대학 지원 최적화 문제

Max Kapur & 홍성필

서울대학교
산업공학과
경영과학/최적화 연구실

May 2, 2022

서론

대학 지원 최적화 문제란 **새로운 조합 최적화 문제**.

예산 제약식 하에서, 다수 확률 변수로 이루어진 포트폴리오의 **기대 최댓값을 최대화**한다.

방법론적 지향:

- **자산배분 최적화 관점:** 가치 (valuation) vs. 가치 (value), 위험/보상 관리, 효율적 투자선.
- **배낭 문제 관점:** 정수 조건, 동적 계획, NP-completeness, 근사 해법.

발표 요약

대학 지원 문제를 **모형화**.

전형적인 인스턴스를 구경.

알마의 문제: 대학 지원 문제의 **특수한 경우**. 최적해가 포함 사슬 관계를 가지므로 다항 시간 탐욕 해법 존재.

엘리스의 문제: **일반 경우**. NP-completeness를 증명하고 **3개의 해법** 제시.

계산 실험을 통해 알고리즘의 효율성을 확인.

모형

대학 지원 과정

대학 지원 전략을 고민하는 학생을 고려하자.

m개의 학교, $\mathcal{C} = \{1 \dots m\}$. j번째 학교의 이름은 c_j.

학생이 c_j에 다니면 그의 효용은 $t_j > 0$이며 대학에 못가면 t_0. Wlog, $t_0 < t_1 \leq \cdots \leq t_m$.

$f_j \in (0, 1]$: 학생이 c_j에 지원할 때, 그의 **합격 확률**. $f_0 = 1$. 합격 결과는 $Z_j \sim \text{Bernoulli}(f_j)$.

$\mathcal{X}$: 학생의 **지원 포트폴리오**, 즉 지원하는 학교의 집합.

$p_j(\mathcal{X})$: 학생이 c_j에 **진학할 확률**. $j = 0 \dots m$에 대해,

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

목적함수와 제약 조건

학생의 목적은 자신이 기대 효용을 최대화하는 것이며 이는 합격하는 학교 중 효용이 **최대**인 학교이다. 목적함수의 "maximax" 형태는 본 모형의 특색.

$$v(\mathcal{X}) = \mathrm{E}\Big[\max\{t_0, \max\{t_j Z_j : j \in \mathcal{X}\}\}\Big]$$
$$= \sum_{j=0}^{m} t_j p_j(\mathcal{X}) = \sum_{j \in \{0\} \cup \mathcal{X}} \Big(f_j t_j \prod_{\substack{i \in \mathcal{X}: \\ i > j}} (1 - f_i) \Big)$$

(논문에서 보이듯이, $t_0 = 0$이라고 가정할 수 있다.)

2가지의 제약 조건 형태를 고려한다.

- **집합 크기 제약**: 한국 정시 입학 과정처럼 지원할 수 있는 학교의 개수가 h로 제한된 "알마의 문제." 다항 시간 해법 존재.
- **지원 비용 예산 조건**: 각 학교의 지원 비용이 g_j이며 학교가 지원에 쓸 수 있는 금액이 H인 "엘리스의 문제." NP-complete.

문제 정의

문제 1 (알마의 문제)

$$\begin{aligned} \text{maximize} \quad & v(\mathcal{X}) = \sum_{j \in \mathcal{X}} \left(f_j t_j \prod_{\substack{i \in \mathcal{X}: \\ i > j}} (1 - f_i) \right) \\ \text{subject to} \quad & \mathcal{X} \subseteq \mathcal{C}, \quad |\mathcal{X}| \leq h \end{aligned}$$

문제 2 (엘리스의 문제)

$$\begin{aligned} \text{maximize} \quad & v(\mathcal{X}) = \sum_{j \in \mathcal{X}} \left(f_j t_j \prod_{\substack{i \in \mathcal{X}: \\ i > j}} (1 - f_i) \right) \\ \text{subject to} \quad & \mathcal{X} \subseteq \mathcal{C}, \quad \sum_{j \in \mathcal{X}} g_j \leq H \end{aligned}$$

모든 $g_j = 1$로 넣으면 알마의 문제가 엘리스의 문제의 특수한 경우임을 알 수 있다.

대학 지원 문제가 왜 어려운가

기대 가치 $f_j t_j$가 가장 큰 학교를 선택하는 '뻔한' 알고리즘은 사실 틀리다. (근사 계수는 알마 문제에 대해 $\frac{1}{h}$, 엘리스 문제에 대해 $\frac{1}{m-1}$.)

정수 비선형 계획으로 표현할 수 있지만 목적함수가 **오목**이 아니다.

$\rightarrow$ 본성적으로 **조합적인** 문제.

전형적인 인스턴스(다음 화면)에서 입학 확률 f_j와 효용 모수 t_j가 서로 **반비례**한다.

$\rightarrow$ 선호도가 높으며 붙기 어려운 "상향" 지원 학교(reach school)와 선호도가 낮으며 붙기 쉬운 "안정" 지원 학교(safety school) 사이의 균형을 고려해야 한다고 의미.

좋은 지원 전략은 **금전적인 가치**를 가진다. 미국 입학 상담가의 시간당 급료는 평균 24만원 (Sklarow 2018).

선행 연구는 작은 $(m = 8)$ 인스턴스만 고려하고 열거법으로 풀었다 (Chao 2014).

[대입 수시 전략] 총 6번의 기회 …‘상향·소신·안정’ 분산 지원하라

중앙일보 | 업데이트 2015.08.26 10:15

지면보기

전민희 기자 구독

대학 최저학력기준 고려해 전략 지원
지난해 같은 전형 합격한 선배 내신 참고
수능 전 대학별고사 보는 곳은 최소화

‘지피지기 백전불태(知彼知己百戰不殆).’ 적을 알고 나를 알면 백 번 싸워도 위태롭지가 않다는 뜻이다. 고대 중국의 병법서인 『손자』에 나온 말이지만 현대사회에서도 여러 가지 분야에서 회자된다. 그중 하나가 대학입시다. 특히 2주 앞으로 다가온 수시모집은 전형 종류가 다양해 ‘적’(모집전형)을 알고, ‘나’(학생)에 대해 파악하는 게 무엇보다 중요하다.

Timothy Peck — May 7, 2021 — 5 — 12th Grade, College Lists

Safety, Target, & Reach Schools: How to Find the Right Ones

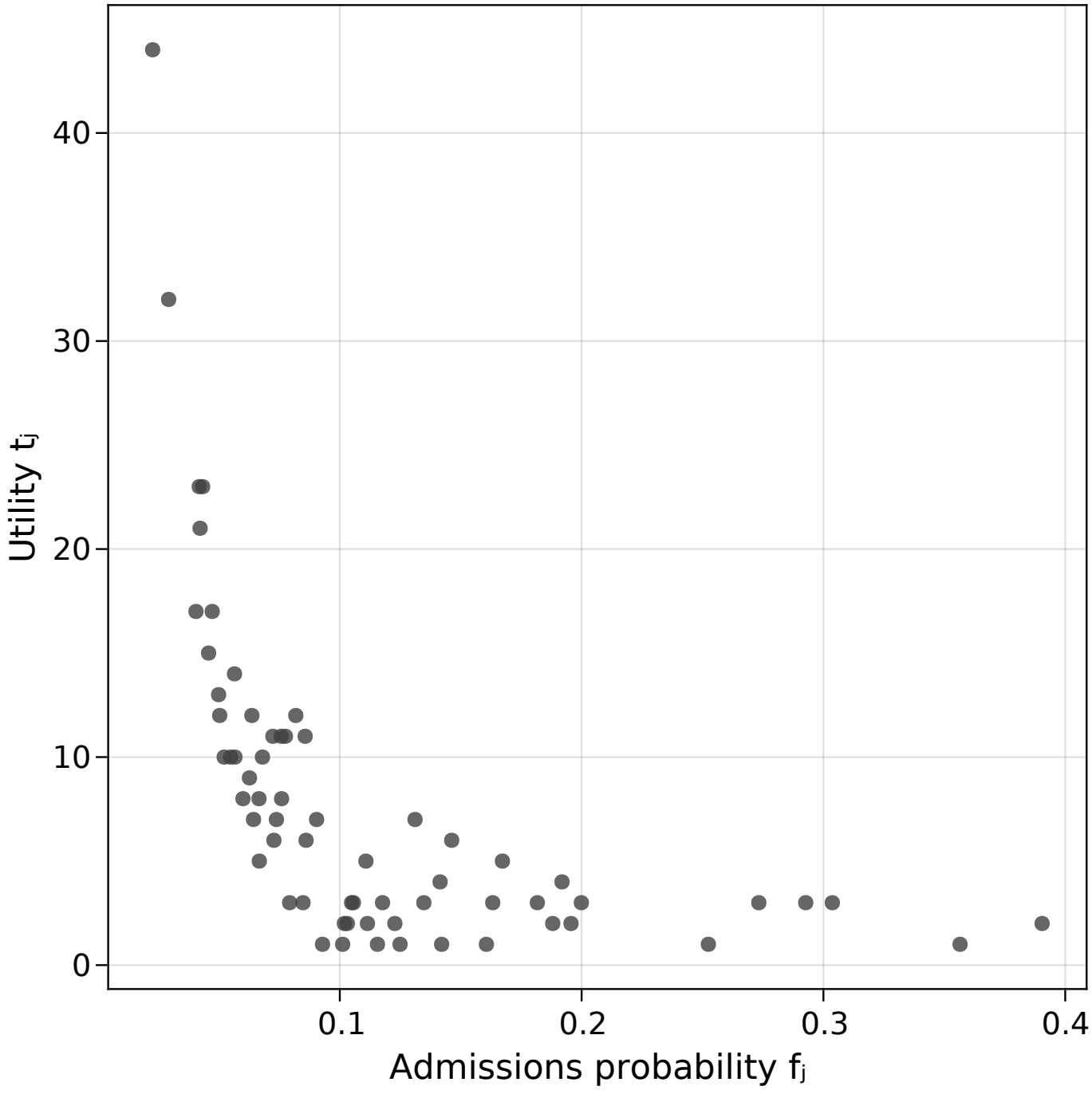
What's Covered:

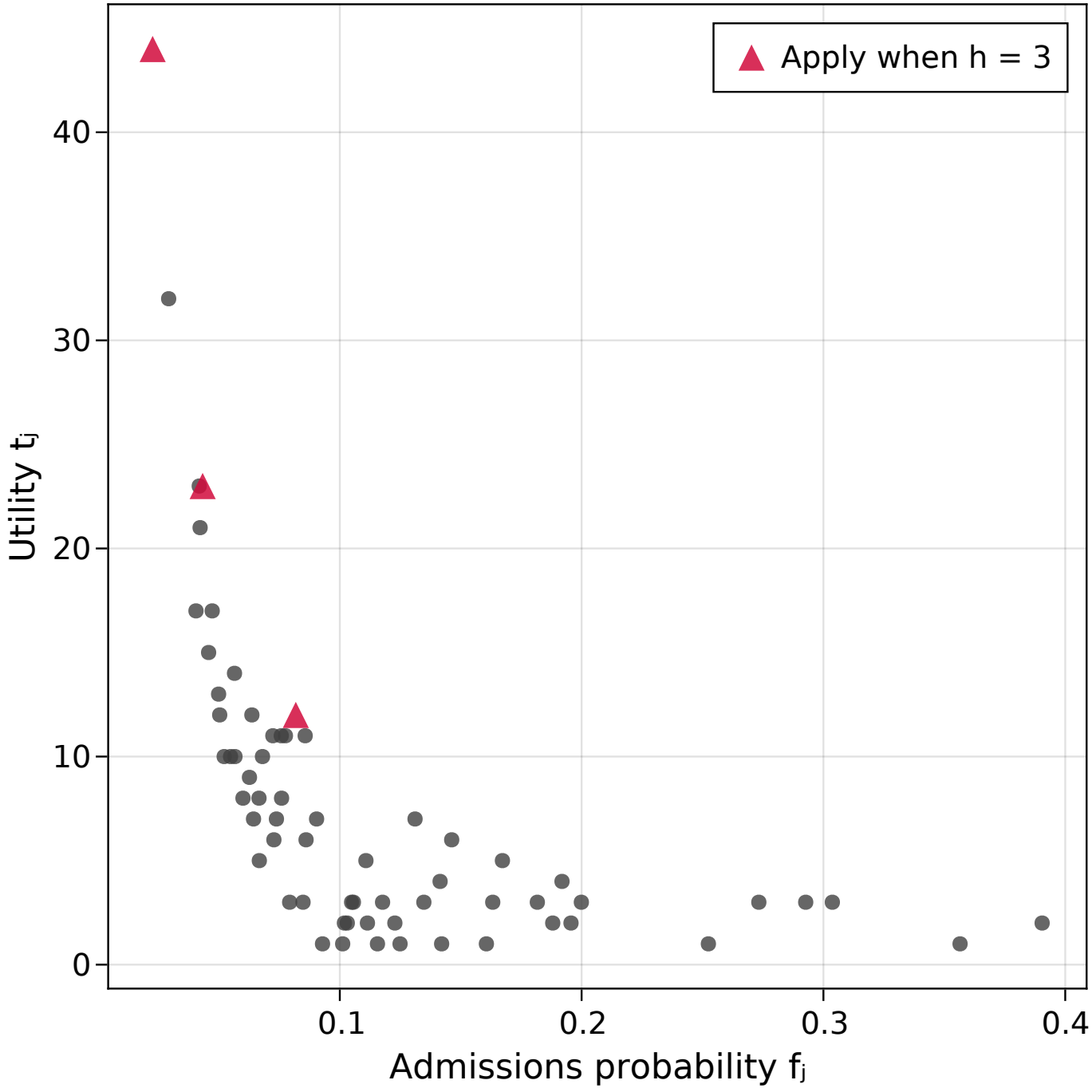
Creating a school list is an important-yet-tricky step in the college application process. A strategically constructed school list weighs your desire to attend reach schools—the institutions you dream about going to—along with safety schools where you're very likely to secure admission. Consequently, the ideal school list is balanced between reach, target, and safety schools, allowing you to shoot for the stars while also ensuring admission into at least one school.

What Are Reach, Target, and Safety Schools?

"Reach," "safety," and "target" are common terms used in college applications to describe the odds a student has of getting accepted at a particular institution. Understanding these terms, and which categories colleges fall into, is a critical step in the application process.

What is a Reach School?


40
30
20
10
0
Utility t_j
0.1
0.2
0.3
0.4
Admissions probability f_j


Apply when h = 3
Utility t_j
Admissions probability f_j
0
10
20
30
40
0.1
0.2
0.3
0.4

알마의 문제

알마 문제 최적해의 포함 사슬 관계

다음 정리는 알마 문제의 최적해가 지원 제한 h로 모수화된 **포함 사슬 관계**를 가진다고 의미한다.

정리 1 (최적 포트폴리오의 포함 사슬 관계)

각 $\mathcal{X}_h$가 지원 제한 h에 대한 최적 포트폴리오며 위의 포함 사슬 관계를 만족하는 포트폴리오 수열 $\{\mathcal{X}_h\}_{h=1}^{m}$가 존재한다.

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

따라서 $v(\mathcal{X})$를 최대화하는 학교를 차례대로 추가하는 탐욕 해법의 타당성이 성립.

이를 좀 더 조정하면 계산 시간을 $O(hm)$으로 절감할 수 있다.

$v(\mathcal{X}_h)$의 오목성

포함 사슬 관계 성질은 알마의 기대 효용이 h의 이산 오목 함수임을 의미한다.

정리 2 (최적 포트폴리오 가치의 h-오목성)

$h = 2 \ldots (m-1)$*에 대해,*

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

이의 따름정리로서 $v(\mathcal{X}_h)$가 $O(h)$ 함수가 되며, 논문에서 타이트한 예를 제시.

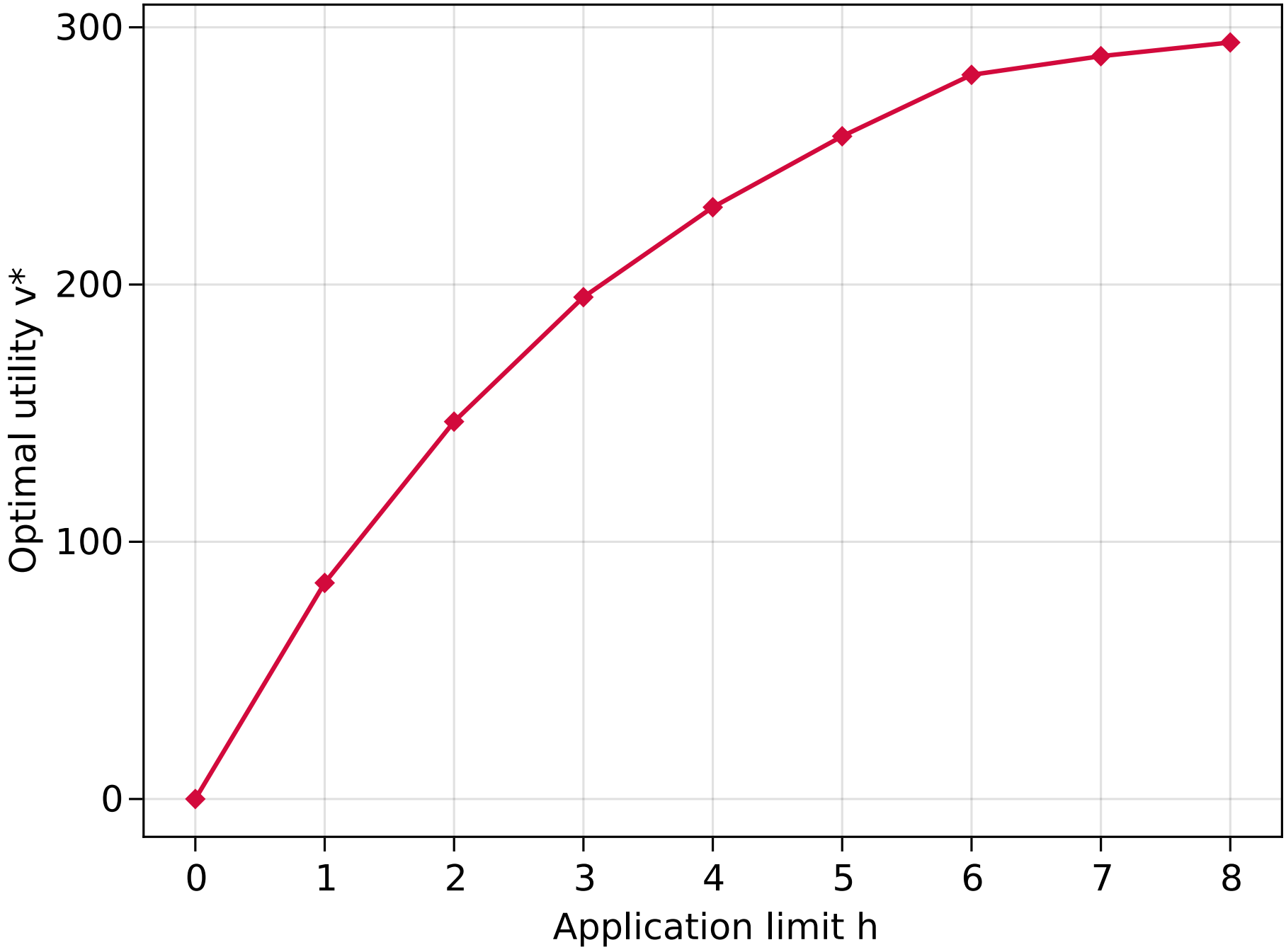
A small example

$m = 8$개의 학교로 이루어진 가상 입학 시장의 대학교 자료와 최적 지원 포트폴리오.

지표 j	**학교** c_j	**합격 확률** f_j	**효용** t_j	**지원 순위**	$v(\mathcal{X}_h)$
1	수성대	0.39	200	4	230.0
2	금성대	0.33	250	2	146.7
3	화성대	0.24	300	6	281.5
4	목성대	0.24	350	1	84.0
5	토성대	0.05	400	7	288.8
6	천왕성대	0.03	450	8	294.1
7	해왕성대	0.10	500	5	257.7
8	명왕성대	0.12	550	3	195.1

포함 사슬 관계 성질에 따라, 지원 제한이 h일 때 최적 포트폴리오는 지원 순위가 h 이하인 h개의 학교로 구성된다.

다음 화면에서 나타나는 곡선은 오목성 성질을 보인다.


300
200
100
0
Optimal utility v*
0
1
2
3
4
5
6
7
8
Application limit h

Ellis's problem

엘리스의 문제를 위한 3개의 알고리즘

$f_j = \varepsilon$으로 넣으면 $v(\mathcal{X})$를 선형 함수에 원하는 만큼 가깝게 만들 수 있다. 배낭 문제가 엘리스의 문제로 변환할 수 있음을 의미하므로 엘리스의 문제는 NP-complete하다.

엘리스 문제를 위해 3가지해법을 제시한다.

- **분지한계법**은 이론적으로 흥미롭지만 현실적 효율성이 낮다.
- 지원 지출액 기반 **동적 계획**. $O(Hm + m \log m)$인 의사 다항 시간 안에 정확한 해를 출력하며, g_j가 작은 정수가 되는 전형적인 인스턴스에 매우 효율적이다.
- 분수값을 내림한 포트폴리오 가치 기반 다른 동적 계획. $O(m^3/\varepsilon)$-시간 안에 $(1-\varepsilon)$-근사 해를 출력하므로 **FPTAS**.

분지한계법

엘리스의 문제는 다음 정수 비선형 계획으로 표현되며 이를 바탕으로 분지한계법을 만든다.

문제 3 (엘리스의 문제를 위한 비선형 정수 계획)

$$\text{maximize} \quad v(x) = \sum_{j=1}^{m} \left(f_j t_j x_j \prod_{i>j} (1 - f_i x_i) \right)$$

$$\text{subject to} \quad \sum_{j=1}^{m} g_j x_j \leq H, \quad x_i \in \{0,1\}^m$$

목적함수는 m차 비오목 다항식 ... 수리적 솔버는 헛수고!

선형 완화 문제

다음 선형 완화 문제를 활용한다.

문제 4 (엘리스의 문제를 위한 선형 완화 문제)

$$\text{maximize} \quad v_{\mathrm{LP}}(x) = \sum_{j=1}^{m} f_j t_j x_j$$

$$\text{subject to} \quad \sum_{j=1}^{m} g_j x_j \leq H, \quad x \in [0,1]^m$$

연속 배낭 문제라고 불리며 쉽게 풀 수 있다 (Dantzig 1957; Balas and Zemel 1980).

포함 사슬 관계의 증명 과정에서 도출한 변수 소거법은 재활용해서 위에 **상한을 더 타이트하게** 조정한다.

이 기반으로 만든 간단한 알고리즘은 작은 ($m \leq 35$) 인스턴스에 괜찮지만 분지 마디를 선택하는 휴리스틱으로 **개선할 여지**가 보인다.

지원 지출액 기반 동적 계획

$\{1, \ldots, j\}$에 속한 학교만 사용하면서 지원 지출액이 h를 넘지 않은 최적 포트폴리오의 가치를 $V[j, h]$라고 하자.

그러면 다음과 같은 **Bellman 식**으로 모든 $V[j, h]$-값은 재귀적으로 계산할 수 있다.

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

이 식의 타당성은 학교를 t_j-값 순서대로 배열함에 의존.

따라서 $V[j, h]$-값으로 표를 채우는 시간이 $O(Hm + m \log m)$이며 이를 참고하면 $\mathcal{X}$를 쉽게 구할 수 있다.

전형적인 인스턴스에서 g_j가 작은 상수이므로 **매우 효율적인** 해법.

포트폴리오 가치 기반 동적 계획

엘리스의 문제는 배낭 문제와 같이, 가치가 가장 높은 포트폴리오의 비용 대신 비용이 가장 낮은 포트폴리오의 가치를 탐색하는 **보완적인 동적 계획**이 존재한다.

포트폴리오의 근사적 가치를 정확도 P로 구성된 고정소수점 십진수(fixed-point decimal)로 나타내자. 다, P는 소수점 뒤에 등장하는 숫자의 수이다. 이때 x를 가장 가까운 고정소수점 십진수로 내림한 것을 $r[x] = 10^{-P}\lfloor 10^P x \rfloor$라고 하자.

임의의 포트폴리오의 가치가 $\bar{U} = \sum_{j \in \mathcal{C}} f_j t_j$를 넘을 수 없다. 따라서 고정소수점 환경에서 발생할 수 있는 포트폴리오 가치로 이루어진 집합 $\mathcal{V}$는 유한 집합이다.

$$\mathcal{V} = \left\{0, 1 \times 10^{-P}, 2 \times 10^{-P}, \ldots, r\left[\bar{U} - 1 \times 10^{-P}\right], r\left[\bar{U}\right]\right\}$$

그러면 $|\mathcal{V}| = \bar{U} \times 10^P + 1$이다.

활용 재귀 관계

$\{1, \ldots, j\}$에 속한 학교만 사용하면 (내림한) 가치가 최소한 v의 포트폴리오 중 지원 비용이 최소한 포트폴리오의 지출액을 $G[j, v]$라고 하자. 이때 다음 반복 관계가 성립한다고 주장한다.

$$G[j,v] = \begin{cases} \infty, & t_j < v \\ \min\{G[j-1,v], g_j + G[j-1, v-\Delta_j(v)]\}, & t_j \geq v \end{cases}$$

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

이제 $G[j, v]$-값으로 표를 채우면 근사적 최적 포트폴리오를 구할 수 있다.

논문에서 $P \leftarrow \left\lceil \log_{10}\left(m^2/\varepsilon \bar{U}\right)\right\rceil$로 넣으면 $(1-\epsilon)$-근사해가 보장되며, 해당 표의 크기가 $O(m^3/\varepsilon)$이므로 이 알고리즘은 **FPTAS**이다.

계산 실험 결과

계산 실험: 알마의 문제

모든 알고리즘은 쥴리아 (Julia, v1.8.0b1) 언어로 구현.

f_j와 t_j가 반비례하는 **가상 인스턴스**를 생성.

각 칸에서 50개의 시장을 생성하고 최적해를 3번 계산한 것 중에 최소 시간을 기록. 표에서 나타나는 값은 평균 (표준편차) 시간이며 단위는 ms.

실험 1: 알마의 문제. $\mathcal{C}$를 저장하는 자료 구현 2가지 비교.

m	**탐욕 알고리즘 (목록 구현)**		**탐욕 알고리즘 (힙 구현)**	
16	0.00	(0.00)	0.01	(0.00)
64	0.03	(0.00)	0.08	(0.02)
256	0.15	(0.01)	0.97	(0.22)
1024	2.31	(0.05)	14.44	(1.66)
4096	37.85	(0.61)	245.74	(17.33)
16384	585.75	(2.11)	4728.59	(552.25)

반복 단계마다 $\arg\max$ 연산이 필요하므로 힙 구현에는 매력이 있지만, 그의 모든 원소를 수정해야 하므로 만회하지 않는다.

계산 실험: 엘리스의 문제

실험 2: 엘리스의 문제. 2개의 정확한 해법 그리고 2개의 정확도로 실행한 FPTAS를 비교.

m	**분지한계법**	**지출액 DP**	**FPTAS,** $\varepsilon = 0.5$	**FPTAS,** $\varepsilon = 0.05$
8	0.04 (0.02)	0.01 (0.00)	0.05 (0.01)	0.21 (0.06)
16	0.22 (0.11)	0.07 (0.01)	0.43 (0.10)	3.15 (0.74)
32	166.20 (422.31)	0.31 (0.05)	2.38 (0.38)	33.38 (11.68)
64	— (—)	1.36 (0.18)	15.32 (2.77)	405.73 (125.98)
128	— (—)	6.52 (0.72)	84.50 (22.59)	2362.19 (1095.11)
256	— (—)	31.23 (1.91)	1085.60 (1186.24)	22129.92 (6588.40)

실험 데이터에서 g_j가 작은 정수이므로 **지원 지출액 기반 동적 계획은 상당한 우위**를 발휘.

FPTAS의 병목 요소는 계산 시간이 아니라 메모리 소모량.

결론

결론

Maximax 형태와 정수 조건 때문에 대학 지원 전략은 **흥미로운 최적화 문제**이며 또한 **실용 가치**가 있다.

본 연수는 탐욕 근사 해법과 최적해의 포함 사슬 관계를 살펴본 **선행 연구의 확장**으로 볼 수 있다 (Fisher et al. 1978; Rozanov and Tamir 2020).

알마 문제의 포함 사슬 관계 성질과 일반 문제의 NP-completeness는 **배낭 문제와 유사한** 결과.

향후 연구 방향:

- 고전적 Markowitz (1952)자산배분 모형처럼, 위험 회피를 명시적으로 다루는 새로운 목적함수.
- 한국 입학 과정의 가나다군과 같은 다각화 제약 조건.
- 동적 계획의 메모리 소모량 절감.

참고 문헌 I

Balas, Egon and Eitan Zemel. 1980. "An Algorithm for Large Zero-One Knapsack Problems." *Operations Research* 28 (5): 1130–54. https://doi.org/10.1287/opre.28.5.1130.

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.

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.

Markowitz, Harry. 1952. "Portfolio Selection." *The Journal of Finance* 7 (1): 77–91. https://www.jstor.org/stable/2975974.

참고 문헌 II

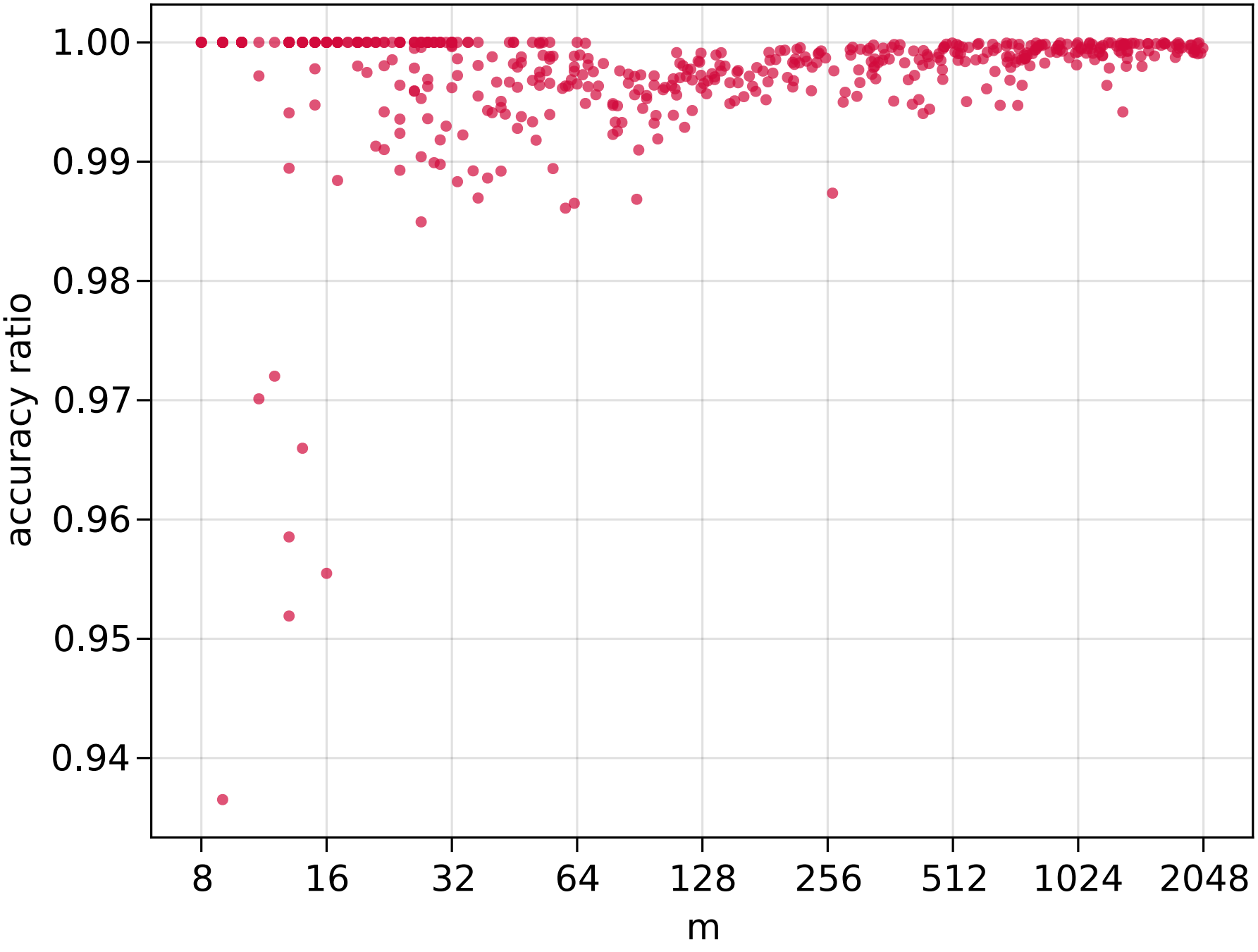
Rozanov, Mark and Arie Tamir. 2020. "The nestedness property of the convex ordered median location problem on a tree." *Discrete Optimization* 36: 100581. https://doi.org/10.1016/j.disopt.2020.100581.

Sklarow, Mark. 2018. *State of the Profession 2018: The 10 Trends Reshaping Independent Educational Consulting.* Technical report, Independent Educational Consultants Association. https://www.iecaonline.com/wp-content/uploads/2020/02/IECA-Current-Trends-2018.pdf.

부록: 알고리즘 요약

알고리즘	문제	제한	정확도	계산 시간
나이브	동일한 지원 비용	없음	$(1/h)$-근사	$O(m)$
탐욕 해법	동일한 지원 비용	없음	정확	$O(hm)$
분지한계법	일반 문제	없음	정확	$O(2^m)$
지출액 동적 계획	일반 문제	g_j 정수	정확	$O(Hm + m \log m)$
FPTAS	일반 문제	없음	$(1-\varepsilon)$-근사	$O(m^3/\varepsilon)$
모의 담금질	일반 문제	없음	0-근사	$O(Nm)$

모의 담금질 휴리스틱: 내역은 특별할 것이 없다. 정확도 실험 결과는 다음 화면.


1.00
0.99
0.98
0.97
0.96
0.95
0.94
accuracy ratio
8
16
32
64
128
256
512
1024
2048
m

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 four algorithms for this more general setup: a branch-and-bound routine, a dynamic program that produces an exact solution in pseudopolynomial time, a different dynamic program that yields a fully polynomial-time approximation scheme, and a simulated-annealing heuristic. Numerical experiments demonstrate the algorithms' accuracy and efficiency.

요약

본 논문은 다수의 확률 변수로 구성된 포트폴리오의 기대 최댓값을 예산 조건 하에서 최대화하는 문제를 고려한다. 이를 대학 지원 최적화 문제라고 부른다. 각 확률 변수의 비용, 즉 각 대학의 지원 비용이 동일한 경우, 최적 포트폴리오는 예산 제약식으로 결정된 포함 사슬 관계 성질을 가짐을 보이고 이를 바탕으로 다항 시간 해법을 제시한다. 대학의 지원 비용이 서로 다른 경우, 문제가 NP-complete 함을 증명한다. 일반적인 문제를 위해 4가지 해법을 제시한다. 분지한계 기반 해법, 의사 다항 시간 안에 정확한 해를 출력하는 동적 계획 해법, 다른 동적 계획 기반으로 완전 다항 시간 근사 해법(fully polynomial-time approximation scheme), 그리고 모의 담금질(simulated annealing)을 이용한 휴리스틱 해법. 수리적 실험을 통하여 알고리즘의 정확도와 효율성을 보인다.