Discrete Applied Mathematics 161 (2013) 258–276

Contents lists available at SciVerse ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Approximation of a maximum-submodular-coverage problem involving spectral functions, with application to experimental designs

Guillaume Sagnol *

Zuse Institut Berlin (ZIB), Takustr. 7, 14195 Berlin, Germany

A R T I C L E   I N F O

Article history:
Received 5 December 2011
Received in revised form 10 July 2012
Accepted 14 July 2012
Available online 11 August 2012

Keywords:
Maximum coverage
Optimal design of experiments
Kiefer's p-criterion
Polynomial-time approximability
Rounding algorithms
Submodularity
Matrix inequalities

A B S T R A C T

We study a family of combinatorial optimization problems defined by a parameter $p \in [0, 1]$, which involves spectral functions applied to positive semidefinite matrices, and has some application in the theory of optimal experimental design. This family of problems tends to a generalization of the classical maximum coverage problem as p goes to 0, and to a trivial instance of the knapsack problem as p goes to 1.

In this article, we establish a matrix inequality which shows that the objective function is submodular for all $p \in [0, 1]$, from which it follows that the greedy approach, which has often been used for this problem, always gives a design within $1 - 1/e$ of the optimum. We next study the design found by rounding the solution of the continuous relaxed problem, an approach which has been applied by several authors. We prove an inequality which generalizes a classical result from the theory of optimal designs, and allows us to give a rounding procedure with an approximation factor which tends to 1 as p goes to 1.

1. Introduction

This work is motivated by a generalization of the classical maximum coverage problem which arises in the study of optimal experimental designs. This problem may be formally defined as follows: given s positive semidefinite matrices $M_1, \ldots, M_s$ of the same size and an integer $N < s$, solve:

$$\max_{I \subset [s]} \operatorname{rank}\left(\sum_{i \in I} M_i\right) \quad \text{s.t. } \operatorname{card}(I) \leq N, \tag{P_0}$$

where we use the standard notation $[s] := \{1, \ldots, s\}$ and $\operatorname{card}(S)$ denotes the cardinality of S. When each M_i is diagonal, it is easy to see that Problem (P_0) is equivalent to a max-coverage instance, by defining the sets $S_i = \{k : (M_i)_{k,k} > 0\}$, so that the rank in the objective of Problem (P_0) is equal to $\operatorname{card}\big(\cup_{i \in I} S_i\big)$.

A more general class of problems arising in the study of optimal experimental designs is obtained by considering a *deformation* of the rank which is defined through a spectral function. Given $p \in [0, 1]$, solve:

$$\max_{\boldsymbol{n} \in \mathbb{N}^s} \varphi_p(\boldsymbol{n}) \quad \text{s.t. } \sum_{i \in [s]} n_i \leq N, \tag{P_p}$$

* Fax: +49 30 84185269.
E-mail address: sagnol@zib.de.

0166-218X/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2012.07.016

where $\varphi_p(\boldsymbol{n})$ is the sum of the eigenvalues of $\sum_{i\in[s]} n_i M_i$ raised to the exponent p: if the eigenvalues of the positive semidefinite matrix $\sum_{i\in[s]} n_i M_i$ are $\lambda_1, \ldots, \lambda_m$ (counted with multiplicities), $\varphi_p(\boldsymbol{n})$ is defined by

$$\varphi_p(\boldsymbol{n}) = \operatorname{trace}\left(\sum_{i\in[s]} n_i M_i\right)^p = \sum_{k=1}^{m} \lambda_k^p.$$

We shall see that Problem (P_0) is the limit of Problem (P_p) as $p \to 0^+$ indeed. On the other hand, the limit of Problem (P_p) as $p \to 1$ is a knapsack problem (in fact, it is the trivial instance in which the ith item has weight 1 and utility $u_i = \operatorname{trace} M_i$). Note that a matrix M_i may be chosen n_i times in Problem (P_p), while choosing a matrix more than once in Problem (P_0) cannot increase the rank. Therefore we also define the binary variant of Problem (P_p):

$$\max_{\boldsymbol{n}} \left\{ \varphi_p(\boldsymbol{n}) : \boldsymbol{n} \in \{0,1\}^s, \sum_{i\in[s]} n_i \leq N \right\}. \qquad (P_p^{bin})$$

We shall also consider the case in which the selection of the ith matrix costs c_i, and a total budget B is allowed. This is the budgeted version of the problem:

$$\max_{\boldsymbol{n}} \left\{ \varphi_p(\boldsymbol{n}) : \boldsymbol{n} \in \mathbb{N}^s, \sum_{i\in[s]} c_i n_i \leq B \right\}. \qquad (P_p^{bdg})$$

Throughout this article, we use the term *design* for the variable $\boldsymbol{n} = (n_1, \ldots, n_s) \in \mathbb{N}^s$. We say that $\boldsymbol{n}$ is a *N-replicated design* if it is feasible for Problem (P_p), a *N-binary design* if $\boldsymbol{n}$ is feasible for Problem (P_p^{bin}), and a *B-budgeted design* when it satisfies the constraints of (P_p^{bdg}).

1.1. Motivation: optimal experimental design

The theory of *optimal design of experiments* plays a central role in statistics. It studies how to best select experiments in order to estimate a set of parameters. Under classical assumptions, the best linear unbiased estimator is given by least square theory, and lies within confidence ellipsoids which are described by a positive semidefinite matrix depending only on the selected experiments. The *optimal design of experiments* aims at selecting the experiments in order to make these confidence ellipsoids as small as possible, which leads to more accurate estimators.

A common approach consists in minimizing a scalar function measuring these ellipsoids, where the function is taken from the class of Φ_p-information functions proposed by Kiefer [16]. This leads to a combinatorial optimization problem (decide how many times each experiment should be performed) involving a spectral function which is applied to the information matrix of the experiments. For $p \in]0, 1]$, Kiefer's Φ_p-optimal design problem is equivalent to Problem (P_p) (up to the exponent $1/p$ in the objective function).

In fact, little attention has been given to the combinatorial aspects of Problem (P_p) in the optimal experimental design literature. The reason is that there is a natural relaxation of the problem which is much more tractable and usually yields very good results: instead of determining the exact number of times n_i that each experiment will be selected, the optimization is done over the fractions $w_i = n_i/N \in [0, 1]$, which reduces the problem to the maximization of a concave function over a convex set (this is the theory of *approximate optimal designs*). For the common case, in which the number N of experiments to perform is large and $N > s$ (where s is the number of available experiments), this approach is justified by a result of Pukelsheim and Rieder [26], who give a rounding procedure to transform an optimal approximate design $\boldsymbol{w}^*$ into an N-replicated design $\boldsymbol{n} = (n_1, \ldots, n_s)$ which approximates the optimum of the Kiefer's Φ_p-optimal design problem within a factor $1 - \frac{s}{N}$.

The present developments were motivated by a joint work with Bouhtou and Gaubert [4,30] on the application of optimal experimental design methods to the identification of the traffic in an Internet backbone. This problem describes an *underinstrumented situation*, in which a small number $N < s$ of experiments should be selected. In this case, the combinatorial aspects of Problem (P_p) become crucial. A similar problem was studied by Song et al. [31], who proposed to use a greedy algorithm to approximate the solution of Problem (P_p). In this paper, we give an approximation bound which justifies this approach. Another question addressed in this manuscript is whether it is appropriate to take roundings of (continuous) approximate designs in the underinstrumented situation (recall that this is the common approach when dealing with experimental design problems in the *overinstrumented* case, where the number N of experiments is large when compared to s).

Appendix A is devoted to the application to the theory of optimal experimental designs; we explain how a statistical problem (choose which experiments to conduct in order to estimate a set of parameters) leads to the study of Problem (P_p), with a particular focus to the *underinstrumented* situation described above. For more details on the subject, the reader is referred to the monographs of Fedorov [9] and Pukelsheim [25].