COMBINATORICA
Bolyai Society – Springer-Verlag

COMBINATORICA **25** (4) (2005) 465–486

A COMPUTATIONALLY MOTIVATED DEFINITION OF PARAMETRIC ESTIMATION AND ITS APPLICATIONS TO THE GAUSSIAN DISTRIBUTION

LEONARD J. SCHULMAN, VIJAY V. VAZIRANI

Received December 20, 2001

We introduce a treatment of parametric estimation in which optimality of an estimator is measured *in probability* rather than in variance (the measure for which the strongest general results are known in statistics). Our motivation is that the quality of an approximation algorithm is measured by the probability that it fails to approximate the desired quantity within a set tolerance. We concentrate on the Gaussian distribution and show that the sample mean is the unique "best" estimator, in probability, for the mean of a Gaussian distribution. We also extend this method to general penalty functions and to multidimensional spherically symmetric Gaussians.

The algorithmic significance of studying the Gaussian distribution is established by showing that determining the average matching size in a graph is #**P**-hard, and moreover approximating it reduces to estimating the mean of a random variable that (under some mild conditions) has a distribution closely approximating a Gaussian. This random variable is (essentially) polynomial time samplable, thereby yielding an FPRAS for the problem.

1. Introduction

The task of estimating a parameter via sampling lies at the heart of numerous algorithms. In particular, this task is central to approximation algorithms for #**P**-hard problems, for instance using the Markov chain Monte Carlo method (MCMC method). These algorithms rely on an *in probability* statement: establishing that the parameter in question has been estimated within certain bounds with "high" probability (probability bounded away from 1/2 suffices). Despite widespread use of this method, so far questions

Mathematics Subject Classification (2000): 68Q15, 68Q25, 68W20, 68W25, 62F25

about the *optimality* of the estimator have not been studied. (In the case of estimation using the MCMC method, the overall running time of the algorithm depends *multiplicatively* on the number of samples obtained from the Markov chain. Hence decreasing this number by resorting to a better estimator would directly benefit the overall running time.)

Such results belong in the area of parametric estimation within the field of statistics. Traditionally, the notion of optimality most studied in this area is minimization of the mean square error of the estimator. Indeed, among the celebrated theorems of statistics is the Cramer–Rao lower bound on the mean square error of an unbiased estimator of a parameter θ, in terms of its Fisher entropy. A key application of this theorem is to show that the sample mean is an optimal unbiased estimator of the mean of a Gaussian from a variable-location, fixed-scale (unit variance) family $\{G_\theta\}$ where $G_\theta(x) = (2\pi)^{-1/2}\exp(-(x-\theta)^2/2)$ [2,13,5,9,3]. Notice, however, that optimality of an estimator in mean square error does not imply optimality in probability. Furthermore, establishing a bound on the mean square error of an estimator is not sufficient for the purposes of deriving an approximation algorithm, for instance for giving a fully polynomial randomized approximation scheme, FPRAS, for a #**P**-hard problem.

The traditional worst case analysis of algorithms has been particularly successful in unraveling algorithmically relevant combinatorial structure in problems, and in designing powerful algorithmic tools to exploit this structure. We adopt this paradigm in our criterion for parametric estimation.

The current paper is an attempt at initiating a theory of parametric estimation in which optimality is measured in probability in the worst case over the possible values of the parameter. Although our original motivation is algorithmic, we believe that this theory will find value in other areas as well. We derive results for the Gaussian distribution. The reason for concentrating on this distribution is twofold. First, the Gaussian distribution lends itself to a very precise analysis. Second, this distribution arises naturally in several computational situations. Here is a particularly striking case.

Consider the problem of computing the average matching size in a given graph. In Theorem 6 we show that exact computation of this parameter is #P-hard. Now, consider the random variable that is the size of a random, uniformly chosen, matching in G. This random variable is (essentially) polynomial time samplable, since there exists an almost uniform generator for matchings in a graph [7]. Therefore, it can be used to estimate the average size of a matching in G; in fact it even leads to an FPRAS for this problem. The algorithm is straightforward: sample this random variable an appro-

priate number of times, depending on the error parameter, and output the mean of these samples.

Under some mild conditions, this random variable has essentially a Gaussian distribution. This follows from an exceptionally strong result of Godsil describing the size distribution of the matchings of a graph [8,6]. Hence, the FPRAS stated above is simply estimating the mean of this Gaussian distribution! Moreover, it is using the mean estimator to do so. We are interested in whether this is the best estimator for the mean of a Gaussian distribution, where "best" is, as is customary for algorithmic analysis, defined in terms of the *probability* of failing to estimate the mean within the desired accuracy on a *worst-case* input. In Section 2 we define this question more precisely, and in Theorem 2 we answer it in the affirmative.

A preliminary version of the results in this paper appeared in [10].

2. The model and our results

Consider a probability density f on the real line with first moment 0 and finite second moment. Form the family of densities $\{f_\theta\}$, which are the translations of f, indexed by their means θ. (So $f_0 = f$.)

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. An *estimator* $S : \mathbb{R}^n \to \mathbb{R}$ is a function, which given n samples $x_1, \ldots, x_n$ gives an estimate of θ. In general, S may be randomized; $P(S(x) = y)$ denotes the probability (density) with which the estimator S outputs y on input x. 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* (over θ) 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 1. We say that estimator T *majorizes* S if for all $\varepsilon > 0$, $Q_T^\varepsilon \geq Q_S^\varepsilon$.

Theorem 2. *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 14 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. Let $\psi : \mathbb{R} \to \mathbb{R}^+$ (where $\mathbb{R}^+$ is the nonnegative reals) be a *penalty function* satisfying the following conditions: (a) ψ is symmetric, (b) $\psi(x)$ is nondecreasing in $|x|$, (c) ψ is not a constant function, and (d) $L = \int_{-\infty}^{\infty} \psi(x) G_0(x) dx < \infty$.

Without loss of generality we may assume that $\psi(0)=0$. Note also that conditions (a) and (b) imply that ψ is measurable.

Define the ψ^{th} *central moment of estimator* S *at* θ to be

$$M_\theta^\psi(S) = \int P(X|\theta) \int_{t\in\mathbb{R}} \psi(t-\theta)\ P(S(X)=t)dtdX.$$

In case $\psi(x)=x^r$, this is simply the r'th central moment of the estimator S at θ.

By an extension of the method of Theorem 2, we show the more general:

Theorem 3. *For the family* $\{G_\theta\}$*, for every* n *and every penalty function* ψ*, the mean estimator minimizes* $\sup_\theta M_\theta^\psi(S)$ *among all estimators* S*.*

We next extend the method to higher dimensions. Let $X=(x_1,\ldots,x_n)$ denote n independent samples in $\mathbb{R}^d$ picked from G_θ^d, the spherically symmetric Gaussian distribution

$$G_\theta^d(z) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\sum_1^d (z(i)-\theta(i))^2\right).$$

Let $\psi:\mathbb{R}^d\to\mathbb{R}^+$ (where $\mathbb{R}^+$ is the nonnegative reals) be a *penalty function* satisfying the following conditions: (a) ψ is spherically symmetric: $\psi(x)=\psi(y)$ if $|x|=|y|$. (b) $\psi(x)$ is nondecreasing in the Euclidean norm $|x|$. (c) ψ is not a constant function. (d) $L=\int_{\mathbb{R}^d}\psi(x)G_0^d(x)dx<\infty$.

These assumptions imply that ψ is "unimodal on lines", meaning that for every set of real parameters $a_1,\ldots,a_d,b_1,\ldots,b_d$, the function $\psi(a_1t+b_1,\ldots,a_dt+b_d)$ is a unimodal function of the real parameter t. The central moment of an estimator S, $M_\theta^\psi(S)$, is defined in $\mathbb{R}^d$ analogously to $\mathbb{R}$. In section 6 we show:

Theorem 4. *For the family* $\{G_\theta^d\}$*, for every* n *and every penalty function* ψ*, the mean estimator minimizes* $\sup_\theta M_\theta^\psi(S)$ *among all estimators* S*.*

Cramer–Rao

Among the celebrated theorems of statistics is the Cramer–Rao lower bound (due independently to Cramer, Rao and Frechet) 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 $\{G_\theta\}$ [2,13,5,9,3]. Theorem 3 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 indeed in any central moment, among all (not only unbiased) estimators. However it says this only about the worst-case θ while Cramer–Rao enables statements about each θ.

Confidence Intervals

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 statement resembling theorem 2 can exist for upper and lower confidence limits simultaneously. Thus one of the contributions of this paper is the introduction of Q_S^ε as a measure of the quality of an estimator S, since this defines a partial order on estimators that is on the one hand, more refined than that defined by commonly used criteria 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 greatest element in the partial order.

3. Estimating average matching size in a graph

We have established our parametric estimation results for the Gaussian distribution. The question naturally arises whether this special case is of aca-

demic interest only or it actually arises in the computation of a natural #P-hard parameter. In this section we show that the latter is the case – the parameter of interest is the average size of a matching in a graph G.

Let L be a language in **NP**, and let M be its associated polynomial time verifier, i.e., given instance I and polynomially bounded "guess" y, $M(I,y)$ is YES iff $I \in L$. The associated *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 consisting of all strings y such that $M(I,y)$ answers YES. Typically the size of S_I is exponential in $|I|$. A parameter of S_I, θ_I, is defined.

For the problem of interest, an instance is an undirected graph, G, and the solution space is the set of matchings (of all sizes) in G. Recall that a *matching* is a set of edges that pairwise do not share any vertices. The empty set is a matching in every graph. The parameter of interest is the average size of a matching in G. Let us denote it by $\mu(G)$.

The interesting case is when $\Pi \in \mathbf{P}$, and when computing θ_I as a function of I is complete for the counting class #**P** introduced by Valiant in [11]. The problem of finding a matching, even a maximum matching, in G is polynomial time solvable. Below we show that the problem of computing $\mu(G)$ exactly is #P-hard.

Remark 5. Observe that $\mu(G)$ has not been defined as the problem of counting the number of solutions to an **NP**problem, and therefore it does not lie in the class #P. However, one can define a complexity class, based on an abstract machine model that outputs the average of the numbers output on its accepting computation paths, that properly contains this problem.

Theorem 6. *The problem of computing $\mu(G)$ is #P-hard.*

Note that $\mu(G)$ is the ratio p/q of two integers each bounded by $2^{|E(G)|}$ (E being the edge set of G). Computation of $\mu(G)$ can be understood to mean either of two things: (a) computation of a pair p', q' such that $p'/q' = p/q$. (b) Computation of any number r such that $|r - p/q| < 2^{-2|E(G)|-1}$. By the theory of continued fractions, p/q is the unique rational with denominator bounded by $2^{|E(G)|}$ within this distance of r. Hence a pair p', q' as in (a) can be computed from r (simply by computing enough of its continued fraction expansion).

Proof. We reduce from the problem of computing $\phi(G)$, the number of matchings in G, demonstrated #P-complete by Valiant [12].

Consider any edge uv of G. Note that

$$\phi(G) = \phi(G - u - v) + \phi(G - uv).$$

Here we have assumed that $\phi(G) = 1$ in case G has no vertices. Denote by $s(G)$ the sum of sizes of all matchings in G; clearly,

$$s(G) = \mu(G)\phi(G). \tag{1}$$

Observe that among matchings of G which do not use uv, the average matching size is $\mu(G-uv) = s(G-uv)/\phi(G-uv)$. Among matchings of G which do use uv, the average matching size is one more than in $G-u-v$, i.e., $(s(G-u-v)+\phi(G-u-v))/\phi(G-u-v)$. Considering the matchings of G according to whether they contain uv, we see that

$$\mu(G) = \frac{s(G-u-v) + \phi(G-u-v) + s(G-uv)}{\phi(G-u-v) + \phi(G-uv)}.$$

or equivalently

$$\phi(G-u-v)[1 + \mu(G-u-v) - \mu(G)] = \phi(G-uv)[\mu(G) - \mu(G-uv)] \tag{2}$$

Observe that either both or neither of the bracketed terms equal 0, since ϕ is always at least 1.

Next we need:

Lemma 7. *In a graph G with at least one edge, there exists an edge uv such that $\mu(G) > \mu(G-uv)$.*

Proof. We will show that a certain weighted average of the values $\mu(G-uv)$, taken over uv, is less than $\mu(G)$. Assign weight $\phi(G-uv)/\sum_{\{a,b\}\subseteq G}\phi(G-ab)$ to the subgraph $G-uv$. For a graph H write equation 1, with M ranging over matchings, as

$$\mu(H) = \frac{\sum_{M\subseteq H}|M|}{\sum_{M\subseteq H}1}.$$

(H as a set represents the edges of the graph.) Now

$$\sum_{\{u,v\}\subseteq G} \frac{\phi(G-uv)}{\sum_{\{a,b\}\subseteq G}\phi(G-ab)}\mu(G-uv)$$
$$= \sum_{\{u,v\}\subseteq G} \frac{\phi(G-uv)}{\sum_{\{a,b\}\subseteq G}\phi(G-ab)}\frac{s(G-uv)}{\phi(G-uv)}$$

$$
\begin{aligned}
&= \frac{\sum_{\{u,v\}\subseteq G} s(G-uv)}{\sum_{\{u,v\}\subseteq G} \phi(G-uv)} \\
&= \frac{\sum_{M\subseteq G} |M|(|G|-|M|)}{\sum_{M\subseteq G}(|G|-|M|)} \\
&= \frac{\sum_{M\subseteq G}|M|}{\sum_{M\subseteq G} 1} \frac{(\sum_{M\subseteq G} 1)(\sum_{M\subseteq G}|M|(|G|-|M|))}{(\sum_{M\subseteq G}|M|)(\sum_{M\subseteq G}(|G|-|M|))} \\
&= \mu(G) \frac{|G|(\sum_{M\subseteq G} 1)(\sum_{M\subseteq G}|M|) - (\sum_{M\subseteq G} 1)^2 \sum_{M\subseteq G} \frac{|M|^2}{\sum_{M\subseteq G} 1}}{|G|(\sum_{M\subseteq G} 1)(\sum_{M\subseteq G}|M|) - (\sum_{M\subseteq G} 1)^2 \left(\sum_{M\subseteq G} \frac{|M|}{(\sum_{M\subseteq G} 1)}\right)^2} \\
&< \mu(G)\,.
\end{aligned}
$$

The inequality follows from the power mean inequality:

$$\sum_{M\subseteq G} \frac{|M|^2}{\sum_{M\subseteq G} 1} > \left(\sum_{M\subseteq G} \frac{|M|}{(\sum_{M\subseteq G} 1)}\right)^2$$

which is strict unless all terms $|M|$ are equal, which in turn is the case only for a graph with no edges. ∎

Now we can describe our polynomial time Turing reduction of the computation of $\phi(G)$ to the computation of $\mu(G)$:

Algorithm which computes $\phi(G)$ on input G.

1. If G has no edges, output $\phi(G) := 1$. Otherwise continue.
2. Compute $\mu(G)$. For every edge uv compute $\mu(G-uv)$. Pick an edge uv for which $\mu(G-uv) < \mu(G)$. Compute $\mu(G-u-v)$.
3. Recursively compute $\phi(G-uv)$.
4. Set $\phi(G-u-v) := \frac{\phi(G-uv)(\mu(G)-\mu(G-uv))}{1+\mu(G-u-v)-\mu(G)}$.
5. Output $\phi(G) := \phi(G-uv) + \phi(G-u-v)$.

Note: in step 4 we applied equation 2 and lemma 7. ∎

We will say that an algorithm A is a fully polynomial randomized approximation scheme (FPRAS) for computing θ_I if for each instance I and error parameter $\varepsilon > 0$,

$$P(|A(I) - \theta_I| \le \varepsilon\theta_I) \ge \frac{3}{4},$$

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

trick": Run algorithm A a number of times and output the median answer. It is easy to show that to achieve a probability of success of $1-\delta$ it suffices to run A $O(\log(1/\delta))$ times. Observe that if the distribution from which A is sampling is Gaussian, then by Theorem 2 the median trick is no better than simply outputting the mean of the samples.

Typically, an 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. It is a consequence of work of Canetti, Even and Goldreich [1], and also implied by the present paper, that generally there is not much more that one can do: Specifically, if all we know about the random variable X_I is its standard deviation $\sigma(X_I)$, then a necessary and sufficient condition for the existence of a FPRAS for $E(X_I)$ is that $\sigma(X_I)/E(X_I) \leq p(|I|)$ for some polynomial p. What if a good bound on $\sigma(X_I)/E(X_I)$ cannot be established? In an interesting result, Dagum et. al. [4] give an algorithm such that the expected number of times it samples X_I in order to estimate $E(X_I)$ (to within a multiplicative factor of $(1+\epsilon)$ with probability $\geq 1-\delta$) is within a constant factor of the optimal, for every random variable X_I that is distributed in $[0,1]$, even though nothing else is known about this random variable.

Let X be the random variable that on a random, uniformly chosen, matching in G is the size of the matching. Clearly, $E(X) = \mu(G)$, and therefore, estimating $\mu(G)$ amounts to estimating the mean of X. Jerrum and Sinclair [7] use the Markov chain Monte Carlo method to give an almost uniform generator for matchings in a graph, thereby showing that a random variable having probability distribution arbitrarily close to that of X is polynomial time samplable. As shown in Theorem 9 this yields an FPRAS for estimating $\mu(G)$.

Under some mild conditions, random variable X has essentially a Gaussian distribution. This follows from an exceptionally strong result of Godsil describing the size distribution of the matchings of a graph [8,6]. Let $G_1, \ldots$ be a family of graphs. Let $\phi_k(G_n)$ be the number of matchings with k edges in G_n and let $\phi(G_n) = \sum_k \phi_k(G_n)$ be the total number of matchings of G_n. Let $\mu(G_n)$ be the average size of a matching of G_n, $\mu(G_n) = (\sum_k k\phi_k(G_n))/\phi(G_n)$. Let $\sigma^2(G_n)$ be the variance of the distribution of sizes of matchings of G_n, $\sigma^2(G_n) = (\sum_k (k-\mu(G_n))^2 \phi_k(G_n))/\phi(G_n)$. Suppose that $\sigma(G_n) \to \infty$. Then:

Theorem 8 (Godsil). *The distribution of matching sizes is asymptotically locally normal, meaning that if we fix any real x and let $n \to \infty$, then*

$$\frac{\phi_k(G_n)}{\phi(G_n)}\sigma(G_n) \to (2\pi)^{-1/2}e^{-x^2/2}$$

for k such that $k - \mu(G_n) \sim x\sigma(G_n)$, where x is fixed.

Theorem 9. *There exists an FPRAS for estimating $\mu(G)$.*

Jerrum and Sinclair [7] have given an almost uniform generator for matchings (of all sizes) in a graph, where an *almost uniform generator* is a randomized polynomial time algorithm $\mathcal{A}$ that for any $\delta > 0$ and graph G outputs a matching satisfying: for each matching M in G,

$$P[\mathcal{A} \text{ outputs } M] \in \left[(1-\delta)\frac{1}{\phi(G)}, (1+\delta)\frac{1}{\phi(G)}\right].$$

Furthermore, the running time of $\mathcal{A}$ is polynomial in n and $\log(1/\delta)$.

Let Y be the random variable that is the size of the matching generated by this generator. Since the error parameter δ can be made inverse exponential in polynomial time,

$$|E(Y) - \mu(G)| \leq \varepsilon\mu(G).$$

Now it suffices to sample Y polynomially many times, in n and $1/\varepsilon$, and output the mean value in order to obtain an FPRAS for $\mu(G)$.

4. The mean is a majorizing estimator for the family $\{G_\theta\}$

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 2. We begin with a fact that substantially simplifies the matter.

Lemma 10. *It suffices to consider the single-sample case.*

Proof. This is because: (a) The mean of several iid Gaussian random variables is also a Gaussian random variable. (b) The mean is a *sufficient statistic* for samples drawn from the family $\{G_\theta\}$. This means that for every θ, the samples $x_1,\ldots,x_n$ drawn from G_θ are independent of θ given $\bar{x}=\frac{1}{n}\sum x_i$, or in other words that there is a distribution $P((x_1,\ldots,x_n)|\bar{x})$ such that

$$P((x_1,\ldots,x_n)|\theta) = P((x_1,\ldots,x_n)|\bar{x})P(\bar{x}|\theta).$$

Consequently the performance of any estimator will be unchanged if, given $x_1,\ldots,x_n$, 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 $x_1',\ldots,x_n'$. The distribution of the lists produced this way is the same as that of the lists $x_1,\ldots,x_n$, whence the conclusion 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 $\bar{x}$. ∎

Now 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 a sample x is picked from the distribution G_θ. (We will call this the finite-α experiment.)

Let $S:\mathbb{R}\to\mathbb{R}$ be an estimator of θ. In general, S may be randomized; $P(S(x)=y)$ denotes the probability (density) with which the estimator S 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

$$\int\int_{S(x)-\varepsilon}^{S(x)+\varepsilon} P(\theta)G_\theta(x)d\theta dx$$

if S is deterministic, and by

$$\int\int_{-\infty}^{\infty} P(S(x)=y)\int_{y-\varepsilon}^{y+\varepsilon} P(\theta)G_\theta(x)d\theta dy dx$$

if S is randomized.

In the single-sample case, the mean estimator is simply the identity estimator $T(x)=x$.

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

Lemma 11. *For $x \in I'_\alpha$,*

$$\int_{-\infty}^{\infty} P(S(x) = y) \int_{y-\varepsilon}^{y+\varepsilon} P(\theta) G_\theta(x) d\theta dy$$

is uniquely maximized for the identity estimator, i.e., $S = T$.

Proof.

$$\int_{y-\varepsilon}^{y+\varepsilon} P(\theta) G_\theta(x) d\theta$$

is uniquely maximized at $y = x$. 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 identity estimator commutes with translation, we find:

Observation 12. $Q_T^\varepsilon = P_T^{\alpha,\varepsilon}$.

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, $x \notin I'_\alpha$. By Lemma 11, we get:

Corollary 13. $P_T^{\alpha,\varepsilon} \geq M(\alpha, \varepsilon) - P(B(\alpha, \varepsilon))$. ∎

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

By Observation 12 and Corollary 13,

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

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

$$Q_T^\varepsilon \geq Q(\varepsilon) - \limsup_{\alpha} P(B(\alpha, \varepsilon)).$$

Now,

$$\begin{aligned}\limsup_{\alpha} P(B(\alpha, \varepsilon)) &\leq \limsup_{\alpha} [P(|\theta| > \alpha - \alpha^{1/2}) + P(x \notin I'_\alpha \mid |\theta| \leq \alpha - \alpha^{1/2})] \\ &\leq 0 + \limsup_{\alpha} P(|x - \theta| > \alpha^{1/2} - \varepsilon).\end{aligned}$$

Since x is normally distributed with variance $1/n$, this is bounded above by

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

Hence $Q_T^\varepsilon \geq Q(\varepsilon)$. ∎

5. Optimality of the mean estimator for $\{G_\theta\}$ with respect to general penalty functions

Proof of theorem 3. The proof of theorem 3 is similar to that of theorem 2. Instead of providing an upper bound on a "benefit function" (e.g., 1 if $|T(x)-\theta|<\varepsilon$ and 0 otherwise), we provide a lower bound on a penalty function (e.g., $|T(x)-\theta|^r$). Again as in theorem 2, because the mean $\bar{x}=\frac{1}{n}\sum x_i$ is a sufficient statistic for the parameter, we may assume that an estimator depends only on the mean; and because the mean has a Gaussian distribution, it suffices to show that the identity estimator performs at least as well (in supremum over all θ of expected penalty) as any other estimator.

Recall that we have defined $L=\int_{-\infty}^{\infty}\psi(x)G(x)dx$ and that $L<\infty$. Note that for any θ, the expected penalty of the identity estimator is L.

Now fix any $\varepsilon>0$. Select α sufficiently large that the following condition is satisfied:

$$\int_{\varepsilon\alpha}^{\infty} G(\theta)\psi(\theta)d\theta < L\varepsilon/2.$$

Let S be an arbitrary estimator. Pick θ uniformly in the interval $[-\alpha,\alpha]$. The expected penalty of S is

$$\frac{1}{2\alpha}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\alpha}^{\alpha} G(x-\theta)\psi(s-\theta)P(s=S(x))d\theta ds dx$$

Now we ignore x's outside the range $[-\alpha,\alpha]$, and treat negative and positive x's separately. For each x we consider the penalty due only to a limited range in θ.

$$(3)\qquad \ldots \geq \frac{1}{2\alpha}\int_{-\infty}^{\infty}\left[\int_{-\alpha}^{0} P(s=S(x))\int_{-\alpha}^{2x+\alpha} G(x-\theta)\psi(s-\theta)d\theta dx + \int_{0}^{\alpha} P(s=S(x))\int_{2x-\alpha}^{\alpha} G(x-\theta)\psi(s-\theta)d\theta dx\right] ds$$

Because of the unimodality and symmetry of ψ, there is a measure μ on the nonnegative reals such that $\psi(z)=\int_0^{|z|}\mu(y)dy$. (Since G is a density function, the contribution to (3) at discontinuities of ψ is 0.) Rewriting the internal expression in the first of the terms in (3) gives, for any x and s:

$$(4)\qquad \int_{-\alpha}^{2x+\alpha} G(x-\theta)\psi(s-\theta)d\theta = \int_0^{\infty}\mu(y)\left[\int_{-\alpha}^{2x+\alpha} G(x-\theta)d\theta - \int_{[s-y,s+y]\cap[-\alpha,2x+\alpha]} G(x-\theta)d\theta\right] dy$$

For any x the second internal integral in (4) is maximized by $s = x$, so we have

$$\begin{aligned}\ldots &\geq \int_0^\infty \mu(y) \left[\int_{-\alpha}^{2x+\alpha} G(x-\theta)d\theta - \int_{\max\{x-y,-\alpha\}}^{\min\{x+y,2x+\alpha\}} G(x-\theta)d\theta \right] dy \\ &= \int_{-\alpha}^{2x+\alpha} \psi(x-\theta)G(x-\theta)d\theta\end{aligned}$$

A similar sequence of steps applies to the second term in (3), bounding its internal expression by

$$\begin{aligned}\ldots &\geq \int_0^\infty \mu(y) \left[\int_{2x-\alpha}^{\alpha} G(x-\theta)d\theta - \int_{\max\{x-y,2x-\alpha\}}^{\min\{x+y,\alpha\}} G(x-\theta)d\theta \right] dy \\ &= \int_{2x-\alpha}^{\alpha} \psi(x-\theta)G(x-\theta)d\theta\end{aligned}$$

Combining these inequalities, we can bound (3) by

$$\begin{aligned}\geq \frac{1}{2\alpha} \Bigg[&\int_{-\alpha}^{0} \int_{-\alpha}^{2x+\alpha} \psi(x-\theta)G(x-\theta)d\theta dx \\ &+ \int_{0}^{\alpha} \int_{2x-\alpha}^{\alpha} \psi(x-\theta)G(x-\theta)d\theta dx \Bigg]\end{aligned}$$

Because of the symmetries of ψ and G these expressions are equal, so

$$= \frac{1}{\alpha} \int_0^\alpha \int_{2x-\alpha}^{\alpha} \psi(x-\theta)G(x-\theta)d\theta dx$$

Now make the change of variables $z = \alpha - x$.

$$\begin{aligned}= \frac{1}{\alpha} \int_0^\alpha \int_{-z}^{z} G(y)\psi(y)dydz &\geq \frac{1}{\alpha} \int_{\varepsilon\alpha}^{\alpha} \int_{-\varepsilon\alpha}^{\varepsilon\alpha} G(y)\psi(y)dydz \\ &= (1-\varepsilon) \int_{-\varepsilon\alpha}^{\varepsilon\alpha} G(y)\psi(y)dy\end{aligned}$$

Now apply the assumption on α:

$$\geq (1-\varepsilon)^2 L.$$

The supremum penalty of S over θ's in the interval $[-\alpha, \alpha]$ is therefore at least $(1-\varepsilon)^2 L$; since ε was arbitrary, this means that the supremum penalty of S over $\theta \in \mathbb{R}$ is at least L, and therefore at least as great as the supremum penalty of the identity estimator. ∎

6. Optimality of the mean estimator for $\{G_\theta^d\}$ with respect to general penalty functions

Proof of theorem 4. The mean $\frac{1}{n}\sum x_i$ is, just as in one dimension, a sufficient statistic for the parameter θ, so we may assume that an estimator is a function only of the mean; and, because the mean has a spherically symmetric Gaussian distribution, it suffices to show that the identity estimator performs at least as well (in supremum over all θ of expected penalty) as any other estimator.

For $x \in \mathbb{R}^d$ and $r \in \mathbb{R}$ let $b(x,r)$ be the open ball of radius r about x; if $r \le 0$, $b(x,r)$ is empty. The boundary of $b(x,r)$ is denoted $\partial b(x,r)$. Recall that we have defined $L = \int_{\mathbb{R}^d} \psi(x)G(x)dx$; that $L < \infty$; and that for any θ, the expected penalty of the identity estimator is L. Let Vol_k denote k-dimensional Lebesgue measure.

Now fix any $\varepsilon > 0$. Select γ large enough that $\int_{b(0,\gamma)} G(\theta)\psi(\theta)d\theta > (1-\varepsilon)L$, and α large enough that $\mathrm{Vol}_d b(0,\alpha-\gamma)/\mathrm{Vol}_d b(0,\alpha) > 1-\varepsilon$.

Let S be an arbitrary estimator. Pick θ uniformly in the ball $b(0,\alpha)$. The expected penalty of S is

$$\frac{1}{\mathrm{Vol}_d b(0,\alpha)} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{b(0,\alpha)} G(x-\theta)\psi(s-\theta)P(s=S(x))\; d\theta\; dx\; ds$$

Now we ignore x's outside the ball $b(0,\alpha)$, and for each x consider only the penalty due to a limited range of θ's.

$$(5)\ \ldots \ge \frac{1}{\mathrm{Vol}_d b(0,\alpha)} \int_{\mathbb{R}^d}\int_{b(0,\alpha)} P(s{=}S(x)) \int_{b(x,\alpha-|x|)} G(x{-}\theta)\psi(s{-}\theta)\, d\theta\, dx\, ds$$

Because ψ is spherically symmetric and increasing away from the origin, there is a measure μ on the nonnegative reals such that $\psi(z) = \int_0^{|z|} \mu(y)dy$. Rewrite the innermost integral in (5) in terms of μ:

$$\begin{aligned}&\int_{b(x,\alpha-|x|)} G(x-\theta)\psi(s-\theta)\; d\theta \\ &\quad = \int_0^\infty \mu(y)\left[\int_{b(x,\alpha-|x|)} G(x-\theta)\; d\theta - \int_{b(x,\alpha-|x|)\cap b(s,y)} G(x-\theta)\; d\theta\right] dy\end{aligned}$$

For any x and y, $\int_{b(x,\alpha-|x|)\cap b(s,y)} G(x-\theta)\,d\theta$ is maximized at $s = x$. So

$$\begin{aligned}\int_{b(x,\alpha-|x|)} G(x-\theta)\psi(s-\theta)\; d\theta &\ge \int_0^\infty \mu(y) \int_{b(x,\alpha-|x|)-b(x,y)} G(x-\theta)\; d\theta\; dy \\ &= \int_{b(x,\alpha-|x|)} G(x-\theta)\psi(x-\theta)\; d\theta\,.\end{aligned}$$

Returning to (5) we see that the expected penalty of S is at least

$$\frac{1}{\mathrm{Vol}_d b(0,\alpha)} \int_{b(0,\alpha)} \int_{b(x,\alpha-|x|)} G(x-\theta)\psi(x-\theta)\, d\theta\, dx\,.$$

Setting $z = \alpha - |x|$, this is

$$\begin{aligned}
&= \frac{1}{\mathrm{Vol}_d b(0,\alpha)} \int_0^\alpha \mathrm{Vol}_{d-1}\partial b(0,\alpha-z) \int_{b(0,z)} G(\theta)\psi(\theta)\, d\theta\, dz \\
&\geq \frac{1}{\mathrm{Vol}_d b(0,\alpha)} \int_\gamma^\alpha \mathrm{Vol}_{d-1}\partial b(0,\alpha-z) \int_{b(0,\gamma)} G(\theta)\psi(\theta)\, d\theta\, dz \\
&= \frac{\mathrm{Vol}_d b(0,\alpha-\gamma)}{\mathrm{Vol}_d b(0,\alpha)} \int_{b(0,\gamma)} G(\theta)\psi(\theta)\, d\theta\, dz \\
&> (1-\varepsilon)^2 L\,.
\end{aligned}$$

The supremum penalty of S over θ's in the ball $b(0,\alpha)$ is therefore at least $(1-\varepsilon)^2 L$; since ε was arbitrary, this means that the supremum penalty of S over $\theta \in \mathbb{R}^d$ is at least L, and therefore at least as great as the supremum penalty of the identity estimator. ∎

7. Uniqueness of the mean as a majoring estimator for $\{G_\theta\}$

We now strengthen Theorem 2 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 $2,\ldots,n+1$ are the samples $x_1,\ldots,x_n$, the first argument is the estimate for θ), that satisfies the following condition: for all measurable sets $U \subseteq \mathbb{R}^n$, $S(\mathbb{R}\times U) = \mu(U)$.

Theorem 14. *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. Again as in theorem 2, it suffices to consider the single-sample case, with T the identity estimator.

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 2, 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

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

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

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

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

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

Lemma 15. *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 $S(\mathbb{R}\times B)=\mu(B)$. 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 $\Delta_S(B)\ge S(K_j)[(\mathcal{N}(\varepsilon)-\mathcal{N}(-\varepsilon))-(\mathcal{N}(\varepsilon+2^j)-\mathcal{N}(-\varepsilon+2^j))]>0$. $\blacksquare$

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 16. *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,

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

Lemma 17. *If $\Delta_S(\mathbb{R})>0$ then there is a set D of finite measure such that $F_S(D;\mathbb{R})<F_T(D;\mathbb{R})$.*

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 16, $F_S(B';B) \leq F_T(B';B) - \Delta_S(B) < F_T(B';B) - \varepsilon$. 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, $F_S(B';\mathbb{R}-B) \leq \varepsilon$. Therefore,

$$F_S(B';\mathbb{R}) \leq F_S(B';B) + \varepsilon < F_T(B';B) < F_T(B';\mathbb{R}).$$

Case (ii): $\Delta_S(\mathbb{R})$ is finite.
Let B be a finite interval such that $g(\Delta_S(\mathbb{R}-B)) \leq \Delta_S(B)/2$, 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 16,

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

we get

$$\begin{aligned} F_S(B';\mathbb{R}) &= F_S(B';B) + F_S(B';\mathbb{R}-B) \\ &\leq |B|(2\mathcal{N}(\varepsilon) - 1) - \Delta_S(B) + F_S(B';\mathbb{R}-B). \end{aligned}$$

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 $F_S(B';\mathbb{R}) \leq |B'|(2\mathcal{N}(\varepsilon)-1) - \Delta_S(B) < |B'|(2\mathcal{N}(\varepsilon)-1) = F_T(B';\mathbb{R})$. However, $\Delta_S(\mathbb{R}-B)$ may be nonzero. This allows estimates to be shifted so as to increase $F_S(B';\mathbb{R})$. The remainder of the argument is devoted to showing that this increase, which we call $DF_S(B';\mathbb{R})$, 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, $DF_S(B';\mathbb{R})$ is $\mathcal{N}(\varepsilon+r) - \mathcal{N}(\varepsilon)$ for $y \leq 2\varepsilon$, provided $0 \leq r \leq y$ (greater values of r contribute less to $F_S(B';\mathbb{R})$); while for $y \geq 2\varepsilon$ $DF_S(B';\mathbb{R})$ is 0 for $0 \leq r \leq y - 2\varepsilon$, and $\mathcal{N}(\varepsilon+r) - \mathcal{N}(y-\varepsilon)$ for $y - 2\varepsilon \leq r \leq y$ (again, greater values of r contribute less to $F_S(B';\mathbb{R})$).

First, we claim that the best gain in $F_S(B';\mathbb{R})$ (greatest value of $DF_S(B';\mathbb{R})$) 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 $y \leq 2\varepsilon$, 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"

$$\frac{\mathcal{N}(\varepsilon + r) - \mathcal{N}(\varepsilon)}{-\mathcal{N}(\varepsilon + r) + \mathcal{N}(\varepsilon) + \mathcal{N}(-\varepsilon + r) - \mathcal{N}(-\varepsilon)} \tag{6}$$

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

$$\frac{\mathcal{N}(\varepsilon + r) - \mathcal{N}(y - \varepsilon)}{-\mathcal{N}(\varepsilon + r) + \mathcal{N}(\varepsilon) + \mathcal{N}(-\varepsilon + r) - \mathcal{N}(-\varepsilon)} \tag{7}$$

it follows that in an optimal estimator the shift used at that range can be no greater. We therefore obtain an upper bound on $DF_S(B';\mathbb{R})$ in the following way: considering only $y \leq 2\varepsilon$, find the shift r_0 such that $DF_S(B';\mathbb{R})$ 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 $DF_S(B';\mathbb{R})$ 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 $DF_S(B';\mathbb{R})$ 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 $DF_S(B';\mathbb{R}) \leq \Delta_S(B)/2$ and therefore (by comparing with the estimator which is equal to the mean outside B), we find that $F_S(B';\mathbb{R}) \leq |B'|(2\mathcal{N}(\varepsilon) - 1) - \Delta_S(B) + DF_S(B';\mathbb{R}) \leq |B'|(2\mathcal{N}(\varepsilon) - 1) - \Delta_S(B)/2 < |B'|(2\mathcal{N}(\varepsilon) - 1) = F_T(B';\mathbb{R})$. ∎

8. Discussion

The main open issue suggested by our work is whether the concept of a majorizing estimator, as well as the techniques we use for the Gaussian family, can be useful in establishing optimality of estimators for other families of distributions.

Regarding the Gaussian distribution we conjecture that the mean estimator T is the unique penalty-minimizing estimator for any nonzero penalty function ψ, i.e., for all nonzero penalty functions ψ, if S is an estimator and there is a measurable set A such that $S(A) \neq T(A)$, then $M^{\psi}(T) < M^{\psi}(S)$.

Another interesting question has to do with the fact that the Cramer–Rao lower bound (on the variance of any unbiased estimator) varies at the parameter values θ, depending on the sensitivity of the parametric family to change about θ. In the same spirit one may ask (for a general parametric family) for a lower bound $p(\theta, \varepsilon)$ on the probability that an estimator of θ falls outside of the interval $(\theta - \varepsilon, \theta + \varepsilon)$. The bound should have the property that $\lim_{\varepsilon \to 0} p(\theta, \varepsilon) = 1$. Some assumption must be made to keep the estimator "honest" (to rule out a constant function for example), such as unbiasedness, or an assumption about the estimator achieving some minimal in-probability performance for some interval length at all θ.

9. Acknowledgments

We wish to thank Prof. D. Blackwell and Prof. C. R. Rao for helping us confirm the status of Theorem 2.

References

[1] R. Canetti, G. Even and O. Goldreich: Lower bounds for sampling algorithms for estimating the average, *Information Processing Letters* **53** (1995), 17–25.

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

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

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

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

[6] C. D. Godsil: Matching behaviour is asymptotically normal, *Combinatorica* **1** (1981), 369–376.

[7] M. Jerrum and A. Sinclair: The Markov chain Monte Carlo method: an approach to approximate counting and integration, in D. Hochbaum, editor, *Approximation Algorithms for NP-hard problems*. PWS Publishing Co., 1995.

[8] L. Lovász and M. D. Plummer: *Matching Theory*, Akadémiai Kiadó, 1986.
[9] C. R. Rao: Information and accuracy attainable in estimation of statistical parameters, *Bull. Cal. Math. Soc.* **37** (1945), 81–91.
[10] L. J. Schulman and V. V. Vazirani: Majorizing estimators and the approximation of #P-complete problems, in *Proceedings of the 31'st STOC*, 1999.
[11] L. G. Valiant: The complexity of computing the permanent, *Theoretical Computer Science* **8** (1979), 189–201.
[12] L. G. Valiant: The complexity of enumeration and reliability problems, *SIAM Journal of Computing* **8(3)** (1979), 410–421.
[13] S. Zacks: *Parametric Statistical Inference*, Pergamon Press, 1981.

Leonard J. Schulman

Department of Computer Science
Caltech
Pasadena CA 91125
USA
schulman@caltech.edu

Vijay V. Vazirani

College of Computing
Georgia Institute of Technology
Atlanta GA 30332-0280
USA
vazirani@cc.gatech.edu