1998 International Computer Symposium
Workshop on Algorithms
December 17-19, 1998, N.C.K.U., Tainan, Taiwan, R.O.C.

DESIGNING PROXIES FOR STOCK MARKET INDICES IS COMPUTATIONALLY HARD

*Ming-Yang Kao**
Department of Computer Science
Yale University
New Haven, CT 06520

Stephen R. Tate†
Department of Computer Science
University of North Texas
Denton, TX 76203

Abstract

In this paper, we study the problem of designing proxies (or portfolios) for various stock market indices based on historical data. We use four different methods for computing market indices, all of which are formulas used in actual stock market analysis. For each index, we consider three criteria for designing the proxy: the proxy must either track the market index, outperform the market index, or perform within a margin of error of the index while maintaining a low volatility. In all twelve cases (all combinations of four indices with three criteria) we show that the problem is NP-hard, and hence most likely intractable.

1 Introduction

Market indices are widely used to track the performance of stocks or to design investment portfolios [1]. This paper initiates a rigorous mathematical study of the computational complexity of the art of designing proxies for such indices. There are several results on selecting such proxies (or portfolios) in an on-line manner (see, for example, [2] and [3]), we look at off-line algorithms for designing proxies based on historical data. In particular, we show that all combinations of three fundamental problems (such as tracking or outperforming a full market index) with four commonly-used indices give NP-complete problems, so are computationally hard.

To formally define market indices, let $\mathcal{B}$ be a set of b stocks in a market. Let $S_{i,t} \geq 0$ be the price of the i-th stock at time t. Let w_i be the number of outstanding shares of the i-th stock. We assume that w_i does not change with time. This paper discusses computational complexity issues regarding four kinds of market indices currently in use [1]. These indices are calculated by the following formulas, which can be multiplied by arbitrary constants to arrive at desired starting index values at time 0.

- The *price-weighted index* of $\mathcal{B}$ at time t is

$$\Phi_1(\mathcal{B}, t) = \frac{\sum_{i=1}^{b} S_{i,t}}{b}. \tag{1}$$

The Dow Jones Industrial Average is calculated in this manner for some $\mathcal{B}$ consisting of thirty stocks.

- The *value-weighted index* of $\mathcal{B}$ at time t is

$$\Phi_2(\mathcal{B}, t) = \frac{\sum_{i=1}^{b} w_i \cdot S_{i,t}}{\sum_{i=1}^{b} w_i \cdot S_{i,0}}.$$

The Standard & Poor's 500 is computed in this way with respect to 500 stocks.

- The *equal-weighted index* of $\mathcal{B}$ at time t is

$$\Phi_3(\mathcal{B}, t) = \sum_{i=1}^{b} \frac{S_{i,t}}{S_{i,0}}.$$

The index published by the Indicator Digest is calculated by this method, involving stocks listed on the New York Stock Exchange.

- The *price-relative index* of $\mathcal{B}$ at time t is

$$\Phi_4(\mathcal{B}, t) = \left(\Pi_{i=1}^{b} \frac{S_{i,t}}{S_{i,0}}\right)^{\frac{1}{b}}.$$

The Value Line Index is computed by this formula.

There are numerous reasons why stock investors and money managers would want to invest in a subset of stocks rather than those of a whole market [1]. For instance, small investors certainly do not have sufficient capital to invest in every stock in the market. Logically, such investors would attempt to choose a small subset of stocks which hopefully can perform roughly as well as or even outperform the market as a whole. They then face difficult trade-offs between returns and risks. For these and other reasons of optimization, we formulate

*Supported in part by NSF Grant CCR-9531028. E-mail: kao-ming-yang@cs.yale.edu

†Supported in part by NSF Grant CCR-9409945 and Texas Advanced Research Program Grant 1997-003594-019. E-mail: srt@cs.unt.edu

three natural computational problems for the design of market indices. Given a market $\mathcal{M}$ consisting of m stocks, we wish to choose a subset $\mathcal{M}_k$ of at most k stocks and calculate an index of $\mathcal{M}_k$, which is called a *k-proxy* of the corresponding index of the whole market $\mathcal{M}$ (we sometimes refer to $\mathcal{M}_k$ as a *portfolio*). Our goal is to choose $\mathcal{M}_k$ so that the resulting k-proxy tracks or outperforms the corresponding index of $\mathcal{M}$. This paper shows that designing proxies for the above four indices based on historical data is computationally hard.

2 Problem Formulations

In this section we formally define three basic problems related to selecting k-proxies, or portfolios.

Problem 1 (*tracking an index*)

> **Input:** A market $\mathcal{M}$ of m stocks, their prices $S_{i,t} \geq 0$ for $t = 0, \ldots, f$, their numbers w_i of outstanding shares, a real $\epsilon_1 > 0$, an integer $k > 0$, and some $j \in \{1,2,3,4\}$ to indicate the desired type of index.
>
> **Output:** A subset $\mathcal{M}_k$ of at most k stocks in $\mathcal{M}$ such that for all $t = 1, \ldots, f$
>
> $$\left|\frac{\Phi_j(\mathcal{M}_k,t)}{\Phi_j(\mathcal{M}_k,0)} - \frac{\Phi_j(\mathcal{M},t)}{\Phi_j(\mathcal{M},0)}\right| \leq \epsilon_1 \cdot \frac{\Phi_j(\mathcal{M},t)}{\Phi_j(\mathcal{M},0)}. \quad (2)$$

Problem 2 (*outperforming an index*)

> **Input:** A market $\mathcal{M}$ of m stocks, their prices $S_{i,t} \geq 0$ for $t = 0, \ldots, f$, their numbers w_i of outstanding shares, a real $\epsilon_2 \geq 0$, an integer $k > 0$, and some $j \in \{1,2,3,4\}$ to indicate the desired type of index.
>
> **Output:** A subset $\mathcal{M}_k$ of at most k stocks in $\mathcal{M}$ such that for all $t = 1, \ldots, f$
>
> $$\frac{\Phi_j(\mathcal{M}_k,t)}{\Phi_j(\mathcal{M}_k,0)} \geq (1+\epsilon_2) \cdot \frac{\Phi_j(\mathcal{M},t)}{\Phi_j(\mathcal{M},0)}. \quad (3)$$

For the final problem, we need a few extra definitions in order to analyze the *volatility* of a set of stocks. Let $\mathcal{B}$ be a set of stocks as defined in §1.

- The *one-period return* of Φ_j for $\mathcal{B}$ at time $t \geq 1$ is

$$R_j(\mathcal{B},t) = \ln \frac{\Phi_j(\mathcal{B},t)}{\Phi_j(\mathcal{B},t-1)}.$$

- The *average return* of Φ_j for $\mathcal{B}$ up to time $t \geq 1$ is

$$\overline{R}_j(\mathcal{B},t) = \frac{\sum_{i=1}^{t} R_j(\mathcal{B},i)}{t}.$$

- The *volatility* of Φ_j for $\mathcal{B}$ up to time $t \geq 2$ is

$$\Delta_j(\mathcal{B},t) = \sqrt{\frac{\sum_{i=1}^{t}\left(R_j(\mathcal{B},i) - \overline{R}_j(\mathcal{B},t)\right)^2}{t-1}}.$$

Problem 3 (*sacrificing return for less volatility*)

> **Input:** A market $\mathcal{M}$ of m stocks, their prices $S_{i,t} \geq 0$ for $t = 0, \ldots, f$, their numbers w_i of outstanding shares, two reals $\alpha, \beta > 0$, an integer $k > 0$, and some $j \in \{1,2,3,4\}$ to indicate the desired type of index.
>
> **Output:** A subset $\mathcal{M}_k$ of at most k stocks in $\mathcal{M}$ such that for all $t = 1, \ldots, f$ and $s = 2, \ldots, f$
>
> $$\frac{\Phi_j(\mathcal{M}_k,t)}{\Phi_j(\mathcal{M}_k,0)} \geq \alpha \cdot \frac{\Phi_j(\mathcal{M},t)}{\Phi_j(\mathcal{M},0)} \quad (4)$$
>
> and
>
> $$\Delta_j(\mathcal{M}_k,s) \leq \beta \cdot \Delta_j(\mathcal{M},s) \quad (5)$$
>
> In this problem, (4) is called the performance bound, and (5) is called the volatility bound.

3 Price-weighted Index

In this section, we consider taking the value of the market and portfolio using a price-weighted index, defined in (1). As given in the problem statements, we use the notation $\Phi_1(\mathcal{M},t)$ to denote the market average at timestep t, and $\Phi_1(\mathcal{M}_k,t)$ to denote the average of the portfolio at that timestep.

3.1 Tracking an index

To solve the problem of tracking the market average, we need to satisfy (2) using function $\Phi_1(\mathcal{B},t)$. We will refer to this bound as the "tracking bound." In the following proofs, we show this by proving an equivalent relation:

$$1-\epsilon \leq \frac{\Phi_1(\mathcal{M},0)}{\Phi_1(\mathcal{M}_k,0)} \cdot \frac{\Phi_1(\mathcal{M}_k,t)}{\Phi_1(\mathcal{M},t)} \leq 1+\epsilon. \quad (6)$$

Theorem 3.1 *Let ϵ be any error bound satisfying $0 < \epsilon < 1$ and specified using $n^{O(1)}$ bits in fixed point notation. Then the tracking problem for a price-weighted index with error bound ϵ is NP-hard.*

In the remainder of this section, we prove this theorem by reduction from the minimum set cover problem. We will use the notation from the minimum cover definition given in the classic book on NP-completeness by Garey and Johnson [4]: C is a collection of subsets of a finite set S, and K is the desired cover size.

Specifically, we want a subcollection $C' \subseteq C$ such that $|C'| \leq K$ and every item $x \in S$ is in some subset from C'.

Let $n = |C|$, and consider making an $n \times |S|$ matrix in which each column corresponds to a fixed item from S, and each row corresponds to a subset $S' \in C$. The element in row i, column j is some given value v_1 if the element in S for that column is in the subset S', and value v_0 if it is not. Then the minimum cover problem can be stated as follows: Is there a set of K rows such that the $K \times |S|$ matrix defined using only those rows has at least one entry with value v_1 in each column?

It makes sense now to consider this $n \times |S|$ matrix as an input to the portfolio selection problem, where each row corresponds to a security and each column corresponds to a timestep, and we are to choose a portfolio of size $k = K$. Selecting a portfolio is then equivalent to selecting the subcollection in the minimum cover problem. A subcollection that is missing some item from S corresponds to a portfolio in which some timestep has all values equal to v_0, and hence the portfolio average at that timestep must be v_0. Ideally, we would select v_0 and v_1 in such a way that the required tracking bound is met if any v_1 values are included in the portfolio, but not if all values are v_0. However, this simple construction has very unpredictable market averages at each time step, so we need a slightly more involved construction.

We will introduce a new row into our matrix called the "adjustment row", and we will select values to adjust the column averages to predictable values. To guarantee that this row is not selected in our portfolio (so selections are made up entirely of rows from the minimum cover problem), we introduce a special column called the "control column" — any selection including our adjustment row will violate the error bound in that column, and no selection excluding that row will violate the bound. In addition, we need to pad the problem out substantially. This is accomplished by including rows that contain value v_0 in every non-control column, which is equivalent to padding the original set cover problem instance with empty subsets added to C. This clearly has no effect on the set cover problem. Finally, we insert a column of all ones to give the $S_{i,0}$ values for the portfolio selection problem. The final matrix contains $m = 3n$ rows, $f = |S| + 1$ columns, and is depicted in Figure 1.

Note that since $S_{i,0} = 1$ for all i, $\Phi_1(\mathcal{M}, 0) = \Phi_1(\mathcal{M}_k, 0) = 1$, and so (6) reduces to just checking that

$$1 - \epsilon \leq \frac{\Phi_1(\mathcal{M}_k, t)}{\Phi_1(\mathcal{M}, t)} \leq 1 + \epsilon.$$

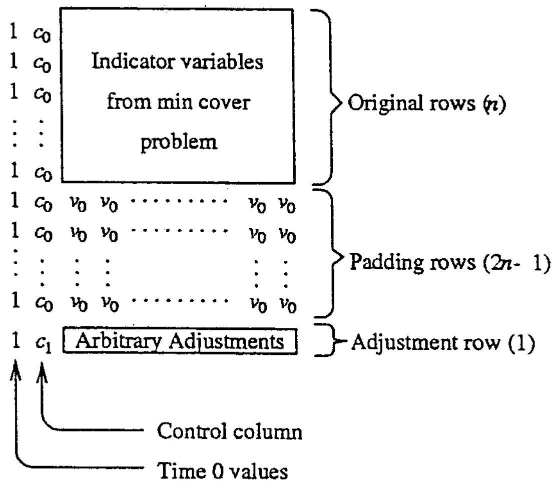

Figure 1: Pictorial depiction of reduction for Theorem 3.1

First we examine properties of the control column, where the values in that column are defined by

$$\begin{aligned} c_0 &= \left\lceil \frac{1-\epsilon}{\epsilon} \right\rceil, \\ c_1 &= c_0 + m. \end{aligned}$$

Lemma 3.1 *The tracking bound is met for the control column if and only if the adjustment row is not included in the portfolio.*

Proof: From the values for c_0 and c_1, it is clear that the average value of the control column is $c_0 + 1$. Since we will be examining the error of approximations relative to this average, we first note that we can bound (due to the ceiling involved in the definition of c_0)

$$\frac{\epsilon}{1+\epsilon} < \frac{1}{c_0 + 1} \leq \epsilon. \tag{7}$$

Any portfolio that does not include the adjustment row has average value c_0, and so we can lower bound the relative error by

$$\frac{\Phi_1(\mathcal{M}_k, t)}{\Phi_1(\mathcal{M}, t)} = \frac{c_0}{c_0 + 1} = 1 - \frac{1}{c_0 + 1} \geq 1 - \epsilon.$$

Since the relative error is clearly less than one, it falls into the acceptable range of values.

On the other hand, if a portfolio *does* include the adjustment row, then the portfolio average is $c_0 + m/k$, and so the relative error is

$$\frac{\Phi_1(\mathcal{M}_k, t)}{\Phi_1(\mathcal{M}, t)} = \frac{c_0 + m/k}{c_0 + 1} = 1 + \frac{m/k - 1}{c_0 + 1}.$$

Due to our padding of the problem, we know that $k \leq m/3$, and so $m/k - 1 \geq 2$. Using this observation and the bound from (7) leads to the conclusion that

$$\frac{\Phi_1(\mathcal{M}_k,t)}{\Phi_1(\mathcal{M},t)} \geq 1 + \frac{2}{c_0+1} > 1 + \frac{2}{1+\epsilon}\epsilon > 1+\epsilon.$$

In other words, any portfolio that includes the adjustment row will not meet the required error bound. Combined with our previous observation, this completes the proof of the lemma. ∎

Next we must define the values v_0 and v_1, and show the equivalence of our portfolio selection instance with the original set cover instance. To do so, define

$$\begin{aligned} \Delta &= \left\lceil \frac{1}{1-\epsilon} \right\rceil, \\ v_0 &= \left\lceil \frac{(k+1)(1-\epsilon)\Delta - 1}{\epsilon} \right\rceil, \\ v_1 &= v_0 + 2k\Delta. \end{aligned}$$

All values in the portfolio selection problem must be non-negative integers, and while these values are clearly integers, are they non-negative? Since $\Delta \geq \frac{1}{1-\epsilon}$, we see that $v_0 \geq \frac{k}{\epsilon} > 0$. Since v_1 is greater than v_0, it too is clearly non-negative.

For column t, if there are M_t rows with value v_1, then the value we use in the adjustment row for that column is

$$A_t = ((k+1)m - 2kM_t)\,\Delta + v_0.$$

The sum down the column is

$$\begin{aligned} &(m - M_t - 1)v_0 + M_t v_1 + A_t \\ &\quad = (m - M_t - 1)v_0 + M_t(v_0 + 2k\Delta) \\ &\qquad + (k+1)m\Delta - 2kM_t\Delta + v_0 \\ &\quad = mv_0 + (k+1)m\Delta, \end{aligned}$$

which means that the column average is $v_0 + (k+1)\Delta$. Notice the independence from t. We make such an adjustment for every column in the matrix.

Is such an adjustment possible? A_t is clearly an integer, and so this is a valid adjustment as long as $A_t \geq 0$. Since $M_t \leq \frac{m}{3}$, we know that $(k+1)m - 2kM_t \geq (k+1)m - 2k\frac{m}{3} = (\frac{k}{3}+1)m$, which is clearly positive, so $A_t \geq 0$. We have demonstrated that such a reduction is possible, so the next thing to demonstrate is the equivalence of the produced portfolio selection instance with the original set cover instance.

Lemma 3.2 *The relative error bound is met if and only if the portfolio contains at least one v_1 value in each column.*

Proof: Let t be an arbitrary column other than the control column, and recall that M_t represents the number of v_1 entries in column t. We first upper bound the approximation ratio for all values of M_t. In particular, we know that the maximum possible portfolio average is $v_1 = v_0 + 2k\Delta$, so we can bound

$$\frac{\Phi_1(\mathcal{M}_k,t)}{\Phi_1(\mathcal{M},t)} \leq \frac{v_0 + 2k\Delta}{v_0 + (k+1)\Delta} = 1 + \frac{(k-1)\Delta}{v_0 + (k+1)\Delta}. \quad (8)$$

We can lower-bound v_0 be removing the ceiling, giving a bound on the last fraction above:

$$\begin{aligned} \frac{(k-1)\Delta}{v_0 + (k+1)\Delta} &\leq \frac{(k-1)\Delta}{\frac{(k+1)(1-\epsilon)\Delta - 1}{\epsilon} + (k+1)\Delta} \\ &= \frac{(k-1)\Delta}{(k+1)\Delta - 1}\epsilon \leq \epsilon, \quad (9) \end{aligned}$$

where the last inequality uses the fact that $\Delta \geq 1$. Combining this with (8) gives $\frac{\Phi_1(\mathcal{M}_k,t)}{\Phi_1(\mathcal{M},t)} \leq 1+\epsilon$, which holds for all values of M_t.

Next, we lower bound the error when at least one row with a v_1 entry is selected (in other words, $M_t \geq 1$). In this case, the portfolio average is at least $v_0 + \frac{1}{k}2k\Delta = v_0 + 2\Delta$, and so we derive

$$\frac{\Phi_1(\mathcal{M}_k,t)}{\Phi_1(\mathcal{M},t)} \geq \frac{v_0 + 2\Delta}{v_0 + (k+1)\Delta} = 1 - \frac{(k-1)\Delta}{v_0 + (k+1)\Delta}.$$

Notice that this results in exactly the same fraction as above, so we can use (9) to give $\frac{\Phi_1(\mathcal{M}_k,t)}{\Phi_1(\mathcal{M},t)} \geq 1-\epsilon$, when at least one row containing v_1 is selected.

What we have shown is that any time at least one row containing v_1 is selected, the portfolio average tracks the total market average within a relative error of ϵ. We next show that this bound is not met when no rows containing v_1 are selected. In this case, the portfolio average is exactly v_0, which results in

$$\frac{\Phi_1(\mathcal{M}_k,t)}{\Phi_1(\mathcal{M},t)} = \frac{v_0}{v_0 + (k+1)\Delta} = 1 - \frac{(k+1)\Delta}{v_0 + (k+1)\Delta}. \quad (10)$$

This last fraction can be bounded by first upper bounding v_0: just remove the ceiling and add 1 (note that this gives a *strict* upper bound). Thus

$$\begin{aligned} \frac{(k+1)\Delta}{v_0 + (k+1)\Delta} &> \frac{(k+1)\Delta}{\frac{(k+1)(1-\epsilon)\Delta - 1 + \epsilon}{\epsilon} + (k+1)\Delta} \\ &= \frac{(k+1)\Delta}{(k+1)\Delta - (1-\epsilon)}\epsilon > \epsilon, \end{aligned}$$

where the last inequality comes from the fact that $\epsilon < 1$. Using this bound in (10) gives $\frac{\Phi_1(\mathcal{M}_k,t)}{\Phi_1(\mathcal{M},t)} < 1-\epsilon$ whenever none of the selected rows contain v_1. In other words, the error bound is not met when no such rows are selected. ∎

As a final note, it is fairly easy to show that all values in the constructed portfolio selection problem have length polynomial in the length of the original set cover problem and the number of bits used to specify ϵ. Therefore, these values form a polynomial time reduction from the set cover problem to the portfolio selection problem, which completes the proof of Theorem 3.1.

3.2 Sacrificing Return for Less Volatility

Next, we will skip Problem 2 and prove a hardness result for Problem 3: sacrificing return for less volatility. In the following section, we will return to problem 2, and show that the hardness of that problem (outperforming an index) follows directly from the results of this section.

As in §3.1, we will show that Problem 3 is NP-complete by reducing the minimum cover problem to this one.

3.2.1 The construction

The main reduction for this proof involves a problem constructed from a minimum cover instance, and this construction is illustrated in Figure 2. This constructed problem is an instance of our portfolio selection problem where the rows represent different securities, the columns represent times, and the values in the matrix represent prices.

In the original minimum cover instance, let $n = |C|$ represent the number of subsets in the input, let $|S|$ represent the size of the overall set, and let K be the number of subsets we are allowed to select. The data from this problem can be encoded into an $n \times |S|$ matrix M, where the values in this matrix are set as follows (v_2 is a value that will be defined shortly):

$$M_{ij} = \begin{cases} v_2 & \text{if subset } i \text{ contains element } j; \\ 0 & \text{otherwise.} \end{cases}$$

We will need a larger matrix in order to complete the reduction, so we embed matrix M into our larger matrix — in Figure 2 the embedded matrix is labeled as the "Coding Region". This gives a portfolio selection problem with m securities, $f = P + |S|$ time steps, and portfolio size $k = K$.

We surround matrix M with various "padding rows" and "padding columns". The number of padding rows and padding columns are defined as follows:

- There are $P + 1$ padding columns, where $P = \max(2(k+1), 2|S|)$.
- The total number of rows is defined in terms of the following constants:

$$q = \lceil \max(1 + (4/\beta), \log_k(2/\alpha)) \rceil ,$$

and

$$B = \lceil \alpha k^q \rceil .$$

The total number of rows is $m = nB$.

The definition of q implies some important properties of the constant B that we note here:

$$B \geq 2; \tag{11}$$

$$B \geq k\alpha \geq \alpha. \tag{12}$$

Finally, from the first part of (12) we can derive

$$\left\lfloor \frac{B}{\alpha} \right\rfloor > \frac{B}{\alpha} \frac{k-1}{k}. \tag{13}$$

All of the first n rows in the padding columns are filled with value v_1, and value v_2 is used in the coding region as previously described. These values are defined in terms of the constant B as follows:

- $v_1 = B - 1$
- $v_2 = k(B - 1)$

Each column may have an "adjustment value", denoted by A_t for column t. Odd numbered columns in the padding region (type-2 columns) do not have an adjustment value, but even numbered columns other than column 0 (type-1 columns) do, and these values are positioned at successively lower rows; therefore, if column t is a type-1 column, then A_t is placed in row $n + \frac{t}{2}$. If we run out of rows before completing this placement, simply put all remaining adjustment values on the last row. Notice that since $P \geq 2(k+1)$ there are at least $k + 1$ type-1 padding columns, and since the number of padding rows is $(m - n) = (nB - n) \geq n \geq k + 1$ (using (11)), there must be at least $k + 1$ distinct rows that contain adjustment values. Columns that cross the coding region (called "coding columns") also have adjustment values, which are all placed on the last row of the matrix (see Figure 2). The adjustment values to be used are defined below, where z_t is the number of zeros in the coding region of column t:

$$A_t = \begin{cases} (m-n)\left(\lfloor \frac{B}{\alpha} \rfloor - 1\right) & \text{if } 0 < t \leq P, t \text{ is even;} \\ (m-n)\left(\lfloor \frac{B}{\alpha} \rfloor - k\right) + z_t \cdot v_2 & \text{if } t > P. \end{cases}$$

Note that the adjustment values in the padding columns are all the same, but the adjustments in the

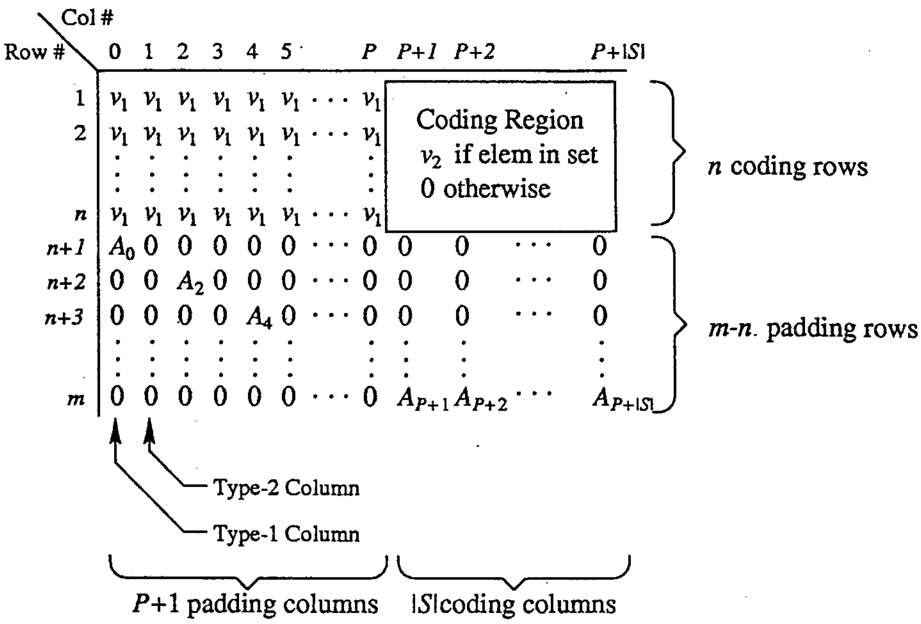

Figure 2: Construction for main reduction

coding region depend on the data in the coding region. Furthermore, (12) guarantees that these adjustment values are all non-negative.

Before analyzing the return and volatility of the constructed portfolio selection problem, we state the following lemma regarding the size of the constructed problem, showing that we have a polynomial reduction.

Lemma 3.3 *If α and β are expressed using $n^{O(1)}$ bits in fixed-point binary notation, and $0 < \alpha \leq n^{O(1)}$ and $\beta = \Omega\left(\frac{\log k}{\log n}\right)$, then the size of the constructed problem (including the size of the values in the matrix) is polynomial in the size of the original minimum cover problem.*

3.2.2 Guarantees on Return

Lemma 3.4 *The performance bound is met for all columns if and only if the selected portfolio contains exactly k items from the coding rows and each coding column has at least one v_2 value from among the selected rows.*

Proof: We will first prove that if the selected portfolio contains exactly k items from the coding rows and each coding column has at least one v_2 value from the selected rows, then the performance bound is met. First consider a padding column t — since the k selected rows are all coding rows, all selected values for any padding column have value v_1, and so the portfolio average for that column is $\Phi_1(\mathcal{M}_k, t) = v_1$. On the other hand, the market average is different for the two types of columns. If column t is a type-1 padding column, then the sum of all the values in the column is

$$\begin{aligned} nv_1 + A_t &= n(B-1) + (m-n)\left(\left\lfloor \frac{B}{\alpha} \right\rfloor - 1\right) \\ &= n(B-1) + (nB-n)\left(\left\lfloor \frac{B}{\alpha} \right\rfloor - 1\right) \\ &= n(B-1) + (B-1)\left(n\left\lfloor \frac{B}{\alpha} \right\rfloor - n\right) \\ &= (B-1)n\left\lfloor \frac{B}{\alpha} \right\rfloor. \end{aligned}$$

Therefore, the market average for column t satisfies

$$\begin{aligned} \Phi_1(\mathcal{M}, t) &= \frac{(B-1)n}{nB}\left\lfloor \frac{B}{\alpha} \right\rfloor = \frac{B-1}{B}\left\lfloor \frac{B}{\alpha} \right\rfloor \quad (14) \\ &\leq \frac{B-1}{B}\frac{B}{\alpha} = \frac{B-1}{\alpha} = \frac{v_1}{\alpha}. \end{aligned}$$

Furthermore, any type-2 padding column has no adjustment value, which makes the market average smaller than a type-1 column. Therefore, for either type of padding column the bound $\Phi_1(\mathcal{M}, t) \leq \frac{v_1}{\alpha}$ is valid, and so it immediately follows that for any padding column t, since $\Phi_1(\mathcal{M}, 0) = \Phi_1(\mathcal{M}_k, 0) = v_1$,

$$\frac{\Phi_1(\mathcal{M}_k, t)}{\Phi_1(\mathcal{M}_k, 0)} \geq \alpha \cdot \frac{\Phi_1(\mathcal{M}, t)}{\Phi_1(\mathcal{M}, 0)}.$$

Therefore, the performance bound is met for all padding columns.

Now consider a coding column t, and recall that we are assuming that at least one v_2 value from column t is included in the portfolio. This means that the portfolio average is $\Phi_1(\mathcal{M}_k, t) \geq v_2/k = v_1$. For the market average, we compute the sum over all values in the column, as we did before, and in this case we get

$$\begin{aligned}
&(n - z_t)v_2 + A_t \\
&= nv_2 - z_t v_2 + (m - n)\left(\left\lfloor \frac{B}{\alpha} \right\rfloor - k\right) + z_t v_2 \\
&= nk(B-1) + (nB - n)\left(\left\lfloor \frac{B}{\alpha} \right\rfloor - k\right) \\
&= nk(B-1) + (B-1)\left(n\left\lfloor \frac{B}{\alpha} \right\rfloor - nk\right) \\
&= (B-1)n\left\lfloor \frac{B}{\alpha} \right\rfloor .
\end{aligned}$$

Similar to the calculation for the padding columns, this gives us

$$\Phi_1(\mathcal{M}, t) = \frac{B-1}{B}\left\lfloor \frac{B}{\alpha} \right\rfloor \leq \frac{B-1}{\alpha} = \frac{v_1}{\alpha}, \qquad (15)$$

which implies that

$$\frac{\Phi_1(\mathcal{M}_k, t)}{\Phi_1(\mathcal{M}_k, 0)} \geq \alpha \cdot \frac{\Phi_1(\mathcal{M}, t)}{\Phi_1(\mathcal{M}, 0)},$$

and so the performance bound is met for the coding columns as well. Therefore we have completed this direction of the proof.

For the other direction, we need to show that any portfolio that meets the performance bound must be made up of exactly k items from the coding rows and each coding column has at least one v_2 value from the selected rows. We first show that any portfolio that meets the performance bound may only use coding rows. By our placement of adjustment values, we noticed before that there are at least $k+1$ distinct padding rows that contain adjustment values. Therefore, there must be at least one type-1 padding column, say column t, that does not have its adjustment value A_t selected as part of the portfolio. Now if all k selections are *not* from the coding rows, then we can bound the portfolio average for column t by

$$\Phi_1(\mathcal{M}_k, t) \leq \frac{(k-1)v_1}{k}.$$

Since this is a type-1 column, (14) gives the market average, and we can further use (13) to conclude that

$$\frac{\Phi_1(\mathcal{M}_k, t)}{\Phi_1(\mathcal{M}_k, 0)} \frac{\Phi_1(\mathcal{M}, 0)}{\Phi_1(\mathcal{M}, t)} \leq \frac{\frac{(k-1)v_1}{k}}{v_1} \frac{v_1}{\frac{(B-1)}{B}\left\lfloor \frac{B}{\alpha} \right\rfloor}$$

$$< \frac{(k-1)(B-1)}{k} \frac{1}{\frac{(B-1)}{B}\frac{B}{\alpha}\frac{k-1}{k}} = \alpha,$$

and so the performance bound would not be met. Therefore, all k row selections must come from the coding rows.

Since we have established that all k selections must come from the coding rows, we will next show that every column in the coding region must have at least one v_2 value among the selected rows. This is, in fact, very easy to see — if no v_2 values are selected in a particular column, then the portfolio average is zero, which cannot meet the performance bound for that column. Therefore, all coding columns must be contain at least one v_2 value, which completes this direction of the proof, and also completes the entire proof. ∎

3.2.3 Guarantees on Volatility

Lemma 3.5 *If the performance bound is met for our constructed portfolio selection problem, then the volatility bound is met as well.*

3.2.4 The main result

Theorem 3.2 *Let α and β be values expressed using $n^{O(1)}$ bits in fixed-point binary notation, and satisfying $0 < \alpha \leq n^{O(1)}$ and $\beta = \Omega\left(\frac{\log k}{\log n}\right)$. Then the problem of sacrificing return for less volatility using the price-weighted index is NP-complete.*

Proof: Follows from Lemmas 3.3, 3.4, and 3.5. ∎

3.3 Outperforming an index

Theorem 3.3 *Let ϵ be any value satisfying $0 < \epsilon < n^c$ for some constant c. Then the problem of outperforming the market average using the price-weighted index with bound ϵ is NP-hard.*

4 Other Indices

For the value-weighted and equal-weighted indices, we will, in fact, use the exact same constructions as in the previous section — the prices in the constructed problem have been selected carefully so that they work using related indices, such as the value-weighted and equal-weighted indices. The results will follow fairly easily from the following lemma.

Lemma 4.1 *Let $\Phi_j(\mathcal{B}, t)$ be an index function where $S_{i,0} = c$ for some constant c implies that*

$$\frac{\Phi_j(\mathcal{B}, t)}{\Phi_j(\mathcal{B}, 0)} = d \cdot \Phi_1(\mathcal{B}, t)$$

for all sets of stocks $\mathcal{B} \subseteq \mathcal{M}$, where d is a constant that does not depend on $\mathcal{B}$ or t, then all of the previous NP-completeness results hold for index $\Phi_j(\mathcal{B}, t)$.

4.1 The Value-Weighted Index

We first apply this lemma to the value-weighted index. For the value-weighted index, we must indicate the weights (the w_i's) in the constructed portfolio selection problem as well as the prices. In all of our constructions, we will pick $w_i = 1$ for all i.

If $S_{i,0} = c$ for some constant c, then for any valid time t and any set of stocks $\mathcal{B}$, using $w_i = 1$ gives

$$\begin{aligned}\Phi_2(\mathcal{B},t) &= \frac{\sum_{i=1}^{b} w_i \cdot S_{i,t}}{\sum_{i=1}^{b} w_i \cdot S_{i,0}} = \frac{\sum_{i=1}^{b} S_{i,t}}{\sum_{i=1}^{b} c} \\ &= \frac{\sum_{i=1}^{b} S_{i,t}}{b\,c} = \frac{1}{c}\Phi_1(\mathcal{B},t).\end{aligned}$$

Furthermore, regardless of $\mathcal{B}$ we have $\Phi_2(\mathcal{B},0) = 1$, and so Lemma 4.1 holds with constant $d = \frac{1}{c}$. The following three theorems are a direct consequence of this Lemma.

Theorem 4.1 *Let ϵ be any error bound satisfying $0 < \epsilon < 1$ and specified using $n^{O(1)}$ bits in fixed point notation. Then the tracking problem for a value-weighted index with error bound ϵ is NP-hard.*

Theorem 4.2 *Let ϵ be any value satisfying $0 < \epsilon < n^c$ for some constant c. Then the problem of outperforming the market average using the value-weighted index with bound ϵ is NP-hard.*

Theorem 4.3 *Let α and β be values expressed using $n^{O(1)}$ bits in fixed-point binary notation, and satisfying $0 < \alpha \leq n^{O(1)}$ and $\beta = \Omega\left(\frac{\log k}{\log n}\right)$. Then the problem of sacrificing return for less volatility using the value-weighted index is NP-complete.*

4.2 The Equal-Weighted Index

If $S_{i,0} = c$ for all i,

$$\begin{aligned}\Phi_3(\mathcal{B},t) &= \sum_{i=1}^{b} \frac{S_{i,t}}{S_{i,0}} = \sum_{i=1}^{b} \frac{S_{i,t}}{c} \\ &= \frac{1}{c}\sum_{i=1}^{b} S_{i,t} = \frac{b}{c}\,\Phi_1(\mathcal{B},t).\end{aligned}$$

It's easy to see that $\Phi_3(\mathcal{B},0) = b$, so

$$\frac{\Phi_3(\mathcal{B},t)}{\Phi_3(\mathcal{B},0)} = \frac{1}{c}\,\Phi_1(\mathcal{B},t),$$

and so Lemma 4.1 applies with constant $d = \frac{1}{c}$. The following three theorems are direct consequences of that Lemma.

Theorem 4.4 *Let ϵ be any error bound satisfying $0 < \epsilon < 1$ and specified using $n^{O(1)}$ bits in fixed point notation. Then the tracking problem for a equal-weighted index with error bound ϵ is NP-hard.*

Theorem 4.5 *Let ϵ be any value satisfying $0 < \epsilon < n^c$ for some constant c. Then the problem of outperforming the market average using the equal-weighted index with bound ϵ is NP-hard.*

Theorem 4.6 *Let α and β be values expressed using $n^{O(1)}$ bits in fixed-point binary notation, and satisfying $0 < \alpha \leq n^{O(1)}$ and $\beta = \Omega\left(\frac{\log k}{\log n}\right)$. Then the problem of sacrificing return for less volatility using the equal-weighted index is NP-complete.*

4.3 The Price-Relative Index

Theorem 4.7 *Let ϵ be any error bound satisfying $0 < \epsilon < 1$ and specified using $O(\log n)$ bits in fixed point notation. Then the tracking problem for a price-relative index with error bound ϵ is NP-hard.*

Theorem 4.8 *Let ϵ be any value satisfying $0 < \epsilon < n^c$ for some constant c. Then the problem of outperforming the market average using the price-relative index with bound ϵ is NP-hard.*

Theorem 4.9 *Let α and β be values expressed using $n^{O(1)}$ bits in fixed-point binary notation, and satisfying $0 < \alpha \leq n^{O(1)}$ and $\beta = \Omega\left(\frac{\log k}{\log n}\right)$. Then the problem of sacrificing return for less volatility using the price-relative index is NP-complete.*

References

[1] G. J. Alexander, W. F. Sharpe, and J. V. Bailey, *Fundamentals of Investments*, Prentice-Hall, Upper Saddle River, NJ, 2nd ed., 1993.

[2] T. M. Cover, *Universal portfolios*, Mathematical Finance, 1 (1991), pp. 1–29.

[3] T. M. Cover and E. Ordentlich, *Universal portfolios with side information*, IEEE Transactions on Information Theory, 42 (1996), pp. 348–363.

[4] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman and Company, New York, 1979.