Majorizing estimators and the approximation of #P-complete problems

Leonard J. Schulman* Vijay V. Vazirani*

Abstract

A key step in counting via sampling is constructing an unbiased estimator, X, for the parameter θ in question, and proving a bound on its second moment, $E(X^2)$. A key application of this method is to obtaining a FPRAS for a #P-complete problem; a FPRAS results if the ratio $r = \frac{E(X^2)^{1/2}}{E(X)}$ is polynomially bounded in the size of the input. We show that if no additional information is available about the distribution of X, then this condition is also *necessary.*

The proof involves establishing a new optimality result in parametric statistics. We introduce the notion of a *majorizing estimator,* a very strict optimality requirement that we need for making worst-case (over inputs) and in-probability (of falling in the desired accuracy range of the parameter θ) statements. We show that for the problem of estimating the mean of a Gaussian distribution (from the variable-location, fixed-scale family $\{G_\theta\}$), the sample mean is a majorizing estimator. An extension of this argument shows that the sample mean is an optimal estimator in every central moment among all estimators. To compare, the celebrated Cramer-Rao lower bound, applied to the family $\{G_\theta\}$, establishes that the sample mean is the optimal estimator in mean square error among all unbiased estimators. We further show that the mean estimator is the *unique* majorizing estimator for $\{G_\theta\}$.

1 Introduction

Counting via sampling has emerged as a fundamental method in the theory of algorithms. For instance, it has been applied successfully to approximately counting the number of solutions for many problems in $\mathbf{P}$ whose solution-counting versions are #$\mathbf{P}$-complete. These include fundamental problems such as counting the number of perfect matchings in dense graphs, determining the fraction of satisfying truth assignments for a DNF formula, and determining the volume of a convex body.

*College of Computing, Georgia Inst. Technology, Atlanta GA 30332-0280.

Proc. 31'st STOC (ACM Symp. Theory Comp.), ACM Press, 1999.

The key step in obtaining these *fully polynomial randomized approximation schemes* (FPRAS; definition below) is constructing an unbiased (or nearly so) estimator, X, for the parameter θ in question (e.g. fraction of truth assignments that satisfy a given DNF formula), and proving a bound on the second moment, $E(X^2)$, of this estimator. The goal is to provide a good *relative approximation* of θ, i.e. an estimate T such that $\theta(1-\varepsilon) \leq T \leq \theta(1+\varepsilon)$ for ε as small as desired.

A quantity of critical interest is the ratio of the square root of the second moment and the first moment of this estimator, $r = \frac{E(X^2)^{1/2}}{E(X)}$. By a straightforward application of Chebychev's inequality, it is easy to see that the mean of $O(\frac{r^2}{\varepsilon^2\delta})$ sample points of this random variable will be in an ε-interval around the quantity being estimated, with probability at least $1-\delta$. So, if r is polynomially bounded in the size of the instance, this yields a FPRAS for the #$\mathbf{P}$-complete problem.

A slight improvement is possible, and is commonly employed, as follows: the mean of $O(\frac{r^2}{\varepsilon^2})$ sample points of the random variable falls in the ε-interval with probability at least 3/4. Now, repeat this process $O(\log \frac{1}{\delta})$ times, and take the median of these means. Since the means are independent and identically distributed, by the Chernoff bound on the tail of the binomial distribution, the median lies in the ε-interval with probability at least $1-\delta$. This reduces the total number of sample points needed to $O(\frac{r^2 \log \frac{1}{\delta}}{\varepsilon^2})$.

This basic method underlies an entire literature of approximation methods (see [5] and [7]). This includes rapidly mixing Markov chains, where the bias can be made vanishingly small as a function of the time for which the Markov chain is run. A prominent problem for which this approach fails to provide a FPRAS is the estimation of 0/1 permanents [4], for which the unbiased estimator does not have a polynomially bounded ratio r.

The importance of this method, and the multitude of problems which still elude solution, motivate the following question: is it possible to use the sample points in a more clever way (rather than simply taking their mean) so that one obtains a FPRAS from an unbiased estimator even though its ratio r is not polynomially bounded? In spite of the fundamental nature of the question, it appears that, to this point, it has not been addressed.

In this paper, we show that if no additional information is available about the distribution of the random variable, beyond a bound on its second moment and the fact that its first moment equals (or is close to) the unknown parameter, then the answer to this question is "No", i.e., the condition that r be polynomially bounded is *necessary and sufficient* for obtaining a FPRAS from an estimator. This is formally stated in Theorem 1 below.

At an informal level, this is perhaps the anticipated answer to the original question of whether there "is anything more clever to be done" with the sample points. Our result demonstrates the correctness of this intuition in a very strong way, in terms of *majorizing estimators* (definition below). As a reminder that intuition cannot substitute for proof, we consider the related question of whether the "median trick" described above (and which has been subjected to no more critical scrutiny than has the mean step) can be improved upon. In this case we believe that the answer is "Yes", and we present a proposal to this effect in section 3.2.

Let Π be a counting problem; the case of greatest interest is that Π is $\#\mathbf{P}$-complete. Let I denote an arbitrary instance of this problem, and let θ_I denote the quantity being estimated. Suppose a polynomial-time samplable random variable X_I is constructed such that it is an unbiased estimator of the quantity θ_I, or else, there is a family of estimators X_I^ℓ such that $|E(X_I^\ell) - \theta_I| \leq \ell^{-c}$ $(c > 0)$. In the latter case a FPRAS for θ_I can be obtained from a FPRAS for $E(X_I^{\varepsilon^{-O(1)}})$, so we only need consider the case of unbiased estimators. (The biased case can occur in Markov chain algorithms; ℓ depends on the time for which the Markov chain is run.)

Theorem 1 *Consider approximation algorithms that can obtain independent samples of the random variable X_I, and are charged one unit of time per sample. The algorithm is given an upper bound on $E(X_I^2)$, and must perform correctly for any distribution consistent with this second moment. A FPRAS of this type exists for θ_I iff*

$$\frac{E(X_I^2)^{1/2}}{E(X_I)} \leq p(|I|)$$

for some polynomial p.

On the one hand, our theorem can be viewed as a negative result, and on the other, it points to a direction worth exploring in order to add to the power of current techniques for approximate counting via sampling: identifying and taking advantage of special features of the probability distribution of the estimator.

Theorem 1 is a corollary of a theorem that belongs, properly, in the field of parametric statistics. Remarkably, it appears to be new to that field. We begin with the setting.

Consider a probability density f on the real line with finite first and second moments; say the first moment is 0. From f, form the family of densities $\{f_\theta\}$, which are the translations of f, indexed by their means θ. (So f_0 has mean 0.)

Now, θ is fixed and unknown to us, and we collect n samples $x_1, \ldots, x_n$ from the density f_θ. We wish to infer an estimate of the parameter θ. For each $\varepsilon > 0$, we are interested in the probability that our estimator $S(x_1, \ldots, x_n)$ falls within distance ε of θ. Furthermore, we are interested in the *worst case* performance of S. For this purpose, let us define the *ε-quality* of estimator S to be

$$Q_S^\varepsilon = \inf_\theta [P(|S - \theta| \leq \varepsilon)].$$

Definition 2 *We say that estimator T* majorizes *S if for all $\varepsilon > 0$, $Q_T^\varepsilon \geq Q_S^\varepsilon$.*

Let G_θ denote the Gaussian density function with mean θ and unit standard deviation. Let $\{G_\theta\}$ denote the variable location, fixed scale family of such densities. We prove:

Theorem 3 *For the family $\{G_\theta\}$, for any given n, the mean estimator, $T(x_1, \ldots, x_n) = \frac{1}{n}\sum x_i$, majorizes every other estimator.*

In Theorem 10 we will further establish that T is the *unique* majorizing estimator.

Let $X = (x_1, \ldots, x_n)$ denote n independent samples picked from G_θ. Define the r^{th} *central moment of estimator S at θ* to be

$$M_\theta^r(S) = \int P(X|\theta) \int_{t \in \mathbb{R}} |t - \theta|^r \; P(S(X) = t) dt dX.$$

In the final version of this paper, by an extension of the method of Theorem 3, we will show:

Theorem 4 *For the family $\{G_\theta\}$, the mean estimator minimizes $\sup_\theta M_\theta^r(S)$ among all estimators S, for every n and every $r > 0$.*

Among the celebrated theorems of statistics is the Cramer-Rao lower bound on the mean squared error of an unbiased estimator of a parameter θ. A key application of that theorem is to show that the sample mean is an optimal unbiased estimator of the mean of a Gaussian from the family $\{G_\theta\}$ [1, 8, 3, 6, 2]. Theorem 4 represents an improvement in the sense in which the mean estimator for the Gaussian is shown to be optimal, as it implies optimality of the mean estimator in mean square (variation), as for any central moment, among all (not only unbiased) estimators.

Discussion:

(1) To prove theorem 1, simply substitute a Gaussian family as the unknown family of distributions with known second moment. Suppose that $\frac{E(X_I^2)^{1/2}}{E(X_I)}$ is greater than any polynomial in $|I|$. Let T be an arbitrary estimator for $\theta = E(X_I)$, and let $\overline{X}$ be the mean of a polynomial number $s(|I|)$ of samples. Then

$$P(|T - \theta| \leq \varepsilon\theta) \leq P(|\overline{X} - \theta| \leq \varepsilon\theta)$$

Now since the standard deviation $E(X_I^2)^{1/2}/s^{1/2}(|I|)$ of $\overline{X}$ is much larger than $\varepsilon\theta = \varepsilon E(X_I)$ (for $\varepsilon \leq 1$), its density function is close to constant in the interval $[\theta(1-\varepsilon), \theta(1+\varepsilon)]$, and the last term is equal, up to a constant factor, to:

$$\frac{\varepsilon\theta s^{1/2}(n)}{E(X_I^2)^{1/2}} = \frac{\varepsilon E(X_I) s^{1/2}(n)}{E(X_I^2)^{1/2}}$$

which tends to 0 for any polynomial s. Hence T is not a FPRAS.

(2) Motivated by the algorithmic applications, we have chosen to measure the quality of an estimator T of a parameter θ by the function $\inf_\theta [P(|T - \theta| \leq \varepsilon)]$. A somewhat "dual" notion is studied in the statistical literature. A *confidence interval* of *level p* is a pair of estimators T_1, T_2 s.t. for every θ, with probability at least p, $T_1 \leq \theta \leq T_2$. Obviously it is desirable that the intervals $[T_1, T_2]$ be as short as possible subject to the confidence level p; this objective is complementary to our goal of maximizing the estimator's probability of falling within a fixed width interval, $\inf_\theta[P(|T-\theta| \leq \varepsilon)]$.

However, in the case of confidence intervals, there is an additional degree of freedom available in "sliding" both ends of the interval without changing the confidence level. While this flexibility is desirable for some applications (e.g. if the penalties for errors in the two directions are unequal), it reduces the extent to which the quality of estimators can be compared. In particular, there does not exist any family of densities $\{f_\theta\}$, and any $0 < p < 1$, for which there is an optimal estimator (in the sense that its confidence intervals are contained within those of any other estimator). (And a statement nearly as strong can be made also for families of distributions which do not arise from densities.) To see this, one has only to consider the two optimal estimators subject to the restrictions that the lower or upper endpoints are at $-\infty$ or $+\infty$. Estimators that are optimal subject to these restrictions are termed "uniformly most accurate upper/lower (respectively) confidence limits"; this appears to be the closest definition in the literature to our notion of a majorizing estimator. However, as just implied, no theorem resembling theorem 3 exists for upper and lower confidence limits. Thus one of the contributions of this paper is the introduction of the measure, Q^ε_S, of the quality of an estimator S, since this is a measure that is on the one hand, much stronger than commonly used measures such as mean squared error; and on the other hand, the resulting partial order on estimators is not so weak as to preclude the existence of a dominating estimator in the partial order.

It is interesting that in seeking to answer our question, which arose very naturally from computional applications, we encountered an elementary, but unexplored, domain within parametric statistics.

(3) Consider the following "maximum likelihood" estimator for parameter θ:

$$S(X) = \arg\max_\theta\{\int_{t=\theta-\varepsilon}^{\theta+\varepsilon} P(X|t)dt\}.$$

Actually, this is a whole class of estimators – depending on how ties are resolved. Is such an estimator always a majorizing estimator? First notice that a maximum likelihood estimator exists for any density f; on the other hand, it is easy to construct a density for which there is no majorizing estimator (any asymmetric density function suffices).

What if we were to restrict to density functions that do admit a majorizing estimator? In a later version of this paper we will describe two density functions: one for which there is a majorizing estimator that is not a maximum likelihood estimator, and another for which there is a maximum likelihood estimator that is not a majorizing estimator. A question that we have not answered is whether for density functions that do admit a majorizing estimator, there is always such an estimator that is also a maximum likelihood estimator.

(4) In the final version of this paper we will discuss the case of estimation of the mean in the family of spherically symmetric multidimensional Gaussian distributions.

2 The computational setting

A *counting problem*, Π, consists of:

- A set of *instances*, D_Π.
- The *size* of instance $I \in D_\Pi$, denoted by $|I|$, is defined as the number of bits needed to write I under the assumption that all numbers occurring in the instance are written in binary.
- A *solution space* S_I, typically of size exponential in $|I|$, is associated with instance I. A parameter of S_I, θ_I, is defined.

As an example, consider the problem of determining the probability of satisfying a given DNF formula, given probabilities of each of its Boolean variables being true.

The interesting case is when $\Pi \in \mathbf{P}$, and when computing θ_I as a function of I is $\#\mathbf{P}$-complete. We will say that an algorithm A is a fully polynomial randomized approximation scheme for computing θ_I if for each instance I and error parameter $\varepsilon > 0$,

$$P(|A(I) - \theta_I|/\theta_I \leq \varepsilon) \geq \frac{3}{4},$$

and the running time of A is polynomially bounded in $|I|$ and $\frac{1}{\varepsilon}$. As stated in the introduction, once this is achieved, the probability of success can be amplified using the "median trick".

Typically, a FPRAS for computing θ_I is constructed as follows: A polynomial time samplable probability distribution is defined on S_I, together with a random variable X_I, which is shown to be an unbiased estimator of θ_I, or nearly so, the error being $< \varepsilon$. A specified number of sample points are picked from the probability distribution, the random variable is computed at these points, and the mean of these values is output. Theorem 1 shows that in general, this is essentially all one can do.

3 The mean is a majorizing estimator for the Gaussian variable-location fixed-scale family

Let T be the mean estimator. Clearly, an arbitrary estimator S may be able to do better than T on certain specific values of θ. We wish to show that even so, in the worst case, T must be doing at least as well as S. An important observation is that T *commutes with translation,* i.e., $T[X+a] = T[X] + a$, where $X + a$ denotes the n samples $(x_1 + a, x_2 + a, \ldots, x_n + a)$. Therefore, its probability of falling within an ε distance of θ, $P(|T-\theta| \leq \varepsilon)$, is independent of θ.

Thus, the worst case performance of T is the same as its performance at any θ. The worst case performance of a general estimator S, however, is difficult to characterize. Instead, we will show that in the limit, the *average performance* of T over a large range of θ's must be at least as good as that of S. This will lead to the majorization result.

Proof of theorem 3:
Let G_θ denote the Gaussian density function with mean θ and unit standard deviation, $G_\theta(x) = (2\pi)^{-1/2}\exp(-(x-\theta)^2/2)$. Consider the following process: Let $\varepsilon > 0$ be fixed. For fixed $\alpha > 0$, θ is picked uniformly at random from the interval $I_\alpha = [-\alpha, \alpha]$, and then n samples $X = (x_1, \ldots, x_n)$ are picked from the distribution G_θ. (We will call this the finite-α experiment.) Let $P(X|\theta) = \prod G_\theta(x_i)$ denote the probability (density) with which X is produced while sampling from G_θ.

Let $S(x_1,\ldots,x_n)$ be an estimator of θ. In general, S may use the flips of a fair coin, i.e., it may be randomized; $P(S(X)=y)$ denotes the probability (density) with which the estimator T outputs y on input X. Let $\varepsilon>0$ be fixed. We will say that S succeeds if $\theta\in[S(X)-\varepsilon,S(X)+\varepsilon]$. The probability of success of S over the entire finite-α experiment is given by

$$\int\int_{S(X)-\varepsilon}^{S(X)+\varepsilon} P(\theta)P(X|\theta)d\theta dX$$

if S is deterministic, and by

$$\int\int_{-\infty}^{\infty} P(S(X)=y)\int_{y-\varepsilon}^{y+\varepsilon} P(\theta)P(X|\theta)d\theta dy dX$$

if S is randomized.

Let $\overline{X}$ denote the arithmetic mean of the n sample points $x_1,\ldots,x_n$.

Lemma 5 *For given X define the function $g(\theta)=P(X|\theta)$. For X such that $\overline{X}\in I_\alpha$:*

- *$g(\theta)$ is maximized for $\theta=\overline{X}$,*
- *$g(\theta)$ is symmetric around $\theta=\overline{X}$, i.e., $g(\overline{X}-\delta)=g(\overline{X}+\delta)$, and*
- *$g(\overline{X}-\delta)$ is monotonically decreasing with $|\delta|$.*

Proof: Since G_θ is a Gaussian density function with mean θ and unit standard deviation,

$$P(X|\theta)=(2\pi)^{-n/2}e^{-\frac{1}{2}\sum_{i=1}^n(x_i-\theta)^2}.$$

Let $y_i=x_i-\overline{X}$, for $1\le i\le n$. Substituting,

$$P(X|\theta)=(2\pi)^{-n/2}e^{-\frac{1}{2}\sum_{i=1}^n(y_i+\overline{X}-\theta)^2}$$

$$=(2\pi)^{-n/2}e^{-\frac{1}{2}(\sum_i y_i^2+n(\overline{X}-\theta)^2+2(\sum_i y_i)(\overline{X}-\theta))}.$$

Since $\sum_i y_i=0$, we get

$$P(X|\theta)=(2\pi)^{-n/2}e^{-\frac{1}{2}(\sum_i y_i^2+n(\overline{X}-\theta)^2)}$$

which is maximized for $\theta=\overline{X}$.

Finally,

$$P(X|\overline{X}-\delta)=P(X|\overline{X}+\delta)=e^{-\frac{1}{2}(\sum_i y_i^2+n\delta^2)},$$

which makes it clear that g is symmetric around $\theta=\overline{X}$, and that $g(\overline{X}-\delta)$ is monotonically decreasing with $|\delta|$.

Let I'_α denote the interval $[-(\alpha-\varepsilon),(\alpha-\varepsilon)]$.

Lemma 6 *For X such that $\overline{X}\in I'_\alpha$,*

$$\int_{-\infty}^{\infty} P(T(X)=y)\int_{y-\varepsilon}^{y+\varepsilon} P(\theta)P(X|\theta)d\theta dy$$

is uniquely maximized for the mean estimator, i.e., the estimator that deterministically outputs $T(X)=\overline{X}$.

Proof: By the fact that θ is picked uniformly at random from the interval I_α, and by Lemma 5,

$$\int_{y-\varepsilon}^{y+\varepsilon} P(\theta)P(X|\theta)d\theta$$

is uniquely maximized at $y=\overline{X}$. The lemma follows. □

For an estimator S, let $P_S^{\alpha,\varepsilon}$ denote the probability of success of S in the finite-α experiment. Since the mean estimator commutes with translation, we find:

Observation 7 $\quad Q_T^\varepsilon=P_T^{\alpha,\varepsilon}$.

Let $M(\alpha,\varepsilon)$ denote the supremum over all estimators S of $P_S^{\alpha,\varepsilon}$. Let $B(\alpha,\varepsilon)$ be the event that, after picking θ at random from I_α and X at random using the distribution G_θ, $\overline{X}\notin I'_\alpha$. By Lemma 6, we get:

Corollary 8

$$P_T^{\alpha,\varepsilon}\ge M(\alpha,\varepsilon)-P(B(\alpha,\varepsilon)).$$

□

Finally, let $Q(\varepsilon)=\sup_S Q_S^\varepsilon$. We wish to show that $Q_T^\varepsilon=Q(\varepsilon)$, thus proving the theorem.

By Observation 7 and Corollary 8,

$$Q_T^\varepsilon=P_T^{\alpha,\varepsilon}\ge\liminf_\alpha M(\alpha,\varepsilon)-\limsup_\alpha P(B(\alpha,\varepsilon)).$$

Since any estimator can be employed without modification in the finite-α experiment, $M(\alpha,\varepsilon)\ge Q(\varepsilon)$. Therefore,

$$Q_T^\varepsilon\ge Q(\varepsilon)-\limsup_\alpha P(B(\alpha,\varepsilon)).$$

Now,

$$\limsup_\alpha P(B(\alpha,\varepsilon))$$

$$\le\limsup_\alpha[P(|\theta|>\alpha-\alpha^{1/2})+P(\overline{X}\notin I'_\alpha\mid|\theta|\le\alpha-\alpha^{1/2})]$$

$$\le 0+\limsup_\alpha P(|\overline{X}-\theta|>\alpha^{1/2}-\varepsilon).$$

Since $\overline{X}$ is normally distributed, this is bounded above by

$$\limsup_\alpha\exp(-n(\alpha^{1/2}-\varepsilon)^2/2)=0.$$

Hence $Q_T^\varepsilon\ge Q(\varepsilon)$. □

3.1 Discrete distributions

Often in an algorithmic problem some extra information is available about the underlying distribution. It is worth looking at some of the most common kinds of extra information available, and ruling out the possibility that they might dramatically change the situation and render theorem 1 irrelevant.

A frequent kind of extra information is that θ is an integer, and the estimator takes on only integer values.

In the same situations, however, it is also the case that the range of these integers is superpolynomial as a function of the complexity parameter n. Thus the following *reduction* typically eliminates the possibility of a substantive difference between the general and integer cases.

Theorem 9 *Let S be an estimator that, for any family of distributions $\{F_\theta\}$ on the integers, with standard deviation bounded by s and integer means θ, has the following property: given n iid samples from a distribution F_θ, with probability at least q, S falls in the interval $[\theta - r, \theta + r]$.*

Then there is an estimator T that, for any family of distributions $\{f_\theta\}$ with standard deviation bounded by $s-1$ and means θ, has the following property: given n iid samples from a distribution f_θ, with probability at least $q - O(n/s)$, T falls in the interval $[\theta - r - 1/2, \theta + r + 1/2]$.

Proof: Take $\{f_\theta\}$ to be the Gaussian family with standard deviation $s-1$.

For any θ form F_θ by setting $F_\theta(x) = \int_{x-1/2}^{x+1/2} f_\theta(y)dy$ for integer x.

We reduce inference of $\{f_\theta\}$ to inference of the family of distributions $\{F_\ell\}$ (ℓ integer).

A particular f_θ is an instance of the original inference problem; it is reduced to the inference problem $F_{\theta'}$ where θ' is the integer in $[\theta - 1/2, \theta + 1/2)$.

If A is an r.v. distributed according to f_θ, and B is an r.v. distributed according to $F_{\theta'}$, then $|E(B) - E(A)| \leq 1/2$ (or 0 in the special case that θ is an integer), and $|\mathrm{StDev}(B) - \mathrm{StDev}(A)| \leq 1$.

The reduction proceeds by rounding each sample point x to the integer in $[x-1/2, x+1/2)$. First suppose θ is integer: the hypothesis in conjunction with the previous paragraph implies that the probability of S falling in the interval $[\theta - r, \theta + r]$ is at least q. If θ is not integer, we relate F_θ to $F_{\theta'}$ by sampling from a coupled distribution, and introducing only an additional error term for the possibility that a sample falls on the uncoupled portion of the distribution: hence the probability of S falling in the interval $[\theta - r - 1/2, \theta + r + 1/2]$ is at least $q - L_1(F_\theta, F_{\theta'}) \cdot (\#\text{ samples}) \geq q - O(n/s)$. (Here L_1 is the variation distance.)

3.2 The next question

In the definition of a FPRAS, once the probability of falling in an ε interval is ensured to be at least $3/4$ (perhaps by taking the mean of m samples), then it can also be enhanced to $1 - \delta$ for any $\delta > 0$ as follows: repeat this experiment $k = O(\log(\frac{1}{\delta}))$ times, and take the median of these values. This process involves using km sample points.

Here is an alternative method: obtain km sample points, and for *every* subset of m of these points, compute their mean. Output the median of these values. Is the probability of this median lying in the "ε-window" about the true value, higher than for the disjoint-subsets scheme?

This is not an improvement in all situations: Bruce Reed and Peter Winkler have observed that in some cases in which the probability of the mean of m samples falling in the ε window is less than $1/2$, this method performs worse than the standard (disjoint subsets) method. However, the important case for approximation algorithms is when the probability of the mean falling in the desired window is large.

If this proposal is successful, it will also be important to determine whether it suffices to examine only polynomially many subsets instead of all $\binom{mk}{m}$.

4 Uniqueness of the mean estimator

We now strengthen Theorem 3 by showing that the mean is the *unique* majorizing estimator for the family $\{G_\theta\}$. This requires a more delicate argument than the earlier theorem. In the earlier case we did not have to rule out an estimator which improved its odds of success at some values of θ, so long as we could rule out its doing better, by an amount bounded away from 0, everywhere; for this purpose it was sufficient to look at a long enough segment of θ's, show that not much benefit could be contributed to this interval by samples from outside of it, and then average uniformly the probability of success within the interval, showing that this average could improve over the mean estimator only by a quantity tending to zero in the length of the interval. There was nothing to prevent the estimator differing from the mean estimator, and indeed improving on the mean estimator locally, so long as it compensated for that change by "importing" estimates toward the values of θ that were "neglected". Now, however, we have to show that if the estimator differs from the mean estimator anywhere, then such a compensation mechanism, while easy to construct in the neighborhood of a small difference, must ultimately fail. The reason for this failure is that the needed compensations in the estimator themselves require compounding compensations, and that this process "diverges".

We begin with some notation: $\mathcal{L}$ is the set of Lebesgue measurable sets in $\mathbb{R}^j$ and μ is the usual Lebesgue measure on $\mathbb{R}^j$ (we write $\mathcal{L}$ and μ regardless of j). For an interval $B \subseteq \mathbb{R}$ we also write $|B| = \mu(B)$. Let $G(y) = (2\pi)^{-1/2}\exp(-y^2/2)$, and let $\mathcal{N}(y) = \int_{-\infty}^{y} G(z)dz$. The uniqueness theorem is proven in the following generality: an estimator S is a measure on $(\mathbb{R}^{n+1}, \mathcal{L})$ (arguments $2, ..., n+1$ are the samples $x_1, ..., x_n$, the first argument is the estimate for θ), that satisfies the following condition: for all measurable sets $A \subseteq \mathbb{R}^n$, $S(\mathbb{R} \times A) = \mu(A)$.

Theorem 10 *If there is a measurable set A such that $S(A) \neq T(A)$ then for every ε, $Q_S^\varepsilon < Q_T^\varepsilon$.*

Proof: We begin with a simple fact that substantially simplifies the matter.

Lemma 11 *It suffices to consider the single-sample case.*

Proof: The mean estimator is a *sufficient statistic* for the family $\{G_\theta\}$. Consequently the performance of any estimator will be unchanged if, given $x_1, ..., x_n$, we first compute the mean $\bar{x} = \frac{1}{n}\sum x_i$, then choose a list of differences $(x_i' - \bar{x})_1^n$ from the same distribution as for the Gaussian (note in particular that the distribution is supported only on lists whose sum is 0), then supply the estimator with the list $x_1', ..., x_n'$. The distribution of such lists is the same as that of the lists $x_1, ..., x_n$, whence the conclustion that the performance is unaffected. Now, the process of substitution followed by application of the estimator, may be viewed jointly as a (randomized) estimator that takes as its input only the mean $\bar{x}$. □

Let T be the deterministic estimator which, given x, guesses that $\theta = x$. More precisely T is the diagonal measure: if $J = \{(t, x) : t = x\}$ and π_2 is the projection of $\mathbb{R}^2$ on its second coordinate then $T(A) = \mu(\pi_2(A \cap J))$.

Define the *ε-quality of estimator S at θ* to be

$$Q_S^\varepsilon(\theta) = \int_{x \in \mathbb{R}} G(x - \theta) \int_{t=\theta-\varepsilon}^{\theta+\varepsilon} dS(t, x).$$

For any $\theta \in \mathbb{R}$, the ε-quality of T at θ is $\mathcal{N}(\varepsilon) - \mathcal{N}(-\varepsilon) = 2\mathcal{N}(\varepsilon) - 1$. For the rest of the discussion, assume that $\varepsilon > 0$ is fixed.

The quantity of interest for us is $Q_S^\varepsilon = \inf_\theta Q_S^\varepsilon(\theta)$. As in the proof of Theorem 3, we will need to consider the *average* performance of S in order to characterize its worst case performance. Thus, for a measurable set B, we will be interested in

$$F_S(B) = \int_{\theta\in B}\int_{x\in\mathbb{R}} G(x-\theta)\int_{t=\theta-\varepsilon}^{\theta+\varepsilon} dS(t,x)d\theta.$$

Let us define this to be the *estimation total for θ in B.* For convenience, let us first express this as a double integral: Let u_B be the characteristic function for B. For $x,t\in\mathbb{R}$, define

$$\alpha(x,t,B) = \int_{t-\varepsilon}^{t+\varepsilon} G(s-x)u_B(s)ds,$$

For instance, if $B=\mathbb{R}$, then this is simply $\mathcal{N}(-x+t+\varepsilon) - \mathcal{N}(-x+t-\varepsilon)$. The reader can now verify that

$$F_S(B) = \int_{x\in\mathbb{R}}\int_{t\in\mathbb{R}} \alpha(x,t,B)dS(t,x).$$

More generally, for two measurable sets B and D, let us define the *estimation total for θ in B due to x in D* to be

$$F_S(B;D) = \int_{x\in D}\int_{t\in\mathbb{R}} \alpha(x,t,B)dS(t,x).$$

A quantity of special interest is the total amount accrued due to x in D, $F_S(\mathbb{R};D)$. Notice that this is maximized by the mean estimator; in particular,

$$F_T(\mathbb{R};D) = \mu(D)(2\mathcal{N}(\varepsilon)-1).$$

Finally, define the *deficit of estimator S on set B,*

$$\Delta_S(B) = \int_{x\in B}\int_{t\in\mathbb{R}} \alpha(x,t,\mathbb{R})(dT(t,x)-dS(t,x)).$$

For the special case of finite measure B this is the same as

$$\Delta_S(B) = F_T(\mathbb{R};B) - F_S(\mathbb{R};B).$$

Lemma 12 *If $S(A)\neq T(A)$ for a measurable set A, then there is a finite interval B for which $\Delta_S(B)>0$.*

Proof: By countable additivity, we may assume that there is a finite interval B such that $A\subseteq\mathbb{R}\times B$. We first claim that $S((\mathbb{R}\times B)-J) > T((\mathbb{R}\times B)-J)$. Clearly, $S(A\cap J)\le T(A\cap J)$. If $S(A\cap J) < T(A\cap J)$, the claim follows since $S(\mathbb{R}\times B) = S(\mathbb{R}\times B) = \mu(B)$. On the other hand, if $S(A\cap J) = T(A\cap J)$, then $S(A-J) > T(A-J) = 0$. Since $T((\mathbb{R}\times B)-J) = 0$, the claim follows again.

Partition $(\mathbb{R}\times B)-J$ into regions $K_j = \{(t,x) : 2^j \le |t-x| < 2^{j+1}\}\cap(\mathbb{R}\times B)$, for each integer j. Again, using countable additivity, there is a j such that $S(K_j)>0$. Then $\Delta_S(B)\ge S(K_j)[(\mathcal{N}(\varepsilon)-\mathcal{N}(-\varepsilon)) - (\mathcal{N}(\varepsilon+2^j)-\mathcal{N}(-\varepsilon+2^j))] > 0$. □

Let B' denote the interval obtained by extending interval B by ε on each side. The next lemma shows that deficit must lead to a smaller estimation total for S (as compared to T).

Lemma 13 *For a finite interval B, $F_T(B';B) - F_S(B';B) \ge \Delta_S(B)$.*

Proof: Observe that $F_T(\mathbb{R};B) = F_T(B';B)$. Furthermore, since $\alpha(x,t,\mathbb{R})\ge\alpha(x,t,B')$ for any $x,t\in\mathbb{R}$, $F_S(\mathbb{R};B)\ge F_S(B';B)$. Therefore,

$$\Delta_S(B) = F_T(\mathbb{R};B) - F_S(\mathbb{R};B) \le F_T(B';B) - F_S(B';B).$$

□

Lemma 14 *If $\Delta_S(\mathbb{R})>0$ then there is set D of finite measure such that $F_S(D) < F_T(D)$.*

Proof: There are two cases:
Case (i): $\Delta_S(\mathbb{R})$ is infinite.
Let B be a finite interval such that $\Delta_S(B)>\varepsilon$. By Lemma 13, $F_S(B';B)\le F_T(B';B)-\Delta_S(B) < F_T(B';B)-\varepsilon$. Clearly, $F_S(B';\mathbb{R}-B)$ is maximized by the estimator that, for each $x\in\mathbb{R}-B$, guesses the closest endpoint of B. It is easy to verify that for such an estimator, $F_S(B';\mathbb{R}-B)\le\varepsilon$. Therefore,

$$F_S(B') \le F_S(B';B)+\varepsilon < F_T(B';B) < F_T(B').$$

Case (ii): $\Delta_S(\mathbb{R})$ is finite.
Let B be a finite interval such that $g(\Delta_S(\mathbb{R}-B))\le\Delta_S(B)/2$, where g, to be defined below, is a monotone increasing continuous function on the nonnegative reals, with $g(0)=0$. Define B' as above.

By Lemma 13,

$$F_S(B';B)\le F_T(B';B)-\Delta_S(B)\le |B|(2\mathcal{N}(\varepsilon)-1)-\Delta_S(B),$$

we get

$$F_S(B') = F_S(B';B)+F_S(B';\mathbb{R}-B)\le$$
$$|B|(2\mathcal{N}(\varepsilon)-1)-\Delta_S(B)+F_S(B';\mathbb{R}-B).$$

In the simplest case, that S is identical to T on $\mathbb{R}-B$, the last term equals $2\varepsilon(2\mathcal{N}(\varepsilon)-1)$ and so $F_S(B')\le|B'|(2\mathcal{N}(\varepsilon)-1)-\Delta_S(B) < |B'|(2\mathcal{N}(\varepsilon)-1) = F_T(B')$. However, $\Delta_S(\mathbb{R}-B)$ may be nonzero. This allows estimates to be shifted so as to increase $F_S(B')$. The remainder of the argument is devoted to showing that this increase, which we call $DF_S(B')$, is less than $\Delta_S(B)$, provided $\Delta_S(\mathbb{R}-B)$ is sufficiently small as specified above.

If, at distance y from B, the estimator is shifted by distance r toward B, then the contribution toward $\Delta_S(\mathbb{R}-B)$ is proportional to $\int_0^r(G(-\varepsilon+s)-G(\varepsilon+s))ds = -\mathcal{N}(\varepsilon+r)+\mathcal{N}(\varepsilon)+\mathcal{N}(-\varepsilon+r)-\mathcal{N}(-\varepsilon)$. Meanwhile, $DF_S(B')$ is $\mathcal{N}(\varepsilon+r)-\mathcal{N}(\varepsilon)$ for $y\le 2\varepsilon$, provided $0\le r\le y$ (greater values of r contribute less to $F_S(B')$); while for $y\ge 2\varepsilon$ $DF_S(B')$ is 0 for $0\le r\le y-2\varepsilon$, and $\mathcal{N}(\varepsilon+r)-\mathcal{N}(y-\varepsilon)$ for $y-2\varepsilon\le r\le y$ (again, greater values of r contribute less to $F_S(B')$).

First, we claim that the best gain in $F_S(B')$ (greatest value of $DF_S(B')$) given the limit on $\Delta_S(\mathbb{R}-B)$ is achieved by a "deterministic" estimator, i.e. one which for any y, places the entire measure on a particular value of r. This is for the following reason. Let the equation

$$-\mathcal{N}(\varepsilon+r)+\mathcal{N}(\varepsilon)+\mathcal{N}(-\varepsilon+r)-\mathcal{N}(-\varepsilon) = z$$

implicitly define r as a function of z, and let h denote the function such that $h(z)$ equals $\mathcal{N}(\varepsilon+r)-\mathcal{N}(\varepsilon)$ for the r corresponding to z. Then calculation shows that for $y\le 2\varepsilon$, h is a convex cap, increasing function, hence a convex combination $\sum_{p_i} h(z_i)$ is maximized, given an upper bound on $\sum_{p_i} z_i$

(the local deficit), by choosing a singular distribution, i.e. a deterministic estimator. A similar argument yields the same conclusion for $y \geq 2\varepsilon$.

Moreover, the ratio of "gain" to "cost"

$$\frac{\mathcal{N}(\varepsilon+r)-\mathcal{N}(\varepsilon)}{-\mathcal{N}(\varepsilon+r)+\mathcal{N}(\varepsilon)+\mathcal{N}(-\varepsilon+r)-\mathcal{N}(-\varepsilon)} \quad (1)$$

does not depend on y, for $y \leq 2\varepsilon$; hence it is optimal to use the same shift r for all $y \leq 2\varepsilon$. Moreover since the ratio is only worse for $y \geq 2\varepsilon$, where it is given by the equation

$$\frac{\mathcal{N}(\varepsilon+r)-\mathcal{N}(y-\varepsilon)}{-\mathcal{N}(\varepsilon+r)+\mathcal{N}(\varepsilon)+\mathcal{N}(-\varepsilon+r)-\mathcal{N}(-\varepsilon)} \quad (2)$$

it follows that in an optimal estimator the shift used at that range can be no greater. We therefore obtain an upper bound on $DF_S(B')$ in the following way: considering only $y \leq 2\varepsilon$, find the shift r_0 such that $DF_S(B')$ is maximized without the deficit exceeding $\Delta_S(\mathbb{R}-B)$. Observe that r_0 is at least as great as the shift used by the optimal estimator for $y \leq 2\varepsilon$ (the optimal estimator may not use all of the deficit on these values of y, and so may not be able to "afford" as great a shift.) Now since r_0 can be at most 2ε, and since the optimal estimator uses a shift of at most r_0 for $y \geq 2\varepsilon$, it follows that the optimal estimator does not introduce any shift at all for any $y > 4\varepsilon$. So we can upper bound $DF_S(B')$ by $8\varepsilon(\mathcal{N}(\varepsilon+r_0)-\mathcal{N}(\varepsilon))$. (A factor of 2 has been introduced to account for both sides of B.)

The equation defining r_0 is $\Delta_S(\mathbb{R}-B) = 4\varepsilon[-\mathcal{N}(\varepsilon+r_0)+\mathcal{N}(\varepsilon)+\mathcal{N}(-\varepsilon+r_0)-\mathcal{N}(-\varepsilon)]$. Let g_1 denote the implicitly defined function on $\mathbb{R}_{\geq 0}$ giving r_0 as a function of $\Delta_S(\mathbb{R}-B)$; note that g_1 is monotone increasing, continuous and that $\lim_{x\to 0} g_1(x) = 0$. Next, let $g_2(x) = 8\varepsilon(\mathcal{N}(\varepsilon+x)-\mathcal{N}(\varepsilon))$; note that g_2 is monotone increasing, continuous and that $\lim_{x\to 0} g_2(x) = 0$. The composite function $g(x) = g_2(g_1(x))$ is an upper bound on $DF_S(B')$ as a function of $\Delta_S(\mathbb{R}-B)$; note that g is monotone increasing, continuous and that $\lim_{x\to 0} g(x) = 0$. This is the function g required at the outset of the proof in the selection of B; and now, using the assumption that $g(\Delta_S(\mathbb{R}-B)) \leq \Delta_S(B)/2$, we find that $DF_S(B') \leq \Delta_S(B)/2$ and therefore (by comparing with the estimator which is equal to the mean outside B), we find that $F_S(B') \leq |B'|(2\mathcal{N}(\varepsilon)-1)-\Delta_S(B)+DF_S(B') \leq |B'|(2\mathcal{N}(\varepsilon)-1)-\Delta_S(B)/2 < |B'|(2\mathcal{N}(\varepsilon)-1) = F_T(B')$. □

5 Open issues

One of the contributions of this paper is the introduction of the notion of a majorizing estimator – this notion deserves further study. For instance, does every distribution possessing a majoring estimator have a deterministic majorizing estimator? Another issue worth resolving is clarifying the relationship of this notion to that of a maximum likelihood estimator (see end of Section 1). Finally, the alternative to the "median trick" suggested in Section 3.2 is also worth studying.

6 Acknowledgments

We wish to thank Prof. D. Blackwell and Prof. C. R. Rao for helping us confirm the status of Theorem 3.

References

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[2] H. Cramer. A contribution to the theory of statistical estimation. *Skandinavisk Aktuarietidskrift*, 29:85–94, 1946.

[3] M. Frechet. sur l'extension de certain evaluations statistique au cas des petit echantillons. *Rev. Inst. Stat.*, 11:182–205, 1943.

[4] N. Karmarkar, R. Karp, R. Lipton, L. Lovász, and M. Luby. A Monte-Carlo algorithm for estimating the permanent. *SIAM Journal on Computing*, 22(2):284–293, April 1993.

[5] R. Motwani and P. Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

[6] C. R. Rao. Information and accuracy attainable in estimation of statistical parameters. *Bull. Cal. Math. Soc.*, 37:81–91, 1945.

[7] A. Sinclair. *Algorithms for Random Generation and Counting: a Markov Chain Approach.* Birkhauser, 1992.

[8] S. Zacks. *Parametric Statistical Inference.* Pergamon Press, 1981.