Computational Geometry 45 (2012) 33–53

Contents lists available at SciVerse ScienceDirect

Computational Geometry: Theory and Applications

www.elsevier.com/locate/comgeo

Maxima-finding algorithms for multidimensional samples: A two-phase approach

Wei-Mei Chen [a,*], Hsien-Kuei Hwang [b], Tsung-Hsi Tsai [b]

[a] *Department of Electronic Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan*
[b] *Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan*

ARTICLE INFO

Article history:
Received 12 August 2010
Accepted 17 August 2011
Available online 23 August 2011
Communicated by K. Clarkson

Keywords:
Maximal points
Pareto optimality
Dominance
Multi-objective optimization
Skyline

ABSTRACT

Simple, two-phase algorithms are devised for finding the maxima of multidimensional point samples, one of the very first problems studied in computational geometry. The algorithms are easily coded and modified for practical needs. The expected complexity of some measures related to the performance of the algorithms is analyzed. We also compare the efficiency of the algorithms with a few major ones used in practice, and apply our algorithms to find the maximal layers and the longest common subsequences of multiple sequences.

1. Introduction

A point $\mathbf{p} \in \mathbb{R}^d$ is said to *dominate* another point $\mathbf{q} \in \mathbb{R}^d$ if the coordinatewise difference $\mathbf{p} - \mathbf{q}$ has only non-negative coordinates and $\mathbf{p} - \mathbf{q}$ is not identically a zero vector, where the dimensionality $d \geqslant 1$. For convenience, we write $\mathbf{q} \prec \mathbf{p}$ or $\mathbf{p} \succ \mathbf{q}$. The non-dominated points in a sample are called the *maxima* or *maximal points* of the sample. Note that there may be two identical points that are both maxima according to our definition of dominance. Such a dominance relation among points has been one of the simplest and widely used partial orders due to its simplicity. We can define dually the corresponding *minima* of the sample by reversing the direction of the dominance relation.

1.1. Maxima in diverse scientific disciplines

Daily lives are full of tradeoffs or multi-objective decision problems with often conflicting factors; the numerous terms appeared in different scientific fields reveal the importance and popularity of maxima in theory, algorithms, applications and practice: maxima (or vector maxima) are sometimes referred to as *non-dominance*, *records*, *outer layers*, *efficiency*, or *non-inferiority* but are more frequently known as *Pareto optimality* or *Pareto efficiency* (with the natural derivative *Pareto front*) in econometrics, engineering, multi-objective optimization, decision making, etc. Other terms used with essentially the same denotation include *admissibility* in statistics, *Pareto front* (and the corresponding notion of *elitism*) in evolutionary algorithms, and *skyline* in database language; see [2,16,24,25] and the references therein and the books [20,22,28] for more information. They also proved useful in many computer algorithms and are closely related to several well-known problems,

* Corresponding author.
E-mail addresses: wmchen@mail.ntust.edu.tw (W.-M. Chen), hkhwang@stat.sinica.edu.tw (H.-K. Hwang), chonghi@stat.sinica.edu.tw (T.-H. Tsai).

0925-7721/$ – see front matter
doi:10.1016/j.comgeo.2011.08.001

including convex hulls, top-k queries, nearest-neighbor search, largest empty rectangles, minimum independent dominating set in permutation graphs, enclosure problem for rectilinear d-gon, polygon decomposition, visibility and illumination, shortest path problem, finding empty simplices, geometric containment problem, data swapping, grid placement problem, and multiple longest common subsequence problem to which we will apply our algorithms later; see [16,52] for more references.

We describe briefly here the use of maxima in the contexts of database language and multi-objective optimization problems using evolutionary algorithms.

Skylines in database queries are nothing but minima. A typical situation where the skyline operator arises is as follows; see [14] for details. Travelers are searching over the Internet for cheap hotels near the beach. Since the two criteria "lower price" and "shorter distance" are generally conflicting with each other and since there are often too many hotels to choose from, one is usually interested in those hotels that are non-dominated according to the two criteria; here dominance is defined using minima. Much time will be saved if the search or sort engine can automatically do this and filter out those that are dominated for database queriers (by, say clicking at the skyline operator). On the other hand, frequent spreadsheet users would also appreciate such an operator, which can find the maxima, minima or skyline of multidimensional data by simple clicks.

In view of these and many other natural applications such as e-commerce, multivariate sorting and data visualization, the skylines have been widely and extensively addressed in recent database literature, notably for low- and moderate-dimensional data, following the pioneering paper [14]. In addition to devising efficient skyline-finding algorithms, other interesting issues include top-k representatives, progressiveness, absence of false hits, fairness, incorporation of preference, and universality. A large number of skyline-finding algorithms have been proposed for various needs; see, for example, [5,14,33,45,48,51,56] and the references therein.

On the other hand, the last decade has witnessed a tremendous growth of interest in the study of multi-objective evolutionary algorithms (MOEAs), where the idea of maxima also appeared naturally in the form of non-dominated solutions (or elites). MOEAs provide a popular approach for multi-objective optimization, which identify the most feasible solutions lying on the Pareto front under various (often conflicting) constraints by repeatedly finding non-dominated solutions based on biological evolutionary mechanisms. These algorithms have turned out to be extremely fruitful in diverse engineering, industrial and scientific areas, as can be witnessed by the huge number of citations many papers on MOEA have received so far. Some popular schemes in this context suggested the maintenance of an explicit archive/elite for all non-dominated solutions found so far; see below and [29,43,47,57,58] and the references therein. See also [19] for an interesting historical overview.

Finally, maxima also arises in a random model for river networks (see [3,10]) and in an interesting statistical estimate called "layered nearest neighbor estimate" (see [11]).

1.2. Maxima, maximal layers and related notions

Maxima are often used for some ranking purposes or used as a component problem for more sophisticated situations. Whatever the use, one can easily associate such a notion to define multidimensional sorting procedures. One of the most natural ways is to "peel off" the current maxima, regarded as the first-layer maxima, and then finding the maxima of the remaining points, regarded then as the second-layer maxima, and so on until no point is left. The total number of such layers gives rise to a natural notion of *depth*, which is referred to as the *height* of the corresponding random, partially ordered sets in [13]. Such a maximal-layer depth is nothing but the length of the longest increasing subsequences in random permutations when the points are selected uniformly and independently from the unit square, a problem having attracted widespread interests, following the major breakthrough paper [4].

On the other hand, the maximal layers are closely connected to chains (all elements comparable) and antichains (all elements incomparable) of partially ordered set in order theory, an interesting result worthy of mention is the following dual version of Dilworth's theorem, which states that the size of the largest chain in a finite partial order is equal to the smallest number of antichains into which the partial order may be partitioned; see, for example, [41] for some applications.

In addition to these aspects, *maximal layers* have also been widely employed in multi-objective optimization applications since the concept was first suggested in Goldberg's book [34]. Based on identifying the maximal layer one after another, Srinivas and Deb [54] proposed the non-dominated sorting genetic algorithm (NSGA) to simultaneously find multiple Pareto-optimal points, which was later on further improved in [23], reducing the time complexity from $O(dn^3)$ to $O(dn^2)$. Jensen [40] then gave a divide-and-conquer algorithm to find the maximal layers with time complexity $O(n(\log n)^{d-1})$; see Section 5 for more details. See also [12,50] and the references cited there for more algorithms for maximal layers.

In the contexts of multi-objective optimization problems, elitism usually refers to the mechanism of storing some obtained non-dominated solutions into an external archive during the process of MOEAs because a non-dominated solution with respect to its current data is not necessarily non-dominated with respect to the whole feasible solutions. The idea of elitism was first introduced in [58] and is regarded as a milestone in the development of MOEAs [19]. Since the effectiveness of this mechanism relies on the size of the external non-dominated set, an elite archive with limited size was suggested to store the truncated non-dominated sets [43,58], so as to avoid the computational costs of maintaining all non-dominated