LA-UR- 08-5083

Approved for public release; distribution is unlimited.

Title: Strong profiling is not mathematically optimal for discovering rare malfeasors

Author(s): William H. Press

Intended for: Nature

Strong profiling is not mathematically optimal for discovering rare malfeasors

William H. Press[1]

[1]*Los Alamos National Laboratory, Los Alamos, NM 87545, USA and Department of Computer Sciences, The University of Texas at Austin, Austin, TX 78703, USA*

In a large population of individuals labeled $j = 1, 2, \ldots, N$, governments attempt to find the rare malfeasor $j = j_\star$ (terrorist, for example[1–4]) by making use of priors p_j that estimate the probability of individual j being a malfeasor. Societal resources for secondary random screening such as airport search or police investigation are concentrated against individuals with the largest priors.[5] We may call this "strong profiling" if the concentration is at least proportional to p_j for the largest values. Strong profiling often results in higher probability, but otherwise innocent, individuals being repeatedly subjected to screening. We show here that, entirely apart from considerations of social policy, strong profiling is not mathematically optimal at finding malfeasors. Even if prior probabilities were accurate, their optimal use would be only as roughly the geometric mean between a strong profiling and a completely uniform sampling of the population.

Racial profiling, as commonly defined[6], occurs when an individual is randomly selected for secondary security screening on the basis of his or her race, ethnicity, nationality, or religion. Secondary screening may take the form of airport luggage search, police investigation, physical search, or other societally sanctioned but personally intrusive actions. What distinguishes racial profiling and related so-called actuarial methods[5] from police investigational methods often perceived as more acceptable is its use of prior probabilities associated with the individual, and not associated with evidence of actual criminal conduct.

For simplicity, assume that there is a single malfeasor $j = j_*$. An omnipotent authoritarian government might enumerate all of the p_j's, $j = 1, \ldots, N$, sort them from largest to smallest value, and then screen individuals in the population, visiting each just once in the order of their probability. This strategy (hereafter, "A") can easily be shown to find the malfeasor with the smallest average number of tests. That number is

$$\mu_A \equiv \sum_{i=1}^{N} i p_{(i)} \qquad (1)$$

where $p_{(i)}$ is the order statistic; that is, $p_{(i)}$ is the ith largest value among the p_j's.

For moral or practical reasons, democratic governments (hereafter, "D") employ strategies of sampling with replacement. That is, individuals are sampled with some sampling probability q_j determined in principle by a public policy. The sampling process is memoryless in that an individual may be sampled more than once, for example, whenever he goes through an airport security checkpoint. In this case, the mean number of tests required to find the malfeasor is evidently $1/q_{j_*}$. The expectation of this over p_j, which we want to minimize subject to $\sum q_i = 1$, is

$$\mu_D = \sum_{j=1}^{N} p_j/q_j \qquad (2)$$

A straightforward minimization with a Lagrange multiplier gives the optimal choice for the q_j's

$$q_j = p_j^{1/2} \Big/ \sum_{i=1}^{N} p_i^{1/2} \qquad (3)$$

and the mean number of tests per found malfeasor,

$$\mu_D = \left(\sum_{j=1}^{N} p_j^{1/2} \right)^2 \tag{4}$$

In words, equation (3) says that individuals should be selected for screening in proportion to the *square root* of their prior probability. This does use the priors, but only weakly: It results in secondary screening being distributed over a much larger segment of the population than would be the case with strong profiling. Although equation (3) should be a well-known result, we are not aware of any published reference earlier than Abagyan and collaborators in a completely different context.[7,8].

It is instructive to compare the optimal result to what one might have guessed to be the obvious answer, namely $q_j = p_j$: screen in proportion to the prior p_j, a strong profiling usually termed "importance sampling"[9]. Substituting into equation (2) yields $\mu = N$. On average, importance sampling tests the full population size before finding the malfeasor. Indeed, this version of strong profiling does no better than random sampling without reference to the priors, which also yields $\mu = N$. The reason that this strong profiling strategy is inefficient is that, on average, it keeps re-testing the same innocent individuals who happen to have large p_j's. The optimal strategy is optimal precisely because it avoids this oversampling.

A figure of merit for the optimal "D" sampling is F.M. $= \mu_D/\mu_A$, the factor by which it is less efficient than the perfect authoritarian strategy "A", above. We can compute F.M. for various assumptions about the distribution of p_j's. Smaller F.M. is better. If the prior probability is concentrated uniformly in some number of individuals N_0 (out of N), then F.M. ≈ 2, independent of N_0. That is, D is only a factor of 2 less efficient than authoritarian A.

Another interesting case is the "scale-free" distribution $p_j \propto 1/j^\alpha$. For $\alpha < 2$, this yields F.M. $\approx 4/(2-\alpha)$. For $\alpha > 2$, it is F.M. $\approx \zeta(\alpha/2)^2/\zeta(\alpha-1)$, where ζ is the Riemann Zeta function. This is $\approx 4/(\alpha-2)$ for α of order unity, and ≈ 1 for large α. In all these cases, for any fixed value α not near 2, we are within a constant factor of strategy A. Only the case $\alpha \approx 2$ gives the unbounded result F.M. $\approx \log N$, which is itself large only logarithmically.

A final case of interest is $p_j \propto \exp(-j^{1/\beta})$. This occurs in cases where the j's are ordered by radius from the origin in a (say) high-dimensional space, and the probability decreases away from the origin either exponentially or as a multivariate normal distribution. It also applies to a mixture of such distributions, and thus to Gaussian mixture models generally. In all such cases β is related to (and increases with) the dimension of the space. One readily calculates,

$$\text{F.M.} = 2^{2\beta+1}\Gamma(1+\beta)^2/\Gamma(1+2\beta) \tag{5}$$

with the limiting cases ≈ 2 as $\beta \rightarrow 0$ and $\approx 2(\pi\beta)^{1/2}$ as β becomes large, a surprisingly modest increase for what might have been thought to be a dimensional explosion of volume.

The idea of sampling by square-root probabilities is quite general and can have many other applications. It applies whenever a "bell-ringer" event must be found by sampling with replacement, but can be recognized when seen. For example, one can thus sample paths through a trellis or hidden Markov model when their number is too large to enumerate explicitly, but one path can be recognized (e.g., by secondary testing) as the desired bell-ringer.

A generalization of the scenario already discussed is the case where the bell-ringer can be recognized, when sampled, only with some probability s_i. In that case (see Supplementary Methods) the optimal sampling q_i (that minimizes the mean number of samples needed to find the bell-ringer) is proportional to $(p_i/s_i)^{1/2}$. The optimal sampling thus expends relatively more samples on the less-likely-to-recognize cases. It is the direct opposite of the proverbial "looking under the lamppost"! This may seem counter-intuitive, but it is correct under the assumptions stated. In rough terms, if you *don't* spend a lot of time "not under the lamppost," then you provide excessive sanctuary for the malfeasor who might be there.

1. Siggins, P. Racial Profiling in an Age of Terrorism, talk at Markkula Center for Applied Ethics, 20 March, 2002; transcript at http://www.scu.edu/ethics/publications/ethicalperspectives/profiling.html

2. Ellmann, S.J. Racial Profiling and Terrorism. New York Law School Law Review **46**, 675–730 (2003).

3. Lund, N. The Conservative Case Against Racial Profiling in the War on Terrorism. Albany Law Review **66**, 329–342 (2003).

4. London, H. Profiling as Needed. Albany Law Review **66**, 343–347 (2003).

5. Harcourt, B.E. Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age. (University of Chicago Press, Chicago, 2007).

6. American Civil Liberties Union, Campaign Against Racial Profiling, at http://www.aclu.org (2008).

7. Abagyan, R.A., and Totrov, M. Ab Initio Folding of Peptides by the Optimal-Bias Monte Carlo Minimization Procedure. J. Computational Physics **151**, 402–421 (1999).

8. Zhou, Y., and Abagyan, R. Efficient Stochastic Global Optimiztion for Protein Structure Prediction. in Rigidity Theory and Applications (M.F. Thorpe and P.M. Duxbury, eds.) (Springer, New York, 2002).

9. Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. Numerical Recipes: The Art of Scientific Computing. Third Edition (Cambridge University Press, New York, 2007), §7.9.1.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The author thanks Martha Minow, Nozer Singpurwalla, Sallie Keller-McNulty, and Richard Garwin for helpful comments and discussion.

Competing Interests The author declares no competing financial interests.

Correspondence Correspondence should be addressed to the author (email: wpress@cs.utexas.edu).

Supplementary Methods for "Strong profiling is not mathematically optimal for discovering rare malfeasors"

William H. Press

1 Mathematical Statement of Problem

There are a large number N of candidate objects numbered $i = 1, \ldots, N$, one of which (an unknown value $i = i_\star$) is the desired "bell-ringer" that can be recognized when examined closely. The difficulty is that N is large, while close looks, sufficient to recognize that $i = i_\star$, are expensive. The problem is how to allocate our expensive looks among all the i's.

If we know nothing else about the object i's, then we must simply go through all the objects in an arbitrary order until we find $i_\star$. On average, we will find it after $N/2$ looks. But what if we have prior information that distinguishes among the objects? Suppose we can estimate a probability that each i is the bell-ringer, $p_i \equiv \text{Prob}(i = i_\star)$. Then the optimal strategy is to order the p_i's from largest to smallest, and then look at objects in that order. The average number of looks required is

$$\mu_s = \sum_{i=1}^{N} i p_{(i)} \tag{1}$$

where $p_{(i)}$ denotes the ith order statistic of the p_i's, that is, the ith largest value. The subscript s on μ_s stands for "sorting". There is no way to do better than this.

However, we are interested in cases when the sorting strategy is, for one reason or another, not feasible, and we must instead randomly sample *with replacement* among the i's. This might occur for reasons of policy (as in the main text) or because N is so large that it is not feasible to enumerate and sort all the p_i's. For example, the latter situation might occur if the "objects" are complex hypotheses, each of which involves different choices of sub-hypotheses. In that case N can be combinatorially large, but still easy to sample from.

2 Biased Sampling by Square Root of Prior

2.1 Derivation

Suppose we sample the i's with probabilities q_i, with

$$\sum_i q_i = 1 \tag{2}$$

We control the q_i's and want to optimize their choice. The probability of missing the bell-ringer exactly on exactly $m \geq 0$ looks, and finding it on the $m + 1$st is $(1 - q_\star)^m q_\star$, where $q_\star \equiv q_{i_\star}$. So the mean number of looks required is

$$\sum_{m=0}^{\infty} (m+1)(1-q_\star)^m q_\star = \frac{1}{q_\star} \tag{3}$$

(an answer that we could have written down by inspection in any case).

The expectation of equation (3) over the p_i's, which we want to minimize, subject to the normalization constraint (2) is thus

$$\mu = \sum_i \frac{p_i}{q_i} \tag{4}$$

Using a Lagrange multiplier λ, we minimize with respect to q_j

$$\mathcal{L} = \sum_i \frac{p_i}{q_i} + \lambda \left(\sum_i q_i - 1 \right) \tag{5}$$

This easily gives

$$q_j \propto p_j^{1/2} = p_j^{1/2} \Big/ \sum_i p_i^{1/2} \tag{6}$$

Equation (6) is a main result underlying the main text. It says that, under the conditions posited, one should sample objects in proportion to the square root of the prior probabilities of their being the bell-ringer. While this should be (and perhaps is in some circles) a well-known result, we are not aware of any reference earlier than 1999 (see main text).

2.2 Performance Comparison to Naive Sampling Strategies

Substituting equation (6) into equation (4) gives, for square-root sampling, the average number of looks

$$\mu_{sr} = \left(\sum_i p_i^{1/2} \right)^2 \tag{7}$$

where subscript *sr* means "square root". It is informative to compare this to the corresponding results for two naive sampling strategies. First, uniform sampling with replacement (ignoring the p_i's):

$$q_i = \frac{1}{N}, \qquad \mu_u = \sum_i \frac{p_i}{1/N} = N \tag{8}$$

Second, sampling in proportion to p_i (what would be called *importance sampling* in the context of Monte Carlo integration): This seems like a natural way of sampling likely objects more heavily. However, it gives

$$q_i = p_i, \qquad \mu_{is} = \sum_i \frac{p_i}{p_i} = N \tag{9}$$

exactly the same as uniform sampling, equation (8)! The reason that importance sampling is far from optimal in for this problem is that it repeatedly revisits the same high probability objects, even after they have been seen to be *not* the bell-ringer.

To get an idea of how much better than (8) or (9) is the optimal result (7), consider first a case where all of the probability is concentrated uniformly in M of the p_i's, with the remaining $N - M$ of the p_i's being zero. Then, one easily gets from (7),

$$\mu_{sr} = M \tag{10}$$

If $M \ll N$, this is $\ll \mu_u$ or μ_{is} equations(8) or (9), and is only on average a factor of 2 worse than the perfect sorting strategy, equation (1).

3 Performance Comparison for Some Other Distributions

3.1 Definition of Figure of Merit (F.M.)

Equations (7) and (1) suggest that we define as a figure of merit for the square-root sampling strategy the ratio of its mean number of looks μ_{rs} to that of the perfect sorting strategy μ_s, that is,

$$\text{F.M.} = \left(\sum_i p_i^{1/2} \right)^2 \Bigg/ \sum_{i=1}^{N} i p_{(i)} \tag{11}$$

Smaller values of F.M. are better. The example of equation (10) can be summarized as

$$\text{F.M.[concentrated uniform]} = 2 \tag{12}$$

As a computational shortcut, note that equation (11) is invariant under scaling the p_i's by a constant factor, so we can test distributions without requiring them to be normalized.

3.2 Power-Law Prior Distributions

Suppose (after sorting into monotonically decreasing order)

$$p_i \propto i^{-\alpha}, \qquad \alpha \geq 0 \tag{13}$$

Consider first the case $0 \leq \alpha < 2$. Approximating the sums by integrals, which is here accurate for large N, gives (using non-normalized p_i's, n.b.!)

$$\begin{aligned}\sum_i ip_{(i)} &\approx \int_1^{N+1} x^{1-\alpha}dx \approx \frac{1}{2-\alpha}(N+1)^{2-\alpha} \\ \left(\sum_i p_i^{1/2}\right)^2 &\approx \left(\int_1^{N+1} x^{-\alpha/2}dx\right)^2 \approx \left(\frac{2}{2-\alpha}\right)^2 (N+1)^{2-\alpha}\end{aligned} \tag{14}$$

which implies

$$\text{F.M.[power law} < 2] \approx \frac{4}{2-\alpha} \tag{15}$$

For exponent α bounded away from 2, this is a constant of order unity (that is, independent of N); so, sampling by the square root of the prior is not much less efficient than the perfect sorting strategy. In the case $\alpha = 0$, incidentally, we recover equation (12).

Next consider the case $\alpha > 2$. The sums are now dominated by small values of i, so we can approximate by extending the sums to infinity. If $\zeta()$ is the Riemann Zeta function, we have $p_{(i)} \approx i^{-\alpha}/\zeta(\alpha)$ and

$$\begin{aligned}\sum_i ip_{(i)} &\approx \frac{\zeta(\alpha-1)}{\zeta(\alpha)} \\ \left(\sum_i p_i^{1/2}\right)^2 &\approx \frac{\zeta(\alpha/2)^2}{\zeta(\alpha)}\end{aligned} \tag{16}$$

which implies

$$\text{F.M.[power law} > 2] \approx \frac{\zeta(\alpha/2)^2}{\zeta(\alpha-1)} \approx \begin{cases} \frac{4}{\alpha-2} & \alpha \to 2 \\ 1 & \alpha \to \infty \end{cases} \tag{17}$$

Figure 1 plots equation (17) for intermediate values of α. We again see that for exponent α bounded away from 2, the figure of merit is a constant of order unity, independent of N.

Finally, for power law distributions, we consider the case $\alpha = 2$. In this case $\sum_i ip_{(i)} \approx \ln N$, while $(\sum_i p_i^{1/2})^2 \approx (\ln N)^2$, so

$$\text{F.M.[power law} = 2] \approx \ln N \tag{18}$$

This is not bounded as $N \to \infty$, but it increases only logarithmically with N.

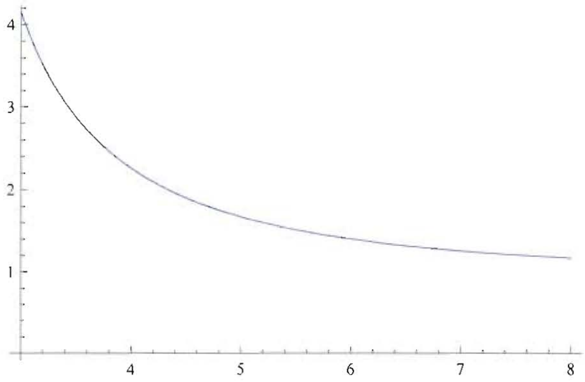

Figure 1: Figure of merit (ordinate) for power law distributions with (abscissa) $\alpha > 2$. The value 1 means "as efficient as the perfecting sorting strategy".

3.3 Exponential Prior Distributions

Suppose now that the monotonically sorted $p_{(i)}$'s have the form

$$p_{(i)} \propto \exp(-ci^{1/d}), \qquad c, d > 0 \tag{19}$$

This is a form that can occur in various contexts, notably where the probability decreases as an exponential or Gaussian away from the origin in a (say) high-dimensional space, and the i's are lattice points (or uniformly distributed) in that space. The form is also relevant to a finite mixture of such forms, i.e., to Gaussian mixture models with a finite number of components. In these cases d is related to the dimension of the space D (e.g., $d = D/2$ for Gaussians).

Approximating by integrals, we have

$$\begin{aligned} \sum_i i p_{(i)} &\approx \int_0^\infty x \exp(-cx^{1/d} dx c^{-2d} d\Gamma(2d) \\ \left(\sum_i p_i^{1/2}\right)^2 &\approx \left(\int_0^\infty \exp(-\tfrac{1}{2}cx^{1/d})dx\right)^2 = 2^{2d}c^{-2d}\Gamma(d+1)^2 \end{aligned} \tag{20}$$

which gives

$$\text{F.M.[exponential]} \approx \frac{2^{2d+1}\Gamma(1+d)^2}{\Gamma(1+2d)} \tag{21}$$

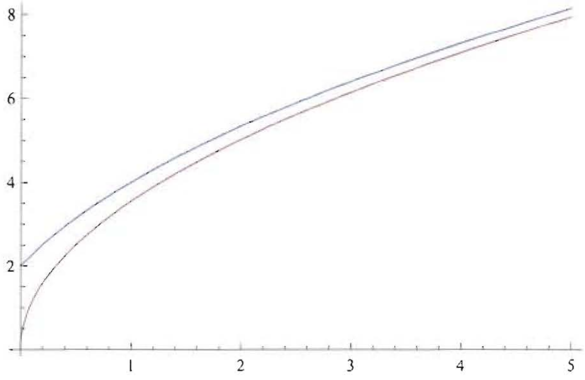

Figure 2: Upper curve: Figure of merit (ordinate) for the exponential distribution equation (19) as a function of the dimension parameter d (abscissa). Lower curve: Asymptotic form $2\sqrt{\pi d}$.

For $d \longrightarrow 0$ this becomes F.M. $\longrightarrow$ 2, recovering the case of equation (12). For large d, equation (21) is $\approx 2\sqrt{\pi d}$. Intermediate values (along with the asymptotic form) are plotted in Figure 2. The figure of merit is approximately independent of the parameter c in all cases.

4 Case of Probabilistic Recognition

Suppose now a slightly different setup: The bell-ringer object is still some $i = i_\star$, but we may not recognize it the first time we look at it. This could be because some additional random condition (not within our control) is required for detection. Suppose that, on each look, the probability that we recognize the bell-ringer is s_i, $i = 1, \ldots, N$; and that (for simplicity) each look is independently random.

4.1 Probabilistic Recognition Perfect Sorting Strategy

The perfect sorting strategy that led to equation (1) is now no longer valid, since it looks at each object only once. What is the optimal sorting strategy now?

If we have looked at object i already m_i times, then its probability of both being the bell-ringer and escaping previous detection is $(1-s_i)^{m_i}p_i$. So the total remaining probability in which the bell-ringer is hiding is

$$P = \sum_i (1-s_i)^{m_i} p_i \tag{22}$$

Now suppose we look next at object j. Then the change in equation (22) is

$$-\Delta P = (1-s_j)^{m_j} p_j - (1-s_j)^{m_j+1} p_j = (1-s_j)^{m_j} s_j p_j \equiv u_{m_j,j} \tag{23}$$

Thus the greedy strategy, which can easily be seen to be also the optimal strategy, is to visit the j's according to the order statistic of the two-dimensional lattice $u_{m,j}$, with $j = 1, \ldots, N$ and integer $m \geq 0$. Denoting that order statistic by $u_{(i)}$, we have

$$\mu_{ddp} = \sum_i i u_{(i)} \tag{24}$$

since one easily checks that

$$\sum_{m,j} u_{m,j} \equiv \sum_i u_{(i)} = 1 \tag{25}$$

4.2 Probabilistic Recognition Square Root Sampling Strategy

We derive the best sampling strategy as before. The mean number of looks to success is $(s_i q_i)^{-1}$, so we want to minimize

$$\mu = \sum_i \frac{p_i}{s_i q_i} \tag{26}$$

Now the same calculation as before gives,

$$q_j = \sqrt{\frac{p_j}{s_j}} \Bigg/ \sum_i \sqrt{\frac{p_i}{s_i}} \tag{27}$$

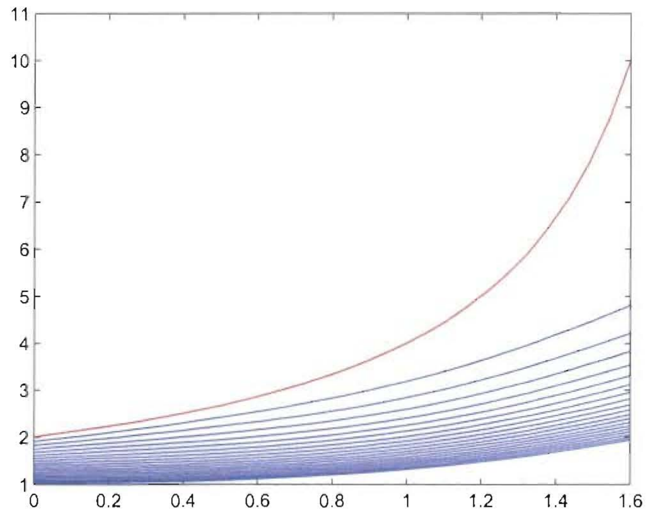

Figure 3: Power-law distributions with denial and deception. The abscissa is the power law exponent α. The ordinate is the figure of merit. Upper curve: Figure of merit for $s_i = 1$ (certain recognition, equation 15). Lower curves: Figures of merit for $s_i = 0.95(0.05)0.05$. Because even the optimal sorting strategy must now resample, the square root sampling strategy is now closer to optimal (F.M.= 1) than before.

and

$$\mu_{dds} = \left(\sum_i \sqrt{\frac{p_i}{s_i}} \right)^2 \tag{28}$$

The figure of merit F.M. is now μ_{dds}/μ_{ddp}.

4.3 Power Law Prior Distributions

We have evaluated the figure of merit for the case $N = 1000$, $s_i =$ constant, and varying powers α as in section 3.2. Results are shown in Figure 3 for $0 < \alpha < 1.6$, and in Figure 4 for $3 < \alpha < 8$. In obtaining these results, equation (28) is evaluated straightforwardly, while the evaluation of equation

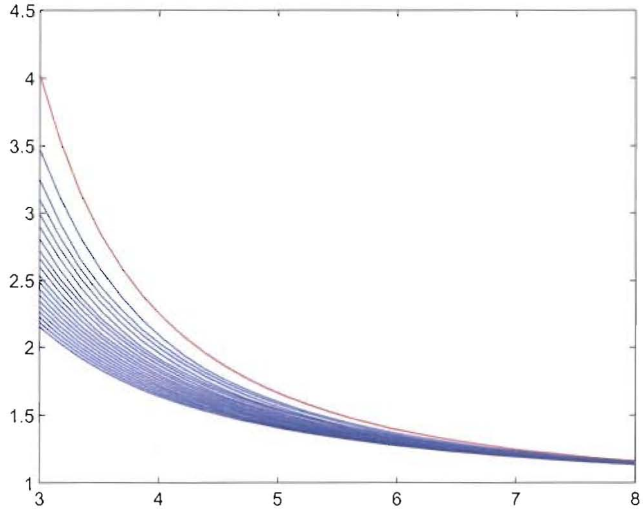

Figure 4: Same as Figure 3, but for larger values of the power law exponent α. The uppermost (red) curve plots the case of certain recognition, equation (17).

(24) requires the use of a heap data structure to iterate efficiently over the (m, j) lattice, given the values of the p_i's and s_i's.

4.4 Exponential Prior Distributions

If the number of values of i with with significant probabilities in equation (19) is not too small, then the figure of merit remains approximately independent of c; so we need compute only for a range of values d, and a range of values s_i (assumed constant over i). The results are shown in Figure 5. One again sees that smaller values of s_i bring the square root sampling strategy closer to optimal (F.M.= 1), since even the optimal strategy must resample many times.

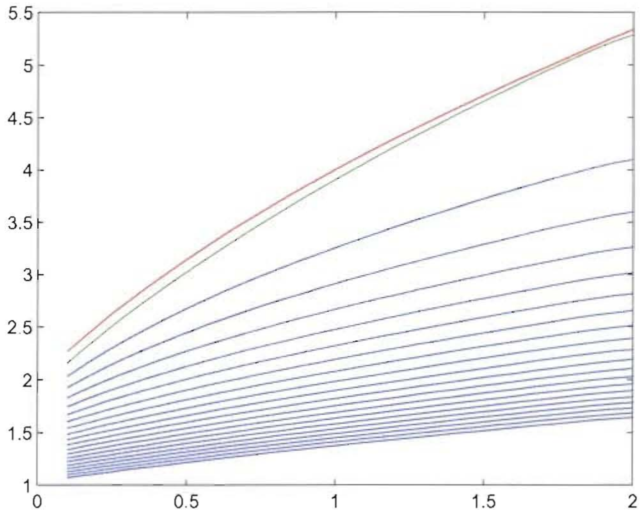

Figure 5: Figure of merit (ordinate) for exponential distributions with probabilistic recognition, as a function of the dimension parameter d (abscissa). The top (red) curve plots equation (21). The slightly lower (green) curve plots numerical results for $s_i = 1$. These differ slightly because of the effect of finite N. The lower (blue) curves plot values of $s_i = 0.95(0.05)0.05$. As before, decreasing values of s_i give increasing efficiency for the square root sampling strategy relative to the perfect sorting strategy.