Stochastic Extreme Value Optimization

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Master of Science

in

Computer Science and Technology

by

Liu Yu

Thesis Supervisor: Assistant Professor Li Jian

January 2017

Stochastic Extreme Value Optimization

by

Liu Yu

Submitted to the Institute for Interdisciplinary Information Sciences
in partial fulfillment of the requirements for the degree of

Master of Science

at

TSINGHUA UNIVERSITY

January 2017

© TSINGHUA UNIVERSITY 2017. All rights reserved.

Author .....
Institute for Interdisciplinary Information Sciences
Oct 28, 2016

Certified by .....
Li Jian
Assistant Professor
Thesis Supervisor

Abstract

In this thesis, we study several stochastic optimization problems in which we want to optimize the expectation of the extreme values, including maximum-element , minimization version of maximum-element and minimum-element . Formally, we obtain the following results:

1. We first study maximum-element . It takes a set of random variables $\{X_i\}_{i=1}^n$ (with known distributions) and a collection of feasible sets (constraint) $\mathcal{F} \subset 2^{[n]}$ as inputs, and outputs a feasible set $S \in \mathcal{F}$ which maximizes the value of $\mathbb{E}[\max_{i \in S} X_i]$. In the minimization version of maximum-element , we try to minimize the value of $\mathbb{E}[\max_{i \in S} X_i]$.

   We obtain a constant approximation for maximum-element if there is a constant approximation for the deterministic problem max-sum(find a feasible set $S \in \mathcal{F}$ which maximizes the value of $\sum_{i \in S} v_i$). Compared with previous results, our approximation works for many combinatorial constraints. Furthermore, we obtain the first PTAS for maximum-element if there is a psuedopolynomial time algorithm for the exact-sum problem. The best previous approximation ratio is $1 - 1/e$ [6], based on submodular maximization.

2. We can obtain the same constant approximation and PTAS for the minimization version of maximum-element , by a slight modification of the above algorithm.

3. We also study the minimum-element problem, which takes a set of random variables $\{X_i\}_{i=1}^n$ with costs $\{c_i\}_{i=1}^n$ and a budget C as inputs, and outputs a set S with total cost at most C which minimizes the value of $\mathbb{E}[\min_{i \in S} X_i]$. If all variables are supported on $\{0; 1; : : : ; m-1\}$, we obtain a bi-criteria algorithm which outputs a $(1+\varepsilon)$-approximation with cost at most $O(\log \log \log m)C$, improving the best previous result which incurs a cost at most $O(\log \log m)C$ [10].

Contents

1 Introduction 1

1.1 Motivation 1

1.2 Related Work 2

1.3 Our Contributions 3

1.3.1 MAXIMUM-ELEMENT 5

1.3.2 MINIMUM-ELEMENT 6

1.4 Structure of this Thesis 7

2 Preliminaries 9

2.1 Terminologies 9

3 MAXIMUM-ELEMENT 13

3.1 Problem Formulation 13

3.2 Hardness 14

3.3 A Constant Approximation 15

3.3.1 Special Case: Bernoulli Distributions 16

3.3.2 General Cases 18

3.4 PTAS 22

3.4.1 Discretization 22

3.4.2 Signatures 25

3.4.3 Enumerating Signatures 27

3.5 Minimization version of maximum-element 28

3.6 MAX_K 31

4 MINIMUM-ELEMENT 35

4.1 Problem Formulation 35

4.2 Hardness 35

4.3 CIP Approach 36

4.4 Our CIP Approach 37

5 Conclusion 45

Bibliography 47

Curriculum Vitae 52

Chapter 1

Introduction

1.1 Motivation

Recent years data sets grow rapidly. They are increasingly collected by mobile devices, cameras, remote sensors, software logs and other cheap and numerous information-sensing devices. Along with the increasing of the size of data collected, uncertainty arises in many applications, including social networks, sensor networks, databases, social science economics and so on. Many applications require supports of methods for dealing with uncertain data. Here we give some examples to show where the uncertainty comes and how we model them.

- Bayesian Optimal Mechanism Design [19]: In economics, mechanism design (in particular auctions) is an important research area. A good mechanism should allocate the items and decides the payment based on the private value of the items of each bidder. However no one knows the exact value except the bidder herself. Hence, we model her value as a random variable following some prior distribution (probably computed from historical data of the bidder), and design mechanisms according to the distributions.

- Sensor Network: Sensor networks spread everywhere and produce large amount of data. Due to physical noise or detection precision, much uncertain data is produced at the same time. In such case, uncertain sensor readings are usually modeled by

probabilistic models. For example, Deshpande et al. [8] modeled data as Gaussian distributions.

Stochastic optimization is a method for tackling uncertainty by modeling uncertainty as probability distributions of inputs. The field was initiated by Dantzig [7]. In the follow decades, it has been studied extensively, and attracts more and more attentions from computer science community in recent years. In my dissertation research, I focus on one class of important problems in this domain: stochastic extreme value optimization.

Stochastic extreme value optimization: In a large variety of practical problems, people concern the extreme value (i.e., the maximum value, the minimum value). Here we list some examples.

- In a first price auction, the price of the item is determined by the highest bid.
- Consider a scenario where one agent wants to produce some object. He delivers the jobs of producing components to different factories. The factory with the latest complete time determines the complete time of the whole production.

In such problems, people want to optimize an extreme value as it affects results significantly. When involving uncertain data, we call such problems "stochastic extreme value optimization".

1.2 Related Work

In the past decades the stochastic combinatorial optimization has been widely and deeply studied. Many combinatorial optimization problems themselves are NP-hard. The problem usually looks harder involving uncertainty and people proved some of them are actually harder. For example, a variant of adaptive stochastic knapsack is PSPACE-hard [3]. While an optimal solution is hard to find, people make efforts to find approximate solutions efficiently. People have developed many techniques to deal with approximating problems in the

field. More and more problems are studied in different models: Adaptive stochastic knapsack [3, 17], stochastic matching [1, 2], stochastic set cover [22, 13], stochastic bin packing [17] and so on. We refer interested reader to [16] for a comprehensive survey.

In some stochastic optimization problems, it is possible to make observation on inputs. Upon one observation of one input item, we can find out the exact value (which is a sample from the given distribution). Observations need extra cost. Hence we should decide which inputs to observe to maximize the improvement using as little cost as possible. Goel et al. [10] modeled such situations and defined a class of stochastic optimization problems called "model-driven optimization", which generalizes the stochastic extreme value optimization with knapsack constraint.

1.3 Our Contributions

First we define two stochastic extreme value optimization problems maximum-element and minimum-element formally here.

Problem 1.3.1 (maximum-element). *Inputs: a set of independent non-negative random variables* $\{\mathsf{X}_i\}_{i=1}^n$ *and a collection of feasible solutions (a combinatorial constraint)* $\mathcal{F} \subset 2^{[n]}$*; Outputs: a set* $\mathsf{S} \in \mathcal{F}$*, which maximizes the value* $\mathbb{E}[\max_{i \in \mathsf{S}} \mathsf{X}_i]$.

Note that $\mathcal{F}$ may contain exponential number of feasible sets and is represented implicitly in the input. For example, each random variable is associated with an edge of a graph and $\mathcal{F}$ is the set of all spanning trees of the graph. Distributions of random variables can be inputted in two ways: lists of supports and corresponding probabilities, which works for discrete random variables; and oracles returning samples following the corresponding distributions, which works for both discrete and continuous random variables. In this thesis, we only consider discrete random variables, leaving the case of continuous random variables as future work.

Problem 1.3.2 (minimum-element). *Inputs: a set of independent non-negative random variables* $\{\mathsf{X}_i\}_{i=1}^n$*, costs* $\{\mathsf{c}_i\}$*, and a budget* C*;*