

A quadratic-time algorithm for the cardinality-constrained college application problem

Max Kapur¹ and Sung-Pil Hong²

^{1,2}Department of Industrial Engineering, Seoul National University

July 14, 2022

Abstract

This paper considers a novel portfolio optimization problem called the college application problem. We show that the objective function is a nondecreasing submodular set function and provide a quadratic-time solution algorithm. We discuss the properties of the optimal solution and their implications for student welfare.

Keywords: combinatorial optimization, submodular maximization, portfolio optimization

Correspondence may be addressed to Max Kapur.

Email: maxkapur@gmail.com

Address: 39-411, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

Contents

1	Introduction	2
1.1	Literature review	3
1.2	Outline	4
2	Preliminaries	4
2.1	The objective function	4
2.2	An elimination technique	5
2.3	Submodularity of the objective	5
3	The greedy algorithm	6
3.1	The nestedness property	6
3.2	Quadratic-time solution	7
4	Discussion: Heuristics and student welfare	8
4.1	The distributional heuristic	8
4.2	The linearization heuristic	9
4.3	Implications for student welfare	10
5	Conclusion	11

1 Introduction

Due to the widespread perception that career earnings depend on college admissions outcomes, the solution of (1) holds monetary value. In the American college consulting industry, students pay private consultants an average of \$200 per hour for assistance in preparing application materials, estimating their admissions odds, and identifying target schools [13]. In Korea, an important revenue stream for admissions consulting firms such as Megastudy (megastudy.net) and Jinhak (jinhak.com) is “mock application” software that claims to use artificial intelligence to optimize the client’s application strategy.

Most quantitative research on college admissions to date has been descriptive in nature. A standard benchmark for logistic regression techniques, for example, involves estimating students’ admissions probabilities from information about their past academic performance [1, 8]. When professional admissions consultants advise their clients on where to apply to college, however, they typically rely on qualitative heuristics, such as balancing applications between safety and reach schools, rather than optimization models [6, 11]. (Admissions consultants refer to schools with low utility and high admissions probability as *safety schools*, and those with high utility and low admissions probability as *reach schools*.)

In the present study, we formulate the student’s college application decision as the following combinatorial optimization problem:

$$\begin{aligned}
 & \text{maximize} && v(\mathcal{X}) = \mathbb{E} \left[\max \{ t_0, \max \{ t_j Z_j : j \in \mathcal{X} \} \} \right] \\
 & \text{subject to} && \mathcal{X} \subseteq \mathcal{C}, \quad |\mathcal{X}| \leq H
 \end{aligned} \tag{1}$$

Here $\mathcal{C} = \{1 \dots m\}$ represents the set of colleges in the market and $h > 0$ is the number of colleges the student can apply to. For $j = 1 \dots m$, let f_j denote the student’s probability of being admitted to school j if she applies and Z_j a random, independent Bernoulli variable with

probability f_j , so that Z_j equals one if she is admitted and zero if not. It is appropriate to assume that the Z_j are statistically independent as long as f_j are probabilities estimated specifically for this student (as opposed to generic acceptance rates). For $j = 0 \dots m$, $t_j \geq 0$ indicates the utility she associates with attending school j , where her utility is t_0 if she does not attend college. Then the student's objective is to maximize the expected utility associated with the best school she is admitted to. Therefore, her optimal college application strategy is given by the solution \mathcal{X} to the problem above, where \mathcal{X} represents the set of schools to which she applies, which we call her application *portfolio*.

The problem can also be applied to other competitive matching games such as job application and marriage proposals. Here the budget constraint may arise due to the time required to communicate with each applicant or a legal limit on the number of applications allowed.

We show that the optimal portfolios for the cardinality-constrained college are *nested* in the budget constraint, meaning that where \mathcal{X}_h denotes the optimal portfolio for budget h , $\mathcal{X}_h \subset \mathcal{X}_{h+1}$. The nestedness property is equivalent to the optimality of the *greedy* algorithm that iteratively adds to \mathcal{X} the school that elicits the greatest increase in the objective value. The proof of the nestedness property yields insight into the properties of the optimal college application strategy: Students who cannot afford to apply to many schools allocate a greater proportion of their application budget to safety schools than students with a large application budget. A classic finding of behavioral psychology is that agents with limited resources tend to make more risk-averse decisions than their wealthier peers, especially with respect to education and career choice [4, 5]. In the case of college application, our model provides a rational basis for this tendency.

Our analysis of the college application problem also yields tight bounds on the marginal value of an additional application, which may be useful in designing admissions markets.

1.1 Literature review

The objective function is a monotonic submodular function in the sense first described by Nemhauser et al. [10]. They showed that the greedy algorithm is asymptotically $(1 - 1/e)$ -optimal for optimizing a monotonic submodular function over a cardinality constraint, and a subsequent result of Nemhauser and Wolsey [9] proved that this approximation ratio is the best achievable by a polynomial-time algorithm. In the special case of college application, however, we show that the same algorithm is exact.

The proof of the optimality of the greedy algorithm rests on the nested structure of the portfolios, a property that is of independent interest. As Rozanov and Tamir [12] articulate, the knowledge that the optima are nested aids not only in computing the optimal solution, but in the implementation thereof under uncertain information. For example, in the United States, many college applications are due at the beginning of November, and it is typical for students to begin working on their applications during the prior summer because colleges reward students who tailor their essays to the target school. However, students may not know how many schools they can afford to apply to until late October. The nestedness property—or equivalently, the validity of a greedy algorithm—implies that even in the absence of complete budget information, students can begin to carry out the optimal application strategy by writing essays for schools in the order that they enter the optimal portfolio.

Fu [3] considered a problem that is reducible to ours: since application is at cost rather than a constraint, the objective function is separable in the cost and therefore must be optimal over portfolios of the given size. Thus, solving her problem reduces to solving m instances of (1).

To the best of our knowledge, the first systematic study of the college application problem was undertaken by Kapur in his master's thesis [7], which considers (1) as well as the more general case of a knapsack constraint. The present study extends Kapur's results with a further discussion of the application strategies recommended in the admissions consulting industry and how these compare to the optimal strategy.

1.2 Outline

This paper has five sections. In section 2, we derive a closed-form expression for the objective function of (1) and show that it is a monotonic submodular function. In section 3, we prove that the optimal portfolios are nested in the budget h , which implies the validity of the greedy algorithm. Further refinement reduces the computation time to $O(hm)$. In section 4, we examine alternative solution mechanisms used in the admissions consulting industry and their performance relative to the optimal solution. A brief conclusion follows in section 5.

2 Preliminaries

For the remainder of the paper, unless otherwise noted, we assume with trivial loss of generality that each $f_j \in (0, 1]$ and $0 < t_1 \leq \dots \leq t_m$. Unless otherwise noted, we assume $t_0 = 0$, an assumption justified presently. Unless otherwise noted, we assume that $t_0 = 0$, a restriction that we justify presently.

2.1 The objective function

First we derive a closed-form expression for the objective function of (1).

We refer to the set $\mathcal{X} \subseteq \mathcal{C}$ of schools to which a student applies as her *application portfolio*. The expected utility the student receives from \mathcal{X} is called its *valuation*. Given an application portfolio, let $p_j(\mathcal{X})$ denote the probability that the student attends school j . This occurs if and only if she *applies* to school j , is *admitted* to school j , and is *rejected* from any school she prefers to j ; that is, any school with higher index. Hence, for $j = 0 \dots m$,

$$p_j(\mathcal{X}) = \begin{cases} f_j \prod_{\substack{i \in \mathcal{X}: \\ i > j}} (1 - f_i), & j \in \{0\} \cup \mathcal{X} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The following proposition follows by computing $v(\mathcal{X}) = \sum_{j=0}^m t_j p_j(\mathcal{X})$.

Proposition 1 (Closed form of portfolio valuation function).

$$v(\mathcal{X}) = \sum_{j=0}^m t_j p_j(\mathcal{X}) = \sum_{j \in \{0\} \cup \mathcal{X}} \left(f_j t_j \prod_{\substack{i \in \mathcal{X}: \\ i > j}} (1 - f_i) \right) \quad (3)$$

Next, we show that without loss of generality, we may assume that $t_0 = 0$.

Lemma 1. *For some $\gamma \leq t_0$, let $\bar{t}_j = t_j - \gamma$ for $j = 0 \dots m$. Then $v(\mathcal{X}; \bar{t}_j) = v(\mathcal{X}; t_j) - \gamma$ for any $\mathcal{X} \subseteq \mathcal{C}$.*

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

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

which completes the proof. \square

2.2 An elimination technique

Now, we present a variable-elimination technique that will prove useful throughout the paper.¹ Suppose that the student has already resolved to apply to school k , and the remainder of her decision consists of determining which *other* schools to apply to. Writing her total application portfolio as $\mathcal{X} = \mathcal{Y} \cup \{k\}$, we devise a function $w(\mathcal{Y})$ that orders portfolios according to how well they “complement” the singleton portfolio $\{k\}$. Specifically, the difference between $v(\mathcal{Y} \cup \{k\})$ and $w(\mathcal{Y})$ is the constant $f_k t_k$.

To construct $w(\mathcal{Y})$, let \tilde{t}_j denote the expected utility the student receives from school j *given* that she has been admitted to school j and applied to school k . For $j < k$ (including $j = 0$), this is $\tilde{t}_j = (1 - f_k)t_j + f_k t_k$; for $j > k$, this is $\tilde{t}_j = t_j$. This means that

$$v(\mathcal{Y} \cup \{k\}) = \sum_{j \in \{0\} \cup \mathcal{Y}} \tilde{t}_j p_j(\mathcal{Y}). \quad (5)$$

The transformation to \tilde{t} does not change the order of the t_j -values. Therefore, the expression on the right side of (5) is itself a portfolio valuation function. In the corresponding market, t is replaced by \tilde{t} and \mathcal{C} is replaced by $\mathcal{C} \setminus \{k\}$. To restore our convention that $t_0 = 0$, we obtain $w(\mathcal{Y})$ by taking $\bar{t}_j = \tilde{t}_j - \tilde{t}_0$ for all $j \neq k$ and letting

$$w(\mathcal{Y}) = \sum_{j \in \{0\} \cup \mathcal{Y}} \bar{t}_j p_j(\mathcal{Y}) = \sum_{j \in \{0\} \cup \mathcal{Y}} \tilde{t}_j p_j(\mathcal{Y}) - \tilde{t}_0 = v(\mathcal{Y} \cup \{k\}) - f_k t_k \quad (6)$$

where the second equality follows from Lemma 1. The validity of this transformation is summarized in the following theorem, where we write $v(\mathcal{X}; \bar{t})$ instead of $w(\mathcal{Y})$ to emphasize that $w(\mathcal{Y})$ is, in form, a portfolio valuation function.

Lemma 2 (Eliminate school k). *For $\mathcal{X} \subseteq \mathcal{C} \setminus \{k\}$, $v(\mathcal{X} \cup \{k\}; t) = v(\mathcal{X}; \bar{t}) + f_k t_k$, where*

$$\bar{t}_j = \begin{cases} (1 - f_k)t_j, & t_j \leq t_k \\ t_j - f_k t_k, & t_j > t_k. \end{cases} \quad (7)$$

Proof. It is easy to verify that (7) is the composition of the two transformations (from t to \tilde{t} , and from \tilde{t} to \bar{t}) discussed above. \square

2.3 Submodularity of the objective

Now, we show that the portfolio valuation function is submodular. This result is primarily of taxonomical interest and may be safely skipped, as our subsequent results do not rely on submodular analysis.

¹We thank Yim Seho for pointing out this useful transformation.

Definition 1 (Submodular set function). Given a ground set \mathcal{C} and function $v : 2^{\mathcal{C}} \mapsto \mathbb{R}$, $v(\mathcal{X})$ is called a *submodular set function* if and only if $v(\mathcal{X}) + v(\mathcal{Y}) \geq v(\mathcal{X} \cup \mathcal{Y}) + v(\mathcal{X} \cap \mathcal{Y})$ for all $\mathcal{X}, \mathcal{Y} \subseteq \mathcal{C}$. Furthermore, if $v(\mathcal{X} \cup \{k\}) - v(\mathcal{X}) \geq 0$ for all $\mathcal{X} \subseteq \mathcal{C}$ and $k \in \mathcal{C} \setminus \mathcal{X}$, $v(\mathcal{X})$ is said to be a *nondecreasing* submodular set function.

Theorem 1. *The college application portfolio valuation function $v(\mathcal{X})$ is a nondecreasing submodular set function.*

Proof. It is self-evident that $v(\mathcal{X})$ is nondecreasing. To establish its submodularity, we apply proposition 2.1.iii of [9] and show that

$$v(\mathcal{X} \cup \{j\}) - v(\mathcal{X}) \geq v(\mathcal{X} \cup \{j, k\}) - v(\mathcal{X} \cup \{k\}) \quad (8)$$

for $\mathcal{X} \subseteq \mathcal{C}$ and $j \neq k \in \mathcal{C} \setminus \mathcal{X}$. By Lemma 2, we can repeatedly eliminate the schools in \mathcal{X} according to (7) to obtain a portfolio valuation function $w(\mathcal{Y})$ with parameter \bar{t} such that $w(\mathcal{Y}) = v(\mathcal{X} \cup \mathcal{Y}) + \text{const.}$ for any $\mathcal{Y} \subseteq \mathcal{C} \setminus \mathcal{X}$. Therefore, (8) is equivalent to

$$w(\{j\}) - w(\emptyset) \geq w(\{j, k\}) - w(\{k\}) \quad (9)$$

$$\iff w(\{j\}) + w(\{k\}) \geq w(\{j, k\}) \quad (10)$$

$$\iff \mathbb{E}[\bar{t}_j Z_j] + \mathbb{E}[\bar{t}_k Z_k] \geq \mathbb{E}[\max\{\bar{t}_j Z_j, \bar{t}_k Z_k\}] \quad (11)$$

which is plainly true. \square

3 The greedy algorithm

3.1 The nestedness property

The optimality of the greedy algorithm for the college application problem rests on the fact the the optimal solution possesses a special structure: An optimal portfolio of size $h + 1$ includes an optimal portfolio of size h as a subset.

Theorem 2 (Nestedness of optimal application portfolios). *There exists a sequence of portfolios $\{\mathcal{X}_h\}_{h=1}^m$ satisfying the nestedness relation*

$$\mathcal{X}_1 \subset \mathcal{X}_2 \subset \dots \subset \mathcal{X}_m \quad (12)$$

such that each \mathcal{X}_h is an optimal application portfolio when the application limit is h .

Proof. By induction on h . Applying Lemma 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

$$v(\mathcal{X}_1) = f_j t_j > v(\{k\}) = f_k t_k \quad (13)$$

and

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

which is a contradiction.

(Inductive step.) Assume that $\mathcal{X}_1 \subset \dots \subset \mathcal{X}_h$, 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, suppose that $v(\mathcal{Y}_h) < v(\mathcal{X}_h)$ and $v(\mathcal{X}_{h+1}) > v(\mathcal{X}_h \cup \{k\})$. Then

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

is a contradiction.

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 $\{k\}$.

Applying Lemma 2, we eliminate school k , transform the remaining t_j -values to \bar{t}_j according to (7), and obtain a function $w(\mathcal{Y}) = v(\mathcal{Y} \cup \{k\}) - f_k t_k$ that grades portfolios $\mathcal{Y} \subseteq \mathcal{C} \setminus \{k\}$ according to how well they complement $\{k\}$. Since $w(\mathcal{Y})$ is itself a portfolio valuation function and $\bar{t}_0 = 0$, the inductive hypothesis implies that $\mathcal{Y}_{h-1} \subset \mathcal{Y}_h$, which completes the proof. \square

3.2 Quadratic-time solution

Applying Theorem 2 yields an efficient greedy 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 Lemma 2. Once school k is added to \mathcal{X} , we eliminate it from the set $\mathcal{C} \setminus \mathcal{X}$ of candidates, and update the t_j -values of the remaining schools according to (7). 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 (7), we eliminate the need to compute $v(\mathcal{X})$ entirely. Moreover, this algorithm does not require the schools to be indexed in ascending order by t_j , which removes the $O(m \log m)$ sorting cost.

Theorem 3 (Validity of Algorithm 1). *Algorithm 1 produces an optimal application portfolio for the cardinality-constrained college application problem in $O(hm)$ time.*

Proof. Optimality follows from the proof of Theorem 2. At each of the h iterations of the main

Algorithm 1: Optimal portfolio algorithm for the college application problem.

Input: Utility values $t \in (0, \infty)^m$, admissions probabilities $f \in (0, 1]^m$, application limit $h \leq m$.

```
1  $\mathcal{C} \leftarrow \{1 \dots m\};$ 
2  $\mathbf{X}, \mathbf{V} \leftarrow$  empty  $h$ -arrays;
3 for  $i = 1 \dots h$  do
4    $k \leftarrow \arg \max_{j \in \mathcal{C}} \{f_j t_j\};$ 
5    $\mathcal{C} \leftarrow \mathcal{C} \setminus \{k\};$ 
6    $\mathbf{X}[i] \leftarrow k;$ 
7   if  $i = 1$  then  $\mathbf{V}[i] \leftarrow f_k t_k$  else  $\mathbf{V}[i] \leftarrow \mathbf{V}[i - 1] + f_k t_k;$ 
8   for  $j \in \mathcal{C}$  do
9     if  $t_j \leq t_k$  then  $t_j \leftarrow (1 - f_k) t_j$  else  $t_j \leftarrow t_j - f_k t_k;$ 
10  end
11 end
12 return  $\mathbf{X}, \mathbf{V}$ 
```

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 $O(hm)$. \square

By running the algorithm with $h = m$, one can obtain the optimal portfolios for *all* h by taking $\mathcal{X}_h = \{\mathbf{X}[1], \dots, \mathbf{X}[h]\}$. Thus, Fu (2014)'s variant of the college application problem

$$\text{maximize } v(\mathcal{X}) - c(|\mathcal{X}|)$$

(where $c(\cdot)$ is an increasing cost function) can be solved in $O(m^2)$ -time by a trivial modification of Algorithm 1: Simply compute

$$h^* = \arg \max_{h=1 \dots m} \{\mathbf{V}[h] - c(h)\}$$

and return \mathcal{X}_{h^*} .

4 Discussion: Heuristics and student welfare

Let us discuss a few intuitive heuristics for the college application problem, how their accuracy compares to the optimal solution, and the implications for student welfare.

4.1 The distributional heuristic

In real-world admissions markets, f_t and t_j correlate inversely because desirable schools attract many applicants. In the admissions consulting industry, schools with high utility and low admissions probability are called reach schools, while schools with low utility and high admissions probability are called safety schools, and those in between are called target schools. According to a common heuristic used by admissions consultants, it is best to apply to a roughly equal mix of target, safety, and reach schools [6, 11]. How does this strategy, which we call the *distributional* heuristic, perform compared to the optimal application strategy?

Note that the distributional heuristic does not specify the boundary between target, safety, and reach schools. However, regardless of how this boundary is specified, we argue that the distributional heuristic is limited by the fact that it does not respond adaptively to the application

budget h . In a market where f_j and t_j are inversely correlated, the optimal application strategy gradually favors reach schools over safety schools as h increases.

To see why large values of h favor a risky portfolio allocation, consider two schools i and k having $t_i < t_k$ and $f_i t_i = f_k t_k = 1$. Thus, the actuarial value of applying *individually* to either school is identical, and a risk-neutral student is indifferent between the safety school i and the reach school k . However, suppose that the student has previously resolved to apply to a target school j having $t_i < t_j < t_k$. It is now optimal to apply to $\mathcal{X} = \{j, k\}$ instead of $\{i, k\}$ as long as

$$\begin{aligned}
0 &\leq v(\{j, k\}) - v(\{i, j\}) \\
&= (1 - f_k)f_j t_j + f_k t_k - (1 - f_j)f_i t_i - f_j t_j \\
&= \left(1 - \frac{1}{t_k}\right)f_j t_j + 1 - (1 - f_j) - f_j t_j \\
&= -\frac{1}{t_k}f_j t_j + f_j \\
\iff \frac{f_j t_j}{t_k} &\leq f_j,
\end{aligned}$$

which is *always* true because $t_j < t_k$. This example shows that however one defines a “target school,” it is always better to pair a target school with a reach school than with a safety school as long as the choices are equivalent in their actuarial value. By the elimination argument, the same result holds with respect to adding a reach school to a portfolio that is otherwise balanced.

4.2 The linearization heuristic

The expected utility associated with applying to school j alone is simply $E[t_j Z_j] = f_j t_j$. It is therefore tempting to adopt the following heuristic:

Definition 2 (Linearization heuristic for the college application problem). Apply to the h schools having the highest expected utility $f_j t_j$.

This algorithm’s computation time is $O(m)$ using the PICK algorithm of [2]. In essence, this heuristic maximizes $E[\sum t_j Z_j]$ as a surrogate for the true objective function $E[\max\{t_j Z_j\}]$. Accordingly, we call it the *linearization heuristic*.

The following theorem says that the linearization heuristic is a $(1/h)$ -approximation algorithm for the college application problem.

Theorem 4. 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 $f_j t_j$. Then $v(\mathcal{T}_h)/v(\mathcal{X}_h) \geq 1/h$.

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

$$\begin{aligned}
v(\mathcal{X}_h) &= E\left[\max_{j \in \mathcal{X}_h} \{t_j Z_j\}\right] \leq E\left[\sum_{j \in \mathcal{X}_h} \{t_j Z_j\}\right] \leq E\left[\sum_{j \in \mathcal{T}_h} \{t_j Z_j\}\right] \\
&= h E\left[\frac{1}{h} \sum_{j \in \mathcal{T}_h} \{t_j Z_j\}\right] \leq h E\left[\max_{j \in \mathcal{T}_h} \{t_j Z_j\}\right] = h v(\mathcal{T}_h)
\end{aligned} \tag{16}$$

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

The following example establishes the tightness of the approximation factor.

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

j	1	\cdots	h	$h+1$	$h+2$	\cdots	$m-1$	m
f_j	1	\cdots	1	ε^1	ε^2	\cdots	ε^{h-1}	ε^h
t_j	1	\cdots	1	ε^{-1}	ε^{-2}	\cdots	$\varepsilon^{-(h-1)}$	ε^{-h}

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

$$v(\mathcal{X}_h) = \sum_{j=h+1}^m \left(f_j t_j \prod_{j'=j+1}^m (1 - f_{j'}) \right) = \sum_{j=1}^h (1 - \varepsilon)^j \approx h. \quad (17)$$

Thus, as ε approaches zero, we have $v(\mathcal{T}_h)/v(\mathcal{X}_h) \rightarrow 1/h$. (The optimality of \mathcal{X}_h follows from the fact that it achieves the upper bound of Theorem 4.)

Example 1 also illustrates that for large application budget, reach schools offer greater marginal utility than safety schools.

4.3 Implications for student welfare

Classical portfolio optimization models in the Markowitz tradition feature risk aversion as an exogenous *parameter*. The optimal portfolio is expressed as a proportion of the total budget, and this allocation is constant for budgets of all sizes. In the college application problem, however, our analysis of the distributional heuristic reveals that even putatively risk-averse students must modulate their risk allocation according to the number of schools to which they can afford to apply. Students with a application budget can tolerate, and indeed benefit, from taking larger risks than those without. From a welfare standpoint, this optimal risk elasticity offers weak evidence for the notion that reducing the costs of college application could improve student equity. On the other hand, dissuading students from applying to too many reach schools may be a desirable feature in an admissions market inasmuch as it eliminates the administrative costs associated with reading low-quality applications.

A further complication arises in an intrinsic feature of submodular set functions such as the objective function of the college application problem—that they exhibit rapidly diminishing returns: The optimality of the greedy algorithm implies that the gain in student utility from increasing the budget from $h-1$ to h is larger than that from increasing it from h to $h+1$.

Theorem 5. For $h = 2 \dots (m-1)$,

$$v(\mathcal{X}_h) - v(\mathcal{X}_{h-1}) \geq v(\mathcal{X}_{h+1}) - v(\mathcal{X}_h). \quad (18)$$

Proof. Applying Theorem 2, we write $\mathcal{X}_h = \mathcal{X}_{h-1} \cup \{j\}$ and $\mathcal{X}_{h+1} = \mathcal{X}_{h-1} \cup \{j, k\}$. By optimality, $v(\mathcal{X}_h) - v(\mathcal{X}_{h-1}) \geq v(\mathcal{X}_{h-1} \cup \{k\}) - v(\mathcal{X}_{h-1})$. By submodularity and nestedness, $v(\mathcal{X}_{h-1} \cup \{k\}) - v(\mathcal{X}_{h-1}) \geq v(\mathcal{X}_h \cup \{k\}) - v(\mathcal{X}_h) = v(\mathcal{X}_{h+1}) - v(\mathcal{X}_h)$. \square

(An elementary proof is provided in [7].) It follows that when \mathcal{X}_h is the optimal h -portfolio for a given market, $v(\mathcal{X}_h)$ is $O(h)$. Example 1, in which $v(\mathcal{X}_h)$ can be made arbitrarily close to h , establishes the tightness of this bound.

5 Conclusion

References

- [1] Mohan S. Acharya, Asfia Armaan, and Aneeta S. Antony. “A Comparison of Regression Models for Prediction of Graduate Admissions”. In: *Second International Conference on Computational Intelligence in Data Science*. 2019. DOI: 10.1109/ICCIDS.2019.8862140.
- [2] Manuel Blum et al. “Time Bounds for Selection”. In: *Journal of Computer and System Sciences* 7.4 (1973), pp. 448–61. DOI: 10.1016/S0022-0000(73)80033-9.
- [3] Chao Fu. “Equilibrium Tuition, Applications, Admissions, and Enrollment in the College Market”. In: *Journal of Political Economy* 122.2 (2014), pp. 225–81. DOI: 10.1086/675503.
- [4] Vanessa Hartlaub and Thorsten Schneider. “Educational Choice and Risk Aversion: How Important Is Structural vs. Individual Risk Aversion?” In: *SOEPpapers on Multidisciplinary Panel Data Research*. 433. 2012. URL: https://www.diw.de/documents/publikationen/73/diw_01.c.394455.de/diw_sp0433.pdf.
- [5] Thomas van Huizen and Rob Alessie. “Risk Aversion and Job Mobility”. In: *Journal of Economic Behavior & Organization* 164 (2019), pp. 91–106. URL: 10.1016/j.jebo.2019.01.021.
- [6] Minhee Jeon. “[College application strategy] Six chances total...divide applications across reach, target, and safety schools”. In: *Jungang Ilbo* (Aug. 26, 2015). URL: <https://www.joongang.co.kr/article/18524069>.
- [7] Max Kapur. “The College Application Problem”. MA thesis. Seoul National University, 2022. URL: <https://github.com/maxkapur/CollegeApplication>.
- [8] Daniel Kibum Lim. “A Simulation Approach to Predicting College Admissions”. MA thesis. University of California Los Angeles, 2013. URL: <https://escholarship.org/uc/item/8r2695n5>.
- [9] George Nemhauser and Laurence Wolsey. “Best Algorithms for Approximating the Maximum of a Submodular Set Function”. In: *Mathematics of Operations Research* 3.3 (1978), pp. 177–88. DOI: 10.1287/moor.3.3.177.
- [10] George Nemhauser, Laurence Wolsey, and Marshall Fisher. “An Analysis of Approximations for Maximizing Submodular Set Functions—I”. In: *Mathematical Programming* 14 (1978), pp. 265–94. DOI: 10.1007/BF01588971.
- [11] Timothy Peck. *Safety, Target, & Reach Schools: How to Find the Right Ones*. 2021. URL: <https://blog.collegevine.com/the-college-list-decoded-safeties-targets-and-reaches/>.
- [12] Mark Rozanov and Arie Tamir. “The Nestedness Property of the Convex Ordered Median Location Problem on a Tree”. In: *Discrete Optimization* 36 (2020), article no. 100581. URL: <https://doi.org/10.1016/j.disopt.2020.100581>.
- [13] Mark Sklarow. *State of the Profession 2018: The 10 Trends Reshaping Independent Educational Consulting*. Tech. rep. Independent Educational Consultants Association, 2018. URL: <https://www.iecaonline.com/wp-content/uploads/2020/02/IECA-Current-Trends-2018.pdf>.