A Formal Analysis of Why Heuristics Work

B. John Oommen*and Luis G. Rueda†

Abstract

Many optimization problems in computer science have been proven to be NP-hard, and hence there are no general polynomial-time algorithms to solve them. Alternatively, they are solved using *heuristics*, providing a *sub-optimal* solution that, hopefully, is arbitrarily close to the *optimal* one. Such problems are found in a wide range of applications, including artificial intelligence, game theory, graph partitioning, database query optimization, etc. Given two heuristics, the question of determining which is superior, has typically demanded a yes/no answer which is often substantiated based on empirical evidence. We have solved the problem of deciding on the *superior* heuristic by using Pattern Classification Techniques (PCT). We prove the following assertion: Given two heuristics H_1 and H_2 used in determining the goal of a particular problem, if the accuracy in obtaining the *optimal* solution by H_1 is greater than that of H_2, then H_1 has a higher probability of leading to the optimal solution than H_2. To the best of our knowledge, this is an open problem; *this unproven conjecture has been the basis for designing numerous algorithms such as the A* algorithm, and its variants.* By formulating the problem from a Pattern Recognition perspective, we use PCT to present a mathematical, rigorous proof of this fact, and show some uniqueness results. The corresponding database query optimization problem has been open for at least two decades, the difficulty of which has partially been due to the hurdles involved in the formulation itself. To validate our proofs, we report empirical results on database query optimization techniques involving a few well-known histogram estimation methods.

1 Introduction

1.1 Overview

The theory of Pattern Recognition (PR) is quite advanced. Numerous books and papers have been written to present a foundational basis for the field [1, 2]. PR is a field of machine intelligence that has been one of the dominant technologies in the last few decades. Broadly speaking, PR is a decision-making process, based on *a priori* and learned knowledge of the classes and objects being recognized. More specifically, the system learns

*Senior Member, IEEE. School of Computer Science, Carleton University, 1125 Colonel By Dr., Ottawa, ON, K1S 5B6, Canada. E-mail: oommen@scs.carleton.ca. Partially supported by NSERC, the Natural Science and Engineering Research Council of Canada.

†School of Computer Science, Carleton University, 1125 Colonel By Dr., Ottawa, ON, K1S 5B6, Canada. E-mail: lrueda@scs.carleton.ca. Partially supported by Departamento de Informática, Universidad Nacional de San Juan, Argentina.

information about the features of a set of classes. Given an object, and this information, the system attempts to recognize the unknown object as belonging to one of the known classes with some arbitrary accuracy.

There are many applications of PR including face and speech recognition, fingerprint identification, character recognition, medical diagnosis, etc. In each of these applications, the information about the classes can be *structural* or *statistical*. In the first case, we deal with the field of structural and syntactic pattern recognition, and in the second case, with the field of statistical pattern recognition.

The statistical information, or *features*, about the classes is represented by random vectors. The procedure of obtaining the features consists of mapping the feature values of each sample to a vector. Feature values, for example, can be the width or the height of a figure, the value of a pixel of an image, etc.

Statistical pattern recognition can also be subdivided into two well defined approaches, *parametric* and *non-parametric*. In the former, the random vectors have a known probability distribution, e.g. normal (or Gaussian), exponential, multinomial, etc. No such model is assumed in a non-parametric case.

As opposed to this, the area of computer science has still quite a few open, unsolved problems. An optimization problem[1] takes an instance which consists of a finite set, an objective function and some feasibility functions, and find a maximum (or a minimum) subject to certain constraints [4]. An *optimal solution* is a feasible solution for which the return is as large as possible (or the cost as small as possible, in which case the goal requires minimization). A *heuristic solution* is an algorithm that attempts to find a certain structure or a solution for which the cost is as close to the optimal as possible. It usually consists of an iterative process of applying one or more heuristics (in general, randomized) based on a certain strategy. Many heuristic strategies have been reported in the literature, including *alpha-beta search* [5], *backtracking*, *hill-climbing* [4], *simulated annealing* [6], *genetic algorithms* [7], *tabu search* [8], *learning automata* [9], etc. How heuristics are used in these methods and in searching, game playing, etc., can be found in [10, 11].

Our result can be crystallized as follows: Given two heuristics, the question of determining which is superior, has typically demanded a yes/no answer which is often substantiated based on empirical evidence. We have solved the problem of deciding on the *superior* heuristic by using Pattern Classification Techniques (PCT). We prove the following assertion: Given two heuristics, H_1 and H_2, used in determining the goal of a particular problem, if the accuracy in obtaining the *optimal* solution by H_1 is greater than that of H_2, then H_1 has a higher probability of leading to the optimal solution than H_2. To the best of our knowledge, this is an open problem. *However, this unproven conjecture has been the basis for designing numerous algorithms such as the A* algorithm, and its variants, in searching and game playing, etc.* [10, 11, 12].

To explain the above, we shall see how heuristic accuracy and solution optimality are fundamentally related. Consider a heuristic, H_1, used to determine the solution to a problem. The question of how the quality of the solution is related to the accuracy of H_1 is usually determined intuitively, and is based on empirical results which depends on the domain in which H_1 is applied. In this paper, we introduce a formal theoretical model to give a *stochastically* positive/negative answer to this relationship. We use a reasonable model for the accuracy of the heuristic, in which the optimal solution is a doubly-exponential random variable. This distribution, which as we shall presently see, is used to approximate the Gaussian distribution, is typically used in failure

[1] Every optimization problem can also be formulated as a decision problem [3].

models, and hence is reasonable in this scenario. In our model, the accuracy of the heuristic is related to the variance of the random variable used to represent it. The analysis for the Gaussian distribution follows thereafter.

If we now consider another heuristic, H_2, whose variance is greater than that of H_1, and whose mean is the same as that of H_1, we have theoretically proven that H_1 is more likely to succeed in obtaining the optimal solution. For this model, we have also proved the uniqueness of the result, and the conditions for which both heuristics lead to coincident probabilities of success.

The doubly exponential distribution is actually meant to be an approximation of the Gaussian distribution, typically used to model errors. However, the algebraic analysis for Gaussian distributions is impossible as there is no closed-form expression for integrating its probability density function. Consequently, we have extended the analysis for the doubly exponential distribution to formulate a reasonable analysis for the Gaussian distribution using numerical integration. By means of this analysis, we have corroborated the validity of our hypothesis for Gaussian distributions also. We believe that our results are quite impressive.

We also provide empirical results on using a few histogram-like estimation methods in database query optimization, which demonstrate the validity of our theoretical analysis.

1.2 Applications

There are many heuristic algorithms that can be used to solve a wide variety of NP-hard problems. Such problems can be found in a wide range of applications spanning the whole spectrum of artificial intelligence, and include game playing and game theory, graph theory, database query optimization, networking, computational geometry, number theoretic problems, parallel processing, etc. The results presented in this paper are general enough to be applicable to any heuristic algorithm and for any particular problem. In this introductory section, we just describe a few of them.

In the area of database query optimization, when more than two tables have to be joined, intermediate join operations are performed to ultimately obtain the final relation. As a result, the same query can be performed by means of different intermediate (join) operations. A simple sequence of join operations that leads to the same final result is called a QEP. Each QEP has associated an internal cost, which depends on the number of operations performed in the intermediate joins. The problem of choosing the best QEP is a combinatorially explosive optimization problem. This problem is currently solved by estimating the query result sizes of the intermediate relations and selecting the most efficient access QEP.

Since the analysis of selecting the best QEP must be done in "real" time, it is not possible to inspect the real data in this phase. Consequently, query result sizes are usually estimated using statistical information about the structures and the data maintained in the database catalogue. This information is used to approximate the distribution of the attribute values in a particular relation. Hence the problem of selecting the best QEP depends on how well that distribution is approximated.

In [13], it has been shown that errors in query result size estimates may increase exponentially with the number of joins. Since current databases and the associated queries increase in complexity, numerous efforts

have being made to devise more efficient techniques that solve the query optimization problem.

Many techniques have been proposed to estimate query result sizes, including histograms, sampling, and parametric techniques [14, 15, 16, 17]. Histograms are the most commonly used form of statistical information. They are incorporated in most of the commercial database systems such as Oracle, Microsoft SQL Server, Teradata, and DB2, which mainly use the Equi-depth histogram. The prominent models of histograms known in the literature are: *Equi-width* [14, 18], *Equi-depth* [15, 16], the *Rectangular Attribute Cardinality Map* (*R-ACM*) [19], the *Trapezoidal Attribute Cardinality Map* (*T-ACM*) [20], and the *V-Optimal Histograms* [13, 21].

Another area in which our model can be used to answer open questions is in the fields of *game theory* and *game playing* [12]. The most widely used structure used to analyze the best possible move and strategy is a *game tree*, whose root node represents the initial status of the board. All possible moves of the first player are the edges from the root to the first level, the edges of each child represent all possible moves of the second player, the opponent. Continuing in the same fashion, the game is played (or rather plans executed) until one of the players wins. The aim is to optimize the moves of the first player based on searching *all* the branches of the tree until the leaves, and perform the best move based on maximizing the reward of the first player and minimizing that of the second one.

There are many techniques used to optimize the moves of the first the player. One of them is the *minimax search algorithm*, which searches over a fixed number of levels of the entire tree, and finds the best moves at each node. This exhaustive searching procedure has a complexity that grows exponentially with the number of nodes of the tree. A more efficient mechanism is the *alpha-beta search* algorithm [5], a heuristic that significantly reduces the number of nodes explored. Both of these assume that the heuristic that they use is advantageous in determining a superior strategy. This is the question that we address in this paper. The model presented in this paper has important consequences in choosing a heuristic. Such a heuristic could be, for example, the cost of a path from the current state to a goal state, which unfortunately is not exactly known, but is estimated using a heuristic. The searching scheme, such as the alpha-beta search and the minimax search algorithm, uses this heuristic to search for a, hopefully, *optimal* path in the game tree.

Another application of our method is in graph theory, for example, in the uniform graph partitioning. Given a complete graph on $2n$ vertices, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, along with a cost function $f : \mathcal{E} \rightarrow \mathbb{Z}^+ \bigcup \{0\}$, the aim is to find a partition whose sum of costs of the individual subsets is minimized. This problem is also known to be NP-Hard and have several applications especially in VLSI design, hydrology, networks, etc. Many heuristic algorithms have been proposed to solve this problem, including simulated annealing, genetic algorithms, learning automata, etc. [4, 22].

When considering a particular heuristic algorithm, we can incorporate different heuristics used to approximate the sum of costs of the individual subsets of a particular partitioning. It is intuitive that a more accurate heuristic is more likely to succeed in finding the optimal solution. However, this is not what happens in all cases. We rather provide a stochastic answer to this question. By means of a rigorous theoretical analysis, we prove that a particular heuristic, which provides more accurate approximations for the sum of costs of the individual subsets, is more likely to obtain the minimal cost for a partitioning than a *less accurate* heuristic.

1.3 Problem Statement

In this paper, we show that this fundamental open problem in computer science, namely that of relating heuristic accuracy with solution optimality, can be solved using the principles of the theory of *pattern classification*. This problem has been (to our knowledge) open. In particular, the corresponding database query optimization problem has been unsolved for more than two decades.

More specifically, we prove the following: Given two heuristics, H_1 and H_2, used in a decision problem, if the accuracy in approximating the optimal solution by H_1 is greater than that of H_2, then H_1 has a higher probability of leading to the optimal solution than H_2.

The importance of the results of this paper is that we show that the answer to the accuracy/optimality question is "stochastically positive". In other words, we prove that although a superior heuristic may not always yield a more optimal solution, the probability that the superior heuristic method yields an optimal solution exceeds the probability that an inferior heuristic yields an optimal solution. This paper thus justifies and gives a formal rigorous basis for why heuristics work.

We analytically prove that under the well-acclaimed models of inaccuracy, the better the accuracy of a heuristic, the greater the probability of it choosing the optimal solution. We have also provided some empirical results related to the field of database query optimization. These results show the superiority of R-ACM over the traditional histogram estimation methods, the Equi-width and the Equi-depth. The empirical results obtained by testing these properties for many of the above histogram methods in random databases show that the R-ACM is significantly superior to both the Equi-width and the Equi-depth schemes.

2 The Relation between Efficiency and Optimality

Consider two heuristics, H_1 and H_2. The probability of correctly estimating a cost value of a particular solution by H_1 and that of estimating a cost value by H_2 are represented by two independent random variables. Clearly, this assumption of independence is valid because there is no reason why the value obtained by one heuristic should affect the value obtained by the second.

For the analysis done below, we work with two models for the error function: the doubly exponential distribution and the normal distribution. In the former, the probability of obtaining a value that deviates from the mean (or true value) falls exponentially as a function of the deviation. The exponential distribution is more typical in reliability analysis and in failure models, and in this particular domain, the question is one of evaluating how reliable the quality of a solution is if only an estimate of its performance is available. More importantly, it is used as an approximation to the Gaussian distribution for reasons which will be clarified momentarily. The Gaussian model is much more difficult to analyze, since there is no closed-form algebraic expression for integrating the probability density function. However, a formal computational proof is included, which confirms our hypothesis.

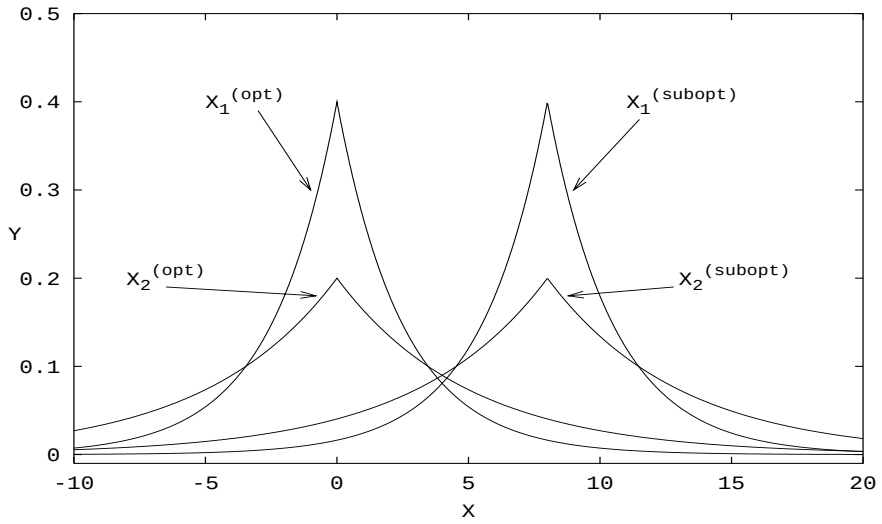

Figure 1: An example of doubly exponential distributions for the random variables $X_1^{(opt)}$, $X_2^{(opt)}$, $X_1^{(subopt)}$ and $X_2^{(subopt)}$, whose parameters are $\lambda_1 = 0.4$ and $\lambda_2 = 0.2$.

2.1 Analysis Using Exponential Distributions

A random variable, X, is said to be *doubly exponentially* distributed with parameter λ if the density function is given by:

$$f_X(x) = \frac{1}{2}\lambda e^{-\lambda|x-c|} \qquad -\infty < x < \infty \ , \tag{1}$$

where $E(X) = c$.

If X is a doubly exponential random variable, it can be shown that:

$$\mathrm{Var}[X] = \frac{2}{\lambda^2} \ . \tag{2}$$

Without loss of generality, if the mean cost of the cost of optimal solution is c_1, by shifting the origin by c_1, we can work with the assumption that the cost of the best solution is 0, which is the mean of these two random variables. The cost of the second best solution is given by another two random variables (one for H_1 and the other one for H_2) whose mean, $c_2 > 0$, is the same for both variables. An example will help to clarify this.

Example 1. Suppose that H_1 chooses the optimal cost value with probability represented by a doubly exponential random variable, $X_1^{(opt)}$, whose mean is 0 and $\lambda_1 = 0.4$. This method also chooses another sub-optimal cost value according to $X_1^{(subopt)}$ whose mean is 8 and $\lambda_1 = 0.4$.

H_2 is another method that chooses the optimal cost value with probability given by $X_2^{(opt)}$ whose parameters are $c_1 = 0$ and $\lambda_2 = 0.2$. It chooses the second sub-optimal cost value with probability given by $X_2^{(subopt)}$ whose parameters are $c_2 = 8$ and $\lambda_2 = 0.2$.

Since $\frac{2}{\lambda_1^2} < \frac{2}{\lambda_2^2}$, this is used to signify that the probability that H_1 chooses a sub-optimal cost value is smaller than that of H_2 choosing the sub-optimal cost value. This scenario is depicted in Figure 1. □

The result depicted above is formalized in the following theorem, *which is the first primary result of this paper, and answers the open question referred to above.* The theorem is formulated in terms of the probabilities that two heuristics make the wrong decision, which we show is inherently related to the probability that these heuristics converge to *sub-optimal* solutions. Observe too that the formulation and proof use techniques typically foreign to database theory, game theory, artificial intelligence, or for that matter any computer science area in which this approach can be applied, but which are fundamental to the theory of PR. The second theorem, extends the results of the first, and shows how the results can be geometrically interpreted.

Theorem 1. Suppose that:

- H_1 and H_2 are two heuristics.
- X_1 and X_2 are two doubly exponential random variables that represent the estimated cost values of the *optimal* solution obtained by H_1 and H_2 respectively.
- X_1' and X_2' are another two doubly exponential random variables representing the estimated cost values of a *non-optimal* solution obtained by H_1 and H_2 respectively.
- $0 = \mathrm{E}[X_1] = \mathrm{E}[X_2] \leq \mathrm{E}[X_1'] = \mathrm{E}[X_2'] = c$.

Let p_1 and p_2 be the probabilities that H_1 and H_2 respectively make the wrong decision. Then,

$$\text{if } \mathrm{Var}[X_1] = \mathrm{Var}[X_1'] = \frac{2}{\lambda_1^2} \leq \frac{2}{\lambda_2^2} = \mathrm{Var}[X_2] = \mathrm{Var}[X_2'], \quad p_1 \leq p_2 \ .$$

Proof. Consider a particular value x. The probability that the value x leads to a wrong decision made by H_1, is given by:

$$\begin{aligned} I_{11} &= \textstyle\int_{-\infty}^{x} \frac{1}{2}\lambda_1 e^{\lambda_1(u-c)}\, du && \text{if } x < c \ , \text{ and} \\ I_{12} &= \textstyle\int_{-\infty}^{c} \frac{1}{2}\lambda_1 e^{\lambda_1(u-c)}\, du + \int_{c}^{x} \frac{1}{2}\lambda_1 e^{-\lambda_1(u-c)}\, du && \text{if } x > c \ . \end{aligned} \tag{3}$$

Solving the integrals, (3) results in:

$$
\begin{aligned}
I_{11} &= \tfrac{1}{2}e^{\lambda_1(x-c)} - \lim_{u\to-\infty}\tfrac{1}{2}e^{-\lambda_1(u-c)} && = \tfrac{1}{2}e^{-\lambda_1(x-c)} \ , \text{ and} \\
I_{12} &= \lim_{u\to-\infty}\tfrac{1}{2}e^{-\lambda_1(-u+c)} + \tfrac{1}{2} - \tfrac{1}{2}e^{-\lambda_1(x-c)} + \tfrac{1}{2} && = 1 - \tfrac{1}{2}e^{-\lambda_1(x-c)} \ .
\end{aligned} \tag{4}
$$

The probability that H_1 makes the wrong decision for *all* the values of x is the following function of λ_1 and c:

$$
p_1 = I(\lambda_1, c) = \int_{-\infty}^{0} I_{11}\frac{1}{2}\lambda_1 e^{\lambda_1 x}\,dx + \int_{0}^{c} I_{11}\frac{1}{2}\lambda_1 e^{-\lambda_1 x}\,dx + \int_{c}^{\infty} I_{12}\frac{1}{2}\lambda_1 e^{-\lambda_1 x}\,dx \ . \tag{5}
$$

which, after applying the distributive law and substituting the values of I_{11} and I_{12}, can be written as:

$$
\int_{-\infty}^{0} \frac{\lambda_1}{4}e^{2\lambda_1 x-\lambda_1 c}\,dx - \int_{0}^{c}\frac{\lambda_1}{4}e^{-\lambda_1 c}\,dx + \int_{c}^{\infty}\left[\frac{\lambda_1}{2}e^{-\lambda_1 x} - \frac{\lambda_1}{4}e^{-2\lambda_1 x+\lambda_1 c}\right]dx \ . \tag{6}
$$

After solving the integrals, (6) is transformed into:

$$
\frac{1}{8}e^{-\lambda_1 c} + \frac{1}{4}\lambda_1 c e^{-\lambda_1 c} + \frac{3}{8}e^{-\lambda_1 c} = \frac{1}{2}e^{-\lambda_1 c} + \frac{1}{4}\lambda_1 c e^{-\lambda_1 c} \ . \tag{7}
$$

Similarly, we do the same analysis for p_2, which is a function of λ_2 and c:

$$
p_2 = I(\lambda_2, c) = \frac{1}{2}e^{-\lambda_2 c} + \frac{1}{4}\lambda_2 c e^{-\lambda_2 c} \ . \tag{8}
$$

We have to prove that:

$$
p_1 = \frac{1}{2}e^{-\lambda_1 c} + \frac{1}{4}\lambda_1 c e^{-\lambda_1 c} \le \frac{1}{2}e^{-\lambda_2 c} + \frac{1}{4}\lambda_2 c e^{-\lambda_2 c} = p_2 \ . \tag{9}
$$

Multiplying both sides by 2, and substituting $\lambda_1 c$ for α_1 and $\lambda_2 c$ for α_2, (9) can be written as follows:

$$
e^{-\alpha_1} + \frac{1}{2}\alpha_1 e^{-\alpha_1} \le e^{-\alpha_2} + \frac{1}{2}\alpha_2 e^{-\alpha_2} \ . \tag{10}
$$

Substituting α_2 for $k\alpha_1$, $\alpha_1 \ge 0$ and $0 < k \le 1$, (10) results in:

$$q_1 = e^{-\alpha_1} + \frac{1}{2}\alpha_1 e^{-\alpha_1} \leq e^{-k\alpha_1} + \frac{1}{2}k\alpha_1 e^{-k\alpha_1} \ = q_2\,. \tag{11}$$

We now prove that $q_1 - q_2 \leq 0$. After applying natural logarithm to both sides of (11) and some algebraic manipulations, $q_1 - q_2 \leq 0$ implies:

$$F(\alpha_1, k) = k\alpha_1 - \alpha_1 + \ln(1 + \frac{1}{2}\alpha_1) - \ln(1 + \frac{1}{2}k\alpha_1) \leq 0\,. \tag{12}$$

To prove that $F(\alpha_1, k) \leq 0$, we use the fact that $\ln x \leq x - 1$. Hence, we have:

$$\begin{aligned}
F(\alpha_1, k) &= \alpha_1(k-1) + \ln\left(\frac{1 + \frac{1}{2}\alpha_1}{1 + \frac{1}{2}k\alpha_1}\right) && (13)\\
&\leq \alpha_1(k-1) + \frac{1 + \frac{1}{2}\alpha_1}{1 + \frac{1}{2}k\alpha_1} - 1 && (14)\\
&= \alpha_1(k-1) + \frac{\alpha_1 - k\alpha_1}{2 + k\alpha_1} && (15)\\
&= \frac{k\alpha_1 + k^2\alpha_1^2 - \alpha_1 - k\alpha_1^2}{2 + k\alpha_1} && (16)\\
&= \frac{\alpha_1(k-1)(k\alpha_1 + 1)}{2 + k\alpha_1} \leq 0\,, && (17)
\end{aligned}$$

because:

(i) $0 < k \leq 1$ and $\alpha_1 \geq 0 \ \Rightarrow \ \alpha_1(k-1) \leq 0$ and $k\alpha_1 + 1 > 0$. Hence $\alpha_1(k-1)(k\alpha_1 + 1) \leq 0$, and

(ii) $0 < k \leq 1$ and $\alpha_1 \geq 0 \ \Rightarrow \ 0 < k\alpha_1 \leq \alpha_1 \ \Rightarrow \ k\alpha_1 + 2 > 2 > 0$.

Hence the theorem. □

The above theorem can be viewed as a "sufficiency result". In other words, we have shown that $q_1 - q_2 \leq 0$ or that $p_1 \leq p_2$. We now show a "necessity result" stated as a uniqueness result. This result states that the function $p_1 \leq p_2$ has its equality ONLY at the boundary condition where the two distributions are exactly identical.

To prove the necessity result, we consider $q_2 - q_1$ which, derived from (11), can be written, as a function of α_1 and k, as:

$$G(\alpha_1, k) = e^{-k\alpha_1} + \frac{1}{2}k\alpha_1 e^{-k\alpha_1} - e^{-\alpha_1} - \frac{1}{2}\alpha_1 e^{-\alpha_1}\,. \tag{18}$$

By examining its partial derivatives we shall show that there are two solutions for equality. Furthermore, when $\alpha_1 \geq 0$ and $0 < k \leq 1$, we shall see that for a given k, there is only one solution, namely $\alpha_1 = 0$ and k,

$0 < k \leq 1$, proving the uniqueness.

Theorem 2. Suppose that $\alpha_1 \geq 0$, $0 < k \leq 1$. Let $G(\alpha_1, k)$ be:

$$G(\alpha_1, k) = e^{-k\alpha_1} + \frac{1}{2}k\alpha_1 e^{-k\alpha_1} - e^{-\alpha_1} - \frac{1}{2}\alpha_1 e^{-\alpha_1}\,. \tag{19}$$

Then $G(\alpha_1, k) \geq 0$, and there are exactly two solutions for $G(\alpha_1, k) = 0$, being: $\{\alpha_1 = -1, k = 1\}$ and $\{\alpha_1 = 0, k\}$.

Proof. We must prove that, as defined in the theorem statement, $G(\alpha_1, k) \geq 0$.

We shall prove that this is satisfied by determining the local minima for $G(.,.)$, where $\alpha_1 \geq 0$ and $0 < k \leq 1$. We first find the partial derivatives of (19) with respect to α_1 and k:

$$\frac{\partial G}{\partial \alpha_1} = -\tfrac{1}{2}ke^{-k\alpha_1} - \tfrac{1}{2}k^2\alpha_1 e^{-k\alpha_1} + \tfrac{1}{2}e^{-\alpha_1} + \tfrac{1}{2}\alpha_1 e^{-\alpha_1} = 0\,, \text{ and} \tag{20}$$

$$\frac{\partial G}{\partial k} = -\tfrac{1}{2}\alpha_1 e^{-k\alpha_1} - \tfrac{1}{2}k\alpha_1^2 e^{-k\alpha_1} = 0\,. \tag{21}$$

We now solve (20) and (21) for α_1 and k. Equation (21) can be written as follows:

$$-\frac{1}{2}\alpha_1 e^{-k\alpha_1} = \frac{1}{2}k\alpha_1^2 e^{-k\alpha_1}\,, \tag{22}$$

which, after canceling some terms results in $k\alpha_1^2 + \alpha_1 = 0$. Solving this equation for α_1, we have: $\alpha_1 = -\frac{1}{k}$ and $\alpha_1 = 0$. Substituting $\alpha_1 = -\frac{1}{k}$ in (20) and canceling some terms, we obtain:

$$\frac{1}{2}e^{-\alpha_1} + \frac{1}{2}\alpha_1 e^{-\alpha_1} = 0\,, \tag{23}$$

which results in the solution to be $\alpha_1 = -1$, and consequently, $k = 1$.

The second root, $\alpha_1 = 0$, indicates that the minimum is achieved for any value of k.

We have thus found two solutions for (20) and (21), $\{\alpha_1 = 0, k\}$ and $\{\alpha_1 = -1, k = 1\}$. Since $\alpha_1 \geq 0$, it means that α_1 can have at least a value of 0, and hence the local minima is in $\{\alpha_1 = 0, k\}$. Substituting these two values in G, we see that $G(\alpha_1, k) = 0$, which is the minimum. Therefore, $G(\alpha_1, k) \geq 0$ for $\alpha_1 \geq 0$ and $0 < k \leq 1$.

Hence the theorem. □

To get a physical perspective of these results, let us analyze the geometric relation of the function G and the heuristics. G is a positive function in the region $\alpha_1 \geq 0$, $0 < k \leq 1$. When $\alpha_1 \to 0$, $G \to 0$. This means that for small values of α_1, G is also small. Since $\alpha_1 = \lambda_1 c$, the value of α_1 depends on λ_1 and c. When

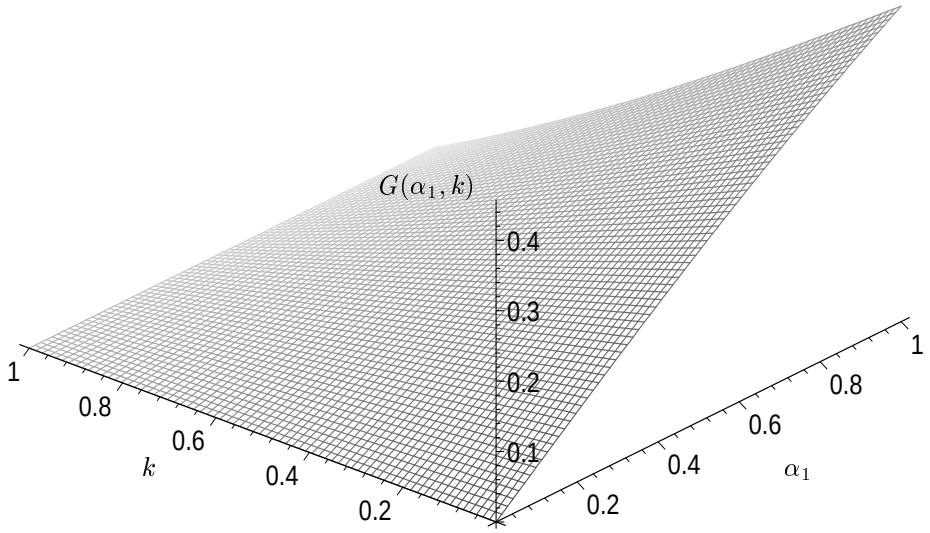

Figure 2: Function $G(\alpha_1, k)$ plotted in the ranges $0 \leq \alpha_1 \leq 1$ and $0 \leq k \leq 1$.

c is small, G is very close to its minimum, 0, and hence both probabilities, p_1 and p_2, are very close. This behavior can be observed in Figure 2, and is the phenomenon observed if the heuristics are both comparable and almost equally efficient.

In terms of histogram methods and in database query optimization, when c is small, the optimal and the sub-optimal QEP are very close. Since histogram methods such as Equi-width and Equi-depth produce a larger error than the R-ACM and the T-ACM, the former are less likely to find the optimal QEP than the latter.

Interpreted alternatively, G is very small when λ_1 is close to 0. This means that $\text{Var}[X_1]$ is very large. Since $\text{Var}[X_1] \leq \text{Var}[X_2]$, $\text{Var}[X_2]$ is also very large, and both are close each other (In Figure 1, we would observe almost flat curves for both distributions). Random variables for histogram methods such as Equi-width and Equi-depth yield similar error estimation distributions with large and similar variances. Hence, the probabilities p_1 and p_2 are quite close, and consequently, similar results are expected for these estimation methods. However, when the heuristics yield widely different estimated costs (as in the case when the new histogram methods, the R-ACM and the T-ACM, are compared to the traditional methods), these effectively imply random variables with smaller variances being compared to random variables with larger variances. In such a case, the value of G is very high – implying that the former would yield superior solutions.

2.2 Analysis Considering Normal Distributions

For the analysis done in this section, we consider that we are given two heuristics, H_1 and H_2, for which the probabilities of choosing optimal or suboptimal solutions are represented by two normally distributed random variables, X_1 and X_2, whose means are μ_1 and μ_2, and whose variances are σ_1^2 and σ_2^2 respectively.

Although the model using normal distributions is more realistic in real life problems, the analysis becomes

impossible because there is no closed-form algebraic expression for integrating the normal probability density function. Alternatively, we have used numerical integration and we have obtained rather representative values for which the implication between efficiency and optimality is again corroborated.

Without loss of generality, if the mean cost value of the optimal solution is μ_1, by shifting the origin by μ_1, we again assume that the cost of the best solution is 0, which is the mean of these two random variables. The cost of the second best solution is given by another two random variables (one for H_1 and the other one for H_2) whose mean, $\mu_2 > 0$, is the same for both variables. We also assume that, by scaling both distributions[2], the variance of H_1 choosing the optimal solution is 1. An example will help to clarify this.

Example 2. Suppose that H_1 chooses the optimal cost value with probability represented by the normal random variable $X_1^{(opt)}$ whose mean is 0 and standard deviation is $\sigma_1 = 1$. This heuristic also estimates another sub-optimal cost value according to $X_1^{(subopt)}$ whose mean is 4 and $\sigma_1 = 1$.

H_2 is another heuristic that estimates the optimal cost value with probability given by $X_2^{(opt)}$ whose parameters are $\mu = 0$ and $\sigma_2 = 1.4$. Another sub-optimal cost value is obtained with probability given by $X_2^{(subopt)}$ whose parameters are $\mu = 4$ and $\sigma_2 = 1.4$.

Observe that $\sigma_1 < \sigma_2$, and hence we are expecting that the probability of H_1 making a wrong decision is smaller than that of H_2. The probability density functions for these four random variables are depicted in Figure 3. Note that, as in the doubly exponential distribution, given a particular value of x, if its probability under $X_1^{(opt)}$ is high, then the area for which H_1 makes the wrong decision (i.e. its cumulative probability under $X_1^{(subopt)}$) is small. Since these two quantities are multiplied and integrated, the final value is smaller than that of H_2, as σ_2 is significantly higher than $\sigma_1 = 1$. This is what we formally show below. □

Result 1. [3]

Suppose that:

- H_1 and H_2 are two heuristics.
- X_1 and X_2 are two normally distributed random variables that represent the cost values of the *optimal* solution obtained by H_1 and H_2 respectively.
- X_1' and X_2' are another two normally distributed random variables representing the cost values of a *non-optimal* solution obtained by H_1 and H_2 respectively.
- $0 = \mathrm{E}[X_1] = \mathrm{E}[X_2] \leq \mathrm{E}[X_1'] = \mathrm{E}[X_2'] = \mu$.

Let p_1 and p_2 be the probabilities that H_1 and H_2 respectively make the wrong decision. Then,

[2] This can be done by multiplying σ_1^2 and σ_2^2 by σ_1^{-2}, and μ_1 and μ_2 by σ_1^{-1}. This is a particular case of the simultaneous diagonalization between d-dimensional normal random vectors for which $d = 1$ [1].

[3] We cannot claim this result as a theorem, since the formal analytic proof is impossible. This is because there is no closed-form expression for integrating the Gaussian probability density function. However, the computational proof that we present renders this to be more than a conjecture.

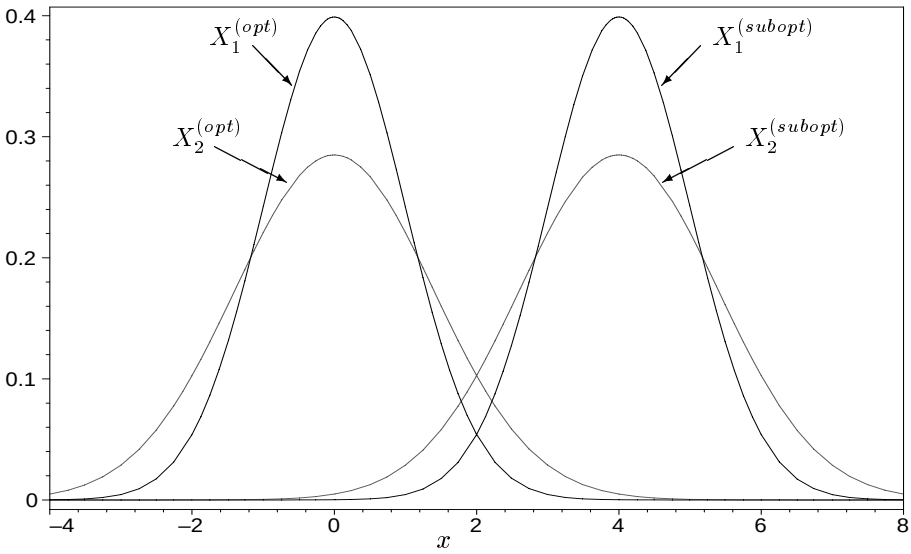

Figure 3: An example showing the probability density function of four normal random variables whose parameters are $\sigma_1 = 1$, $\sigma_2 = 1.4$, $\mu_1 = 0$, and $\mu_2 = 4$.

$$\text{if } \mathrm{Var}[X_1] = \mathrm{Var}[X_1'] = \sigma_1^2 \leq \sigma_2^2 = \mathrm{Var}[X_2] = \mathrm{Var}[X_2'], \quad p_1 \leq p_2 \ .$$

Computational Proof. To achieve this proof, we proceed by doing the same analysis that we did for the doubly exponential distributions. If we consider a particular value x, the probability that x leads to a wrong decision made by H_1, is given by:

$$I_1 = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(u-\mu)^2}{2\sigma_1^2}} \, du \, . \tag{24}$$

The probability that H_1 makes the wrong decision for *all* the values of x is obtained by integrating the function resulting from multiplying every value of I_1 for each x and the probability density function of $X_1^{(opt)}$, which results in:

$$p_1 = \int_{-\infty}^{\infty} I_1 \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} \, dx \ . \tag{25}$$

Similarly, p_2 can also be expressed as follows:

$$p_2 = \int_{-\infty}^{\infty} I_2 \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{x^2}{2\sigma_2^2}} \, dx \ \ , \tag{26}$$

$\sigma_1 \rightarrow$ σ_2 $\downarrow$	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00	10.00
1.00	1.0000									
2.00	33.6276	1.0000								
3.00	73.9210	2.1982	1.0000							
4.00	102.5081	3.0483	1.3867	1.0000						
5.00	122.1988	3.6339	1.6531	1.1921	1.0000					
6.00	136.2472	4.0516	1.8431	1.3291	1.1150	1.0000				
7.00	146.6138	4.3599	1.9834	1.4303	1.1998	1.0761	1.0000			
8.00	154.7078	4.6006	2.0929	1.5092	1.2660	1.1355	1.0552	1.0000		
9.00	161.0448	4.7891	2.1786	1.5710	1.3179	1.1820	1.0984	1.0410	1.0000	
10.00	166.1716	4.9415	2.2480	1.6211	1.3598	1.2196	1.1334	1.0741	1.0318	1.0000

Table 1: Ratio between the probability of making the wrong decision for two normally distributed random variables whose standard deviations are σ_1 and σ_2.

where I_2 is obtained in the same way as in (24) for the distribution with variance σ_2^2.

Since there is no closed-form algebraic expression for integrating the normal probability density function, no analytical solution for proving that $p_1 \leq p_2$ can be formalized.

Alternatively, we have invoked a computational analysis by calculating these integral for various representative values of σ_1 and σ_2 by using the trapezoidal rule. The values of $G = \frac{p_2}{p_1} \geq 1$ (i.e. for $1 \leq \sigma_1 \leq 10$ and $1 \leq \sigma_2 \leq 10$, where $\sigma_1 \leq \sigma_2$) are depicted in Table 1 in the form of a *lower-diagonal* matrix. All the values of the *upper-diagonal* matrix (not shown here) are less than unity. Note that by making the value of $\sigma_1 = 1$, the analysis reduces to the first and second columns of this table. For example, if $\sigma_1 = 1$ and $\sigma_2 = 2$, $\frac{p_2}{p_1} \approx 33.6276$. For more neighboring values of σ_1 and σ_2, e.g. $\sigma_1 = 9$ and $\sigma_2 = 10$ ($\sigma_1 = 1$ and $\sigma_2 \approx 1.2345$ after scaling), $\frac{p_2}{p_1} \approx 1.0318$, which is very close to unity. The ratio for $\sigma_1 = 1$ and $\sigma_2 = 10$ is much bigger, i.e. more than one hundred times. □

In order to get a better perspective of the computational analysis, we study the behavior of the function $G = \frac{p_2}{p_1}$. Using the values of G given in Table 1, we have plotted this function in the three-dimensional space as $G(\sigma_1, \alpha_1)$, where $\alpha_1 = k\sigma_1$, $1 \leq k \leq 10$. The plot is depicted in Figure 4. In order to enhance the visualization of G, we have approximated it by using the regression utilities of the symbolic mathematical software package Maple V [23]. When $k = 1$, the surface lies on the $z = 0$ plane, in the form of a straight line $x = y$ (labelled "$k = 1$ or $\sigma_1 = \sigma_2$" in the figure). This is the place in which G reaches its minimum, when both heuristics have identical variances. When k is larger (i.e. $k = 10$), the function G becomes much larger (up to 166 in in Table 1). This clearly shows the importance of the variance in deciding on a heuristic.

When it concerns histograms in database query optimization, when k is small, it implies that the optimal and sub-optimal QEP are very close. Therefore, histogram methods like Equi-width and Equi-depth are less likely to find the optimal QEP, since they produce larger errors than histogram approximation methods such as the R-ACM and the T-ACM. The latter produce very small errors, and hence, when comparing any of them with the Equi-width or Equi-depth, we will have a much larger value of k. This will be reflected in our

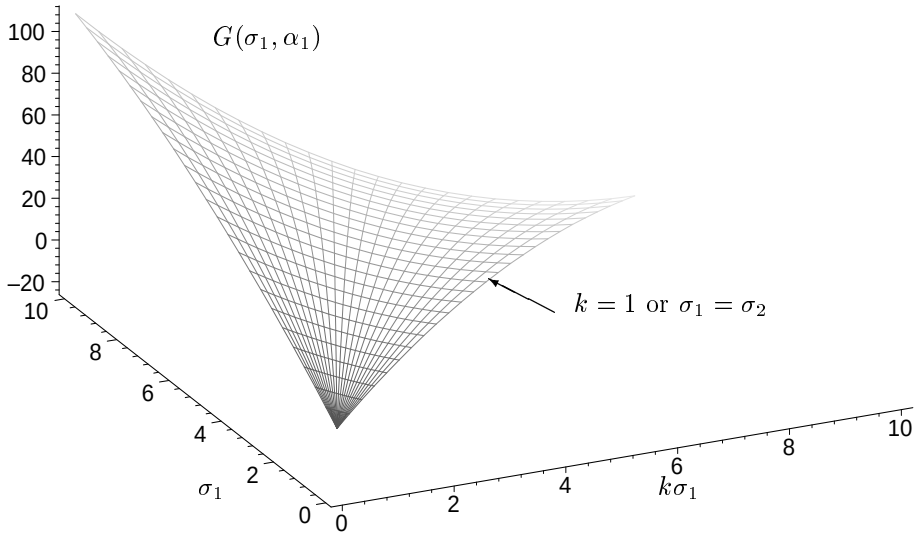

Figure 4: Function $G(\sigma_1, k\sigma_1)$ plotted in the ranges $1 \leq \sigma_1 \leq 10$ and $1 \leq k\sigma_1 \leq 10$, where $\sigma_2 = k\sigma_1$.

empirical results presented in the next section.

3 Empirical Results

In order to provide practical evidence of the theoretical results presented above[4], we have performed some simulations in database query optimization. In the experiments we have conducted four independent runs (the details of which can be found in [24]). In each run, 100 random databases were generated. Each database was composed of six relations, each of them having six attributes. Each relation was populated with 100 tuples.

The efficiency of R-ACM was compared with that of the Equi-width and the Equi-depth after performing these simulations using 50 values per attribute. We set the number of bins for the Equi-width and the Equi-depth to be 22. In order to be impartial with the evaluation, we set the number of bins for the R-ACM to be *approximately half* of that of the Equi-width and the Equi-depth, because the former needs twice as much storage as that of the latter.

The simulation results obtained from 400 independent runs, used to compare the efficiency of the R-ACM with that of the Equi-width and that of the Equi-depth, are given in Table 2. The column labeled "R-ACM-W" is the number of times that R-ACM is better than Equi-width. The column labelled "Equi-width" indicates the number of times in which the Equi-width obtains a better QEP than that of the R-ACM. Similarly, the column labelled "R-ACM-D" represents the number of times that the R-ACM yields better solutions than Equi-depth, and the column labelled "Equi-depth" is the number of times in which the Equi-depth is superior to the R-ACM. The last row, the total of each column, gives us the evidence that the R-ACM is superior to

[4]The empirical results presented in this paper are not intended to compare the various histogram methods: Equi-width, Equi-depth, R-ACM, T-ACM, V-optimal, etc. The experimental results submitted are merely included to demonstrate that the theoretically proven results can be experimentally justified.

Simulation	R-ACM-W	Equi-width	R-ACM-D	Equi-depth
1	26	12	35	12
2	24	15	42	13
3	35	11	46	8
4	29	15	46	8
Total	**114**	**53**	**169**	**41**

Table 2: Simulation results for the R-ACM, Equi-width, and Equi-depth, after optimizing a query on 400 randomly generated databases.

Equi-width in more than twice as much, and the R-ACM is better than Equi-depth by a factor of about four.

4 Conclusions

The theory of PR is quite developed, and has many applications. We believe that this theory can be used to prove unsolved results in various other fields. In particular, we have applied pattern classification techniques to solve a fundamental open problem in computer science relating heuristic accuracy and solution optimality.

In this paper, we have discussed the efficiency of using heuristics for optimization problems and resolved an open problem, which has been (to our knowledge) open for at least twenty years. The problem describes how the accuracy of a heuristic relates to the quality of the solution obtained. The efficiency has been quantified by means of the probability of a heuristic choosing the optimal solution.

We have shown analytically (using a reasonable model of accuracy, namely the doubly exponential distribution for errors) that as the accuracy of a heuristic increases, the probability of it leading to a superior solution also increases. This result is quite general, and can be applied to artificial intelligence, game theory, graph partitioning, etc. In particular, for the field of database query optimization, we have highlighted that for histogram methods that produce errors with similar variances (the Equi-width and the Equi-depth), the query processing results are also quite similar. However, we have also shown that the R-ACM and the T-ACM, which produce error with smaller variances than the traditional methods, yield better query optimization plans more often.

Due to the constraints involved in deriving a closed-form expression for integrating the normal probability density function, we have presented a computational analysis of the accuracy/optimality problem for the Gaussian distribution. Our analysis has also validated the result that heuristics producing smaller errors lead more often to optimal solutions.

Finally, we have provided evidence of the theoretical contributions by means of the empirical results in database query optimization obtained from evaluating the Equi-width, the Equi-depth, and the R-ACM on randomly generated databases. These results show that the R-ACM provides superior solutions more than twice as many times as the Equi-width, and more than four times as often as the Equi-depth. More detailed empirical results including the design of random databases and random queries in these random databases can be found in [24].

References

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

[2] A. Webb, *Statistical Pattern Recognition.* New York: Oxford University Press Inc., 1999.

[3] M. Garey and D. Johnson, *Computers and Intractability : A Guide to the Theory of NP-Completeness.* W H Freeman & Co., 1979.

[4] D. Kreher and D. Stinson, *Combinatorial Algorithms : Generation, Enumeration, and Search.* CRC Press, 1998.

[5] D. Levy, *How Computers Play Chess.* New York: Computer Science Press, 1991.

[6] E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing.* John Wiley & Son, 1989.

[7] M. Michell, *An Introduction to Genetic Algorithms* . MIT Press, 1998.

[8] F. Glover and M. Laguna, *Tabu Search.* Kluwer Academic Pub, 1997.

[9] K. Narendra and M. Thathachar, *Learning Automata: An Introduction.* Englewood Cliffs, New Jersey: Prentice Hall, 1989.

[10] N. Nilsson, *Artificial Intelligence : A New Synthesis.* Morgan Kaufmann Publishers, 1998.

[11] E. Rich and K. Knight, *Artificial Intelligence.* McGraw Hill, 2nd ed., 1991.

[12] G. Romp, *Game Theory: Introduction & Applications.* Oxford University Press, 1997.

[13] Y. Ioannidis and S. Christodoulakis, "On the propagation of errors in the size of join results," in *Proceedings of the ACM-SIGMOD Conference*, pp. 268–277, 1991.

[14] R. P. Kooi, *The optimization of queries in relational databases.* PhD thesis, Case Western Reserve University, 1980.

[15] G. Piatetsky-Shapiro and C. Connell, "Accurate estimation of the number of tuples satisfying a condition," in *Proceedings of ACM-SIGMOD Conference*, pp. 256–276, 1984.

[16] M. Muralikrishna and D. Dewitt, "Equi-depth histograms for estimating selectivity factors for multi-dimensional queries," in *Proceedings of ACM-SIGMOD Conference*, pp. 28–36, 1988.

[17] M. Mannino, P. Chu, and T. Sager, "Statistical profile estimation in database systems," in *ACM Computing Surveys*, vol. 20, pp. 192–221, 1988.

[18] S. Christodoulakis, "Estimating selectivities in data bases," in *Technical Report CSRG-136*, (Computer Science Dept, University of Toronto), 1981.

[19] B. J. Oommen and M. Thiyagarajah, "The Rectangular Attribute Cardinality Map: A New Histogram-like Technique for Query Optimization," in *International Database Engineering and Applications Symposium, IDEAS'99*, (Montreal, Canada), pp. 3–15, August 1999.

[20] B. J. Oommen and M. Thiyagarajah, "On the Use of the Trapezoidal Attribute Cardinality Map for Query Result Size Estimation," in *IDEAS-2000, the 2000 International Database Engineering and Applications Symposium*, (Yokohama, Japan), pp. 236–242, September 2000.

[21] W. Poosala, *Histogram Based Estimation Techniques in Databases.* PhD thesis, University of Wisconsin - Madison, 1997.

[22] B. Oommen and T. D. S. Croix, "Graph Partitioning Using Learning Automata," *IEEE Transactions on Computers*, vol. 45, no. 2, pp. 195–208, 1995.

[23] E. Deeba and A. Gunawardena, *Interactive Linear Algebra with MAPLE V.* Springer, 1997.

[24] B. J. Oommen and L. Rueda, "The Efficiency of Modern-day Histogram-like Techniques for Query Optimization," (Submitted for publication).