Series on
Applied Mathematics
Volume 19

PARTITIONS

Optimality and Clustering
Vol. I: Single-Parameter

Frank K Hwang
Uriel G Rothblum

World Scientific

Series on
Applied Mathematics
Volume 19

PARTITIONS

Optimality and Clustering
Vol. I: Single-Parameter

Frank K Hwang

National Chiao-Tung University, Taiwan

Uriel G Rothblum

Technion, Israel

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data
Hwang, Frank.
Partitions : optimality and clustering (v. 1. Single-parameter) / by Frank K. Hwang & Uriel G. Rothblum.
p. cm. -- (Series on applied mathematics ; v. 19)
Includes bibliographical references and index.
ISBN-13 978-981-270-812-0 (hardcover : alk. paper)
ISBN-10 981-270-812-X (hardcover : alk. paper)
1. Partitions (Mathematics). I. Title. II. Rothblum, Uriel G.
QA165 .H83 2011
512.7'3
2010038315

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Printed in Singapore by World Scientific Printers.

PARTITIONS

Optimality and Clustering
Vol. I: Single-Parameter

SERIES ON APPLIED MATHEMATICS

---

*Published**

Vol. 4 Computation of Differential Equations and Dynamical Systems
eds. K. Feng and Z.-C. Shi

Vol. 5 Numerical Mathematics
eds. Z.-C. Shi and T. Ushijima

Vol. 6 Machine Proofs in Geometry
by S.-C. Chou, X.-S. Gao and J.-Z. Zhang

Vol. 7 The Splitting Extrapolation Method
by C. B. Liem, T. Lü and T. M. Shih

Vol. 8 Quaternary Codes
by Z.-X. Wan

Vol. 9 Finite Element Methods for Integrodifferential Equations
by C. M. Chen and T. M. Shih

Vol. 10 Statistical Quality Control — A Loss Minimization Approach
by D. Trietsch

Vol. 11 The Mathematical Theory of Nonblocking Switching Networks
by F. K. Hwang

Vol. 12 Combinatorial Group Testing and Its Applications (2nd Edition)
by D.-Z. Du and F. K. Hwang

Vol. 13 Inverse Problems for Electrical Networks
by E. B. Curtis and J. A. Morrow

Vol. 14 Combinatorial and Global Optimization
eds. P. M. Pardalos, A. Migdalas and R. E. Burkard

Vol. 15 The Mathematical Theory of Nonblocking Switching Networks
(2nd Edition)
by F. K. Hwang

Vol. 16 Ordinary Differential Equations with Applications
by S. B. Hsu

Vol. 17 Block Designs: Analysis, Combinatorics and Applications
by D. Raghavarao and L. V. Padgett

Vol. 18 Pooling Designs and Nonadaptive Group Testing — Important Tools
for DNA Sequencing
by D.-Z. Du and F. K. Hwang

Vol. 19 Partitions: Optimality and Clustering
Vol. 1: Single-parameter
by F. K. Hwang and U. G. Rothblum

*For the complete list of the volumes in this series, please visit
http://www.worldscibooks.com/series/sam_series.shtml

Dedicated to our wives –

Aileen and Naomi

Preface

The need of optimal partition arises from many real-world problems involving the distribution of limited resources to many users. An important example is the "clustering" problem whose goal is to partition the data into clusters to minimize the intra-cluster distances and to maximize the inter-cluster distances. This book is the first attempt to collect all theoretical development on the topic of optimal partition to a single source. In fact, it does much more than simply collecting the results; it provides a general framework to unify these results and presents them in an organized and simplified way. There are also many new results in this book.

Because of its size, this book is partitioned (optimally, we hope) into two volumes. Volume 1 focuses on single-parameter partition problem where each element in the set to be partitioned is represented by a single parameter, i.e., a 1-dimensional point. We develop basic theory and methods to attack the optimal partition problem as a foundation to solve the multi-parameter partition problem in Volume 2. However, partition points in a multi-dimensional space have some fundamental differences from partitioning points in a line. For example, points in a line can be linearly ordered, but no such natural order exists for points in high-dimensional space, forcing the development of theory and methods unique to Volume 2.

We also collect a set of optimal-partition problems which have been discussed in the literature in the first chapter of Volume 1, and use the results in this book to give solutions to these problems (or explaining why they can't be solved) in the last chapter of Volume 2.

Earlier and shorter versions of this book have been taught thrice by the first author as a 1-year graduate course in National Chiao-Tung University. Many students were able to do research and to solve problems which were open then but now a part of the literature cited in this book.

In particular, we thank Dr. Yuchi Liu, Dr. Hongbin Chen, Dr. Feihuang Chang and Dr. Huilan Chang for proof-reading some parts of the book and making suggestions.

Special thanks are due to Ms. Lillian Bluestein for persistent help in typing and retyping many versions of chapters of both volumes of our book. Last, but not least, we thank Ron Rothblum for technical advice about TeX-setting that helped us overcome many burdens in shaping up the final outcome.

Contents

Chapter 1

Formulation and Examples

The first chapter is devoted to the introduction of a framework of reference for partition problems. In particular, partition problems are classified by three main characteristics: (i) the set of partitions over which an optimization problem takes place, (ii) the number of characteristics associated with each of the partitioned elements, and (iii) the objective function that is to be optimized. Further, 15 examples from diverse areas are introduced to demonstrate the expressive power of partition problems. Some of these examples are known as NP-hard, implying that the development of efficient solution methods is unlikely. Still, we show in Chapter 12 how the theory we develop can be used to solve most of these examples rather efficiently. Some of the examples we mention have broad modelling potential that is useful to describe complicated situations. The description of these examples in the current chapter focuses on basic scopes of the models. More general or more complicated variants of the models will be provided in Chapter 12 (along with the corresponding solution methods).

1.1 Formulation and Classification of Partitions

Consider a finite set $\mathcal{N}$ of distinct positive integers (for most of our development $\mathcal{N} = \{1, \ldots, |\mathcal{N}|\}$). A *partition of* $\mathcal{N}$ is a finite collection of sets $\pi = (\pi_1, \ldots, \pi_p)$ where $\pi_1, \ldots, \pi_p$ are pairwise disjoint nonempty sets whose union is $\mathcal{N}$. In this case we refer to p as the *size of* π, and to the sets $\pi_1, \ldots, \pi_p$ as the *parts of* π. Further, if $n_1, \ldots, n_p$ are the sizes of $\pi_1, \ldots, \pi_p$, respectively, we define the *shape of* π as the vector $(n_1, \ldots, n_p)$; of course, in this case $\sum_{j=1}^{p} n_j = |\mathcal{N}|$. We sometimes add prefix "p-" or "$(n_1, \ldots, n_p)$-" to explicitly express the size or shape of a partition, referring to a *p-partition of $\mathcal{N}$* or to an *$(n_1, \ldots, n_p)$-partition of $\mathcal{N}$*. Further, for brevity, we frequently omit the reference to the set $\mathcal{N}$ as the partitioned set, and simply refer to *partitions.* In our development, we sometimes require that partitions' parts are nonempty while at other times this requirement is relaxed.

At times, we restrict attention to the set of all partitions or to the set of all partitions whose size or shape satisfies prescribed restrictions; we refer, respectively, to *open*, *constrained-size* and *constrained-shape sets of partitions.* If the restrictions on the size or shape are expressed by prescribing a single element, then we refer to as *single-size* or *single-shape sets of partitions* and if the restrictions are in terms of lower and upper bounds on the size or shape we refer to as *bounded-size* or *bounded-shape sets of partitions*, respectively. Note that all of the above classes of sets of partitions can be treated as special cases of a constrained-shape class. Their respective names simply emphasize the kind of constraints on the shapes. For example, the class of p-partitions collects those partitions with p (nonempty) parts, and the class of open partitions collects all partitions without restriction on the number of (nonempty) parts. Thus, we will treat the constrained-shape class as the most general class. Still, the general framework of constrained-shape sets of partitions does not appear in the forthcoming development and whenever constrained-shape partition problems are mentioned, all shapes have the same size; consequently, we shall use the terminology "constrained shape" for set of partitions with restricted shapes that have the same size (though formally, these are *single-size constrained-shape sets of partitions*). Sometimes, we casually use the above adjectives that describe sets of partitions to partitions that belong to the corresponding given sets.

Let the (partitioned) set $\mathcal{N}$ be given and let $n \equiv |\mathcal{N}|$. When a single-size set of partitions is considered, the prescribed single-size is given as a positive integer p. Given a positive integer p and a set Γ of positive integer-vectors $(n_1, \ldots, n_p)$, each satisfying $\sum_{j=1}^{p} n_j = n$, let Π^{Γ} be the corresponding constrained shape partitions, that is, all partitions with shape in Γ. In particular, if Γ consists of a single vector $(n_1, \ldots, n_p)$, we use the notation $\Pi^{(n_1,\ldots,n_p)}$ for (the single-shape set of partitions) Π^{Γ}. Also, if L and U are nonnegative integer p-vectors satisfying $L \leq U$ and $\sum_{j=1}^{p} L_j \leq n \leq \sum_{j=1}^{p} U_j$, we let $\Gamma^{(L,U)}$ denote the set of nonnegative integer-vectors $(n_1, \ldots, n_p)$ satisfying $L_j \leq n_j \leq U_j$ for $j = 1, \ldots, p$; in this case, we use the notation $\Pi^{(L,U)}$ for (the bounded-shape set of partitions) $\Pi^{\Gamma^{(L,U)}}$. (The restrictions on L and U assure that $\Gamma^{(L,U)}$ and $\Pi^{(L,U)}$ are nonempty.) Note that single-size and single-shape sets of partitions are instances of bounded-shape sets obtained, respectively, by setting $L_j = 1$ and $U_j = n$ for all j or $L_j = U_j = n_j$ for all j. Similarly, open and single-size sets partitions are instances of bounded-size sets. The hierarchy of the classification of partitions is summarized in Table 1.1.1.

Table 1.1.1: Classification of sets of partitions

open	constrained-size	constrained-shape (size given)
	bounded-size	bounded-shape (size given)
	single-size	single-shape

A *multiset* is a group of elements where each element is allowed to have multiple occurrence. The formal notation of a multiset has double brackets, e.g., $\{\{1, 1, 2, 2, 3\}\}$, or is given as a bracketed list of distinct elements with superscripts designating their duplications, e.g., $\{1^2, 2^2, 3\}$. However, at times, we abuse notation and use single brackets, e.g., $\{1, 1, 2, 2, 3\}$.

It is implicitly assumed in the above definitions that the parts of partitions are distinguishable. But, in some applications the parts are indistinguishable and can be permuted without any restrictions. Thus, we also consider unlabeled partitions. Specifically, an *unlabeled partition of* $\mathcal{N}$ is a finite collection of sets $\pi = \{\pi_1, \ldots, \pi_p\}$ where the π_j's are as above. Again, we refer to p and to the sets $\pi_1, \ldots, \pi_p$ as the *size* and the *parts of* π, respectively. Further, if $n_1, \ldots, n_p$ are the sizes of $\pi_1, \ldots, \pi_p$, respectively, we define the *shape of* π as the multiset $\{\{n_1, \ldots, n_p\}\}$; again, we must have that $\sum_{j=1}^{p} n_j = |\mathcal{N}|$. In the literature, unlabeled partitions are commonly referred to as *allocations*. While we reserve the term "partitions" for labeled ones, we sometimes refer to *labeled partitions* (when potential ambiguity may arise).

We apply the same classification to sets of unlabeled partitions as we do to sets of labeled ones; see Table 1.1.1. Single-shape and bounded-shape sets of unlabeled partitions are defined, respectively, by a multiset $\{\{n_1, \ldots, n_p\}\}$ or a multiset of pairs $\{\{(L_1, U_1), \ldots, (L_p, U_p)\}\}$; the specification or the bounds on the sizes of the parts of partitions then hold for some labeling of the parts. Frequently, as parts of unlabeled partitions are indistinguishable, sizes and bounds of unlabeled partitions are uniform, namely all parts have the same size and a multiset of bounds $\{\{(L_1, U_1), \ldots, (L_p, U_p)\}\}$ consists of p identical pairs.

1.2 Formulation and Classification of Partition Problems over Parameter Spaces

In this section, we introduce the framework for partition problems over parameter spaces which are the main goal of this book. Specifically, a *partition problem* concerns the selection of a partition π out of a given set Π of partitions so as to optimize (that is, minimize or maximize) an objective function F that is defined over Π.

We assume throughout that each element i of the partitioned set $\mathcal{N}$ is *associated* with a vector $A^i \in \mathbb{R}^d$ where d is a fixed positive integer (independent of i); we refer to the coordinates of A^i as *parameters* or *characteristics* associated with i. The vectors $A^1, \ldots, A^n$ are part of the data of the problem and are given in the form of a real $d \times n$ matrix A. For a subset S of $\mathcal{N} = \{1, \ldots, n\}$, A^S is the submatrix of A consisting of the columns of A indexed by S, ordered as in A. Also, we use "bars" over matrices, to denote the multiset consisting of their columns, for example, a subset S of $\mathcal{N} = \{1, \ldots, n\}$, $\overline{A^S}$ is the set of columns of A^S, accounting for multiplicities.

An *objective function* $F(\cdot)$ that is to be maximized (or a cost function that is to be minimized). It associates a value $F(\pi)$ to each (feasible) p-partition π and this value depends on the parameters of the elements that are assigned to each part. In the most general case we consider, for each positive integer v, a column-symmetric function $h_v : \mathbb{R}^{d\times v} \to \mathbb{R}^d$, defined over multisets of v d-vectors, functions $g_j : \mathbb{R}^d \to \mathbb{R}^m$, $j = 1, \ldots, p$ and a function $f_p : \mathbb{R}^{d\times p} \to \mathbb{R}^d$. Then the value $F(\pi)$ associated with partition π having shape $(n_1, \ldots, n_p)$ is given by

$$F(\pi) = f_p\left(g_1[h_{n_1}(\overline{A^{\pi_1}})], \ldots, g_p[h_{n_p}(\overline{A^{\pi_p}})]\right) . \qquad (1.2.1)$$

The functions h_{n_j} can, in fact, depend on the location within the variables of f_p, that is, on the index j; also, the functions g_j may depend on n_j. In many common applications, each of the functions h_{n_j} is the summation function, in which we refer to the corresponding problems as *sum-partition problems.* When h_v or g_j is independent of the indexing parameter, we drop the index. Also, when referring to partitions of common size we drop the subscript "p" of f_p. We call f_p *additive* if f_p is the sum function. It is also possible to consider partition problems where the domain of the functions h_{n_j} consists of ordered lists (and the functions h_{n_j} are not symmetric).

For a more concise form of sum-partition problems, we introduce some notation. For a $d \times n$ real matrix A and a p-partition $\pi = (\pi_1, \ldots, \pi_p)$ of

$\mathcal{N}$, we define the π*-summation-matrix of* A, denoted A_π, by

$$A_\pi \equiv \left[\sum_{t\in\pi_1} A^t, \ldots, \sum_{t\in\pi_p} A^t\right] \in \mathbb{R}^{d\times p}, \tag{1.2.2}$$

where the empty sum is defined to be 0 (here and elsewhere in this book). When each of the g_j's is the identity over $\mathbb{R}^d$, the objective function F associates with partition π the value $F(\pi)$ with the representation

$$F(\pi) = f_p(A_\pi) \tag{1.2.3}$$

(as was already mentioned, when the optimization over partitions concerns only partitions of fixed size p, the dependence of the functions f_p on p is suppressed and we refer only to a real-valued function f on $\mathbb{R}^p$).

Of particular interest is the case where all A^i's have a common coordinate, say the first one, and it equals 1 for each A^i. In this case row 1 of A_π is the shape of π. It follows that (1.2.3) allows for the objective function F to depend on the shape of partitions. Of course, for single-shape problems, the part-sizes (that is, the coefficients of the shape) are fixed and can be viewed as parameters of the objective function.

In summary, the three major characteristics by which partition problems are classified are:

(1) The family of partitions Π over which the function F (with representation as in (1.2.1) or (1.2.3)) is considered and optimized: Using the classification of families of partitions provided in Table 1.1.1, we shall refer to *open, constrained-size, bounded-size, single-size, constrained-shape, bounded-shape, and single-shape partition problems*; of course, there is a natural hierarchy of this characteristic: single-shape and single-size are instances of bounded-shape which is an instance of constrained-shape. In addition, we refer to *relaxed-size problems* as single-size problems which allow for empty parts. Further, the description of the set Π of (feasible) partitions has to specify whether empty parts are allowed or are prohibited.

(2) The number of parameters associated with each of the partitioned elements: We shall refer to *single-parameter problems*, *two-parameter problems* and *multi-parameter problems*.

(3) The objective (cost) function F as expressed by (1.2.1): Adjectives like "sum-," "max-" or "mean-" of partition problems reflect properties of h_v while properties of f, like "linear," "convex," "Schur convex" and "separable," reflect properties of f_p; e.g., we refer to sum-partition problems with f Schur convex.