lambda \sqrt{\frac{2Var(X)}{\mu_t^2 t} \cdot (\ln \frac{1}{\delta} + \ln \ln t)} \cdot (1 + o(1)).$$

5 Lower Bound for Bernoulli variables

In this section we give a tight lower bound on the efficiency of anytime frequency estimators.

Definition 15 *For a fixed (implicit) λ, function ϵ_{min} is defined as:*

$$\epsilon_{min}(\vec{X}_t, \delta) \stackrel{\text{def.}}{=} \sqrt{\frac{2\lambda}{\text{freq}(\vec{X}_t) \cdot t} \cdot \ln(1/\delta)} \ .$$

Theorem 16 *Let $X_1, X_2, \ldots, X_t, \ldots$ be Bernoulli variables with $\mu = E[X_t] = p > 0$. For any $\lambda < 1$, any correct anytime estimator $A = (\hat{\mu}, \epsilon)$, any $p < 1$, any $\delta < 1$, and any sufficiently large t, the event $\epsilon(\vec{X}_t, \delta) \geq \epsilon_{min}(\vec{X}_t, \delta)$ occurs with probability at least $\delta' = \delta^{\sqrt{\lambda}} - \delta^{4\lambda} - 2\delta$, if the probability is taken over all vectors $\vec{X}_t = (X_1, \ldots, X_t)$.*

Note that $\delta' > 0$ for all δ sufficiently small w.r.t. λ. So the following is a corollary to this theorem.

Corollary 17 *For any $\lambda < 1$ and any pair $A = (\hat{\mu}, \epsilon)$, if $\epsilon(\vec{X}_t, \delta) < \epsilon_{min}(\vec{X}_t, \delta)$ with probability 1, then A is not a correct anytime estimator.*

Proof of Theorem 16. The essence of the proof is as follows: We choose two probabilities p_1 and p_2 such that p_1 and p_2 differ (multiplicatively) by more than ϵ. We show that if $A = (\hat{\mu}, \epsilon)$ is such that $\epsilon < \epsilon_{min}$ with probability δ', then A fails with probability δ to distinguish between the cases $\mu = p_1$ and $\mu = p_2$, hence it is not a correct estimator.

Fix $\lambda < 1$, then fix a pair $A = (\hat{\mu}, \epsilon)$, any $p_1 < 1$ and any δ. Choose t large w.r.t. δ and p_1. Define ϵ^* as the value satisfying $\epsilon^* = \epsilon_{min}(\vec{X}_t, \delta)$ for some vector $\vec{X}_t$ with $\mathrm{freq}(\vec{X}_t) = (1 - \epsilon^*/\sqrt{\lambda})\, p_1$. This is a recursive but correct definition: all such vectors $\vec{X}_t$ provide the same ϵ^*, and, because of the form of ϵ_{min}, such an ϵ^* exists for all sufficiently large t.

Define also $p_2 = (1 + 2\epsilon^*)p_1$. We assume that t is so large that $(1 + \epsilon^*)p_1 \cong (1 - \epsilon^*)p_2$, $\epsilon^* \ll 1$, and $p_2 < 1$. We will use repeatedly the fact that

$$\epsilon^* \cong \sqrt{\frac{2\lambda}{p_1 t} \ln \frac{1}{\delta}} \qquad \text{, therefore} \quad \exp(-(1/2)(\epsilon^*)^2 p_1 t) \cong \delta^\lambda.$$

Let M be the set of sequences $\vec{X}_t$ such that $\mathrm{freq}(\vec{X}_t) \in [p_1, p_2]$. Partition M into two sets M_1 and M_2 as follows: If $\hat{\mu}(\vec{X}_t, \delta) < (1 + \epsilon^*)p_1 (= (1 - \epsilon^*)p_2)$, put $\vec{X}_t$ in M_1, otherwise put it in M_2.

Suppose w.l.o.g. that $|M_1| \leq |M_2|$, so that M_2 contains at least half the strings in M. Assuming this, fix $\mu = p_1$, i.e., from now on A is assumed to run under $\mu = p_1$ (if $|M_1| > |M_2|$, fix $\mu = p_2$ instead).

Let S_1, S_2, and S_3 be the sets of $\vec{X}_t$ such that $\epsilon(\vec{X}_t, \delta) \geq \epsilon^*$, $\epsilon(\vec{X}_t, \delta) \geq \epsilon_{min}(\vec{X}_t, \delta)$, and $\epsilon_{min}(\vec{X}_t, \delta) \geq \epsilon^*$, respectively. We theorem states precisely that $\Pr[S_2] \geq \delta'$. Note that $S_1 \subseteq S_2 \cup S_3$, so

$$\Pr[S_2] \geq \Pr[S_1] - \Pr[S_3].$$

We prove the theorem by showing an upper bound on $\Pr[S_3]$, then a lower bound on $\Pr[S_1]$.

Upper bound on $\Pr[S_1]$: By the definition of ϵ^* and Chernoff bound we have

$$\begin{aligned} \Pr[S_3] &= \Pr[\,\epsilon_{min}(\vec{X}_t, \delta) \geq \epsilon^*\,] = \Pr[\,\mathrm{freq}(\vec{X}_t) \leq (1 - \epsilon^*/\sqrt{\lambda})p_1\,] \\ &\leq \exp(-(1/2)(\epsilon^*/\sqrt{\lambda})^2 p_1 t) \cong (\delta^\lambda)^{1/\lambda} = \delta. \end{aligned}$$

Lower bound on $\Pr[S_2]$: The argument is as follows: we will show presently that $\Pr[M_2]$ is quite large if $\vec{X}_t$ is generated using $\mu = p_1$. Then note that for input sequences $\vec{X}_t \in M_2$ we have $\hat{\mu}(\vec{X}_t, \delta) > (1 - \epsilon^*)p_1 t \cong (1 - \epsilon_{min}(\vec{X}_t, \delta))p_1 t$. That is, on all sequences $\vec{X}_t \in M_2$, $\hat{\mu}(\vec{X}_t, \delta)$ is not a correct $\epsilon_{min}(\vec{X}_t, \delta)$-approximation

of $\mu = p_1$. But $A = (\hat{\mu}, \epsilon)$ is a correct estimator, thus $\hat{\mu}(\vec{X}_t, \delta)$ is allowed *not* to be a correct $\epsilon(\vec{X}_t, \delta)$-approximation of $\mu = p_1$ with probability at most δ. Therefore, we must have $\epsilon(\vec{X}_t, \delta) \geq \epsilon_{min}(\vec{X}_t, \delta)$ with probability at least $\Pr[M_2] - \delta$, or in other words $\Pr[S_2] \geq \Pr[M_2] - \delta$.

In order to upper-bound $\Pr[M_2]$, recall that $M = M_1 \cup M_2$ and $|M_2| \geq |M_1|$. Among all possible subsets of M containing about half the strings in M, the one having minimum probability when $\mu = p_1$ is the one containing all strings whose number of 1s is in the interval $[(1+\epsilon^*)p_1 t, (1+2\epsilon^*)p_1 t] (= [(1-\epsilon^*)p_2 t, p_2 t])$. Assume that M_2 is precisely this set, i.e., assume the worst case for M_2. Define $\gamma = 1/\sqrt{\lambda} > 1$, let S_t be $X_1 + \ldots + X_t$, and apply Lemmas 8 and 11: If t is sufficiently large,

$$\begin{aligned}
\Pr[M_2] &= \Pr[S_t \in [(1+\epsilon^*)p_1 t, (1+2\epsilon^*)p_1 t]] \\
&= \Pr[S_t \geq (1+\epsilon^*)p_1 t] - \Pr[S_t \geq (1+2\epsilon^*)p_1 t] \\
&\geq \exp(-\frac{1}{2}\gamma\,(\epsilon^*)^2 p_1 t) - \exp(-\frac{1}{2}\,(2\epsilon^*)^2 p_1 t) \\
&\cong (\delta^\lambda)^\gamma - (\delta^\lambda)^4 = \delta^{\sqrt{\lambda}} - \delta^{4\lambda}.
\end{aligned}$$

Putting everything together,

$$\Pr[S_2] \geq \Pr[S_1] - \Pr[S_3] \geq \Pr[M_2] - \delta - \Pr[S_3] \geq \delta^{\gamma\lambda} - \delta^{4\lambda} - \delta - \delta = \delta'.$$

as desired. ∎ (Theorem 16)

It is routine to verify that the proof still holds if δ is not constant but $\delta = \delta(t) = \delta_0 / \ln t$, for some fixed δ_0. This will be used in order to prove the next lower bound, which is tight for anytime estimators.

Theorem 18 *Let $X_1, \ldots, X_t, \ldots$ be Bernoulli variables with $\mu = E[X_t] = p$. For all $\lambda < 1$, if if*

$$\epsilon(\vec{X}_t, \delta) \leq \sqrt{\frac{2\,\lambda}{\mathrm{freq}(\vec{X}_t) \cdot t} \cdot (\ln\ln t + \ln(8/\delta))}$$

holds for all $\vec{X}_t$ and δ, then no pair $A = (\hat{\mu}, \epsilon)$ is a correct anytime estimator.

Proof. *(Sketch)* Observe first that $\ln\ln t + \ln(8/\delta) = \ln(1/(\delta/8\ln t))$. Fix $\lambda < 1$ and some large t for a moment, then let ϵ_t^* be $\epsilon_{min}(\vec{X}_t, 4\delta/\ln t)$.

Call B_t the negation of the event "$(1-\epsilon_t)\mu \leq \hat{\mu}_t \leq (1+\epsilon_t)\mu$" and C_t the negation of the event "$(1-\epsilon_t^*)\mu \leq \hat{\mu}_t \leq (1+\epsilon_t^*)\mu$". Observe that $\Pr[B_t]$ is at least $\Pr[C_t]$ minus the probability that ϵ and ϵ^* differ by a suitably defined small factor. The latter event occurs if $\mathrm{freq}(\vec{X}_t)$ and μ differ by again a small factor, which occurs with probability at most $\delta/(4\ln t)$ if A is a correct estimator.

On the other hand, by Theorem 16, the probability that event C_t occurs is at most $(\delta/4\ln t)' - (\delta/4\ln t)$ for all t sufficiently large. Therefore,

$$\Pr[B_t] > \Pr[C_t] - (\delta/4\ln t) \cong (\delta/4\ln t)^{\sqrt{\lambda}} \geq 2\delta/\ln t$$

for all sufficiently large t. Now choose a sequence $t_1, t_2, \ldots, t_k, \ldots$ sufficiently spaced out that all events B_{t_k} are independent (up to a negligibly small amount). Then, by Lemma 7, for all K sufficiently large

$$\Pr[\,\exists k : K \leq k \leq 2K : \neg B_{t_k}\,] \geq \delta$$

and A is not a correct anytime estimator. ∎

In fact A will fail quite frequently. By an argument as in the proof of Theorem 5, one can show that within each interval $[N, N^d]$ there is probability δ that it fails at least once.

For more general random variables X, the same technique should suffice to prove a lower bound provided one can show both an upper and a lower bound on the concentration of S_n. That is, one needs the analog of Chernoff (or Bernstein) inequalities and the analog of Theorem 11 for X. This is all the information on X that is used in the proof of the lower bound.

6 Future Work

The following are two questions for future work.

One, as already mentioned, in some applications one is interested in approximating quantities other than the average of the trials. For example, in [4], an application to Boosting is described where the quantity of interest is the amount by which the average exceeds 1/2. For application to decision tree induction (in C4.5 style), one may be interested in approximation the approximation of the entropy of certain sequences of trials.

Second, and probably more important, is dealing with sequences of trials that, while still independent, are not identically distributed. More Precisely, we have in mind the situation where the random process generating the trials slowly varies over time. We would like estimators that track such changes while still being approximately correct at all times. There seems to be a tradeoff between using only the most recent trials (hence, having little basis for estimation) and using least recent data (hence, loosing sensitivity to time changes). This question seems particularly acute in the Data Stream model [1], where a central assumption is that data may change over time.

Acknowledgements

We thank Osamu Watanabe for his hospitality at the Tokyo Institute of Technology where some questions addressed in this paper were formulated. We thank Gábor Lugosi for telling us about Lemma 6, which greatly simplified a previous proof. Finally, we thank Marco Minozzo for sending us a copy of his work [20].

References

[1] B. Babcock, S. Babu, M. Datar, R. Motwani, J. Widom: "Models and issues in Data Stream systems". *Proc. of the 2002 ACM Symp. on Principles of Database Systems (PODS 2002),* 2002.

[2] P. Dagum, R. Karp, M. Luby, S. Ross: "An optimal algorithm for monte carlo estimation", *SIAM J. Comput.* 29(5), 1484–1496, 2000.

[3] C. Domingo, R. Gavaldà, Osamu Watanabe. "Practical algorithms for on-line sampling". *Proc. First International Conference on Discovery Science (DS'98).* Springer-Verlag Lecture Notes in Artificial Intelligence 1532 (1998), 150-161.

[4] C. Domingo, R. Gavaldà, Osamu Watanabe. "Adaptive sampling methods for scaling up knowledge discovery algorithms". *Data Mining and Knowledge Discovery* 6 (2002), 131–152.

[5] P. Domingos, G. Hulten: "Mining high-speed data streams". *Proc. 6th Intl. Conference on Knowledge Discovery in Databases*, ACM Press, pp.71–80, 2000.

[6] P. Domingos, G. Hulten: "A general method for scaling up machine learning algorithms and its applications to clustering". *Proc. 8th Intl. Conference on Machine Learning*, Morgan Kaufmann, pp.106–113, 2001.

[7] W. Feller: *An Introduction to Probability Theory and its Applications* (3rd Edition). John Wiley & Sons, 1968.

[8] R. Gavaldà, O. Watanabe: "Sequential sampling algorithms: Unified analysis and lower bounds". *Proc. 1st Intl. Symposium on Stochastic Algorithms: Foundations and Applications (SAGA'01).* Springer-Verlag Lecture Notes in Computer Science 2264 (2001), 173-187.

[9] R. Gavaldà: "Clustering and classification in large datasets by adaptive sampling". In preparation.

[10] B.K. Ghosh, M. Mukhopadhyay, P.K. Sen: *Sequential Estimation.* Wiley, 1997.

[11] A. Khintchine: "Über einen Satz der Wahrscheinlichhkeitsrechnung". *fundamenta Mathematicae* **6** (1924), 9–20.

[12] A.N. Kolmogorov: "Das Gesetz des iterierten Logarithmus". *Mathematische Annalen* **101** (1929), 126–135.

[13] G. Lugosi: *Concentration-of-measure inequalities.* Lecture notes, 2004. `http://www.econ.upf.es/~lugosi/anu.ps`

[14] G. Hulten, P. Domingos: "Mining complex models from arbitrarily large databases in constant time". *Proc. SIGKDD02*, 2002.

[15] G. Hulten, L. Spencer, P. Domingos: "Mining time-changing data streams". *Proc. KDD'01 Conference*, 2001.

[16] J. Kivinen, H. Mannila. "The power of sampling in knowledge discovery" *Proceedings of the ACM SIGACT-SIGMOD-SIGACT Symposium on Principles of Database Theory* (1994), 77–85.

[17] R.J. Lipton, J.F. Naughton, D.A. Schneider, and S. Seshadri: "Efficient sampling strategies for relational database operations". *Theoretical Computer Science* **116** (1993), 195–226.

[18] R.J. Lipton and J.F. Naughton: "Query size estimation by adaptive sampling". *Journal of Computer and System Sciences* **51** (1995), 18–25.

[19] J.F. Lynch: "Analysis and application of adaptive sampling". *Journal of Computer and System Sciences* **66** (2003), 2–19. Preliminary version in PODS'2000.

[20] M. Minozzo: "Purely game-theoretic random sequences: I. Strong Law of Large Numbers and Law of the Iterated Logarithm". *Theory of Probability & Its Applications* **44**:3 (2000), 511–522.

[21] A. Rényi: *Probability Theory.* North Holland, 1970.

[22] T. Scheffer and S. Wrobel, A sequential sampling algorithm for a general class of utility criteria, in *Proc. of the 6th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, ACM Press, 2000, to appear.

[23] T. Scheffer, S. Wrobel: "Finding the most interesting patterns in a database quickly by using sequential sampling" *Journal of Machine Learning Research* 3 (2002) 833-862.

[24] T. Scheffer, S. Wrobel: "A scalable constant-memory sampling algorithm for pattern discovery in large databases". *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery and Data Mining*, 2002.

[25] H. Toivonen: "Sampling large databases for association rules". *Proceedings of the 22nd International Conference on Very Large Databases* (1996), 134–145.

[26] A. Wald: *Sequential Analysis.* John Wiley & Sons, 1947.

Appendix: Proof of Lemmas

Proof of Lemma 6. For $k \in \{1 \ldots n\}$, let A_k be the event

$$S_1 < x \wedge S_2 < x \wedge \ldots \wedge S_{k-1} < x \wedge S_k \geq x,$$

let B_k be the event "$S_n - S_k > -\sqrt{2\sigma^2 n}$", and A the event "$S_n \geq x - \sqrt{2\sigma^2 n}$". Observe that

$$1 - \Pr[B_k] \leq \Pr[|S_n - S_k| \geq \sqrt{2\sigma^2 n}].$$

By Chebyshev's inequality,

$$1 - \Pr[B_k] \leq \frac{n-k}{2n} \leq 1/2$$

so $\Pr[B_k] \geq 1/2$. Now use that A_k and B_k are independent, that $A_i B_i$ and $A_j B_j$ are disjoint, and that $A_k B_k$ implies A:

$$\begin{aligned}
\Pr[\max_k\{S_k\} \geq x\,] &= \sum_k \Pr[A_k] =\leq 2\sum_k \Pr[A_k]\Pr[B_k] \\
&= 2\sum_k \Pr[A_k]\Pr[B_k] = 2\sum_k \Pr[A_k B_k] \\
&= 2\Pr[\bigcup_k A_k B_k] \leq \Pr[A].
\end{aligned}$$

■ (Lemma 6)

Proof of Lemma 7. Use independence and the fact that $1-2x \leq \exp(-2x) \leq 1-x$ for all $x \in (0, 3/4)$:

$$\begin{aligned}
&\Pr[\,\exists k : k_0 \leq k \leq \alpha \cdot k_0 : B_k\,] = 1 - \Pr[\forall k : k_0 \leq k \leq \alpha \cdot k_0 : \neg B_k] \\
= \;& 1 - \prod_{k=k_0}^{\alpha k_0} (1 - \Pr[B_k]) \geq 1 - \prod_{k=k_0}^{\alpha k_0} \left(1 - \frac{2\delta}{k \cdot (\alpha - 1)}\right) \\
\geq \;& 1 - \prod_{k=k_0}^{\alpha k_0} \left(1 - \frac{2\delta}{k_0 \cdot (\alpha - 1)}\right) \geq 1 - \left(1 - \frac{2\delta}{k_0 \cdot (\alpha - 1)}\right)^{(\alpha-1)k_0} \\
\geq \;& 1 - \exp(-2\delta) \geq \delta\,.
\end{aligned}$$

■ (Lemma 7)

Proof of Lemma 9. By the definition of ϵ, we have

$$\frac{1}{\gamma} \cdot \epsilon^2 n \geq 2\beta\sigma^2 \quad \text{and} \quad \frac{\gamma - 1}{\gamma} \cdot \epsilon^2 n \geq 2\beta \frac{\epsilon M}{3}.$$

Adding both inequalities we have

$$\epsilon^2 n \geq 2\beta(\sigma^2 + \frac{\epsilon M}{3}).$$

Then by Lemma 8

$$\Pr[\,|S_n| \geq \epsilon n] \leq 2\,\exp\left(-\frac{1}{2}\,\frac{\epsilon^2 n}{\sigma^2 + \epsilon M/3}\right) \leq 2\exp(-\beta).$$

■ (Lemma 9)

Lemma 19 *For all p, n, and all ϵ such that $\epsilon \leq 1$ and $(1+\epsilon)p \leq 1$,*

$$\begin{aligned}\Pr\,[\,S_n = (1+\epsilon)pn\,] \;&\geq\; \frac{1}{\sqrt{2\pi pqn}}\cdot \\ &\cdot\quad \exp\left(-\frac{1}{2}\,\frac{1}{q}\,\epsilon^2 pn - (\frac{1}{2}\,\frac{p^2}{q^2} - c)\,\epsilon^3 pn\right),\end{aligned}$$

for $c = -3/2 + 2\ln(2) \cong 0.1137$.

Proof of Lemma 19. By using Stirling's approximation, we have

$$\begin{aligned}\Pr[\,S_n \;&=\; (1+\epsilon)pn\,] = \binom{n}{(1+\epsilon)pn}\,p^{(1+\epsilon)pn}\,q^{(1-(1+\epsilon)p)n} \\ &\cong\; \frac{1}{\sqrt{2\pi pqn}}\cdot(1+\epsilon)^{-(1+\epsilon)pn}\cdot\left(\frac{1-(1+\epsilon)p}{q}\right)^{-(1-(1+\epsilon)p)n} \\ &=\; \frac{1}{\sqrt{2\pi pqn}}\cdot \\ &\qquad\cdot\exp\left(-\sum_{i\geq 2}\frac{1}{i(i-1)}\left((-1)^i + (p/q)^{i-1}\right)\epsilon^i pn\right).\end{aligned}$$

Using that $\epsilon < 1$ and $(1+\epsilon)p \leq 1$ ($\Leftrightarrow \epsilon q/p \leq 1$), and

$$\begin{aligned}&\sum_{i\geq 2}\frac{1}{i(i-1)}\left((-1)^i + (p/q)^{i-1}\right)\epsilon^i pn \\ =\;& \frac{1}{2}\left(1+\frac{p}{q}\right)\epsilon^2 pn + \frac{1}{6}\left(-1+\frac{p^2}{q^2}\right)\epsilon^3 pn + \\ &\left(\sum_{i\geq 4}\frac{(-1)^i}{i(i-1)}\,\epsilon^{i-3} + \frac{p^2}{q^2}\sum_{i\geq 4}\frac{1}{i(i-1)}\,(p\epsilon/q)^{i-3}\right)\epsilon^3 pn \\ \leq\;& \frac{1}{2}\,\frac{1}{q}\,\epsilon^2 pn + \frac{1}{6}\left(-1+\frac{p^2}{q^2}\right)\epsilon^3 pn + \\ &\left(\sum_{i\geq 4}\frac{(-1)^i}{i(i-1)}\,1^{i-3} + \frac{p^2}{q^2}\sum_{i\geq 4}\frac{1}{i(i-1)}\,1^{i-3}\right)\epsilon^3 pn \\ =\;& \frac{1}{2}\,\frac{1}{q}\,\epsilon^2 pn + \frac{1}{6}\left(-1+\frac{p^2}{q^2}\right)\epsilon^3 pn + \left((-4/3 + 2\ln 2) + \frac{1}{3}\,\frac{p^2}{q^2}\right)\epsilon^3 pn,\end{aligned}$$

and the lemma follows. ∎ (Lemma 19)

Proof of Lemma 11. Let μ be λ^{-1}. Define as $s = \sqrt{2\pi pqn}$. Observe that

$$\begin{aligned}
\Pr[\, S_n \geq (1+\epsilon)pn \,] &\geq \sum_{i=0}^{s} \Pr[\, S_n = (1+\epsilon)pn + i \,] \\
&\geq s \cdot \Pr[\, S_n = (1+\epsilon)pn + s \,] \\
&= s \cdot \Pr\left[S_n = \left(1 + \epsilon + \sqrt{2\pi q/pn}\right) pn \right] .
\end{aligned}$$

Then apply Lemma 19 assuming enough on n and ϵ, namely that $\epsilon + \sqrt{2\pi q/pn} \leq \min\{1, q/p\}$ and we obtain

$$\left(\frac{1}{2}\frac{p^2}{q^2} - c\right)(\epsilon + \sqrt{2\pi q/pn}) \leq R\frac{1}{2}\frac{1}{q}(\mu - 1) .$$

The last condition is implied by $\epsilon \leq \sqrt{2\pi q/pn}$ and $n \geq \sqrt{2\pi}/q(\mu-1)^2$. ∎ (Lemma 11)

Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

Research Reports - 2004

- LSI-04-1-R : *Automatic Generation of Polynomial Loop Invariants: Algebraic Foundations*, Rodríguez, E. and Kapur, D.
- LSI-04-2-R : *Comparison of Methods to Predict Ozone Concentration* , Orozco, J.
- LSI-04-3-R : *Towards the definition of a taxonomy for the cots product´s market* , Ayala, Claudia P.
- LSI-04-4-R : *Modelling Coalition Formation over Time for Iterative Coalition Games*, Mérida-Campos, C. and Willmott, S.
- LSI-04-5-R : *Illegal Agents? Creating Wholly Independent Autonomous Entities in Online Worlds*, Willmott, S.
- LSI-04-6-R : *An Analysis Pattern for Electronic Marketplaces*, Queralt, A. and Teniente, E.
- LSI-04-7-R : *Exploring Dopamine-Mediated Reward Processing through the Analysis of EEG-Measured Gamma-Band Brain Oscillations*, Vellido, A. and El-Deredy, W.
- LSI-04-8-R : *Studying Embedded Human EEG Dynamics Using Generative Topographic Mapping*, Vellido, A. and El-Deredy, W. and Lisboa, P.J.G.
- LSI-04-9-R : *Similarity and Dissimilarity Concepts in Machine Learning*, Orozco, J.
- LSI-04-10-R : *A Framework for the Definition of Metrics for Actor-Dependency Models*, Quer, C. and Grau, G. and Franch, X.
- LSI-04-11-R : *QM: A Tool for Building Software Quality Models*, Carvallo, J.P. and Franch, X. and Grau, G. and Quer, C.
- LSI-04-12-R : *COSTUME: A Method for Building Quality Models for Composite COTS-based Software Systems*, Carvallo, J.P. and Franch, X. and Grau, G. and Quer, C.
- LSI-04-13-R : *Enabling Collaboration in Virtual Reality Navigators*, Theoktisto, V. and Fairén, M. and Navazo, I.
- LSI-04-14-R : *DesCOTS: A Software System for Selecting COTS Components*, Carvallo, J.P. and Franch, X. and Grau, G. and Quer, C.
- LSI-04-15-R : *Evaluation and symmetrisation of alignments obtained with the Giza++ software*, Lambert, P. and Castell, N.
- LSI-04-16-R : *A note on the use of topology extensions for provoking instability in communication networks*, Blesa, M.J.
- LSI-04-17-R : *An ISO/IEC-compliant Quality Model for ER Diagrams*, Costal, D. and Franch, X.
- LSI-04-18-R : *A Case Study on Pruning General Ontologies for the Development of Conceptual Schemas* , Conesa, J.
- LSI-04-19-R : *Adding Efficient and Reliable Access Paths to the JCF*, Marco, J. and Franch, X.

- LSI-04-20-R : *Exploiting Simple Corporate Memory in Iterative Coalition Games*, Mérida-Campos, C. and Willmott, S.
- LSI-04-21-R : *On the Semantics of Operation Contracts in Conceptual Modeling* , Queralt, A. and Teniente, E.
- LSI-04-22-R : *Complexity issues on bounded restrictive H-coloring*, Díaz, J. and Serna, M. and Thilikos, D.M.
- LSI-04-23-R : *Chromatic number in random scaled sector graphs*, Díaz, J. and Sanwalani, V. and Serna, M. and Spirakis, P.
- LSI-04-24-R : *Bounds on the bisection width for random d-regular graphs*, Díaz, J. and Serna, M. and Wormald, N.C.
- LSI-04-25-R : *Open Source environment to define constraints in route planning for GIS-T*, Pérez, L. and Silveira, A. da M.
- LSI-04-26-R : *A basic repository of operations for the refinement of general ontologies*, de Palol, X.
- LSI-04-27-R : *Tetrahedral mesh subdivision based on underlying volume data*, Rodríguez, L. and Navazo, I. and Vinacua, A.
- LSI-04-28-R : *The Price of Connectedness in Expansions*, Fomin, F.V. and Fraigniaud, P. and Thilikos, D.M.
- LSI-04-29-R : *Smaller kernels for hitting set problems of constant arity*, Nishimura, N. and Ragde, P. and Thilikos, D.M.
- LSI-04-30-R : *Searching Spatial Sense in the Ontological World: Discovering Spatial Objects*, Morocho, V and Pérez, L. and Saltor, F.
- LSI-04-32-R : *Implementation considerations of an Expert System to assess Stream Water Quality management*, Cabanillas, D. and Willmott, S.
- LSI-04-33-R : *Multisided patches*, Pla, N. and Vigo, M. and Cotrina, J.
- LSI-04-34-R : *SVMTool: A general POS tagger generator based on Support Vector Machines*, Giménez, J. and Màrquez, Ll.
- LSI-04-35-R : *A distributed and mobile component system based on the ambient calculus*, Mylonakis, N. and Orejas, F.
- LSI-04-36-R : *Developing Competitive HMM PoS Taggers Using Small Training Corpora*, Padró, M. and Padró, Ll.
- LSI-04-37-R : *The AlignmentSet Toolkit*, Lambert, P.
- LSI-04-38-R : *Integración de Fuentes de Datos espaciales: análisis e implementación de una Ontología de términos espaciales: Primera Parte - - Creación de una Ontología*, Ramos, Erik G.
- LSI-04-39-R : *Integración de Fuentes de Datos espaciales: análisis e implementación de una Ontología de términos espaciales: Segunda Parte - - Evaluación de similaridades*, Ramos, Erik G.
- LSI-04-40-R : *Kernels on Structured Domains*, Valentín, L.

- LSI-04-41-R : *Determining the Structural Events that May Violate an Integrity Constraint*, Cabot, J. and Teniente, E.
- LSI-04-42-R : *Review of Statistical Word Alignment Techniques* , Lambert, P.
- LSI-04-43-R : *Algoritmos geneticos en el problema de la solucion deseadaÒptimizacion de parametros*
, Barreiro, E. and Joan-Arinyo, R. and Luzón, M.V.
- LSI-04-44-R : *Generative Topographic Mapping as a constrained mixture of Student t-distributions: Theoretical developments* , Vellido, A.
- LSI-04-45-R : *Adapting Agent Communication Languages for Web Service to Web Service Communication*, Willmott, S. and Fernández-Peña, F. O. and Mérida-Campos, C. and Con-stantinescu, I.
- LSI-04-49-R : *A brief on constraint solving*, Hoffmann, C.M. and Joan-Arinyo, R.
- LSI-04-50-R : *Missing data imputation through Generative Topographic Mapping as a mixture of t-distributions: Theoretical developments*, Vellido, A.
- LSI-04-51-R : *A two-tiered Methodology for Metamodel Extension Applied to UML 1ì*, Franch, X. and Ribó, J. M.
- LSI-04-52-R : *Virtual reality for prostate gland cryosurgery*, Joan-Arinyo, R.
- LSI-04-53-R : *High level communication functionalities for wireless sensor networks*, Àlvarez, C. and Díaz, J. and Petit, J. and Rolim, J. and Serna, M.
- LSI-04-55-R : *Pure Nash equilibria in games with a large number of actions*, Àlvarez, C. and Gabarró, J. and Serna, M.
- LSI-04-56-R : *An Optimal Anytime Estimation Algorithm*, Gavaldà, R.

Hardcopies of reports can be ordered from:

Núria Sanchez
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Campus Nord, Mòdul C6
Jordi Girona Salgado, 1-3
03034 Barcelona,Spain
nurias@lsi.upc.es

See also the Departament WWW pages, http://www.lsi.upc.es/