A Simple Analysis of the Error Probability of Two-point Based Sampling

K. Gopalakrishnan
Department of Computer Science
Wichita State University
Wichita KS 67260

D. R. Stinson
Department of Computer Science and Engineering
University of Nebraska - Lincoln
Lincoln NE 68588

Abstract

The *two-point based sampling technique* was introduced by Chor and Goldreich with the purpose of reducing the amount of randomness needed in randomized algorithms. It has since found numerous applications in theoretical computer science. In this note, we point out a connection between two-point based sampling and the combinatorial structures known as orthogonal arrays. We also present an elementary combinatorial analysis of two-point based sampling, which leads to an improved bound on the error probability.

keywords: Derandomization, Two-Point Based Sampling, Orthogonal Arrays.

1 Introduction

There are many problems for which no fast deterministic algorithm is known, but which can be solved efficiently using randomized algorithms. Many examples are found in computational number theory. A good survey of randomization techniques is presented in [3].

There are two basic types of randomized algorithms viz., Las Vegas algorithms and Monte Carlo algorithms. Here we will be concerned exclusively with Monte Carlo algorithms. For any problem instance I, a Monte Carlo algorithm always gives an answer, but the answer may be incorrect with some probability ϵ. Monte Carlo algorithms are usually assumed to be algorithms for decision problems, and we will also make that assumption here.

A Monte Carlo algorithm is defined to be *yes-biased* if the following conditions are satisfied:

1. If the instance I is a no-instance, then the algorithm answers "no".

2. If the instance I is a yes-instance, then the probability that the algorithm answers "yes" is at least $1 - \epsilon$, where $\epsilon \geq 0$ is some fixed constant.

Hence, if the algorithm answers "yes", then we know that the answer is correct. However, if the algorithm answers "no", then there is the possibility that the answer may be incorrect. A *no-biased* Monte Carlo algorithm is defined in the obvious way and there is no loss of generality in considering only yes-biased Monte Carlo algorithms.

A yes-biased Monte Carlo algorithm can be viewed as a two-stage procedure in which first a "sample point" is chosen at random and next a deterministic procedure is applied to the sample point. This view is usually known as the *witness paradigm*. In the generic case the sample point is an element r chosen from some specified universe U, and the deterministic procedure can be viewed as computing a zero-one valued function $f(I, r)$. If $f(I, r) = 0$ then the algorithm answers "no" and if $f(I, r) = 1$ then the algorithm answers "yes". The function f should satisfy the following: If I is a no-instance, then $f(I, r) = 0$ for all $r \in U$; and if I is a yes-instance, then the number of values r for which $f(I, r) = 1$ is at least $(1 - \epsilon)|U|$.

2 Two-Point Based Sampling

One reason why a biased algorithm is useful is that the error probability can be made as small as desired by repeated application of the algorithm. Suppose we have a yes-biased algorithm, and we apply the algorithm k times. If we get at least one "yes" answer we can quit, since the instance I must be a yes-instance. On the other hand, if I is a yes instance, then the probability of getting k "no" answers is at most ϵ^k, which can be made arbitrarily small by taking k large enough.

However, the conclusion that the error probability can be made exponentially small is based on the assumption that the sequence of random numbers used in the successive trials are independent and uniformly distributed. When a randomized algorithm is implemented in actual practice, one always uses a pseudo-random number generator which produces a sequence whose later values come from the previous values in a *deterministic* fashion. This invalidates the assumption of independence and might cause one to regard results about randomized algorithms with suspicion.

In view of this objection, many researchers attempted to construct pseudo-random number generators for which provable bounds could be obtained on the probability that none of the k successive values produced is a witness. A simple and effective method known as *two-point sampling* was developed by Chor and Goldreich [1], based on an observation by Joffe [4]. This technique has found a variety of applications in theoretical computer science and the reader is referred to [6] for a collection of references on applications of the two-point sampling technique.

The Chor-Goldreich generator is as follows: Assume $U = \mathbb{Z}_p$, where p is prime. Choose two random independent elements $a, b \in \mathbb{Z}_p$ with uniform probability. Compute the residues

$$r_i = a + i \times b \bmod p,$$

for $1 \leq i \leq k$. These residues are then used as the random numbers in the k trials of the Monte Carlo algorithm.

It is not too hard to see that the r_i's are *pairwise independent uniformly distributed random variables*. In other words, $p(r_i = x) = 1/p$ for all $1 \leq i \leq k$ and for all $x \in \mathbb{Z}_p$, and

$$p(r_i = x, r_j = y) = p(r_i = x) \times p(r_j = y) = \frac{1}{p^2}$$

for all $1 \leq i < j \leq k$ and for all $x, y \in \mathbb{Z}_p$.

By using Chebyshev's inequality and the fact that the r_i's are pairwise independent and uniformly distributed, Chor and Goldreich prove that the probability that none of the r_i's is a witness is at most

$$\frac{\epsilon}{(1-\epsilon)k}.$$

In the context of random pattern testing of VLSI chips, Spencer ([5]) developed a pseudo-random pattern generator based on the idea of two-point sampling. Here, $U = GF(2^m)$, $\alpha \in GF(2^m)$ is a primitive element, and $a, b \in GF(2^m)$ $(b \neq 0)$ are chosen indpendently at random. Then compute the field elements

$$r_i = a + \alpha^{i-1} \times b,$$

for $1 \leq i \leq k$. This generator can be efficiently implemented as a modification of the classical linear feedback shift register. It is also amenable to realistic analysis and in fact Spencer proved a bound on the error probability using the Bonferroni inequalities, again using the fact that the r_i's are uniformly distributed and pairwise independent. His bound is

$$1 - \frac{2(1-\epsilon)k}{1 + \lceil (1-\epsilon)k \rceil} + \frac{(1-\epsilon)^2 k^2}{\lceil (1-\epsilon)k \rceil (1 + \lceil (1-\epsilon)k \rceil)}.$$

3 Two-point Sampling and Orthogonal Arrays

In this section, we point out the connection between orthogonal arrays and two-point based sampling. An *orthogonal array* $OA(k, n)$ is an $n^2 \times k$ array of n symbols, say A, such that in any two columns of A, each of the n^2 ordered pairs of (not necessarily distinct) symbols occurs in exactly one row of A.

Orthogonal arrays provide a convenient method of obtaining pairwise independent uniformly distributed random variables which can be used for two-point sampling.

This is done as follows. Consider an $OA(k,n)$, A, on a symbol set U. Generate $2\log n$ random bits and use them to index a row of A. For $1 \leq i \leq k$, suppose we define the random variable r_i to be the entry in column i of the chosen row. Then the k random variables are easily seen to be uniformly distributed and pairwise independent. We will test the Monte Carlo algorithm on the k values r_i $(1 \leq i \leq k)$ in the given row of A.

There is a classical construction for $OA(q,q)$, for any prime power q, which is essentially obtained from the Desarguesian projective plane of order q. The rows of the *Desarguesian OA*, A, are indexed by $GF(q) \times GF(q)$, the columns are indexed by $GF(q)$, the entries of A are defined as

$$A((a,b),c) = a + bc,$$

$a, b, c \in GF(q)$.

It is easy to see that any k_1 columns of any $OA(k,n)$ form an $OA(k_1,n)$. The Chor-Goldreich generator is obtained by taking columns $1,\ldots,k$ of a Desarguesian $OA(p,p)$ (p prime). Similarly, the Spencer generator is constructed as columns $1,\alpha,\ldots,\alpha^{k-1}$ of the Desarguesian $OA(2^m,2^m)$, where α is a primitive element of $GF(2^m)$ (omitting those rows in which $b=0$).

4 A Combinatorial Analysis

We now present an elementary combinatorial analysis of the two-point sampling technique. We assume an orthogonal array A is used as in the previous section, and we proceed to calculate a bound on the error probability. Let U denote the universe of sample points, where $|U| = n$. Let $S \subseteq U$ be the set of witnesses or "good" sample points, where

$$|S| = (1-\epsilon)n = m.$$

Let x_i denote the number of rows of A in which there are exactly i occurrences of elements from S. Call a row of the matrix a *bad row* if none of the elements in the row is a witness. Then the error probability is simply the probability that the randomly selected row is a bad row. Hence, for a given S, the error probability is

$$err(S) = \frac{x_0}{n^2}.$$

Note that we actually do not know the set S explicitly. The only thing that we know is that $|S| = m$. So the actual error probability is bounded above by

$$err = \max\{err(S) : S \subseteq U, |S| = m\}.$$

We first derive three simple equations from the elementary properties of orthogonal arrays. Since an $OA(k,n)$ has n^2 rows, we have

$$\sum_{i=0}^{n} x_i = n^2.$$

Next, we can count the number of occurrences of good points in the array in two ways. There are exactly x_i rows in which there are i occurrences of good points. In any column of A, each point occurs exactly n times. So, the number of occurrences of good points in a column is nm. As there are k columns, the total number of occurrences of good points in A is knm. This yields the following equation:

$$\sum_{i=0}^{n} i x_i = knm.$$

Similarly, we can count the number of occurrences of pairs of good points which lie in the same row in two ways. In any row in which there are i occurrences of good points, there will be $i(i-1)$ occurrences of pairs of good points. On the other hand, if we look at any two columns of A, the number of occurrences of pairs of good points which lie in the same row is m^2. (This is because any particular pair of good points occurs exactly once in any given pair of columns.) Now the two columns can be selected in $k(k-1)$ ways and so the total number of occurrences is $k(k-1)m^2$. This yields the following equation:

$$\sum_{i=0}^{n} i(i-1)x_i = k(k-1)m^2.$$

To compute the error probability, we need an upper bound on x_0. Equivalently, we need a lower bound on $\sum_{i=1}^{n} x_i$, subject to the constraints

$$\begin{aligned}\sum_{i=1}^{n} i x_i &= knm \\ \sum_{i=1}^{n} i(i-1)x_i &= k(k-1)m^2.\end{aligned}$$

Now, let z be any real number. Then clearly

$$\begin{aligned}0 &\leq \sum_{i=1}^{n} (i-z)^2 x_i \\ &= \sum_{i=1}^{n} (i^2 - 2zi + z^2) x_i \\ &= \sum_{i=1}^{n} i^2 x_i - 2z \sum_{i=1}^{n} i x_i + z^2 \sum_{i=1}^{n} x_i \\ &= k(k-1)m^2 + knm - 2zknm + z^2 \sum_{i=1}^{n} x_i\end{aligned}$$

It follows that

$$\sum_{i=1}^{n} x_i \geq \frac{2knmz - knm - k(k-1)m^2}{z^2}.$$

The right hand side of the above equation is maximized when

$$z = \frac{n + (k-1)m}{n}.$$

Hence we get

$$\sum_{i=1}^{n} x_i \geq \frac{kmn^2}{n + (k-1)m}.$$

Equivalently, we have

$$x_0 \leq n^2 - \frac{kmn^2}{n + (k-1)m}.$$

Hence we get the following bound on the error probability:

$$err \leq 1 - \frac{km}{n + (k-1)m}.$$

Since $m = n(1-\epsilon)$, we have that

$$err \leq \frac{\epsilon}{1 + (k-1)(1-\epsilon)}.$$

Summarizing, we have the following.

Theorem 4.1 *The error probability of the two-point based sampling technique for a universe U of size n, using an orthogonal array $OA(k, n)$, is at most*

$$\frac{\epsilon}{1 + (k-1)(1-\epsilon)}.$$

5 Comparison of the Bounds

We have discussed three bounds on error probabilities:

(**GS**) $$\frac{\epsilon}{1 + (k-1)(1-\epsilon)}$$

(**CG**) $$\frac{\epsilon}{k(1-\epsilon)}$$

(**S**) $$1 - \frac{2(1-\epsilon)k}{1 + \lceil (1-\epsilon)k \rceil} + \frac{(1-\epsilon)^2 k^2}{\lceil (1-\epsilon)k \rceil (1 + \lceil (1-\epsilon)k \rceil)}$$

Note however, that if $k(1-\epsilon)$ is an integer, then the Spencer bound, (**S**), becomes

(**S′**) $$\frac{1}{1 + k(1-\epsilon)}$$

To simplify the arithmetic, we will use the bound ($\mathbf{S'}$) in our subsequent discussion. Now, it is not difficult to prove that

$$\begin{array}{rcl} (\mathbf{GS}) & \leq & (\mathbf{CG}) \\ (\mathbf{GS}) & \leq & (\mathbf{S'}) \\ (\mathbf{S'}) & \leq & (\mathbf{CG}) \quad \Leftrightarrow \quad k \leq \frac{\epsilon}{(1-\epsilon)^2} \end{array}$$

That is, the new bound ($\mathbf{GS}$) is always as least as strong as ($\mathbf{S'}$) and ($\mathbf{CG}$).

Here are some numbers to illustrate the bounds:

ϵ	k	($\mathbf{GS}$)	($\mathbf{CG}$)	($\mathbf{S'}$)
.9	80	.1011	.1125	.1111
.9	90	.0909	.1000	.1000
.9	100	.0825	.0900	.0909
.5	2	.3333	.5000	.5000
.5	5	.1667	.2000	.2857
.5	10	.0909	.1000	.1667

Finally, in Figure 1, we graph the three bounds for the value $\epsilon = .9$. For this value of ϵ, the crossover point of ($\mathbf{S'}$) and ($\mathbf{CG}$) occurs when $k = 90$.

6 Optimality of the New Bound

Finally, we discuss the optimality of the ($\mathbf{GS}$) bound. From the proof of Theorem 4.1, it is easy to see that the ($\mathbf{GS}$) bound is met with equality if and only if

$$z = \frac{n + (k-1)m}{n}$$

is an integer, and (further) $x_i \neq 0$ only if $i = 0, z$. It then follows that

$$x_0 = \frac{n^2(n-m)}{n + (k-1)m}$$

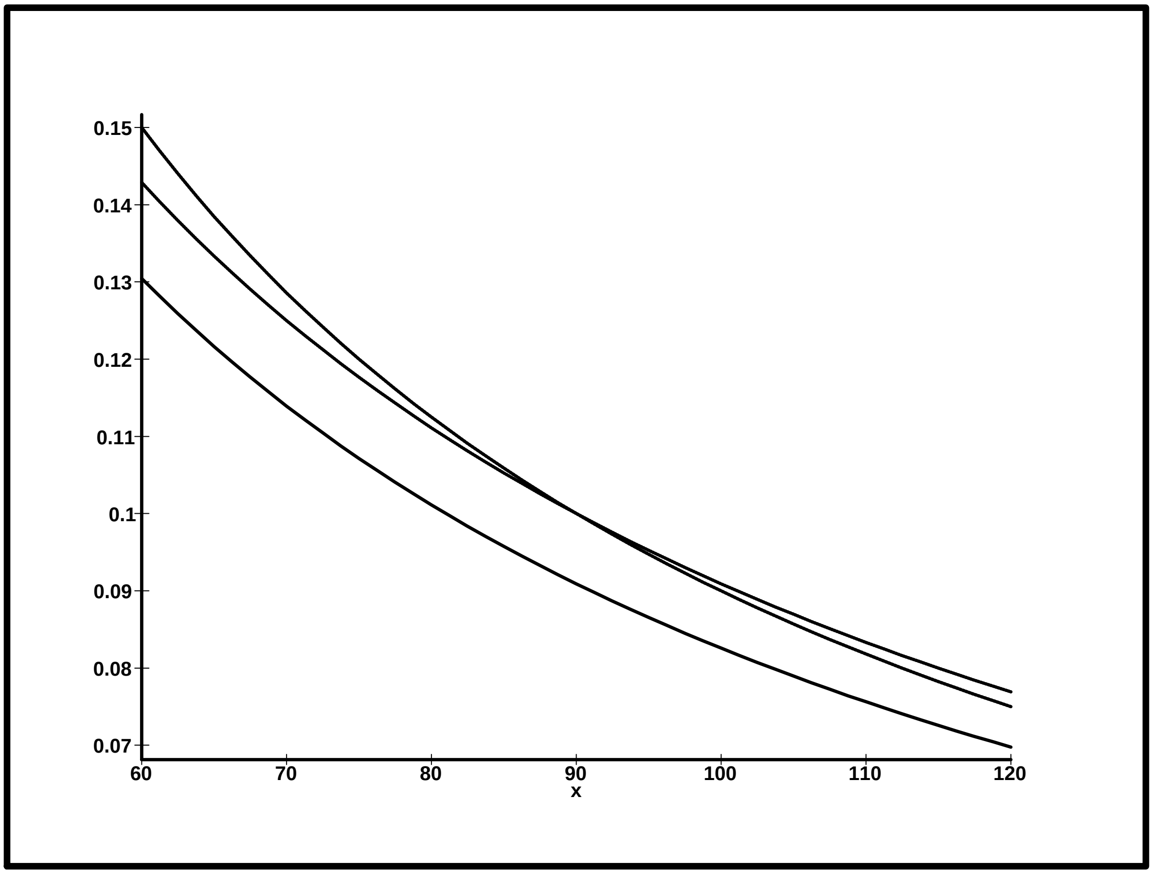
and

$$x_z = \frac{n^2 km}{n + (k-1)m}.$$

In fact, it is possible to construct orthogonal arrays containing subsets S for which these values are attained. Thus, our new bound is in some sense "best possible", given that an OA is being used as a generator for two-point based sampling. The construction uses *maximal arcs* from finite geometry.

We require a couple of definitions. A *projective plane* of *order* q is a pair $(X, \mathcal{B})$, where X consists of q^2+q+1 elements (called *points*) and $\mathcal{B}$ consists of q^2+q+1 subsets of X (called *lines*), such that every line contains $q+1$ points, every point

Figure 1: Comparison of bounds for $\epsilon = .9$



is on $q+1$ lines, and every pair of points are contained in a unique line. Projective planes are known to exist for prime power orders.

A *maximal* $\{w;\ell\}$*-arc* in a projective plane $(X,\mathcal{B})$ of order q is a subset Z of w points with the property that

$$|Z \cap B| \in \{0, \ell\}$$

for all lines $B \in \mathcal{B}$. By elementary counting, it follows that

$$w = (\ell - 1)(q + 1) + 1.$$

Now, suppose that Z is a maximal $\{w;\ell\}$-arc in a projective plane $(X,\mathcal{B})$ of order q. Let $z \in X\backslash Z$, and let $B_1,\ldots,B_k$ be the lines through z that meet Z, where $k = w/\ell$. Denote

$$X_1 = \bigcup_{j=1}^{k} (B_j\backslash\{z\}).$$

Let U be a set of size q and let $S \subset U$ have size ℓ. Now define any mapping $\phi : X_1 \to U$ satisfying the properties

$$\{\phi(x) : x \in B_j\} = U$$

and

$$\{\phi(x) : x \in B_j \cap Z\} = S,$$

for all j, $1 \leq j \leq k$. Finally, we define an $OA(k,q)$ on symbol set U, as follows. Let $A_1, \ldots, A_{q^2}$ be the lines of the projective plane that do not contain z. Now for $1 \leq i \leq q^2$ and $1 \leq j \leq k$, define the entry in row i and column j of the OA to be $\phi(A_i \cap B_j)$.

It is easy to see that we have constructed an $OA(k,q)$ by this recipe, and also that every row of the OA contains either 0 or ℓ points of S. In fact, it can be verified that the (**GS**) bound is met with equality, where $n = q$, $k = q + 1 - q/\ell$ and $m = \ell$.

It has been shown by Denniston [2] that there exists a maximal $\{2^{r+s}+2^r-2^s; 2^r\}$-arc in a projective plane of order 2^s, for all $0 \leq r \leq s$. The discussion above thus proves the following result:

Theorem 6.1 *Suppose* $0 \leq r \leq s$, $n = 2^s$, $k = 2^s + 1 - 2^{s-r}$, *and* $\epsilon = 1 - 2^{r-s}$. *Then there exists an* $OA(k,n)$ *containing a subset* S *of cardinality* $(1-\epsilon)n$ *such that the resulting error probability of the two-point sampling technique is*

$$2^{-r} - 2^{-s} = \frac{\epsilon}{1 + (k-1)(1-\epsilon)}.$$

Acknowledgments

This research was supported by NSF grants CCR-9121051 and CCR-9402141. The first author would like to thank Sunil Shende for introducing him to derandomization techniques.

References

[1] B. Chor and O. Goldreich, On the Power of Two-Point Based Sampling. *Journal of Complexity*, **5** (1989), 96–106.

[2] R. H. F. Denniston, Some Maximal Arcs in Finite Projective Planes. *Journal of Combinatorial Theory A*, **6** (1969), 317–319.

[3] R. Gupta, S. A. Smolka, and S. Bhaskar, On Randomization in Sequential and Distributed Algorithms. *ACM Computing Surveys*, **26** (1994), 7–86.

[4] A. Joffe, On a Set of Almost Deterministic k-Independent Random Variables. *Annals of Probability*, **2** (1974), 161–162.

[5] T. H. Spencer, Provably Good Pattern Generators for a Random Pattern Test. *Algorithmica*, **11** (1994), 429-442.

[6] A. Wigderson, The Amazing Power of Pairwise Independence. In *Proc. 26th ACM Symposium on the Theory of Computing*, 1994, pp. 645–647.