Learning the Empirical Hardness of Optimization Problems:
The Case of Combinatorial Auctions

Kevin Leyton-Brown, Eugene Nudelman, and Yoav Shoham*

Computer Science Department, Stanford University
Stanford CA 94305
{kevinlb;eugnud;shoham}@cs.stanford.edu

Abstract. We propose a new approach for understanding the algorithm-specific empirical hardness of $\mathcal{NP}$-Hard problems. In this work we focus on the empirical hardness of the winner determination problem—an optimization problem arising in combinatorial auctions—when solved by ILOG's CPLEX software. We consider nine widely-used problem distributions and sample randomly from a continuum of parameter settings for each distribution. We identify a large number of distribution-nonspecific features of data instances and use statistical regression techniques to learn, evaluate and interpret a function from these features to the predicted hardness of an instance.

1 Introduction

It is no secret that particular instances of $\mathcal{NP}$-Hard problems can be quite easy to solve in practice. In recent years researchers in the constraint programming and artificial intelligence communities have studied the *empirical* hardness of individual instances or distributions of $\mathcal{NP}$-Hard problems, and have often managed to find simple mathematical relationships between features of the problem instances and the hardness of the problem. The majority of this work has focused on decision problems: that is, problems that ask a yes/no question of the form, "Does there exist a solution meeting the given constraints?". The most successful approach for understanding the empirical hardness of such problems—taken for example in [3, 1]—is to vary some parameter of the input looking for a hard-easy-hard transition corresponding to a phase transition in the solvability of the problem. This approach uncovered the famous result that 3-SAT instances are hardest when the ratio of clauses to variables is about 4.3; it has also been applied to other decision problems such as quasigroup completion [7]. Another approach rests on a notion of backbone [17, 1], which is the set of solution invariants.

* We would like to acknowledge contributions by Rámon Béjar, Carla Gomes, Henry Kautz, Bart Selman, Lyle Ungar and Ioannis Vetsikas.

P. Van Hentenryck (Ed.): CP 2002, LNCS 2470, pp. 556–572, 2002.

1.1 Empirical Hardness of Optimization Problems

Some researchers have also examined the empirical hardness of optimization problems, which ask a real-numbered question of the form, "What is the best solution meeting the given constraints?". These problems are clearly different from decision problems, since they always have solutions and hence cannot give rise to phase transitions in solvability. One way of finding hardness transitions related to optimization problems is to transform them into decision problems of the form, "Does there exist a solution with the value of the objective function $\geq x$?" This approach has yielded promising results when applied to MAX-SAT and TSP. Other experimentally-oriented work includes the extension of the concept of backbone to optimization problems [24], although it is often difficult to define for arbitrary problems and can be costly to compute.

A second approach is to attack the problem analytically rather than experimentally. For example, Zhang performed average case theoretical analysis of particular classes of search algorithms [25]. Though his results rely on independence assumptions about the branching factor and heuristic performance at each node of the search tree that do not generally hold, the approach has made theoretical contributions—describing a polynomial/exponential-time transition in average-case complexity—and shed light on real-world problems. Korf and Reid predict the average number of nodes expanded by a simple heuristic search algorithm such as A* on a particular problem class by making use of the distribution of heuristic values in the problem space [14]. As above, strong assumptions are required: e.g., that the branching factor is constant and node-independent, and that edge costs are uniform throughout the tree.

Both experimental and theoretical approaches have sets of problems to which they are not well suited. Existing experimental techniques have trouble when problems have high-dimensional parameter spaces, as it is impractical to manually explore the space of all relations between parameters in search of a phase transition or some other predictor of an instance's hardness. This trouble is compounded when many different data distributions exist for a problem, each with its own set of parameters. Theoretical approaches are also difficult when the input distribution is complex or is otherwise hard to characterize, but they also have other weaknesses. They tend to become intractable when applied to complex algorithms, or to problems with variable and interdependent edge costs and branching factors. Furthermore, they are generally unsuited to making predictions about the empirical hardness of individual problem instances, instead concentrating on average (or worst-case) performance on a class of instances.

1.2 Our Methodology

Some optimization problems do not invite study by existing experimental *or* theoretical approaches: problems characterized by a large number of apparently relevant features, the existence of many, highly parameterized distributions, significant variation in edge costs throughout the search tree and the desirability