How to apply to college

Max Kapur

February 25, 2022

Abstract

This paper considers the maximization of the expected maximum value of a portfolio of random variables subject to a budget constraint. We refer to this as the optimal college application problem. When each variable's cost, or each college's application fee, is identical, we show that the optimal portfolios are nested in the budget constraint, yielding an exact polynomial-time algorithm. When colleges differ in their application fees, we show that the problem is NP-complete. We provide two dynamic programs for this more general problem. The first produces an exact solution in pseudopolynomial time, and the second yields a fully polynomial-time approximation scheme.

Contents

1 Introduction

We consider the optimal college application strategy for a single student. As Chao (2014) remarked, this represents a somewhat subtle portfolio optimization problem. The traditional Markowitz model trades off the expected value across all assets with a risk term, obtaining a concave maximization problem with linear constraints. But college applicants maximize the observed value of their *best* asset: If a student is admitted to her jth choice, then she is indifferent to whether or not she gets into her $(j+1)$th choice. As a result, the valuation function that students maximize is *convex* in the expected utility associated with individual applications. Risk management is implicit in the college application problem because, in a typical admissions market, college preferability is negatively correlated with competitiveness. Thus, students must negotiate a tradeoff between highly attractive, selective "reach schools" and less preferable "safety schools" where admission is a safer bet. Finally, the combinatorial nature of the college application problem makes it difficult to solve using the gradient-based techniques used in continuous portfolio optimization.

Chao estimated her model (which considers application as a *cost* rather than a constraint) by clustering the schools so that $m = 8$, a scale at which enumeration is possible. However, subject to certain assumptions on the quality of the data available to students in their decision-making process, an optimal application portfolio for a single student can be computed in time polynomial in h and m, as we show presently.

2 Model

3 Homogeneous application costs

Consider a college admissions market with m schools, $\mathcal{C} = \{1 \dots m\}$. The jth school is named c_j. By government regulation, students are allowed to apply to no more than h schools. (In the Korean case, $m = 202$ and $h = 3$.) We consider the optimal application strategy for a single student, whom we will call Alma.

For $j = 1 \dots m$, let $t_j > 0$ denote the utility that Alma receives from attending c_j, and let f_j denote the probability that she is admitted if she applies. Let the random variable Z_j equal one if Alma gets into c_j and zero otherwise. We assume that Alma's admissions outcome at each school is independent of her outcome at the other schools.[1] Thus Z is a vector of independent Bernoulli variables with probabilities given by f. Let c_0 denote Alma's outcome if she does not get into any college, with utility t_0 and $f_0 = 1$. Without loss of generality, we assume that $t_{j-1} \leq t_j$ for $j = 1 \dots m$.(We may assume that $t_0 \leq t_1$ because schools for which $t_j < t_0$ can be trivially excluded from consideration.)

Let $\mathcal{X} \subseteq \mathcal{C}$ denote the set of schools to which Alma applies, called her *application portfolio.*The expected utility Alma receives from $\mathcal{X}$ is called the portfolio's *valuation*.

Definition 1 (Portfolio valuation function). $v(\mathcal{X}) = \mathrm{E}\left[\max\{t_j Z_j : j \in \mathcal{X}\}\right]$.

[1] This assumption is appropriate when f gives the admissions probabilities *specifically* for Alma, as estimated from her academic abilities. It is inappropriate to take a school's overall acceptance rate, which reflects the admissions probability for a *generic* student, as f_j.

It is helpful to define the random variable $X = \max\{t_j Z_j : j \in \mathcal{X}\}$ as the utility achieved by the portfolio, so that $v(\mathcal{X}) = \mathrm{E}[X]$.

Let $p_j(\mathcal{X})$ denote the probability that Alma attends c_j. Alma attends c_j if and only if she *applies* to c_j, is *admitted* to c_j, and is *rejected* from any school she prefers to c_j; that is, any school with higher index. Hence, for $j = 0 \dots m$,

$$p_j(\mathcal{X}) = \begin{cases} f_j \prod_{\substack{j' \in \mathcal{X}: \\ j' > j}} (1 - f_{j'}), & j \in \{0\} \cup \mathcal{X} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$\tag{2}$$

where the empty product equals one. The following proposition follows immediately.

Proposition 1 (Closed form of portfolio valuation function)**.**

$$v(\mathcal{X}) = \sum_{j \in \{0\} \cup \mathcal{X}} \Big(t_j f_j \prod_{\substack{j' \in \mathcal{X}: \\ j' > j}} (1 - f_{j'}) \Big) \tag{3}$$

Next, we show that without loss of generality, we may assume that $t_0 = 0$ (or any constant).

Theorem 1. *Let $\bar{t}_j = t_j - \gamma$ for $j = 0 \dots m$. Then $v(\mathcal{X}; \bar{t}_j) = v(\mathcal{X}; t_j) - \gamma$ regardless of $\mathcal{X}$.*

Proof. By definition, $\sum_{j=0}^{m} p_j(\mathcal{X}) = \sum_{j \in \{0\} \cup \mathcal{X}} p_j(\mathcal{X}) = 1$. Therefore

$$v(\mathcal{X}; \bar{t}_j) = \sum_{j \in \{0\} \cup \mathcal{X}} \bar{t}_j p_j(\mathcal{X}) = \sum_{j \in \{0\} \cup \mathcal{X}} (t_j - \gamma) p_j(\mathcal{X}) \tag{4}$$

$$= \sum_{j \in \{0\} \cup \mathcal{X}} t_j p_j(\mathcal{X}) - \gamma = v(\mathcal{X}; t_j) - \gamma \tag{5}$$

which completes the proof. □

Definition 2 (Alma's problem)**.** Alma's optimal college application portfolio is given by the solution to the following integer nonlinear program:

$$\begin{aligned} \text{maximize} \quad & v(\mathcal{X}) \sum_{j \in \{0\} \cup \mathcal{X}} \Big(t_j f_j \prod_{\substack{j' \in \mathcal{X}: \\ j' > j}} (1 - f_{j'}) \\ \text{subject to} \quad & \mathcal{X} \subseteq \mathcal{C}, \quad |\mathcal{X}| \leq h \end{aligned} \tag{6}$$

3.1 Approximation properties of a naïve solution

Notice that for a given school c_j, the expected utility associated with applying to c_j is simply $\mathrm{E}[t_j Z_j] = t_j f_j$. It is therefore tempting to adopt the following greedy strategy, which turns out to be inoptimal.

Definition 3 (Naïve algorithm for Alma's problem)**.** Apply to the h schools having the highest expected utility $t_j f_j$.

The basic error of this algorithm is that it maximizes $\mathrm{E}\left[\sum t_j Z_j\right]$ instead of $\mathrm{E}\left[\max\{t_j Z_j\}\right]$. The latter is what Alma is truly concerned with, since in the end she can attend only one school. The following example shows that the naïve algorithm can produce a suboptimal solution.

Example 1. Suppose $m = 3$, $q = 2$, and

$$\begin{aligned} t &= (70, 80, 90) \\ f &= (0.4, 0.4, 0.3) \\ \implies t * f &= (28, 32, 27) \end{aligned}$$

The naïve algorithm picks $\mathcal{T} = \{1, 2\}$ with

$$v(\mathcal{T}) = 70(0.4)(1 - 0.4) + 80(0.4) = 48.8$$

But $\mathcal{X} = \{2, 3\}$ with

$$v(\mathcal{X}) = 80(0.4)(1 - 0.3) + 90(0.3) = 49.4$$

is the optimal solution.

In fact, the naïve algorithm is a $(1/h)$-approximation algorithm for Alma's problem, as expressed in the following theorem.

Theorem 2. *When the application limit is h, let $\mathcal{X}_h$ denote the optimal portfolio, and $\mathcal{T}_h$ the set of the h schools having the largest values of $t_j f_j$. Then $v(\mathcal{T}_h)/v(\mathcal{X}_h) \geq 1/h$.*

Proof. Because $\mathcal{T}_h$ maximizes the quantity $\mathrm{E}\big[\sum_{j \in \mathcal{T}_h} \{t_j Z_j\}\big]$, we have

$$\begin{aligned} v(\mathcal{X}_h) &= \mathrm{E}\big[\max_{j \in \mathcal{X}_h} \{t_j Z_j\}\big] \leq \mathrm{E}\big[\sum_{j \in \mathcal{X}_h} \{t_j Z_j\}\big] \leq \mathrm{E}\big[\sum_{j \in \mathcal{T}_h} \{t_j Z_j\}\big] \\ &= h\,\mathrm{E}\big[\tfrac{1}{h} \sum_{j \in \mathcal{T}_h} \{t_j Z_j\}\big] \leq h\,\mathrm{E}\big[\max_{j \in \mathcal{T}_h} \{t_j Z_j\}\big] = h\, v(\mathcal{T}_h) \end{aligned} \tag{7}$$

where the final inequality follows from the concavity of the $\max\{\}$ operator. $\square$

The following example establishes the tightness of the approximation factor.

Example 2. Pick any h and let $m = 2h$. For a small constant $\varepsilon \in (0, 1)$, let

$$\begin{aligned} t &= \Big(\underbrace{1, \ldots, 1}_{h}, \underbrace{\varepsilon^{-1}, \varepsilon^{-2}, \ldots, \varepsilon^{-(h-1)}, \varepsilon^{-h}}_{h}\Big) \\ \text{and} \quad f &= \Big(\underbrace{1, \ldots, 1}_{h}, \underbrace{\varepsilon^{1}, \varepsilon^{2}, \ldots, \varepsilon^{h-1}, \varepsilon^{h}}_{h}\Big) \end{aligned}$$

Since all $t_j f_j = 1$, the naïve algorithm can choose $\mathcal{T}_h = \{1, \ldots, h\}$, with $v(\mathcal{T}_h) = 1$. But the optimal solution is $\mathcal{X}_h = \{h + 1, \ldots, m\}$, with

$$v(\mathcal{X}_h) = \sum_{j=h+1}^{m} \Big(t_j f_j \prod_{j'=j+1}^{m} (1 - f_{j'})\Big) = \sum_{j=1}^{h} (1 - \varepsilon)^j \approx h$$

Thus, as ϵ approaches zero, we have $v(\mathcal{T}_h)/v(\mathcal{X}_h) \to 1/h$. (The optimality of $\mathcal{X}_h$ follows from the fact that it achieves the upper bound of Theorem 7.)

Hope is not lost. We can still find the optimal solution in time polynomial in h and m, as we will now show.

3.2 The nestedness property

It turns out that the solution to Alma's problem possesses a special structure: An optimal portfolio of size $h+1$ includes an optimal portfolio of size h as a subset. To show this result, we need to introduce some new notation. For an application portfolio $\mathcal{X}$, the random variable $X = \max\{t_j Z_j : j \in \mathcal{X}\}$ represents the utility achieved by the portfolio, so that $v(\mathcal{X}) = \mathrm{E}[X]$.

Theorem 3 (Nestedness of optimal application portfolios)**.** *Let $\mathcal{X}_h$ denote Alma's optimal application portfolio when the application limit is h. If each $\mathcal{X}_h$ is unique, then*

$$\mathcal{X}_1 \subset \mathcal{X}_2 \subset \cdots \subset \mathcal{X}_m \tag{8}$$

If the optimal portfolios are not unique, then there is a sequence of optimal portfolios satisfying the above.

Proof. By induction on h. Applying Theorem 1, we assume that $t_0 = 0$.

(Base case.) First, we will show that $\mathcal{X}_1 \subset \mathcal{X}_2$. To get a contradiction, suppose that the optima are $\mathcal{X}_1 = \{j\}$ and $\mathcal{X}_2 = \{k, l\}$, where we may assume that $t_k \leq t_l$. Optimality requires that

$$v(\mathcal{X}_1) = f_j t_j > v(\{k\}) = f_k t_k \tag{9}$$

and

$$\begin{aligned} v(\mathcal{X}_2) = f_k(1-f_l)t_k + f_l t_l &> v(\{j,l\}) && (10)\\ &= f_j(1-f_l)t_j + (1-f_j)f_l t_l + f_j f_l \max\{t_j, t_l\} && (11)\\ &\geq f_j(1-f_l)t_j + (1-f_j)f_l t_l + f_j f_l t_l && (12)\\ &= f_j(1-f_l)t_j + f_l t_l && (13)\\ &\geq f_k(1-f_l)t_k + f_l t_l = v(\mathcal{X}_2) && (14) \end{aligned}$$

which is a contradiction.

(Inductive step.) Assume that $\mathcal{X}_1 \subset \cdots \subset \mathcal{X}_h$, and we will show $\mathcal{X}_h \subset \mathcal{X}_{h+1}$. Let $k = \arg\max\{t_k : k \in \mathcal{X}_{h+1}\}$ and write $\mathcal{X}_{h+1} = \mathcal{Y}_h \cup \{k\}$.

Suppose $k \notin \mathcal{X}_h$. To get a contradiction, assume that $v(\mathcal{Y}_h) < v(\mathcal{X}_h)$. Then

$$\begin{aligned} v(\mathcal{X}_{h+1}) &= v(\mathcal{Y}_h \cup \{k\})\\ &= (1-f_k)v(\mathcal{Y}_h) + f_k t_k\\ &< (1-f_k)v(\mathcal{X}_h) + f_k \mathrm{E}\big[\max\{t_k, X_h\}\big]\\ &= v(\mathcal{X}_h \cup \{k\}) \end{aligned} \tag{15}$$

contradicts the optimality of $\mathcal{X}_{h+1}$.

Now suppose that $k \in \mathcal{X}_h$. We can write $\mathcal{X}_h = \mathcal{Y}_{h-1} \cup \{k\}$, where $\mathcal{Y}_{h-1}$ is some portfolio of size $h-1$. It suffices to show that $\mathcal{Y}_{h-1} \subset \mathcal{Y}_h$. By definition, $\mathcal{Y}_{h-1}$ (respectively, $\mathcal{Y}_h$) maximizes the function $v(\mathcal{Y} \cup \{k\})$ over portfolios of size $h-1$ (respectively, h) that do not include k. That is, $\mathcal{Y}_{h-1}$ and $\mathcal{Y}_h$ are the optimal *complements* to the singleton portfolio $\{k\}$.

We will use the function $w(\mathcal{Y})$ to grade portfolios $\mathcal{Y} \subseteq \mathcal{C} \setminus \{k\}$ according to how well they complement $\{k\}$. To construct $w(\mathcal{Y})$, let $\tilde{t}_j$ denote the expected utility Alma receives from school c_j *given* that she has been admitted to c_j and applied to c_k. For $j < k$, including $j = 0$, this is $\tilde{t}_j = t_j(1 - f_k) + t_k f_k$; for $j > k$, this is $\tilde{t}_j = t_j$. This means that

$$v(\mathcal{Y} \cup \{k\}) = \sum_{j \in \{0\} \cup \mathcal{Y}} \tilde{t}_j p_j(\mathcal{Y}) \tag{16}$$

The transformation to $\tilde{t}$ does not change the order of the t_j-values. Therefore, the expression on the right side of (16) is itself a portfolio valuation function. In the corresponding market, t is replaced by $\tilde{t}$ and $\mathcal{C}$ is replaced by $\mathcal{C} \setminus \{k\}$. Now, we obtain $w(\mathcal{Y})$ through one more transformation: Define $\bar{t}_j = \tilde{t}_j - \tilde{t}_0$ so that $t_0 = 0$ and let

$$w(\mathcal{Y}) = \sum_{j \in \{0\} \cup \mathcal{Y}} \bar{t}_j p_j(\mathcal{Y}) = \sum_{j \in \{0\} \cup \mathcal{Y}} \tilde{t}_j p_j(\mathcal{Y}) - \tilde{t}_0 = v(\mathcal{Y} \cup \{k\}) - t_k f_k \tag{17}$$

where the second equality follows from Theorem 1. This identity says that the optimal complements to $\{k\}$, given by $\mathcal{Y}_{h-1}$ and $\mathcal{Y}_h$, are themselves optimal portfolios of size $h-1$ and h for the market whose objective function is $w(\mathcal{Y})$. Since $\bar{t}_0 = 0$ in the latter market, the inductive hypothesis implies that $\mathcal{Y}_{h-1} \subset \mathcal{Y}_h$, which completes the proof.[2] □

3.3 Polynomial-time solution

Applying the result above yields an efficient algorithm for the optimal portfolio: Start with the empty set and add schools one at a time, maximizing $v(\mathcal{X} \cup \{k\})$ at each addition. Sorting t is $O(m \log m)$. At each of the h iterations, there are $O(m)$ candidates for k, and computing $v(\mathcal{X} \cup \{k\})$ is $O(h)$ using (3); therefore, the time complexity of this algorithm is $O(h^2 m + m \log m)$.

We reduce the computation time to $O(hm)$ by taking advantage of the transformation from the inductive step in the proof of Theorem 3. Once school k is added to $\mathcal{X}$, we remove it from the set $\mathcal{C} \setminus \mathcal{X}$ of candidates, and update the t_j-values of the remaining schools according to the following transformation:

$$\bar{t}_j = \begin{cases} t_j(1 - f_k), & t_j \le t_k \\ t_j - t_k f_k, & t_j > t_k \end{cases} \tag{18}$$

It is easy to verify that this is the composition of the two transformations (from t to $\tilde{t}$, and from $\tilde{t}$ to $\bar{t}$) given in the proof. Now, the *next* school added must be the optimal singleton portfolio in the modified market. But the optimal singleton portfolio consists simply of the school with the highest value of $f_j \bar{t}_j$. Therefore, by updating the t_j-values at each iteration according to (18), we eliminate the need to compute $v(\mathcal{X})$ entirely. Moreover, this algorithm does not require the

[2] We thank Yim Seho for discovering this critical transformation.

schools to be indexed in ascending order by t_j, which removes the $O(m \log m)$ sorting cost.

The algorithm below outputs a list X of the h schools to which Alma should apply. The schools appear in the order of entry such that when the algorithm is run with $h = m$, the optimal portfolio of size h is given by $\mathcal{X}_h = \{\texttt{X[1]}, \dots, \texttt{X[h]}\}$. The entries of the list V give the valuation thereof.

Algorithm 1: Optimal portfolio algorithm for Alma's problem.

```
Data: Utility values t ∈ [0, ∞)^m, admissions probabilities f ∈ [0, 1]^m, application
      limit h ≤ m.
C ← {1 ... m};
X, V ← empty lists;
for i = 1 ... h do
    k ← arg max_{j∈C} {f_j t_j};
    C ← C \ {k};
    append!(X, k);
    if i = 1 then append!(V, f_k t_k) else append!(V, V[i − 1] + f_k t_k);
    for j ∈ C do
        if t_j ≤ t_k then t_j ← t_j(1 − f_k) else t_j ← t_j − f_k t_k;
    end
end
return X, V
```

Theorem 4 (Validity of Algorithm 1)**.** *Algorithm 1 produces an optimal application portfolio for Alma's problem in $O(hm)$ time.*

Proof. Optimality follows from the proof of Theorem 3. Suppose $\mathcal{C}$ is stored as a list. Then at each of the h iterations of the main loop, finding the top school costs $O(m)$, and the t_j-values of the remaining $O(m)$ schools are each updated in unit time. Therefore, the overall time complexity is $O(hm)$. □

In our numerical experiments, we found it effective to store $\mathcal{C}$ as a binary max heap rather than a list. The heap is ordered according to the criterion $i \geq j \iff f_i t_i \geq f_j t_j$. Nominally, using a heap increases the cost of the main loop from $O(hm)$ to $O(hm \log m)$ because the heap is rebalanced when each t_j-value is updated. However, typical problem instances do not achieve this upper bound because the order of the $f_j t_j$-values changes only slightly between iterations. The cost of updating each t_j-value can be reduced to unit time using a Fibonacci heap (Fredman and Tarjan 1987), yielding the same overall computation time.

3.4 Properties of the optimal portfolios

The nestedness property implies that Alma's expected utility is a concave function of h.

Theorem 5 (Optimal portfolio valuation concave in h)**.** *For $h = 2 \dots (m-1)$,*

$$v(\mathcal{X}_h) - v(\mathcal{X}_{h-1}) \geq v(\mathcal{X}_{h+1}) - v(\mathcal{X}_h) \tag{19}$$

Proof. We will prove the equivalent expression $2v(\mathcal{X}_h) \geq v(\mathcal{X}_{h+1}) + v(\mathcal{X}_{h-1})$. Applying Theorem 3, we write $\mathcal{X}_h = \mathcal{X}_{h-1} \cup \{j\}$ and $\mathcal{X}_{h+1} = \mathcal{X}_{h-1} \cup \{j, k\}$. Define the random variables X_i as above. If $t_k \leq t_j$, then

$$\begin{aligned}
2v(\mathcal{X}_h) &= v(\mathcal{X}_{h-1} \cup \{j\}) + v(\mathcal{X}_{h-1} \cup \{j\}) \\
&\geq v(\mathcal{X}_{h-1} \cup \{k\}) + v(\mathcal{X}_{h-1} \cup \{j\}) \\
&= v(\mathcal{X}_{h-1} \cup \{k\}) + (1 - f_j) v(\mathcal{X}_{h-1}) + f_j \operatorname{E}[\max\{t_j, X_{h-1}\}] \\
&= v(\mathcal{X}_{h-1} \cup \{k\}) - f_j v(\mathcal{X}_{h-1}) + f_j \operatorname{E}[\max\{t_j, X_{h-1}\}] + v(\mathcal{X}_{h-1}) \\
&\geq v(\mathcal{X}_{h-1} \cup \{k\}) - f_j v(\mathcal{X}_{h-1} \cup \{k\}) + f_j \operatorname{E}[\max\{t_j, X_{h-1}\}] + v(\mathcal{X}_{h-1}) \\
&= (1 - f_j) v(\mathcal{X}_{h-1} \cup \{k\}) + f_j \operatorname{E}[\max\{t_j, X_{h-1}\}] + v(\mathcal{X}_{h-1}) \\
&= v(\mathcal{X}_{h-1} \cup \{j, k\}) + v(\mathcal{X}_{h-1}) \\
&= v(\mathcal{X}_{h+1}) + v(\mathcal{X}_{h-1})
\end{aligned} \tag{20}$$

The first inequality follows from the optimality of $\mathcal{X}_h$, while the second follows from the fact that adding k to $\mathcal{X}_{h-1}$ can only increase its valuation.

If $t_k \geq t_j$, then the steps are analogous:

$$\begin{aligned}
2v(\mathcal{X}_h) &= v(\mathcal{X}_{h-1} \cup \{j\}) + v(\mathcal{X}_{h-1} \cup \{j\}) \\
&\geq v(\mathcal{X}_{h-1} \cup \{k\}) + v(\mathcal{X}_{h-1} \cup \{j\}) \\
&= (1 - f_k) v(\mathcal{X}_{h-1}) + f_k \operatorname{E}[\max\{t_k, X_{h-1}\}] + v(\mathcal{X}_{h-1} \cup \{j\}) \\
&= v(\mathcal{X}_{h-1}) - f_k v(\mathcal{X}_{h-1}) + f_k \operatorname{E}[\max\{t_k, X_{h-1}\}] + v(\mathcal{X}_{h-1} \cup \{j\}) \\
&\geq v(\mathcal{X}_{h-1}) - f_k v(\mathcal{X}_{h-1} \cup \{j\}) + f_k \operatorname{E}[\max\{t_k, X_{h-1}\}] + v(\mathcal{X}_{h-1} \cup \{j\}) \\
&= v(\mathcal{X}_{h-1}) + (1 - f_k) v(\mathcal{X}_{h-1} \cup \{j\}) + f_k \operatorname{E}[\max\{t_k, X_{h-1}\}] \\
&= v(\mathcal{X}_{h-1}) + v(\mathcal{X}_{h-1} \cup \{j, k\}) \\
&= v(\mathcal{X}_{h-1}) + v(\mathcal{X}_{h+1})
\end{aligned} \tag{21}$$

□

It follows that when $\mathcal{X}_h$ is the optimal h-portfolio, for a given market, $v(\mathcal{X}_h)$ is $O(h)$. Example 2, in which $v(\mathcal{X}_h)$ can be made arbitrarily close to h, establishes the tightness of this bound.

To wrap up, we provide an example showing that if the entries of Z are dependent, then the optimal solution may violate the nestedness property of Theorem 3.

Example 3. Let $t = (3, 3, 4)$, $Z_1 \sim \text{Bernoulli}(0.5)$, $Z_2 = 1 - Z_1$, and $Z_3 \sim \text{Bernoulli}(0.5)$. Then it is easy to verify that the unique optimal portfolios are $\mathcal{X}_1 = \{3\}$ and $\mathcal{X}_2 = \{1, 2\}$.

4 Heterogeneous application costs

In this section, we turn to the more general problem in which the constant g_j represents the *cost* of applying to c_j and the student, whom we now call Ellis, has a *budget* of H to spend on college applications. Applying Theorem 1, we assume $t_0 = 0$ and disregard c_0 throughout. We also assume with trivial loss of generality that all $g_j \leq H$.

Definition 4 (Ellis's problem)**.** Ellis's optimal college application portfolio is given by the solution to the following integer nonlinear program:

$$\begin{aligned} \text{maximize} \quad & v(x) = \sum_{j=1}^{m} x_j t_j f_j \prod_{j'=j+1}^{m} (1 - f_{j'} x_{j'}) \\ \text{subject to} \quad & \sum_{j=1}^{m} g_j x_j \leq H \\ & x_j \in \{0, 1\}, \quad j = 1 \ldots m \end{aligned} \tag{22}$$

The optima for Ellis's problem are not necessarily nested, nor is the number of schools in the optimal portfolio necessarily increasing in H. For example, if $f = (0.5, 0.5, 0.5)$, $t = (1, 1, 1003)$, and $g = (1, 1, 3)$, then it is evident that the optimal portfolio for $H = 2$ is $\{1, 2\}$ while that for $H = 3$ is $\{3\}$.

4.1 NP-completeness

In fact, Ellis's problem is NP-complete, as we will show by a transformation from the binary knapsack problem, which is known to be NP-complete (Garey and Johnson 1979).

Definition 5 (Decision form of knapsack problem)**.** An *instance* consists of a set $\mathcal{B}$ of m objects; utility values $u_j \in \mathbb{N}$ and weight $w_j \in \mathbb{N}$ for each $j \in \mathcal{B}$; and target utility $U \in \mathbb{N}$ and knapsack capacity $W \in \mathbb{N}$. The instance is called a *yes-instance* if and only if there exists a set $\mathcal{B}' \subseteq \mathcal{B}$ having $\sum_{j \in \mathcal{B}'} u_j \geq U$ and $\sum_{j \in \mathcal{B}'} w_j \leq W$.

Definition 6 (Decision form of Ellis's problem)**.** An *instance* consists of an instance of Ellis's problem and a target valuation V. The instance is called a *yes-instance* if and only if there exists a portfolio $\mathcal{X} \subseteq \mathcal{C}$ having $v(\mathcal{X}) \geq V$ and $\sum_{j \in \mathcal{X}} g_j \leq H$.

Theorem 6. *The decision form of Ellis's problem is NP-complete.*

Proof. It is obvious that the problem is in NP.

Consider an instance of the knapsack problem, and we will construct an instance of Ellis's problem that is a yes-instance if and only if the corresponding knapsack instance is a yes-instance. Without loss of generality, we may assume that the the objects in $\mathcal{B}$ are indexed in increasing order of u_j, that each $u_j > 0$, and that the knapsack instance admits a feasible solution other than the empty set.

Let $U_{\max} = \sum_{j \in \mathcal{B}} u_j$ and $\delta = 1/mU_{\max} > 0$, and construct an instance of Ellis's problem with $\mathcal{C} = \mathcal{B}$, $H = W$, all $f_j = \delta$, and $t_j = u_j/\delta$ for all j. Clearly, $\mathcal{X} \subseteq \mathcal{C}$ is feasible for Ellis's problem if and only if it is feasible for the knapsack instance. Now, we observe that if $\mathcal{X}$ is nonempty,

$$\begin{aligned}\sum_{j\in\mathcal{X}} u_j &= \sum_{j\in\mathcal{X}} t_j f_j > \sum_{j\in\mathcal{X}}\Big(t_j f_j \prod_{\substack{j'\in\mathcal{X}:\\ j'>j}} (1-f_{j'})\Big) = v(\mathcal{X}) \\ &= \sum_{j\in\mathcal{X}}\Big(u_j \prod_{\substack{j'\in\mathcal{X}:\\ j'>j}} (1-\delta)\Big) \geq (1-\delta)^m \sum_{j\in\mathcal{X}} u_j \\ &\geq (1-m\delta)\sum_{j\in\mathcal{X}} u_j \geq \sum_{j\in\mathcal{X}} u_j - m\delta U_{\max} = \sum_{j\in\mathcal{X}} u_j - 1\end{aligned} \tag{23}$$

This means that given a nonempty portfolio $\mathcal{X}$, its value for the corresponding knapsack instance is the smallest integer greater than $v(\mathcal{X})$. That is, $\sum_{j\in\mathcal{X}} u_j \geq U$ if and only if $v(\mathcal{X}) \geq U-1$. Taking $V = U-1$ completes the transformation and concludes the proof. □

An intuitive extension of the greedy algorithm for Alma's problem is to iteratively add to $\mathcal{X}$ the school k for which $[v(\mathcal{X}\cup\{k\}) - v(\mathcal{X})]/g_k$ is largest. However, this algorithm offers no guaranteed approximation ratio.

4.2 Pseudopolynomial-time dynamic program

In this subsection, we provide an algorithmic solution to Ellis's problem that runs in $O(Hm + m\log m)$ time and $O(Hm)$ space. The algorithm resembles the dynamic programming algorithm for the binary knapsack problem. Because we cannot assume that $H \leq m$ (as was the case in Alma's problem), this represents a pseudopolynomial-time solution (Dantzig 1957; Garey and Johnson 1979; *Wikipedia*, s.v. "Knapsack problem").

With minimal loss of generality, we assume that $g_j \in \mathbb{N}$ for $j = 1\dots m$ and $H \in \mathbb{N}$. For $j = 0\dots m$ and $h = 0\dots H$, let $\mathcal{X}[j,h]$ denote the optimal portfolio using only the schools $\{1,\dots,j\}$ and costing no more than h, and let $V[j,h] = v(\mathcal{X}[j,h])$. It is clear that if $j=0$ or $h=0$, then $\mathcal{X}[j,h] = \varnothing$ and $V[j,h] = 0$. For convenience, we also define $V[j,h] = -\infty$ for all $h < 0$.

For the remaining indices, $\mathcal{X}[j,h]$ either contains j or not. If it does not contain j, then $\mathcal{X}[j,h] = \mathcal{X}[j-1,h]$. On the other hand, if $\mathcal{X}[j,h]$ contains j, then its value is $(1-f_j)v(\mathcal{X}[j,h]\setminus\{j\}) + f_j t_j$. This requires that $\mathcal{X}[j,h]\setminus\{j\}$ make optimal use of the remaining budget over the remaining schools; that is, $\mathcal{X}[j,h] = \mathcal{X}[j-1,h-g_j]\cup\{j\}$. From these observations, we obtain the following Bellman equation for $j = 1\dots m$ and $h = 1\dots H$:

$$V[j,h] = \max\big\{V[j-1,h], (1-f_j)V[j-1,h-g_j] + f_j t_j\big\} \tag{24}$$

with the convention that $-\infty\cdot 0 = -\infty$. The corresponding optimal portfolios can be computed by observing that $\mathcal{X}[j,h]$ contains j if and only if $V[j,h] > V[j-1,h]$. The optimal solution is given by $\mathcal{X}[m,H]$. The algorithm below performs these computations and outputs the optimal portfolio $\mathcal{X}$.

Algorithm 2: Dynamic program for Ellis's problem with integral application costs.

Data: Utility values $t \in [0, \infty)^m$, admissions probabilities $f \in [0, 1]^m$, application costs $g \in \mathbb{N}^m$, budget $H \in \mathbb{N}$.

Index schools in ascending order by t;

function $V(j, h)$ **do**
- **if** $h < g_j$ **then return** $-\infty$;
- **else if** $j = 0$ *or* $h = 0$ **then return** 0;
- **else return** $\max\{V(j-1, h), (1-f_j)V(j-1, h-g_j) + f_j t_j\}$;

end

$h \leftarrow H$;

$\mathcal{X} \leftarrow \emptyset$;

for $j = m, m-1, \dots, 1$ **do**
- **if** $V(j-1, h) < V(j, h)$ **then**
  - $\mathcal{X} \leftarrow \mathcal{X} \cup \{j\}$;
  - $h \leftarrow h - g_j$;
- **end**

end

return $\mathcal{X}$

Theorem 7 (Validity of Algorithm 2)**.** *Algorithm 2 produces an optimal application portfolio for Ellis's problem in $O(Hm + m \log m)$ time and $O(Hm)$ space.*

Proof. Optimality follows from the foregoing discussion. Sorting by t_j is $O(m \log m)$. The number of indices at which $V(j, h)$ is nontrivially defined is $O(Hm)$. To prevent $V(j, h)$ from being evaluated more than once at a given index, its values can be stored in a dictionary as they are computed. □

4.3 Fully polynomial-time approximation scheme

As with the knapsack problem, Ellis's problem admits a complementary dynamic program that iterates on the value of the cheapest portfolio instead of on the cost of the most valuable portfolio. We will use this algorithm as the basis for a fully polynomial-time approximation scheme for Ellis's problem. Here we assume only that is each t_j is integral. We will represent the approximate value of each portfolio using a fixed-point decimal with a precision of P, where P is the number of digits to retain after the decimal point. Since t_m is a trivial upper bound on the value of any portfolio, the set of possible portfolio valuations is now finite: $\mathcal{V} = \{0, 1 \times 10^{-P}, 2 \times 10^{-P}, \dots, t_m - 1 \times 10^{-P}, t_m\}$. Its cardinality is $|\mathcal{V}| = Pt_m$. Let $r[x] = \lfloor 10^P x \rfloor 10^{-P}$ denote the value of x rounded down to its nearest fixed-point representation.

For integers $0 \leq j \leq m$ and $v \leq t_m$, let $\mathcal{W}[j, v]$ denote the least expensive portfolio using only the schools $\{1, \dots, j\}$ and having value at least v, if such a portfolio exists. Let $G[j, v] = r\big[v(\mathcal{W}[j, v])\big]$ denote the fixed-point representation of the corresponding value, where $G[j, v] = \infty$ if $\mathcal{W}[j, v]$ does not exist. It is clear that if $v \leq 0$, then $\mathcal{W}[j, v] = \emptyset$ and $G[j, h] = 0$.

If $j = 0$ and $v > 0$, then $G[j, h] = \infty$. For the remaining indices (where $j, v > 0$), we claim that

$$G[j, v] = \begin{cases} \infty, & t_j < v \\ \min\bigl\{G[j-1, v], g_j + G[j-1, v-\Delta]\bigr\}, & t_j \geq v \end{cases} \tag{25}$$

$$\text{where} \qquad \Delta_j(v) = \begin{cases} r\left[\frac{f_j}{1-f_j}(t_j - v)\right], & f_j < 1 \\ \infty, & f_j = 1 \end{cases} \tag{26}$$

In the $t_j < v$ case, any feasible portfolio must be composed of schools with utility less than v, and therefore its valuation can not equal v, meaning that $\mathcal{W}[j, v]$ is undefined. In the $t_j \geq v$ case, the first argument to $\min\{\}$ says simply that omitting j and choosing $\mathcal{W}[j-1, v]$ is a feasible choice for $\mathcal{W}[j, v]$. If, on the other hand, $j \in \mathcal{W}[j, v]$, then

$$v(\mathcal{W}[j, v]) = (1 - f_j)v(\mathcal{W}[j, v] \setminus \{j\}) + f_j t_j \tag{27}$$

Therefore, the subportfolio $\mathcal{W}[j, v] \setminus \{j\}$ must have a value of at least $v - \Delta$, where Δ satisfies $v = (1 - f_j)(v - \Delta) + f_j t_j$. When $f_j < 1$, the solution to this equation is $\Delta = \frac{f_j}{1-f_j}(t_j - v)$. By rounding this value down, we ensure that the true valuation of $\mathcal{W}[j, v]$ is *at least* $v - \Delta$.

When $f_j = 1$, then the singleton $\{j\}$ has $v(\{j\}) \geq v$, so

$$G[j, v] = \min\bigl\{G[j-1, v], g_j\bigr\} \tag{28}$$

Defining $\Delta = \infty$ in this case ensures that $G[j-1, v-\Delta] = 0$ as required.

Once $G[j, v]$ has been calculated at each index, the corresponding portfolio can be found by applying the observation that $\mathcal{W}[j, v]$ contains j if and only if $G[j, v] < G[j-1, v]$. Then an approximate solution to Ellis's problem is obtained by applying this observation to the approximate optimal objective value $\max\{w : G[m, w] \leq H\}$.

Algorithm 3: Polynomial-time approximation scheme for Ellis's problem.

Data: Utility values $t \in \mathbb{N}^m$, admissions probabilities $f \in (0,1]^m$, application costs $g \in (0,\infty)^m$, budget $H \in (0,\infty)^m$, tolerance $\varepsilon \in (0,1)$.

```
Index schools in ascending order by t;
P ← ⌈log_10 (m/ε f_m t_m)⌉;
function G(j, v) do
    if v ≤ 0 then return 0;
    else if j = 0 or t_j < v then return ∞;
    else
        return min{ G(j − 1, v), g_j + G(j − 1, v − Δ_j(v)) }
    end
end
v* ← max{w : G(m, w) ≤ H};
v ← v*;
X ← ∅;
for j = m, m − 1, ..., 1 do
    if G(j, v) < ∞ and G(j, v) < G(j − 1, v) then
        X ← X ∪ {j};
        v ← v − Δ_j(v);
    end
end
return X
```

Theorem 8 (Validity of Algorithm 3)**.** *Algorithm 3 produces a $(1-\varepsilon)$-optimal application portfolio for Ellis's problem.*

Proof. Because $\Delta_j(v)$ is rounded down, if $j \in \mathcal{W}[j,v]$, then the true value of $(1-f_j)v\big(\mathcal{W}[j-1, v-\Delta_j(v)]\big) + f_j t_j$ may exceed the fixed-point value v of $\mathcal{W}[j,v]$, but not by more than 10^{-P}. This error accumulates with each school added to $\mathcal{W}$, but the number of additions is at most m. Therefore, where $v^*(\mathcal{W})$ denotes the fixed-point valuation of $\mathcal{W}$ as recorded in the algorithm,

$$v(\mathcal{W}) - v^*(\mathcal{W}) \leq m10^{-P} \tag{29}$$

We can define $v^*(\mathcal{X})$ analogously as the fixed-point valuation of $\mathcal{X}$ when its elements are added in index order and its value is updated and rounded down at each addition in accordance with (27). For the same reason as above, $v(\mathcal{X}) - v^*(\mathcal{X}) \leq m10^{-P}$.

Finally, by the trivial assumption that all $g_j \leq H$, we know $v(\mathcal{X}) \geq v(\{m\}) = f_m t_m$. From this we obtain

$$\begin{aligned} v(\mathcal{W}) &\geq v^*(\mathcal{W}) \geq v^*(\mathcal{X}) \geq v(\mathcal{X}) - m10^{-P} \\ &\geq \left(1 - \frac{m10^{-P}}{f_m t_m}\right) v(\mathcal{X}) \geq (1-\varepsilon)\, v(\mathcal{X}) \end{aligned} \tag{30}$$

which completes the proof. □

Theorem 9 (Computational complexity of Algorithm 3)**.** *Algorithm 3 uses $O(m/\varepsilon + m \log m)$ time and $O(m/\varepsilon)$ space, where D is the product of the denominators of the f_j-values.*

Proof. In the worst case, the number of indices at which $G(j, v)$ is evaluated is $O(|\mathcal{V}|m)$. Since

$$|\mathcal{V}| = t_m 10^P = t_m 10^{\left\lceil \log_{10}(m/\varepsilon f_m t_m) \right\rceil} \leq \text{const.}/\varepsilon \tag{31}$$

is $O(1/\varepsilon)$, adding the $O(m \log m)$-time required to sort t yields the result. By recording the values of $G(j, v)$ in a dictionary as they are computed, we prevent redundant recursive calls at a cost of $O(|\mathcal{V}|m) = O(m/\varepsilon)$ space. □

Since these bounds are polynomial in m and $1/\varepsilon$, Algorithm 3 is a fully polynomial-time approximation scheme for Ellis's problem.

5 References

Budish, Eric. 2011. "The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes." *Journal of Political Economy* 119 (6): 1061–1103. https://doi.org/10.1086/664613.

Dantzig, George B. 1957. "Discrete-Variable Extremum Problems." *Operations Research* 5 (2): 266–88.

Fisher, Marshall, George Nemhauser, and Laurence Wolsey. 1978. "An analysis of approximations for maximizing submodular set functions—I." *Mathematical Programming* 14: 265–94.

Fredman, Michael Lawrence and Robert Tarjan. 1987. "Fibonacci heaps and their uses in improved network optimization algorithms." *Journal of the Association for Computing Machinery* 34 (3): 596–615.

Fu, Chao. 2014. "Equilibrium Tuition, Applications, Admissions, and Enrollment in the College Market." *Journal of Political Economy* 122 (2): 225–81. https://doi.org/10.1086/675503.

Garey, Michael and David Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* New York: W. H. Freeman and Company.

Othman, Abraham, Eric Budish, and Tuomas Sandholm. 2010. "Finding Approximate Competitive Equilibria: Efficient and Fair Course Allocation." In *Proceedings of 9th International Conference on Autonomous Agents and Multiagent Systems.* New York: ACM. https://dl.acm.org/doi/abs/10.5555/1838206.1838323.

Rozanov, Mark and Arie Tamir. 2020. "The nestedness property of the convex ordered median location problem on a tree." *Discrete Optimization* 36: 100581. https://doi.org/10.1016/j.disopt.2020.100581.

Vazirani, Vijay. 2001. *Approximation Algorithms.* Berlin: Springer.