An Optimal Anytime Estimation Algorithm

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

May 3rd, 2005

Abstract

An *anytime algorithm* is one that produces a reliable answer at all times, instead of only when it halts. We present an anytime algorithm that approximates the average value of a potentially infinite sequence of independent, identically distributed trials: at every time step, the algorithm produces an estimation and an uncertainty margin; with high probability, the estimations are *all* within their claimed margins of the true average value, and the uncertainty margins tend to 0 as the number of trials grows. Furthermore, we prove that the algorithm is optimal in the sense that no other anytime algorithm can satisfy these three conditions and produce asymptotically smaller uncertainty margins. The key ingredient in our results is a bound on the supremum of the deviations of partial sums of an infinite sequence of trials, which can be seen as a non-limit analog of the classical Law of the Iterated Logarithm.

*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 the EU Network of Excellence PASCAL.

1 Introduction

The following situation occurs in a wide range of applications: an algorithm has access to a sequence of random trials $X_1, X_2, \ldots, X_t, \ldots$, and is required to approximate an unknown quantity μ underlying the generation of the trials. The X_t may be the outcome of experiments designed on purpose to estimate μ, or may come from an external source of data whose distribution we want to approximate. In this paper, we consider the case in which the X_i are independent and identically distributed and $\mu = E[X_t]$.

Dagum, Karp, Luby, and Ross [1] design an optimal algorithm for what we call the *one-shot* estimation task. A one-shot estimator receives parameters ϵ and δ as inputs, reads the trials in sequence, and in finite time stops producing a number $\hat{\mu}$ such that,

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

The algorithm by Dagum *et al.* achieves this goal and is optimal, in the sense that every possible one-shot algorithm for the same problem must see as many trials as their algorithm before producing such a number $\hat{\mu}$.

This algorithm can be applied as a black box in a wide variety of approximation tasks (see the references in [1]). In many other situations, however, an algorithm that works in a one-shot way is not what is required. Rather, we would like an algorithm that provides approximations of μ while it is running, even if these approximation are not very accurate at first. Accuracy should improve as the running time increases. The algorithm might stop at some point when some prespecified accuracy has been reached, or continue forever, possibly achieving perfect accuracy in the limit.

We look for anytime algorithms for the estimation task such that with high probability they are *at all times* correct. This is stronger that requiring that, for any *fixed* time step, the answer is correct with high probability: The algorithm produces (infinitely) many answers so, unless proven otherwise, small probabilities of error at every step might accumulate into a large probability of *sometime* making an error. The stronger condition may be required in situations where a single incorrect answer may lead us to a drastic decision, potentially catastrophic or very costly.

We present an anytime algorithm for the problem of estimating $\mu = E[X_t]$ as above, when the X_t have a bounded and known range. The algorithm reads a parameter δ and runs forever; at step t, it reads trial X_t and produces a

pair $(\hat{\mu}_t, \epsilon_t)$, where $\hat{\mu}_t$ is the current approximation of μ and ϵ_t is an uncertainty margin for $\hat{\mu}_t$. The infinite series of pairs $(\hat{\mu}_t, \epsilon_t)$ are *all* correct with probability $1 - \delta$, i.e.,

$$\Pr[\ \forall t:\ (1-\epsilon_t)\cdot\mu \le \hat{\mu}_t \le (1+\epsilon_t)\cdot\mu\ ] \ge 1-\delta \tag{1}$$

and, furthermore, ϵ_t tends to 0 with probability 1.

Since the algorithm never stops, it makes no sense to use "stopping time" as a measure of efficiency. Rather, we take as a measure of efficiency the rate at which ϵ_t tends to 0. We show that our algorithm is optimal for a large class of distributions generating the trials X_t. That is, for any algorithm producing pairs $(\hat{\mu}_t, \epsilon_t)$ and satisfying the condition above, the rate of convergence to 0 of the sequence of ϵ_t is asymptotically as large as that in our algorithm, up to constant factors.

It follows from our results that the task of estimating μ in an anytime way is provably slightly more costly than the one-time estimation task. Indeed, for any t let Υ_t be

$$\Upsilon_t = \frac{\sigma^2}{\mu^2 t}.$$

where σ^2 is the variance of the X_i, and let $t = t(\epsilon, \delta)$ be the running time of the algorithm by Dagum *et al.* [1]. Then for sufficiently small ϵ, t satisfies with probability $1 - \delta$

$$\epsilon^2 = O\left(\Upsilon_t \ln \frac{1}{\delta}\right).$$

This is what one would expect from the Central Limit Theorem or equivalently from Bernstein's inequality. The anytime algorithm we present satisfies, with probability $1 - \delta$, that at all times t

$$\epsilon_t^2 = O\left(\Upsilon_t \left(\ln\ln \Upsilon_t + \ln \frac{1}{\delta}\right)\right),$$

i.e., a factor of $\ln\ln \Upsilon_t$ larger. Since our algorithm is optimal up to constant factors, this additional factor is the necessary and sufficient overhead for being anytime with respect to being one-shot.

The estimator algorithm itself is quite simple, as is the one in [1] – which is to be expected, since after all it is only computing an average. We believe that the main contributions of the paper are 1) the probabilistic tool required for its analysis and to prove the lower bound, and 2) the realization that there is a nontrivial, although slight, additional cost for being correct at all times

instead of being correct most of the times, a point we have not seen addressed in the literature. This is somehow surprising given the intense research e.g. on Data Stream algorithms, which are most of the times thought of as anytime algorithms.

The paper is organized as follows: In Section 1.1 we describe a few situations where anytime estimators may be helpful or required. In Section 2 we define more formally the estimation problems we consider, state known results for this task, and sketch the kind of large deviation bound required for anytime estimation. In Section 3 we prove such a bound together with an essentially matching lower bound for all variables for which Bernstein's inequality is tight. the combination of upper and lower bound is a non-limit version of the Law of the Iterated Logarithm. In Section 4 we present the optimal anytime estimator algorithm for a wide class of distributions, which is shown to be optimal in Section 5. Finally, in Section 6 we indicate some possible improvements and future work.

1.1 Motivation

In general, *anytime* algorithms are those that can (1) run for an arbitrary amount of time, unknown at the start, (2) are able to produce an approximate answer about the problem they solve at any point of its execution, and (3) statistically, the quality of the answer improves as the running time increases. Alternatively, an anytime algorithm can be viewed as trading off computation time for quality of answers.

A few of the areas where there has been research on anytime algorithms are reasoning, planning and scheduling, decision making, speech parsing, database transaction processing, resource allocation, networking, constraint satisfaction, and biocomputing. The author's own interest arises from work that applies sequential sampling to the design efficient approximate algorithms data mining and machine learning [3, 2, 5, 9, 10, 16, 17, 18]. Algorithms for the Data Stream model are naturally anytime algorithms, although the strong requirement we place on their behavior (Equation (1)) has never been formulated explicitly in the Data Stream literature to the best of our knowledge.

Typically, the need for anytime algorithms in these areas arises in one of the following scenarios. One, the result of one algorithm is required by another task that can be started as soon an approximate answer is available, and that can later on incorporate a more accurate answer. Another, the al-

gorithm is trying to solve a computationally hard problem for which an exact or very accurate solution will require a long time, but the resolution method provides intermediate answers that improve over time; this occurs, e.g., in branch-and-bound algorithms. In a third scenario the algorithm has only sequential access to its input, and it would be useful to provide approximate answers based on the part of the input seen at any given moment. This is the case, for example, if the input is a sequence of items in some fairly random ordering, so that after seeing a number of items one can guess statistical properties of the whole input sequence. Our results are relevant to anytime algorithms in this scenario.

2 Estimators and Previous Results

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

Let $X_1, X_2, \ldots, X_t, \ldots$ 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. In this paper we consider only the case F is the average function, denoted avg so $\mu = E[X]$. This is also the case considered in [1]. Other F are of interest, e.g., in data mining, and are considered in [2, 5].

Nothing is known initially on the X_t except that they take values in a known finite interval $[a, b]$ with probability 1.

The following two definitions are essentially those in [1] and [5], 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 $X_1, X_2, \ldots$

We 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. The most natural choice for $\hat{\mu}$ is the average avg, but others are possible.

Some known results about the efficiency of estimators are:

1. [1] There is a one-shot estimator $A = (\text{avg}, \epsilon)$ whose expected stopping time is
$$O\left(\frac{\sigma^2}{\epsilon^2\mu^2}\,\ln\frac{1}{\delta}\right).$$
No one-shot estimator can have asymptotically smaller stopping time.

2. For the case of Bernoulli trials, essentially optimal algorithms are also given in [11, 12].

3. [2, 5] When the X_t are i.i.d. Bernoulli variables, there is a function ϵ such that the pair $A = (\text{avg}, \epsilon)$ is a correct anytime estimator algorithm and the value ϵ_t is (in probability)
$$\epsilon_t = O\left(\sqrt{\frac{1}{\mu\, t}\cdot(\ln\frac{1}{\delta} + \ln\frac{1}{\mu\, t})}\right).$$

4. ([5], also follows from [1]) If $A = (\text{avg}, \epsilon)$ is a correct anytime estimator for Bernoulli variables, 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) and (2) determine the efficiency of one-shot estimators, and (3) and (4) provide upper and lower bounds on the efficiency of anytime Bernoulli estimators. Observe that the bounds in (3) and (4) do not match, and furthermore that (4) applies only to estimators whose estimation is the average function; while this is a natural choice, it requires proof that no better choices exist.

Our results in this paper close the gap between (3) and (4) and extend them to non-Bernoulli variables. We provide matching upper and lower bounds for all random variables to which Bernstein's inequality applies and is tight, and for all choices of functions $\hat{\mu}$ (not only the average).

We now motivate the main technical tool to prove our upper and lower bound. 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 $\epsilon(n, \delta)$ can we give so that

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

holds? (Note: $\epsilon(n, \delta)$ may also depend on $X_1, \ldots, X_t$, as well as on μ and other statistics of X; we omit this dependence for conciseness).

Bounds such as Chernoff's, Hoeffding's, or Bernstein's provide a bound θ that applies to any *fixed* n. To obtain an ϵ that satisfies the requirement above, define $\epsilon(n, \delta) = \theta(n, \delta/(n+1)^2)$. Then, using the union bound,

$$\begin{aligned}\Pr[\,\exists n \quad :\quad & |S_n - \mu\, n| \leq \epsilon(n,\delta)\cdot \mu\, n\,] \\ \leq\ & \sum_n \Pr[\,|S_n - \mu\, n| \leq \theta(n,\delta/(n+1)^2)\cdot \mu\, n\,] \leq \sum_n \frac{\delta}{(n+1)^2} \leq \delta.\end{aligned}$$

This is the trick used in [2, 5] to derive an anytime algorithm which is a log factor less efficient than a one-shot algorithm[1]. But [2, 5] leave open whether this additional log factor is necessary or just the result of a sloppy probabilistic analysis. Certainly, using the union bound in this way 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 $n+1$.

In the next section, we provide a finer probabilistic analysis that takes these dependences into account. The result should be compared with the

[1]By considering only ns which are powers of some number $k \geq 1$, this factor is reduced to $\log\log$ in [2]. But the algorithm obtained in this way is not anytime in our sense, since it will only provide an estimation at time steps that are polynomially apart.

classical Law of the Iterated Logarithm (LIL), attributed to Khintchine [6] and Kolmogorov [7]; see also [4, 13, 15]. This law describes the behavior in the limit of the supremum of the deviations of the partial sums of infinitely many variables. Our result is, in fact, a non-limit version of the LIL because it quantifies the deviations that one can expect at every finite n, in the same way that Bernstein's inequality can be regarded as a non-limit version of the Central Limit Theorem. On the other hand, it is sufficiently strong so that the LIL as in [4] can be recovered from it easily. Our proof in fact follows the proof of the LIL in [4], but carefully quantifying every limit statement in that proof.

3 Large Deviation Bounds for Suprema of Partial Sums

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 3 *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*

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

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

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

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

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

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 4 *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*

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

For every $\lambda \in (0,1)$ 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

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

Proofs of these theorems are given in Appendix A.

Next, we apply the two previous theorems to specific types of random variables. We use Bernstein's inequality, a very general large deviation bound that can be proved for sums of bounded random variables. We use the following form of Bernstein's inequality ([8], see also [15]).

Lemma 5 *(Bernstein's inequality) Let X_1, X_2, ..., X_n, ... be i.i.d. real-valued random variables with $a \le X_i \le b$, $E[X_i] = \mu$, $\sigma^2 = Var(X_i)$, and $S_n = \sum_{i=0}^{n} X_i$. Then for any $\epsilon > 0$ and every n,*

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

From Theorem 3 and Bernstein's inequality, one can prove:

Theorem 6 *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 every $\lambda > 1$ there is a constant c such that if $\epsilon(n,\delta)$ is defined as*

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

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

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

(An ad-hoc version of Bernstein's inequality, derived from it, is more convenient to prove this theorem. It is proved in Appendix A as Lemma 13.)

Next, we show that this bound is tight for a large class of random variables, namely, those for which Bernstein's inequality is also tight in the following sense.

Definition 7 *Let $X_1, \ldots, X_n, \ldots,$ be a collection of i.i.d. bounded random variables, with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$. We say that $X_1,\ldots,X_n,\ldots$ is* Bernstein-tight *with constant $c > 0$ if for every sufficiently small ϵ and every n,*

$$\Pr[\,|S_n - \mu| > \epsilon\, n\,] \geq \exp\left(-c\,\frac{\epsilon^2 n}{\sigma^2}\right)\,.$$

For example, collections of i.i.d. Bernoulli variables are Bernstein-tight with any constant $c > 1/2$; this is equivalent to saying that Chernoff's bounds are tight. In fact, only very mild conditions (such as finite third moment) are required to prove that a collection of random variables is Bernstein-tight; see, e.g., [14]. Applying Theorem 4 to Bernstein-tight variables, one obtains:

Theorem 8 *Let $X_1, X_2, \ldots X_n, \ldots$ be Bernstein-tight with constant c, $E[X_i] = \mu$, and $Var(X_i) = \sigma^2$, and let S_n be $X_1 + \ldots + X_n$. For every $\lambda < c$ there is a constant d such that if $\epsilon(n,\delta)$ is defined as*

$$\epsilon(n,\delta) \;=\; \sqrt{\frac{\lambda\,\sigma^2\,\ell(n,\delta)}{\mu^2 n}\cdot(\ln\ln n + \ln(1/\delta))},$$

then for any N sufficiently large

$$\Pr\Big[\;\exists n \in [N, N^d] : |S_n - \mu\, n| > \epsilon(n,\delta)\,\mu\, n\;\Big] \geq \delta.$$

4 An Anytime Estimator Algorithm

In this section we describe our optimal estimation algorithm. Its pseudocode is given in Figure 1. The algorithm accesses trials X_1, X_2, ..., X_t, ... in sequence. No information about the X_t is initially known other than their range $[a, b]$.

In Step 2, a constant λ is fixed arbitrarily, and we should think of the algorithm as parametrized by λ. In Steps 3 and 4, by a c-approximation of a quantity x we mean some $\hat{x}$ such that $(1-c)x \leq \hat{x} \leq (1+c)x$. These steps can be implemented using simple version of the estimator in [1] that obtains a constant-factor approximation of an unknown quantity without any variance information. While steps 1-4 are executed, we can assume that the algorithm outputs pairs (μ_t, ϵ_t) that trivially satisfy the requirement of an anytime algorithm (for example, making ϵ_t very large). After step 4, the algorithm performs a constant amount of work per iteration.

algorithm *AnytimeEstimator*$_\lambda$ *(*δ*)*

1. input $\delta \in (0,1)$;
2. fix constants $\lambda > 1$ and $d = d(\lambda) < 1$;
3. μ_0 := a c-approximation of the average μ;
4. σ_0^2 := a c-approximation of $\max\{\sigma^2, \mu\}$;
5. let t be the time in which step 4 terminates;
   that is, $X_1, \ldots, X_t$ have been read so far;
6. compute $\mu_t := \text{avg}(X_1, \ldots, X_t)$;
7. compute $\epsilon_t = \epsilon(t, \delta, \lambda, \mu_t, \mu_0, \sigma_0^2, a, b)$
   (by a formula given in the analysis);
8. output (μ_t, ϵ_t);
9. $t := t + 1$; go to 6;

Figure 1: The Optimal Anytime Estimator Algorithm

Theorem 9 *There is a formula for* ϵ_t *such that for every* $\lambda > 1$ *and every* $[a, b]$*,*
(1) *AnytimeEstimator*$_\lambda$ *is a correct anytime estimator for any collection of random variables* $\{X_t\}_t$ *with range* $[a, b]$*; that it is, it satisfies*

$$\Pr[\ \forall t:\ (1-\epsilon_t)\cdot\mu \le \hat{\mu}_t \le (1+\epsilon_t)\cdot\mu\ ] \ge 1-\delta.$$

(2) Furthermore, if $\mu = E[X_t]$ *and* $\sigma^2 = Var[X_t]$*, with probability* δ*, for all* t

$$\epsilon_t \le \sqrt{\frac{2\lambda\sigma^2}{\mu^2\, t}\cdot(\ln\frac{1}{\delta} + \ln\ln t)}\cdot(1+g_\lambda(t))$$

for a function $g_\lambda(t) = o(1)$*.*

Proof. (Sketch). Define ϵ_t as

$$\epsilon_t = \max\left\{\sqrt{\frac{2\lambda'\sigma_0^2\,\ell(t,\delta,c)}{\mu_0^2 t}},\ \frac{\lambda'}{\lambda'-1}\cdot\frac{2(b-a)\ell(t,\delta,c)}{3\mu_0 t}\right\} + \sqrt{\frac{2\sigma_0^2}{\lambda' t}}$$

where $\lambda' < \lambda$, $\ell(t,\delta,c) = \ln\ln t + \ln(1/\delta) + c$ and c is given for λ by Theorem 6. Since μ_0 and σ_0^2 are d-approximations of μ and σ, choosing $d(\lambda)$ sufficiently small with respect to λ' and λ, one has

$$\epsilon_t \le \max\left\{\sqrt{\frac{2\lambda\sigma^2\,\ell(t,\delta,c)}{\mu^2 t}},\ \frac{\lambda}{\lambda-1}\cdot\frac{2(b-a)\ell(t,\delta,c)}{3\mu t}\right\} + \sqrt{\frac{2\sigma^2}{t}}$$

and, by Theorem 6, part (1) of the theorem holds. Part (2) holds because the first term inside the max dominates for large t and because the running time of the constant approximation algorithm in [1] used in steps 3-4 is asymptotically smaller. ∎

The algorithm, as presented here, waits for steps 1-4 to terminate before starting to output sensible pairs (μ_t, ϵ_t). Let us note that a more streamlined version of the algorithm does not wait for steps 3-4 to terminate and produce a valid approximation, but rather at every t uses the observed $X_1, \ldots, X_t$ to guess approximations to μ and σ, and uses these to compute ϵ_t. This algorithm is more "anytime" in spirit, but its presentation is more involved and is omitted here for space reasons.

5 A Lower Bound

In this section we state the lower bound on the efficiency of anytime estimators. The proof is given in Appendix B. For simplicity, we state it for variables with range $[0, 1]$ and $\sigma^2 = \mu = E[X_t]$. A simple scaling argument extends it to other variables of bounded range.

Theorem 10 *Let $X_1, \ldots, X_t, \ldots$ be Bernstein-tight variables with $X_t \in [0, 1]$, $\mu = E[X_t]$ and $\sigma^2 = Var[X_t] = \mu_t$. There exists $\lambda < 1$ such that if*

$$\epsilon(\vec{X}_t, \delta) \leq \sqrt{\frac{\lambda}{\text{avg}(\vec{X}_t)\, t} \cdot (\ln \ln t + \ln(1/\delta))}$$

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

In fact one can show that such an A will fail quite frequently: for some constant d, it will fail at least once within each interval $[N, N^d]$ with probability δ.

Since, for large t, $\text{avg}(\vec{X}_t)$ is close to μ with high probability, this theorem implies that asymptotically

$$\epsilon_t = \Omega\left(\sqrt{\frac{\sigma^2}{\mu^2 t}\ (\ln \ln t + \ln(1/\delta))}\right)$$

for every anytime estimator for Bernstein-tight variables with range $[0, 1]$.

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 [2], an application to Boosting is described where the quantity of interest is the amount by which the average exceeds 1/2.

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. This question seems particularly acute in the Data Stream model, 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 11, which greatly simplified a previous proof. Finally, we thank Marco Minozzo for sending us a copy of his work [13].

References

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

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

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

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

[5] 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.

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

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

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

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

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

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

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

[13] 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.

[14] S. V. Nagaev: "Lower bounds on large deviation probabilities for sums of independent random variables". *Theory of Probability & Its Applications* **46**:1 (2002), 79–102.

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

[16] 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.

[17] 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.

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

Appendix A: Proofs of Theorems 3 and 4

To prove Theorems 3 and 4 we use the following two lemmas, whose proofs are given later. The following is an easy variation of Lemma E in [15] (Ch. VII, §4), and is proved for self-containment.

Lemma 11 *[15] 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 12 *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 3. Fix $\lambda > 1$ and for each $k \ge 1$, 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

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

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

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

where the first inequality holds because $\theta_n n$ is increasing, the second by Lemma 11, and the third and fourth by definition of θ_n and $n_{k+1} \cong \lambda n_k$. By assumption on ϵ we have

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

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 3)

Proof of Theorem 4. 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 12,

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

which proves inequality (3) and, so, the theorem. ∎ (Theorem 4)

Proof of Lemma 11. 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 11)

Proof of Lemma 12. 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 12)

To prove Theorem 6, it is convenient to rephrase Bernstein's inequality as follows:

Lemma 13 *Under the conditions of Lemma 5, 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

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

Proof. 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 5

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

■

Appendix B: Proof of Theorem 10

We use the following definition:

Definition 14 *For a vector $\vec{X}_t = (X_1, \ldots, X_t)$, function ϵ_{min} is defined as:*

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

Observe that $\epsilon_{min}(\vec{X}_t, \delta)$ is the same for all vectors $\vec{X}_t$ with the same average.

In order to prove Theorem 10, we prove first this more technical result:

Theorem 15 *Let $\{X_t\}_t$ be Bernstein-tight, and $E[X_t] = Var(X_t) = \mu > 0$. For some constant $\lambda > 0$, any correct anytime estimator $A = (\hat{\mu}, \epsilon)$, any $\delta < \delta(\lambda)$, and any sufficiently large t, the event $\epsilon(\vec{X}_t, \delta) \geq \lambda \cdot \epsilon_{min}(\vec{X}_t, \delta)$ occurs with probability at least δ^λ, if the probability is taken over all vectors $\vec{X}_t = (X_1, \ldots, X_t)$.*

The following is a corollary to this theorem.

Corollary 16 *For some $\lambda > 0$ and any pair $A = (\hat{\mu}, \epsilon)$, if $\epsilon(\vec{X}_t, \delta) < \lambda \cdot \epsilon_{min}(\vec{X}_t, \delta)$ with probability 1, then A is not a correct anytime estimator on variables $\{X_t\}_t$ as above.*

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 is used in the proof of Theorem 10.

Proof of Theorem 15. Informally, the argument is as follows: Fix an estimator $A = (\hat{\mu}, \epsilon)$. For sufficiently large t, we pick up two distributions X and Y with expected values μ_1 and μ_2, where μ_2 exceeds μ_1 by about $2\lambda\epsilon_{min}$. We show that the set of sequences with average above μ_2 has noticeable probability both when generated according to μ_1 and to μ_2. If $\epsilon < \lambda\epsilon_{min}$ too often, then $\hat{\mu}$ cannot be within ϵ of both μ_1 and μ_2 too often. So A has to be biased towards μ_1 or μ_2, and then it will be wrong on the other distribution.

More formally, fix a Bernstein-tight random variable X generating the sequence $\{X_t\}$. Fix then the estimator $A = (\hat{\mu}, \epsilon)$, so that A can, for example, "know" the range of X, but then has to be a correct estimator for all variables

with the same range as X. Let μ_1 be $E[X]$ and c be the constant witnessing that X is Bernstein-tight. We choose $\lambda > 0$ as small w.r.t. c as required later in the proof. Let $\Pr_X[A]$ denote the probability of event A when the trials X_t are generated according to X.

Choose t sufficiently large. Let ϵ^* be the value satisfying $\epsilon^* = \lambda\epsilon_{min}(\vec{X}_t, \delta)$ for some vector $\vec{X}_t$ with $\text{avg}(\vec{X}_t) = (1+\epsilon^*)\mu_1$. 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 if t is sufficiently large. Observe that if t is large enough,

$$\epsilon^* \cong \lambda\sqrt{\frac{1}{\mu_1 t}\ln\frac{1}{\delta}}, \quad \text{so} \quad \exp((\epsilon^*)^2\mu_1 t) \cong \delta^{\lambda^2}.$$

We define next two sets M and R using ϵ^*. Let M be the set of sequences $\vec{X}_t = (X_1, \ldots, X_t)$ such that $\text{avg}(\vec{X}_t) \in [(1+3\epsilon^*)\mu_1, (1+3\beta\epsilon^*)\mu_1]$, where $\beta^2/2 > 10c$. Again, we assume that t is sufficiently large so that the interval above is properly contained in the range of X. Using Bernstein inequality and Bernstein-tightness,

$$\begin{aligned}
\Pr_X[M] &= \Pr_X[\text{avg}(\vec{X}_t) \in [(1+3\epsilon^*)\mu_1, (1+3\beta\epsilon^*)\mu_1] \\
&= \Pr_X[\text{avg}(\vec{X}_t) \geq [(1+3\epsilon^*)\mu_1] - \Pr_X[\text{avg}(\vec{X}_t) > (1+3\beta\epsilon^*)\mu_1] \\
&\geq \exp(-c\cdot(3\epsilon^*)^2\mu_1 t) - \exp(-(1/2)\beta^2(3\epsilon^*)^2\mu_1 t)] \cong \delta^{9c\lambda^2} - \delta^{9\beta^2\lambda^2/2} \\
&\geq \delta^{9c\lambda^2} - \delta^{10c\lambda^2} \geq 2\delta^\lambda,
\end{aligned}$$

if $c\lambda^2 \ll \lambda$, i.e., $\lambda \ll 1/c$, and δ is sufficiently small w.r.t. λ.

Let R (for "risky") be the set of sequences $\vec{X}_t$ such that $\epsilon_t(\vec{X}_t, \delta) < \epsilon^*$, and R^c its complement. Our goal is to show that $\Pr_X[R^c] \geq \delta^\lambda$. Note that, for $X_t \in M$, the condition $\epsilon_t(\vec{X}_t, \delta) < \epsilon^*$ is equivalent to $\epsilon_t(\vec{X}_t, \delta) < \lambda\epsilon_{min}(\vec{X}_t, \delta)$ by the definition of M and ϵ^*.

Now, let μ_2 be $(1+3\epsilon^*)\mu_1$. We assume that ϵ^* is so small that $(1+\epsilon^*)\mu_1 < (1-\epsilon^*)\mu_2$. Let Y be a random variable with the following properties: (1) It has the same range as X, (2) $E[Y] = \mu_2$, and (3) For any $B \subseteq M$, $\Pr_Y[B] \geq \Pr_X[B]$; here $\Pr_Y[B]$ denotes the probability of event B when trials X_t are generated from Y. We will show later that Y does exist.

Let $\vec{X}_t$ be in $M \cap R$. We say that $\vec{X}_t$ is in Case 1 if $\hat{\mu}(\vec{X}_t, \delta) \geq (1+\epsilon^*)\mu_1$, and in Case 2 otherwise. If $\vec{X}_t$ is in Case 1, since it is in R, then $\hat{\mu}(\vec{X}_t, \delta) > (1+\epsilon(\vec{X}_t, \delta))\mu_1$, so A gives a wrong ϵ_t-approximation of μ_1 on $\vec{X}_t$. If $\vec{X}_t$ is in Case 2, then $\hat{\mu}(\vec{X}_t, \delta) \leq (1+eps^*)\mu_1 < (1-\epsilon^*)\mu_2 \leq (1-\epsilon(\vec{X}_t, \delta))\mu_2$, so A gives

a wrong ϵ_t-approximation to μ_2. Suppose $\Pr_X[M \cap R] > 2\delta$. Then either $\Pr_X[\text{Case 1}] > \delta$ or $\Pr_X[\text{Case 2}] > \delta$, hence $\Pr_Y[\text{Case 2}] > \delta$ by property (3) of Y. In the first case, A is an incorrect estimator, as witnessed by X. In the second case, A is an incorrect estimator, as witnessed by Y. So we must have $\Pr_X[M \cap R] \leq 2\delta$.

To conclude the argument, observe that $\Pr_X[R^c] \geq \Pr_X[M \cap R^c] = \Pr_X[M] - \Pr_X[M \cap R] \geq 2\delta^\lambda - 2\delta \geq \delta^\lambda$, for all δ sufficiently smaller than λ.

It remains to argue that a random variable Y as claimed exists. Let d be the density function of X, α any value in $[0, 1]$, and define d_α as follows. For $u < \mu_2$, $d_\alpha(u) = \alpha d(u)$, and for $u \geq \mu_2$, $d_\alpha(u) = \beta d(u)$, where $\beta = \beta(\alpha) > 1$ is a constant that makes this a valid density function. Let Y_α have density d_α. Clearly, Y has the same range as X (condition (1)) and for any set $B \subseteq [\mu_2, \infty)$, $\Pr_{Y_\alpha}[B] \geq \Pr_X[B]$ (condition (3)). Also, for $\alpha = 1$, $E[Y_\alpha] = \mu_1 < \mu_2$, and for $\alpha = 0$, $E[Y_\alpha] \geq \mu_2$ since $Pr_X[X \in [\mu_2, \infty]] > 0$. Therefore, there is an $\alpha \in [0, 1]$ such that $E[Y_\alpha] = \mu_2$, i.e., this Y_α satisfies condition (2). A similar reweighting works for discrete distributions. ∎ (Theorem 15)

Proof of Theorem 10. *(Sketch)* Observe first that $\ln \ln t + \ln(1/\delta) = \ln(1/(\delta/\ln t))$. Let $\lambda > 0$ be given by Theorem 15. Fix a 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 $\text{avg}(\vec{X}_t)$ 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 15, the probability that event C_t occurs is at least $(\delta/4 \ln t)^\lambda$ for all t sufficiently large. Therefore,

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

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, as in the proof of Theorem 4. Then, by Lemma 12, 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. ∎ (Theorem 10)