Approximation Algorithms for Distributionally Robust Stochastic Optimization

by

André Linhares Rodrigues

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Combinatorics and Optimization

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Nicole Megow Professor, Institute of Computer Science, Universität Bremen
Supervisor:	Chaitanya Swamy Professor, Department of Combinatorics and Optimization, University of Waterloo
Internal Member:	Jochen Könemann Professor, Department of Combinatorics and Optimization, University of Waterloo
Internal Member:	Laura Sanità Associate Professor, Department of Combinatorics and Optimization, University of Waterloo
Internal-External Member:	Lap Chi Lau Associate Professor, Cheriton School of Computer Science, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The results in this thesis are chiefly based on the publication [85], which is joint work with my supervisor Chaitanya Swamy. I have made a major contribution toward this publication.

Abstract

Two-stage stochastic optimization is a widely used framework for modeling uncertainty, where we have a probability distribution over possible realizations of the data, called scenarios, and decisions are taken in two stages: we take first-stage actions knowing only the underlying distribution and before a scenario is realized, and may take additional second-stage recourse actions after a scenario is realized. The goal is typically to minimize the total expected cost. A common criticism levied at this model is that the underlying probability distribution is itself often imprecise. To address this, an approach that is quite versatile and has gained popularity in the stochastic-optimization literature is the *two-stage distributionally robust stochastic model*: given a collection D of probability distributions, our goal now is to minimize the maximum expected total cost with respect to a distribution in D.

There has been almost no prior work however on developing approximation algorithms for distributionally robust problems where the underlying scenario collection is discrete, as is the case with discrete-optimization problems. We provide frameworks for designing approximation algorithms in such settings when the collection D is a ball around a central distribution, defined relative to two notions of distance between probability distributions: Wasserstein metrics (which include the L_1 metric) and the L_∞ metric. Our frameworks yield efficient algorithms even in settings with an *exponential* number of scenarios, where the central distribution may only be accessed via a *sampling oracle*.

For distributionally robust optimization under a Wasserstein ball, we first show that one can utilize the *sample average approximation* (SAA) method—solve the distributionally robust problem with an empirical estimate of the central distribution—to reduce the problem to the case where the central distribution has a polynomial-size support, and is represented explicitly. This follows because we argue that a distributionally robust problem can be reduced in a novel way to a standard two-stage stochastic problem with bounded inflation factor, which enables one to use the SAA machinery developed for two-stage stochastic problems. Complementing this, we show how to approximately solve a fractional relaxation of the SAA problem (i.e., the distributionally robust problem obtained by replacing the original central distribution with its empirical estimate). Unlike in two-stage {stochastic, robust} optimization with polynomially many scenarios, this turns out to be quite challenging. We utilize a variant of the ellipsoid method for convex optimization in conjunction with several new ideas to show that the SAA problem can be approximately solved provided that we have an (approximation) algorithm for a certain max-min problem that is akin to, and generalizes, the k-max-min problem—find the worst-case scenario consisting of at most k elements—encountered in two-stage robust optimization. We ob-

tain such an algorithm for various discrete-optimization problems; by complementing this via rounding algorithms that provide *local* (i.e., per-scenario) approximation guarantees, we obtain the *first* approximation algorithms for the distributionally robust versions of a variety of discrete-optimization problems including set cover, vertex cover, edge cover, facility location, and Steiner tree, with guarantees that are, except for set cover, within O(1)-factors of the guarantees known for the deterministic version of the problem.

For distributionally robust optimization under an L_∞ ball, we consider a fractional relaxation of the problem, and replace its objective function with a proxy function that is pointwise close to the true objective function (within a factor of 2). We then show that we can efficiently compute approximate subgradients of the proxy function (for a certain notion of approximate subgradients introduced by Shmoys and Swamy [114]), provided that we have an algorithm for the problem of computing the t worst scenarios under a given first-stage decision, given an integer t. We can then approximately minimize the proxy function via a variant of the ellipsoid method by [114], and thus obtain an approximate solution for the fractional relaxation of the distributionally robust problem. Complementing this via rounding algorithms with local guarantees, we obtain approximation algorithms for distributionally robust versions of various covering problems, including set cover, vertex cover, edge cover, and facility location, with guarantees that are within O(1)-factors of the guarantees known for their deterministic versions.

Acknowledgements

First, I would like to thank Chaitanya Swamy for being such a dedicated supervisor. Our weekly meetings have been a constant source of inspiration and encouragement, and have been invaluable for the completion of this thesis. I am also thankful for his very detailed and constructive advice and feedback regarding academic writing and talks, as well as other academic and career matters.

I would also like to thank the other members of the examining committee, namely Nicole Megow, Jochen Könemann, Laura Sanità, and Lap Chi Lau, for devoting time to reading this thesis and providing helpful feedback.

Thank you to the other C&O professors from whose courses I have benefited, and with whom I had the pleasure to collaborate, especially Ricardo Fukasawa, Jochen Könemann, Joseph Cheriyan, and Laura Sanità. I am also grateful to the members of the department's administrative team for their assistance, particularly Melissa, Carol, and Megan; and to my fellow graduate students for their friendship.

Finally, I am deeply grateful to my parents for their unwavering and unconditional support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we introduce the model studied in this thesis in Section 1.1, and give a high-level overview of our main contributions in Section 1.2 (we postpone a more detailed overview to Section 3.3). In Section 1.3, we list some basic mathematical definitions, notation, and conventions that we use throughout this thesis. In Section 1.4, we outline the contents of each of the following chapters.

1.1 The distributionally robust stochastic optimization framework

In practical applications of optimization, one often encounters problems involving uncertain parameters—for example, parameters that cannot be measured exactly, or that depend on future events that cannot be predicted with certainty. A naive approach for tackling such problems consists in computing estimates of all the uncertain parameters, and solving the deterministic problem obtained by utilizing these estimates in lieu of the real (uncertain) parameters. Unsurprisingly, this approach can lead to unsatisfactory (i.e., low-quality or even infeasible) decisions. Therefore, developing models that incorporate uncertainty in the parameters is crucial for making satisfactory decisions in such settings.

An important and widely used model is the *two-stage recourse model*, wherein we seek to take actions in two stages. In the first stage, we make a *here-and-now decision* x, using only limited information (or no information at all) on the uncertain parameters. In the second stage, a *scenario* A is revealed (i.e., the values of the uncertain parameters become known), and we can make a *recourse decision* z^A in order to satisfy the requirements imposed in this

scenario. The second-stage actions are typically costlier than the corresponding first-stage actions, as they may entail making decisions in rapid reaction to the observed scenario (e.g., deploying resources with smaller lead time).

An oft-cited prototypical example is *two-stage facility location*, wherein we need to decide where to set up facilities to serve clients, in the face of uncertain client demands. We can open some facilities initially, given only limited information about demands; after a specific demand pattern is realized, we can take additional recourse actions such as opening more facilities, incurring their recourse costs. Different objective functions can be adopted, depending in particular on the level of risk aversion of the decision maker and on the granularity of the information on the uncertain parameters that is made available to them. This choice leads to several variants of the two-stage recourse model; two popular models are two-stage {robust, stochastic} optimization. We briefly define these models and discuss some issues that may arise when utilizing them. We then define the model that is the focus of this thesis, namely *two-stage distributionally robust stochastic optimization*, which mitigates these issues by *interpolating* between robust and stochastic optimization. We defer a discussion of previous work related to these models to Chapter 2.

Two-stage robust optimization. If no information on the likelihood of the different scenarios is available, or if the decision maker is highly averse to risk, a suitable model is *two-stage robust optimization*, wherein we seek to minimize the total cost in the worst-case scenario. Formally, we consider the problem

$$\min_{\left(x,\{z^A\}_{A\in\mathcal{A}}\right)} \left\{ (\text{cost of } x) + \max_{A\in\mathcal{A}} \left\{ \text{cost of } z^A \right\} \right\},$$

where $\mathcal{A}$ denotes the set of all possible scenarios. While this model can be appealing in the circumstances mentioned above, it is in some cases overly cautious, leading to severely undesirable decisions. One potential issue is that the existence of a single catastrophic scenario may force us to take first-stage actions that are unhelpful for all the remaining scenarios, and this may occur even if such a scenario is extremely unlikely.

Two-stage stochastic optimization. If some information on the likelihood of the different scenarios is available, a less conservative approach is to leverage this information and seek instead a decision that is desirable *in expectation*. This gives rise to another prominent model, namely *two-stage stochastic optimization*, wherein we seek to minimize the total expected cost incurred (with respect to the scenario realized). That is, if the scenario is

modeled as a random variable drawn from $\mathcal{A}$ according to a probability distribution p, then we consider the problem

$$\min_{(x,\{z^A\}_{A\in\mathcal{A}})} \left\{(\text{cost of } x) + \mathbb{E}_{A\sim p}\big[\text{cost of } z^A\big]\right\},$$

where $\mathbb{E}_{A\sim p}[\cdot]$ denotes the expectation when A is chosen according to p. A significant issue that occurs when using this model to capture real-world applications, which is a common source of criticism, is that the probability distribution p modeling the uncertainty is itself often imprecise. Usually, one computes a distribution p based on some historical data. While historical data may serve as a reasonably accurate representation of the behavior of the underlying unknown parameters, it is not sufficient to give an *exact* characterization of this behavior. For example, a scenario that occurs with extremely low (but positive) probability is unlikely to be observed in the historical data, so an empirical distribution computed in this way would be likely to incorrectly assign probability zero to such a scenario. As another example, suppose we observe a sequence of N^2 coin tosses, such that in each batch of N coin tosses the number of heads is between $0.49N$ and $0.51N$. Then, while it is possible that this sequence arose from a coin without bias (i.e., the probability of heads is equal to 0.5), all we can really say is that there is a range of biases concentrated around 0.5 under which the above statistic is likely.[1]

Two-stage distributionally robust stochastic optimization. The issues encountered in {robust, stochastic} optimization that we discussed above motivate the study of models that leverage information on the likelihood of the different scenarios, but do not assume knowledge of the exact underlying distribution. As mentioned before, one usually models the distribution to be statistically consistent with some historical data, so we really have a *collection of probability distributions*, and a more robust approach is to *hedge against the worst-possible probability distribution in this collection*. (Note that the worst-possible distribution depends on the chosen first-stage decision x.) This gives rise to the model that is the focus of this thesis, namely *two-stage distributionally robust stochastic optimization*. The setup is similar to that of the two-stage stochastic model, but we now have a collection D of probability distributions, which we refer to as the *ambiguity set*; our goal is to minimize the maximum expected total cost with respect to a distribution in D.

[1] More precisely, for any confidence level $\delta > 0$, there is some $\varepsilon > 0$ (depending on δ and N) such that for any coin with bias $p \in [0.5 - \varepsilon, 0.5 + \varepsilon]$, the probability of seeing the above statistics from a coin of bias p is at least $1 - \delta$.

That is, we consider the problem

$$\min_{\left(x,\{z^A\}_{A\in\mathcal{A}}\right)} \left\{ (\text{cost of } x) + \sup_{p\in D} \mathbb{E}_{A\sim p}\big[\text{cost of } z^A\big] \right\}.$$

Distributionally robust stochastic (DRS) optimization is a versatile approach dating back to Scarf [103] that has regained interest recently in the Operations-Research literature, where it is sometimes called *data-driven* or *ambiguous* stochastic optimization (see, e.g., [16, 43, 52, 126], and the references therein).

The two-stage DRS model also serves to nicely interpolate between the extremes of: (a) two-stage stochastic optimization, which optimistically assumes that the underlying distribution p is known precisely (which can be captured by setting $D = \{p\}$); and (b) two-stage robust optimization, which abandons the distributional view and seeks to minimize the maximum cost incurred in a scenario, thereby adopting the overly cautious approach of being robust against *every* possible scenario, regardless of how likely it is for a scenario to materialize (this can be captured by letting $D = \{\text{all distributions over } \mathcal{A}\}$, where $\mathcal{A}$ is the scenario collection in the two-stage robust problem; alternatively, we could take D to be the collection of all distributions over $\mathcal{A}$ concentrated at a single scenario $A \in \mathcal{A}$). Both extremes can lead to suboptimal decisions: with robust optimization, the presence of a single scenario, however unlikely, may lead to decisions that are undesirable for all other scenarios; with stochastic optimization, the optimal solution for a specific distribution p could be quite suboptimal even for a "nearby" distribution q, as illustrated by the following example.

Example: consider an instance of *two-stage set cover* with a single element e and a single set $S = \{e\}$. Suppose that the first-stage and second-stage costs of buying S are 1 and $\frac{M}{\varepsilon}$ respectively, where $\varepsilon \in (0, 1]$ and $M \gg 1$. The collection of scenarios is $\mathcal{A} := \{\emptyset, \{e\}\}$; a scenario specifies which elements must be covered. Let p be the probability distribution with $p_\emptyset = 1$ and $p_{\{e\}} = 0$, and q be the probability distribution with $q_\emptyset = 1 - \varepsilon$ and $q_{\{e\}} = \varepsilon$. The optimal solution under p is to not buy S in the first stage; this incurs cost $(1 - \varepsilon) \cdot 0 + \varepsilon \cdot \frac{M}{\varepsilon} = M$ under q, but the optimal solution under q is to buy S in the first stage and incur cost 1. This shows that even if $\|p - q\| \leq \mathrm{O}(\varepsilon)$, we can find instances where an optimal decision for p is undesirable under q.

1.2 Our contributions

Despite the modeling benefits and popularity of DRS optimization, to our knowledge, there has been almost no prior work on developing approximation algorithms for *discrete two-stage DRS problems*, and, more generally, for two-stage DRS problems with a *discrete* underlying scenario set (as is the case in discrete optimization). (The exception is Agrawal, Ding, Saberi, and Ye [2], which we discuss in Section 2.3; peripherally related is Wu, Du, and Xu [130], who consider a DRS version of facility location where the uncertainty only affects the costs and not the constraints, which yields a much simpler and more restrictive model.) In this thesis, we provide frameworks for designing efficient approximation algorithms in such settings. In this section, we give a high-level overview of our contributions. We refer the reader to Section 3.3 for a more precise and detailed exposition of our main results and the techniques we use to obtain them, as well as a summary of the approximation factors we obtain for various applications.

We develop *general frameworks* for designing approximation algorithms for discrete two-stage DRS optimization problems where the ambiguity set D is a ball around a central distribution $\mathring{p}$ under some metric L over probability distributions; that is, we have $D = \{p : L(\mathring{p}, p) \leq r\}$, where $r > 0$ is the radius of the ball. We consider three choices for the metric L: (i) the L_∞ metric, defined by $L_\infty(p, q) := \max_{A \in \mathcal{A}} |p_A - q_A|$; (ii) the $\frac{1}{2}L_1$ metric (also known as the *total-variation distance*), defined by $\frac{1}{2}L_1(p, q) := \frac{1}{2}\sum_{A \in \mathcal{A}} |p_A - q_A|$; and (iii) Wasserstein metrics, a rich class of metrics obtained by lifting an underlying scenario metric $\ell : \mathcal{A} \times \mathcal{A} \to \mathbb{R}_+$ to a metric over distributions, which includes the $\frac{1}{2}L_1$ metric as a special case (see Definition 3.1). Our results hold under the *bounded-inflation assumption*, which roughly speaking encodes that each first-stage action has a corresponding second-stage action whose cost is at most λ times higher, for a given *inflation factor* $\lambda \geq 1$. The cardinality of the scenario collection $\mathcal{A}$ may be very large, even *exponential in the input size*; the central distribution $\mathring{p}$ may only be accessed via a *sampling oracle*.

DRS optimization under a Wasserstein ball. For a first-stage decision x and a scenario A, let $g(x, A)$ denote the minimum cost incurred in extending x to a feasible solution for A, if we allow *fractional second-stage decisions*. For a broad class of problems, we relate the approximability of a discrete two-stage DRS optimization problem under a Wasserstein ball (where the Wasserstein metric is defined relative to a scenario metric $\ell : \mathcal{A} \times \mathcal{A} \to \mathbb{R}_+$) to the approximability of the problem of computing

$$g(x, y, A) := \max_{A' \in \mathcal{A}} \{g(x, A') - y \cdot \ell(A, A')\},$$

given an integer first-stage decision x, a number $y \geq 0$, and a scenario A. Informally, our main result (see Theorem 3.6) is that we can compute an $O(\alpha\beta_1\beta_2\rho)$-approximate solution for a discrete DRS problem in $\mathsf{poly}(\text{input size}, \lambda)$ time as long as we have the following three ingredients:

(i) a *second-stage (LP-relative) α-approximation algorithm*, which is an algorithm that given an integer first-stage decision x and a scenario A, computes an integer second-stage decision z^A of cost at most $\alpha \cdot g(x, A)$ such that $\big(x, z^A\big)$ is feasible for scenario A;

(ii) a (β_1, β_2)-*approximation algorithm for computing* $g(x, y, A)$, which is an algorithm that given a tuple (x, y, A) computes a scenario $\overline{A} \in \mathcal{A}$ such that

$$g\big(x, \overline{A}\big) - y \cdot \ell\big(A, \overline{A}\big) \geq \max_{A' \in \mathcal{A}} \left\{ \frac{1}{\beta_1} g(x, A') - \beta_2 y \cdot \ell(A, A') \right\} ;$$

and

(iii) a *local ρ-approximation algorithm*, which is an algorithm that rounds a fractional first-stage decision to an integer one while incurring at most a ρ-factor blow-up in the first-stage cost, and in the cost of each scenario.

The proof of the result above has two main components, which are of independent interest. The first component (see Chapter 4) is a sample-average-approximation (SAA) result for DRS optimization under a Wasserstein ball, which reduces the discrete DRS problem (with a central distribution $\mathring{p}$ given by a sampling oracle) to a collection of SAA problems with fractional second-stage decisions, wherein $\mathring{p}$ is replaced with an empirical estimate $\widehat{p}$, constructed using $\mathsf{poly}(\text{input size}, \lambda)$ samples. The second component (see Chapter 5) is an approximation algorithm for the SAA problems; this is obtained via a variant of the ellipsoid method, using ingredients (ii) and (iii). Combining these two components, we obtain an integer first-stage decision, and fractional second-stage decisions, which we then convert into integer second-stage decisions using ingredient (i).

DRS optimization under an L_∞ ball. Our main result for discrete DRS optimization under an L_∞ ball (see Theorem 3.9) is that, for a broad class of problems, we can compute an $O(\rho)$-approximate solution in $\mathsf{poly}\big(\text{input size}, \lambda, \frac{1}{r}\big)$ time, as long we have the following ingredients:

(i) an algorithm that given a fractional first-stage decision and a number $t \leq \min\big\{|\mathcal{A}|, \frac{1}{r}\big\}$, computes the t worst scenarios under the given first-stage decision in $\mathsf{poly}(\text{input size}, t)$ time; and

(ii) a local ρ-approximation algorithm.

The proof of this result is presented in Chapter 7, and relies on a variant of the ellipsoid method based on approximate subgradients by Shmoys and Swamy [114].

Applications. By applying the frameworks mentioned above for DRS optimization under a Wasserstein ball or an L_∞ ball, and furnishing the ingredients required by them, we obtain the *first* approximation results for DRS versions of a variety of problems, including set cover, edge cover, vertex cover, facility location, and Steiner tree. Tables 3.1 and 3.2 show the approximation factors that we obtain for various choices of the scenario collection $\mathcal{A}$ and of the scenario metric ℓ (in the Wasserstein setting).

1.3 Basic definitions, notation, and conventions

In this section we state some basic definitions, notation, and conventions that will be used throughout the thesis.

We use $a := b$ to denote the fact that a is defined as b. For a set S, we denote by 2^S the collection of all subsets of S. For a vector $u \in \mathbb{R}^S$, indexed by S, we denote by $\|u\| := \sqrt{\sum_{e \in S} u_e^2}$ its Euclidean norm, and by $\mathsf{supp}(u) := \{e \in S : u_e \neq 0\}$ its support. For an integer n, we let $[n] := \{1, 2, \ldots, n\}$ if $n \geq 1$, and $[n] := \emptyset$ otherwise.

We denote by $\log_b a$ the logarithm of a to base b. We often only care about the asymptotic growth of an expression; in such cases, the base is not relevant and we may omit it. We denote by $\ln a$ the natural logarithm of a.

We denote by $\Pr[E]$ the probability of an event E. We denote by $\mathbb{E}_{\xi \sim p}[v]$ the expectation of an expression v when a random variable ξ is randomly chosen according to a probability distribution p. When the distribution of the random variable ξ is clear from the context, we may simply write $\mathbb{E}[v]$.

Consider a real-valued function $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^n$. A vector $d \in \mathbb{R}^n$ is a *subgradient of f at $u \in D$* if we have $f(v) \geq f(u) + d^\intercal(v - u)$ for every $v \in D$. It is well known that a convex function has a subgradient at every point in the relative interior of its domain (see, e.g., Theorem 23.4 of Rockafellar [99]). We say that f is *K-Lipschitz continuous* (for some nonnegative number K) if we have $|f(u) - f(v)| \leq K \|u - v\|$ for every $u, v \in D$. The smallest number K (if any) for which this holds is called the *Lipschitz constant* of f. The following well known result shows that the existence of subgradients

with bounded Euclidean norm implies a bound on the Lipschitz constant of f (see, e.g., Claim 4.11 of Shmoys and Swamy [114]).

Lemma 1.1. *Let $f : D \to \mathbb{R}$ be a convex function, where $D \subseteq \mathbb{R}^n$. Suppose that for every point $u \in D$ there exists a subgradient of f at u with Euclidean norm at most K. Then f is K-Lipschitz continuous.*

For parameters $(n_1, \ldots, n_k)$, we say that an expression $f(n_1, \ldots, n_k)$ is $\mathsf{poly}(n_1, \ldots, n_k)$ if there exist (absolute) constants $c_1, c_2 > 0$ such that $f(n_1, \ldots, n_k) \leq (n_1 n_2 \ldots n_k)^{c_1}$ for every $n_1, \ldots, n_k \geq c_2$. We also write $f(n_1, \ldots, n_k) = \mathsf{poly}(n_1, \ldots, n_k)$ with the same meaning.

This thesis deals mostly with NP-hard problems, so we focus on designing *approximation algorithms*. For an instance I of an optimization problem P, we denote by $\mathrm{OPT}(I)$ its optimal value (if it exists). An *α-approximate solution* for I (where $\alpha \geq 1$) is a feasible solution that attains objective value:

- *at most* $\alpha \cdot \mathrm{OPT}(I)$, if P is a minimization problem; and
- *at least* $\frac{1}{\alpha} \cdot \mathrm{OPT}(I)$, if P is a maximization problem.

An *α-approximation algorithm* for P is an algorithm that, given any instance of P, computes in polynomial time an α-approximate solution for it. We also use the following broader terminology to allow additive approximation, and algorithms whose running time depends on additional parameters other than the input size. An *approximate solution* for an instance I is a feasible solution whose objective value is within some bounded multiplicative factor and/or additive term of $\mathrm{OPT}(I)$. An *approximation algorithm* for P is an algorithm that, given any instance I of P, computes an approximate solution for I.

1.4 Organization of the thesis

In Chapter 2, we give an overview of relevant previous work on robust, stochastic, and distributionally robust stochastic optimization. In Chapter 3, we formally define the DRS optimization model that we study and the class of problems to which our frameworks apply. We then give a summary of the main results we obtain for DRS optimization under a Wasserstein ball or an L_∞ ball (see Section 3.3), and prove some preliminary results. The next two chapters contain the two components of our framework for DRS optimization under a Wasserstein ball: Chapter 4 presents an SAA result, reducing the original problem

(with a central distribution given by a sampling oracle) to a collection of SAA problems with an explicit central distribution; Chapter 5 shows how to approximately solve the SAA problems. In Chapter 6, we show how to combine the two components to obtain our main result for DRS optimization under a Wasserstein ball (Theorem 3.6). We then apply our framework to obtain approximation algorithms for DRS versions of various problems under a Wasserstein ball, namely set cover, edge cover, vertex cover, facility location, and Steiner tree. In Chapter 7, we present our framework for DRS optimization under an L_∞ ball, and prove our main result in this setting (Theorem 3.9). We then apply our framework to obtain approximation algorithms for DRS versions of various applications under an L_∞ ball, such as set cover, edge cover, vertex cover, and facility location. In Chapter 8, we discuss some directions for future research.

Chapter 2

Background

In this chapter, we give an overview of relevant previous work. We first consider robust and stochastic optimization in Sections 2.1 and 2.2 respectively; while these two models are not the focus of this thesis, our frameworks for distributionally robust stochastic (DRS) optimization build upon various ideas that were originally developed for them. We give an overview of previous work on DRS optimization in Section 2.3. There are various other models that (like the DRS model) allow interpolating between robust and stochastic optimization; we mention some of them in Section 2.4. Finally, in Section 2.5 we state some classical inequalities that we use throughout this thesis.

2.1 Robust optimization

The study of single-stage robust optimization dates back to Falk [45], Soyster [120], and Thuente [125], who considered linear programs with uncertain constraints or objective function. Starting in the late 1990s, there has been a vast amount of work on various robust models that apply to both continuous and discrete problems, with various types of uncertainty sets, such as ellipsoidal and polyhedral (see, e.g., [8, 9, 10, 14, 15, 39, 41]). For a comprehensive treatment of robust optimization, we refer the reader to the textbook by Ben-Tal, El Ghaoui, and Nemirovski [4] and the survey by Bertsimas, Brown, and Caramanis [11].

In the remainder of this section we give an overview of previous work on the *two-stage robust optimization model*, which we described in Section 1.1. The study of two-stage (and more generally, multistage) robust optimization (also known as *adjustable* robust optimization) was initiated by Ben-Tal, Goryashko, Guslitzer, and Nemirovski [5]; for a more

comprehensive treatment of this model, we refer the reader to Delage and Iancu [31] and the references therein. In the remainder of this section, we focus on the *demand-robust model* introduced by Dhamdhere, Goyal, Ravi, and Singh [33] (see also Goyal [56]). Whereas earlier models imposed a fixed set of constraints (whose coefficients may be uncertain), in the demand-robust model we are given a collection of constraints whose coefficients are known exactly, but only an uncertain subset of these constraints needs to be satisfied. More formally, we are given a collection of constraints indexed by a set U, and a scenario is a subset of U. When a scenario $A \subseteq U$ is realized, the pair of first-stage and second-stage decisions $\left(x, z^A\right)$ must satisfy all the constraints indexed by elements of A (and constraints indexed by $U \setminus A$ may be ignored). For instance, in two-stage robust set cover, U is a ground set of elements, and a scenario $A \subseteq U$ indicates the set of elements to be covered in that scenario; the constraints encode that in every scenario A, the combination of first-stage and second-stage decisions $\left(x, z^A\right)$ (which indicate the sets that are picked in the first stage and in scenario A respectively) should cover all the elements of A. In addition to the uncertainty in the collection of constraints that must be satisfied, this model also incorporates uncertainty in the objective function: if scenario A is realized, the cost of each first-stage decision increases by a factor $\lambda_A \geq 1$ in the second-stage (in some settings, the factor λ_A is required to be uniform across all scenarios).

Earlier works in approximation algorithms for demand-robust optimization considered the setting wherein the collection of scenarios $\mathcal{A}$ is given explicitly as part of the input (and hence the number of scenarios is polynomial in the input size). For example, Dhamdhere, Goyal, Ravi, and Singh [33] give LP-based approximation algorithms for demand-robust shortest path, Steiner tree, vertex cover, facility location, minimum cut, and minimum multi-cut; Golovin, Goyal, and Ravi [55] give improved approximation factors for demand-robust shortest path and minimum cut; Chen, Megow, Rischke, and Stougie [25] give LP-based approximation algorithms for a class of demand-robust scheduling problems.

A significantly more challenging setting is the one where the scenario collection $\mathcal{A}$ is given implicitly, and its size may be *exponential in the input size*. The first results in this setting were obtained by Feige, Jain, Mahdian, and Mirrokni [46], who give approximation algorithms for demand-robust versions of set cover, edge cover, and vertex cover in the *k-bounded setting*, wherein the scenario collection $\mathcal{A}$ is comprised of all subsets of cardinality at most k of the ground set U (this scenario collection is specified implicitly by the pair (U, k)). Their results are obtained by reducing a certain convex-program relaxation of the problem to the *fractional* k-max-min *problem*: find the worst possible scenario given $x = 0$ as the first-stage decision, when we allow *fractional second-stage decisions*. More precisely, the fractional k-max-min problem asks to find the set $A \subseteq U$ with cardinality at most k for which the minimum cost of a fractional second-stage decision z^A that satisfies all

the constraints for scenario A is as large as possible. The authors also consider *integer* k-max-min problems, wherein one seeks a scenario for which the minimum cost of an *integer* solution is as large as possible. They prove that {fractional, integer} k-max-min {vertex cover, edge cover, set cover} are APX-hard, and give approximation algorithms for these problems by drawing a connection between them and online versions of the underlying optimization problem.

Khandekar, Kortsarz, Mirrokni, and Salavatipour [80] expanded the collection of results known for demand-robust problems in the k-bounded setting, by designing approximation algorithms in this setting for Steiner tree, Steiner forest on a tree, and facility location.

Gupta, Nagarajan, and Ravi [60] give a framework for obtaining approximation algorithms for demand-robust combinatorial problems in the k-bounded setting and for k-max-min problems, under the assumption that the inflation factor λ_A is uniform across all the scenarios. Using this framework, they obtain improved approximation factors for k-bounded demand-robust Steiner tree and set cover, and the first approximation algorithms for k-bounded demand-robust Steiner forest, minimum cut, and multicut. In a companion paper, Gupta, Nagarajan, and Ravi [59] give approximation algorithms for demand-robust problems wherein the scenario collection $\mathcal{A}$ is defined as the collection of subsets of a ground set U that satisfy a series of knapsack and matroidal constraints. (This generalizes the k-bounded setting, since the collection of subsets of U of cardinality at most k is the collection of independent sets of a uniform matroid.)

2.2 Stochastic optimization

The study of stochastic optimization dates back to Dantzig [29]; although there is a vast amount of literature on this field (see, e.g., [18, 97, 101, 110] and the references therein), its study from an approximation-algorithms perspective is relatively recent. Various approximation results have been obtained in the two-stage stochastic model (which was introduced in Section 1.1) over the last 15 years in the CS and Operations-Research (OR) literature. In this section, we give an overview of relevant previous work from an approximation-algorithms perspective. For a more comprehensive overview, we refer the readers to the surveys by Romeijnders, Stougie, and Vlerk [100], Shi [112], and Swamy and Shmoys [122].

There are multiple ways of specifying the underlying probability distribution p for a stochastic problem. Perhaps the most natural approach is the *explicit-distribution model*, wherein p is represented as a collection of pairs $\{(A, p_A)\}_{A \in \mathsf{supp}(p)}$ specifying the probabilities of the scenarios in the support of p.

For some applications, the scenario collection $\mathcal{A}$ may be extremely large, making it impractical to specify the distribution p explicitly. To handle such cases, one must resort to *implicit* representations of the distribution p. Two common approaches are the *independent activation model* and the *black-box model*, which allow encoding two-stage problems with *exponentially many scenarios*. In both models, the scenarios are subsets of a given ground set U. In the *independent-activation model*, the scenario realized includes each element $e \in U$ independently with some given probability q_e. We can therefore specify the distribution p implicitly by specifying the values $\{q_e\}_{e \in U}$. In the *black-box model*, the central distribution p can only be accessed via a very limited interface called a *sampling oracle*. Each time we query the oracle, it randomly generates and returns a scenario $A \in \mathcal{A}$, according to the distribution p. (Note that querying the oracle simply means requesting a scenario; we do not need to convey any information to the oracle.) When measuring the running time of an algorithm in this model, sampling from the oracle is considered an elementary operation. Note that the black-box model encompasses a much broader class of distributions, since, unlike the independent-activation model, it can account for correlations among the activation of the various elements of U. Moreover, note that algorithms for the black-box model can also be used for problems in the explicit-distribution model and in the independent-activation model, since if we are given a problem in those two models, it is straightforward to simulate a sampling oracle for the underlying distribution.

In the explicit-distribution model, approximation algorithms are known for two-stage stochastic versions of various problems, such as the service-provision problem considered in [36], maximum-weight matching [82], minimum-cost bipartite matching [79], shortest path [74, 98], set cover [98], vertex cover [74, 98], bin packing [74, 98], facility location [98], minimum-cost flow [74], Steiner tree [64, 65, 74], single-sink network design [64], minimum spanning tree [34], and scheduling problems [25, 113].

In the independent-activation model, approximation results are known for two-stage stochastic versions of shortest path [74], vertex cover [62, 74], bin packing [74], minimum-cost flow [74], Steiner tree [62, 74], Steiner forest [48, 62], facility location [62], traveling-salesman problem [116], and minimum-cost bipartite matching [79].

In the remainder of this section, we focus on the black-box model, which is the most relevant for this thesis. From a theoretical viewpoint, a question that has attracted considerable attention is that of computing a near-optimal solution for a two-stage stochastic problem with a black-box distribution using a *polynomially bounded* number of samples.

Gupta, Pál, Ravi, and Sinha [62] give sample-complexity bounds under the *constant-inflation assumption*: every first-stage action has a corresponding second-stage action that is *exactly* λ times more costly, where $\lambda \geq 1$ is a given constant. In this setting, they

obtained approximation algorithms using only $\mathsf{poly}(\lambda)$ samples for two-stage stochastic rooted Steiner tree, vertex cover, and facility location. These results are obtained via the *boosted-sampling* framework, which yields combinatorial approximation algorithms for two-stage stochastic problems utilizing approximation algorithms of a certain type for the deterministic version of the underlying problem. Fleischer, Könemann, Leonardi, and Schäfer [48] and Gupta and Pál [61] give approximation algorithms for two-stage stochastic *unrooted* Steiner tree with constant inflation using the boosted-sampling framework.

Shmoys and Swamy [114] give a framework for computing near-optimal solutions for a broad class of two-stage stochastic linear programs under the *bounded-inflation assumption*: we are given a factor $\lambda \geq 1$ such that every first-stage action has a corresponding second-stage action that is at most λ times more costly in every scenario. Note that the inflation of the cost of a first-stage action is now allowed to depend on the action *and* on the scenario. In this setting, the authors provide a *fully polynomial randomized approximation scheme*: given any $\varepsilon > 0$ one can compute a $(1+\varepsilon)$-approximate solution for the stochastic LP using $\mathsf{poly}\big(\text{input size}, \lambda, \frac{1}{\varepsilon}\big)$ samples. This result is obtained via a variant of the ellipsoid method based on approximate subgradients (which we discuss in Section 3.4.1). Combining this result with a suitable rounding scheme that rounds a fractional first-stage decision to an integer one while increasing the cost incurred in the first stage and in each scenario by at most a bounded factor, the authors obtain approximation algorithms for two-stage stochastic versions of a variety of combinatorial-optimization problems including set cover, vertex cover, and facility location. The authors also show that the dependence of the number of samples on the inflation factor λ is unavoidable for two-stage stochastic set cover in the black-box model. We note that our framework for discrete DRS optimization utilizes the same type of rounding algorithms as [114].

A common approach for solving two-stage stochastic problems in the black-box model is the *sample-average-approximation (SAA) method*: sample some number of scenarios from the distribution p, use them to compute an empirical estimate $\widehat{p}$ of p, and solve the stochastic problem with $\widehat{p}$ as the underlying distribution instead of p. We refer to the stochastic problems with distributions p and $\widehat{p}$ as the *original problem* and the *SAA problem* respectively. Earlier works show that optimal solutions to the SAA problem converge to optimal solutions of the original problem as the number of samples increases (see Shapiro [108]). Various works provide numerical experiments demonstrating the effectiveness of the SAA method in practice [84, 102, 105, 127].

A first result regarding the theoretical effectiveness of the SAA method was obtained by Kleywegt, Shapiro, and Homem-de-Mello [81]. For a two-stage stochastic problem with a finite first-stage decision set X, they show the following result under mild assumptions:

given $\eta > 0$, if we construct $\widehat{p}$ using $\mathsf{poly}\left(\log|X|, \frac{1}{\eta}, \sigma\right)$ independent samples, then with high probability any optimal solution for the SAA problem is a near-optimal solution for the original problem (within an η term). The term σ that appears in the sample size is a quantity that bounds the variance of a certain random variable that depends on the second-stage costs, and may be exponentially large even for well-structured stochastic problems (see Shmoys and Swamy [114]). Shapiro and Nemirovski [111] show that this bound is tight for general two-stage stochastic problems, and give SAA results for two-stage stochastic problems wherein the set of first-stage decisions is a continuous set $\mathcal{P} \subseteq \mathbb{R}^m$, by applying the result of Kleywegt, Shapiro, and Homem-de-Mello [81] to a gridding of $\mathcal{P}$.

Swamy and Shmoys [124] show that their approximate-subgradient machinery from [114] can also be used to obtain an SAA result for the class of two-stage LPs considered therein: given any $\varepsilon > 0$, any optimal solution for an SAA problem constructed using $\mathsf{poly}\left(\text{input size}, \lambda, \frac{1}{\varepsilon}\right)$ samples is a $(1+\varepsilon)$-approximate solution for the original problem.

Nemirovski and Shapiro [92] provide an alternative proof for a special case of the SAA result by Swamy and Shmoys [124], namely two-stage stochastic fractional set cover. Charikar, Chekuri, and Pál [24] give additional SAA results for two-stage stochastic problems with bounded inflation, based on the framework by Kleywegt, Shapiro, and Homem-de-Mello [81] and Shapiro [108]. They show that for a broad class of two-stage stochastic problems with a finite first-stage decision set X, any optimal solution of an SAA problem constructed using $\mathsf{poly}\left(\log|X|, \lambda, \frac{1}{\varepsilon}\right)$ samples is a $(1+\varepsilon)$-approximate solution for the original problem. Whereas for the class of problems considered by Swamy and Shmoys [124] the SAA problem can be solved exactly (since it is an LP), it is not always possible to efficiently solve the SAA version of a problem in the class considered by Charikar, Chekuri, and Pál [24]. To circumvent this difficulty, [24] also developed techniques for converting *approximate* (rather than optimal) solutions for SAA problem(s) into approximate solutions for the original problem.

Other two-stage stochastic problems for which approximation algorithms have been developed in the black-box model include stochastic minimum-spanning tree [34], Steiner forest [58], traveling-salesman problem [104], and scheduling problems [25, 113]. Approximation results are also known for *multistage stochastic* versions of various covering problems (see, e.g., Byrka and Srinivasan [20], Gupta, Pál, Ravi, and Sinha [63], and Swamy and Shmoys [123]).

2.3 Distributionally robust stochastic optimization

In this section, we give an overview of previous work on the distributionally robust stochastic (DRS) model, which we introduced in Section 1.1. The DRS model was introduced by Scarf [103] in the context of an inventory-control problem, as an alternative to the classical stochastic model, to address the issue that in practice one typically does not have a probability distribution that precisely describes the behavior of the uncertain parameters. When adopting the classical stochastic model, one typically optimizes decisions with respect to a distribution that is inferred from historical data, which may lead to decisions that perform poorly with respect to the actual underlying distribution. This phenomenon, referred to as "overfitting", "optimizer's curse", "postdecision surprise", or "error maximization effect", is discussed for example by Brown [19], Harrison and March [69], Michaud [90], and Smith and Winkler [118]. The DRS model circumvents this issue by hedging against the worst-possible probability distribution in a collection of distributions; this collection typically consists of distributions that are statistically consistent with some historical data. As noted before, the DRS model also allows to interpolate between robust optimization and stochastic optimization, avoiding the risks of overconservatism and overfitting that are present in these two models.

The DRS model (re)gained interest recently in the Operations-Research literature, where it is sometimes called *data-driven* or *ambiguous* stochastic optimization, and has been used in a variety of areas and applications such as inventory control [96, 103, 131, 133], portfolio selection [30, 32, 40, 54], healthcare [89], vehicle routing [22], the traveling-salesman problem [21], facility location [130], and machine learning [50, 106, 107].

Most of the DRS optimization literature, including the seminal work of Scarf [103], considers *moment-based ambiguity sets* (see, e.g., [17, 30, 32, 35, 53, 88, 109, 128]). In this setting, the ambiguity set consists of all distributions whose (typically first and second) moments have a specified value, or are constrained to being in a specified convex set.

Another popular family of ambiguity sets is that of *distance-based ambiguity sets*; our work falls in this category. In this setting, the ambiguity set is defined as a ball around a central distribution, with respect to some notion of distance among distributions. Various notions of distance have been considered, such as Wasserstein metrics [21, 43, 50, 51, 52, 132], ϕ-divergence [3, 6, 72], and the Prohorov metric [42].

A third category of ambiguity sets is that of *hypothesis-test based ambiguity sets*. In this setting, the ambiguity set consists of all distributions that pass a certain type of hypothesis test, when given a certain historical data (see, e.g., Bertsimas, Gupta, and Kallus [12, 13] and Chen, Lin, and Xu [26]).

Various works give *non-polynomial time* algorithms for DRS optimization problems [86, 87] and tractable approximate reformulations for special cases [52, 53, 67, 128]. Another research direction is obtaining tractable *exact* reformulations for certain classes of DRS problems, under various {linearity, convexity, concavity} assumptions on the objective function (see, e.g., Delage and Ye [32], Esfahani and Kuhn [43], Gao and Kleywegt [52], Mehrotra and Zhang [88], and Wiesemann, Kuhn, and Sim [128]). However, in most cases these results apply only to *continuous scenario spaces*. Moreover, to the best of our knowledge, with the exception of Agrawal, Ding, Saberi, and Ye [2], which we discuss below, there are no prior approximation algorithms for discrete two-stage DRS optimization problems when the number $|\mathcal{A}|$ of possible scenarios is finite, but exponentially large (even if the ambiguity set is defined as a ball centered at a distribution with polynomial-size support).

Various works propose and analyze (theoretically and/or experimentally) algorithms for constructing a suitable ambiguity set given historical data (see, e.g., [12, 32, 43, 106]). In addition to the tractability of the resulting DRS problem, one typically wants the ambiguity set to be "small" and contain the true underlying probability with high probability, so as to avoid the overconservatism that typically arises in the classical robust model, and ensure that the DRS problem gives guarantees on the quality of the solutions with respect to the true underlying distribution. Various works have advocated the use of a Wasserstein ball around an empirical distribution for this purpose (see, e.g., Esfahani and Kuhn [43], Gao and Kleywegt [52], Van Parys, Esfahani, and Kuhn [126], and Zhao and Guan [132]), but there are no results proving polynomial bounds on the number of samples needed in order to produce provably good results. Note that these works, by definition, consider the setting where the central distribution has polynomial-size support. The distributionally robust setting has also been considered for chance-constrained problems; see, e.g., Erdoğan and Iyengar [42] and the references therein.

The work of Agrawal, Ding, Saberi, and Ye [2] in the CS literature on correlation gap can be interpreted as studying DRS discrete-optimization problems, but in the moment-based setting, where the ambiguity set is the collection of distributions that agree with some given expected values; the correlation gap quantifies the worst-case ratio of the DRS objective when one chooses the optimal decisions with respect to the distribution in the ambiguity set that treats all random variables as independent, versus the optimum of the DRS problem. The authors prove various $O(1)$ bounds on the correlation gap for submodular functions and subadditive functions admitting suitable cost shares.

2.4 Other models interpolating between robust and stochastic optimization

Finally, we discuss a few other models that are more peripherally related to the topic of this thesis, but of a somewhat similar spirit in that they pursue goals that are intermediate between the robust and stochastic settings. Byrka and Srinivasan [20], So, Zhang, and Ye [119], and Swamy [121] consider extensions of the classical stochastic model that incorporate risk aversion. In the context of online algorithms, Esfandiari, Korula, and Mirrokni [44] and Mirrokni, Gharan, and Zadimoghaddam [91] give online algorithms for allocation problems that are simultaneously competitive both in a random input model and in an adversarial input model. Finally, we note that our distributionally robust setting can be seen to be in a similar spirit as a recent focus in algorithmic mechanism design, where one does not assume precise knowledge of the underlying distribution; rather one (implicitly) has a collection of distributions, and one seeks to design mechanisms that work for every distribution in this collection (see, e.g., Huang, Mansour, and Roughgarden [73]).

2.5 Some classical inequalities

Some of the classical inequalities used in this thesis admit multiple non-equivalent statements. In the interest of precision, we state below which versions we use.

Theorem 2.1 (Markov's inequality)**.** *Let X be a nonnegative random variable. Then for every $t > 0$ we have*

$$\mathbf{Pr}[X \geq t] \leq \frac{\mathbb{E}[X]}{t} \ .$$

Theorem 2.2 (Jensen's inequality [77])**.** *Let $X \in \mathbb{R}^n$ be a random variable, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a concave function. Then*

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X]) \ .$$

Theorem 2.3 (Hoeffding's inequality [71])**.** *Let $X_1, \ldots, X_N$ be independent real-valued random variables in the range $[a, b]$, where $a < b$, and let $\overline{X} := \frac{1}{N} \sum_{i \in [N]} X_i$. For every $\eta \geq 0$, we have*

$$\mathbf{Pr}\left[\left|\overline{X} - \mathbb{E}\left[\overline{X}\right]\right| > \eta\right] \leq 2 \exp\left(-\frac{2N\eta^2}{(b-a)^2}\right) \ .$$

The following result follows easily from Hoeffding's inequality.

Corollary 2.4. *Let X be a real-valued random variable in the range $[a, b]$. Given any $\eta > 0$ and $\delta \in (0, 1]$, there exists $N_0 = \mathsf{poly}\left(\frac{b-a}{\eta}, \log \frac{1}{\delta}\right)$ such that the following holds. Let $\overline{X}$ be an empirical estimate of X computed using $N \geq N_0$ independent samples. Then with probability at least $1 - \delta$ we have*

$$\left|\overline{X} - \mathbb{E}[X]\right| \leq \eta \ .$$

Proof. Let $X_1, \ldots, X_N$ be independent samples of X, and let $\overline{X} = \frac{1}{N} \sum_{i \in [N]} X_i$ be the empirical estimate of X computed using those samples. Note that $\mathbb{E}\left[\overline{X}\right] = \mathbb{E}[X]$. Using Hoeffding's inequality (Theorem 2.3), we obtain

$$\mathbf{Pr}\left[\left|\overline{X} - \mathbb{E}[X]\right| > \eta\right] = \mathbf{Pr}\left[\left|\overline{X} - \mathbb{E}\left[\overline{X}\right]\right| > \eta\right] \leq 2 \exp\left(-\frac{2N\eta^2}{(b-a)^2}\right) \ .$$

Therefore we have $\left|\overline{X} - \mathbb{E}[X]\right| \leq \eta$ with probability at least $1 - \delta$ as long as the number of samples N satisfies $2 \exp\left(-\frac{2N\eta^2}{(b-a)^2}\right) \leq \delta$. Solving this inequality for N yields

$$N \geq \frac{(b-a)^2}{2\eta^2} \ln \frac{2}{\delta} = \mathsf{poly}\left(\frac{b-a}{\eta}, \log \frac{1}{\delta}\right) \ . \qquad \square$$

Chapter 3

Two-stage distributionally robust stochastic optimization

In this chapter, we lay down the foundations for our study of two-stage distributionally robust optimization. In Section 3.1, we formally define the model that we study. In Section 3.2, we define the broad class of problems to which our frameworks apply. In Section 3.3, we give an overview of the main results we obtain for DRS optimization under a Wasserstein ball or an L_∞ ball, including tables showing the approximation factors we obtain for various applications. In Section 3.4, we prove some preliminary results regarding the optimization of functions over the set of (integer or fractional) first-stage decisions, and techniques for converting fractional solutions into integer ones (while incurring a bounded increase in the objective value).

3.1 Formal model description

We study the following *two-stage distributionally robust stochastic (DRS) optimization* model. We are given an underlying finite set $\mathcal{A}$ of scenarios, and a collection D of probability distributions over $\mathcal{A}$, called the *ambiguity set*. Decisions are taken in two stages. In the first stage, before a scenario is realized, we have at our disposal a finite set $X \subseteq \mathbb{R}^m_+$ of possible decisions. Selecting a first-stage decision $x \in X$ incurs a cost $c^\intercal x$, where $c \in \mathbb{R}^m_+$ is a given cost vector. In the second stage, after a scenario $A \in \mathcal{A}$ is realized, we have at our disposal a finite set $Z \subseteq \mathbb{R}^n_+$ of possible decisions. Selecting a second-stage decision $z^A \in Z$ incurs a nonnegative cost denoted by (cost of z^A). For every scenario $A \in \mathcal{A}$, there is a

corresponding set $F(A) \subseteq X \times Z$ of feasible solutions; the combination of the first-stage decision and the second-stage decision must satisfy $\left(x, z^A\right) \in F(A)$. Our goal is to solve the problem

$$\min_{\substack{x \in X, z \in Z^{\mathcal{A}}: \\ \left(x, z^A\right) \in F(A) \quad \forall A \in \mathcal{A}}} \left\{ c^\intercal x + \sup_{p \in D} \mathbb{E}_{A \sim p}\left[\text{cost of } z^A\right] \right\} . \tag{DRSO}$$

One natural setting to consider is the one where the ambiguity set is a ball $D = \{p : L(\mathring{p}, p) \leq r\}$ of probability distributions over $\mathcal{A}$ around a central distribution $\mathring{p}$. Here, L denotes a metric over probability distributions, and $r > 0$ is the radius of the ball. While the choice of the metric L is an application-dependent modeling decision, we would like D to contain distributions that are "reasonably similar" to $\mathring{p}$, and exclude completely unrelated distributions, as the latter could lead to overly conservative decisions, à la robust optimization.

Two natural choices for L are the L_∞ metric, defined by $L_\infty(p, q) := \max_{A \in \mathcal{A}} |p_A - q_A|$, and the $\frac{1}{2}L_1$ metric, defined by $\frac{1}{2}L_1(p, q) := \frac{1}{2}\sum_{A \in \mathcal{A}} |p_A - q_A|$, which is also known as the *total-variation distance*. A significantly more refined way of comparing probability distributions is to see if they spread their probability mass on "similar" scenarios. Wasserstein distances capture this viewpoint crisply, and lift an underlying *scenario metric* to a metric over distributions.

Definition 3.1 (Wasserstein (a.k.a. transportation or earth-mover) distance). The Wasserstein distance between two probability distributions p and q over $\mathcal{A}$ is defined with respect to an underlying scenario metric $\ell : \mathcal{A} \times \mathcal{A} \to \mathbb{R}_+$. A *flow* or *transportation plan* from p to q is a vector $\gamma \in \mathbb{R}_+^{\mathcal{A} \times \mathcal{A}}$ such that: (i) $\sum_{A' \in \mathcal{A}} \gamma_{A,A'} = p_A$ for every scenario $A \in \mathcal{A}$; and (ii) $\sum_{A \in \mathcal{A}} \gamma_{A,A'} = q_{A'}$ for every scenario $A' \in \mathcal{A}$. The *Wasserstein distance* between p and q, denoted by $L_{\mathrm{W}}(p, q)$, is the minimum value of $\sum_{A,A'} \gamma_{A,A'} \ell(A, A')$ over all flows from p to q.

Note that if ℓ is a {symmetric, asymmetric, pseudo}-metric,[1] then so is L_{W}. Also, note that $\frac{1}{2}L_1$ is the Wasserstein metric with respect to the *discrete scenario metric* ℓ^{disc}, defined by $\ell^{\mathsf{disc}}(A, A') = 1$ if $A \neq A'$, and 0 otherwise. As we show in Chapters 4–6, our results hold even when ℓ is not a metric, but only satisfies nonnegativity, and $\ell(A, A) = 0$ for all $A \in \mathcal{A}$.

[1] The distance function ℓ is a *pseudometric* if it satisfies the triangle inequality and $\ell(A, A) = 0$ for every $A \in \mathcal{A}$, but $\ell(A, A')$ is allowed to be 0 for $A \neq A'$.

Settings where the ambiguity set D is a ball with respect to some metric over distributions arise naturally when one tries to infer a scenario distribution from observed data (see, e.g., [42, 43, 132])—hence, the moniker *data-driven optimization*—and it has been argued that defining D using the Wasserstein distance has various benefits (see, e.g., [43, 52, 126, 132]).

We would like to be able to handle settings where the number of scenarios in the collection $\mathcal{A}$ is extremely large, possibly even exponential in the size of the underlying combinatorial-optimization problem. In such settings, it is impractical and wasteful to assume that scenario-specific information (e.g., scenario probabilities, pairwise scenario distances in the Wasserstein case, and the feasibility conditions for each scenario) is explicitly specified in the input and/or output. Instead, we will assume a suitable oracle model for specifying portions of the input (or output) that involve scenario-dependent data, and we use the term *input size* to denote the encoding size of the data that does not depend on $\mathcal{A}$. That is, the input size, denoted by $\mathcal{I}$, measures the encoding size of the underlying deterministic problem, along with the first-stage and second-stage costs and the radius r of the ball D. We adopt the *black-box model* (already discussed in the context of two-stage stochastic optimization in Section 2.2), wherein the central distribution $\mathring{p}$ is specified via a sampling oracle that allows one to sample a scenario A from the underlying distribution; when we sample a scenario A, we get to know any scenario-specific data.[2]

Moreover, as is typically the case when specifying a combinatorial-optimization problem, the first-stage and second-stage decision sets are not explicitly specified, but are implicit from the semantic description of the problem. Similarly, in the Wasserstein setting, the pairwise scenario distances $\{\ell(A, A')\}_{A,A' \in \mathcal{A}}$ are not specified explicitly; instead, we assume that given any pair of scenarios (A, A'), we can compute $\ell(A, A')$ in $\mathsf{poly}(\mathcal{I})$ time.

An example: two-stage distributionally robust stochastic facility location (DRSFL). As an illustrative example, consider the following distributionally robust facility location problem (DRSFL).[3] We have a metric space $\left(\mathcal{F} \cup \mathcal{C}, \{w_{ij}\}_{i,j \in \mathcal{F} \cup \mathcal{C}}\right)$, where $\mathcal{F}$ is a set of facilities, and $\mathcal{C}$ is a set of clients. A scenario is a subset of $\mathcal{C}$ indicating the set of clients

[2] An important stepping stone used to obtain results in the black-box setting is the setting where the central distribution $\mathring{p}$ has moderate-size support and is represented explicitly by the collection of pairs $\{(A, \mathring{p}_A)\}_{A \in \mathsf{supp}(\mathring{p})}$. In this setting, the input size also includes the encoding size of this collection of pairs (this is in contrast with the black-box model, wherein the central distribution $\mathring{p}$ does not contribute to the input size).

[3] More examples can be found in Chapter 6.

that need to be served in that scenario. (Note that we can model integer demands by creating colocated clients.) Two common choices for the scenario collection are $\mathcal{A} = 2^{\mathcal{C}}$ (the *unrestricted setting*) and $\mathcal{A} = \{A \subseteq \mathcal{C} : |A| \leq k\}$ (the *k-bounded setting*). We may open facilities of $\mathcal{F}$ in either stage. The first-stage and second-stage opening costs are given by vectors $f^{\mathrm{I}}, f^{\mathrm{II}} \in \mathbb{R}_+^{\mathcal{F}}$ respectively. In scenario $A \subseteq \mathcal{C}$, we need to assign every client $j \in A$ to a facility $i^A(j)$ that has been opened either in the first stage or in the second stage (in scenario A). The goal is to minimize

$$\sum_{i \text{ opened in stage I}} f_i^{\mathrm{I}} + \sup_{p \in D} \mathbb{E}_{A \sim p} \left[\sum_{i \text{ opened in scenario } A} f_i^{\mathrm{II}} + \sum_{j \in A} w_{i^A(j) j} \right].$$

Here the input size $\mathcal{I}$ is the encoding size of $\big(\mathcal{F}, \mathcal{C}, w, f^{\mathrm{I}}, f^{\mathrm{II}}, r\big)$. In addition to L being the L_∞ or $\frac{1}{2}L_1$ metrics, we can consider various ways of defining a scenario metric ℓ in terms of the underlying assignment-cost metric w to capture that two scenarios involving demand locations in the same vicinity are deemed similar; lifting these scenario metrics to Wasserstein metrics over distributions yields a rich class of two-stage DRS facility-location models. For instance, we can define the *asymmetric metric* $\ell_\infty^{\mathsf{asym}}(A, A') := \max_{j' \in A'} w(j', A)$, where $w(j', A) := \min_{j \in A} w_{j'j}$, which measures the maximum separation between clients in A' and locations in A (the resulting Wasserstein metric L_{W} will now be an asymmetric metric over distributions). Other natural scenario metrics include the asymmetric metric $\ell_1^{\mathsf{asym}}(A, A') := \sum_{j' \in A'} w(j', A)$, and the symmetrizations of these asymmetric metrics: $\ell_\infty^{\mathrm{sym}}(A, A') := \max\{\ell_\infty^{\mathsf{asym}}(A, A'), \ell_\infty^{\mathsf{asym}}(A', A)\}$, and $\ell_1^{\mathrm{sym}}(A, A') := \max\{\ell_1^{\mathsf{asym}}(A, A'), \ell_1^{\mathsf{asym}}(A', A)\}$.

We refer the reader to Section 6.7 for a formal explanation of how an instance of DRSFL can be modeled by the generic DRS problem (DRSO). We remark that our framework is not restricted to the choices of scenario collections and scenario metrics mentioned above; instead, it applies more generally to any scenario collection and any scenario metric, provided we have access to approximation algorithms for suitable problems (see Theorems 3.6 and 3.9).

Recall that a solution for problem (DRSO) consists of a first-stage decision $x \in X$, along with a second-stage decision $z^A \in Z$ for each scenario $A \in \mathcal{A}$. Since the scenario collection $\mathcal{A}$ may have exponential size, returning the output in an explicit fashion is not viable. To bypass this issue, we will focus on obtaining *two-stage algorithms*.

Definition 3.2. A *two-stage algorithm* for problem (DRSO) is a pair of algorithms $\mathsf{Alg} := \big(\mathsf{Alg}^{\mathrm{I}}, \mathsf{Alg}^{\mathrm{II}}\big)$ such that:

- $\mathsf{Alg}^{\mathrm{I}}$ computes a first-stage decision $x \in X$; and
- $\mathsf{Alg}^{\mathrm{II}}$ receives as input a scenario $A \in \mathcal{A}$, and computes a second-stage decision $z^A \in Z$ such that $\left(x, z^A\right) \in F(A)$.

Since $\mathsf{Alg}^{\mathrm{II}}$ needs to know the first-stage decision x, we assume that it is only called after $\mathsf{Alg}^{\mathrm{I}}$. We allow $\mathsf{Alg}^{\mathrm{II}}$ to utilize not only the first-stage decision x, but also other data computed by $\mathsf{Alg}^{\mathrm{I}}$. We define the running time of Alg as the sum of the running times of $\mathsf{Alg}^{\mathrm{I}}$ and $\mathsf{Alg}^{\mathrm{II}}$.

For example, for DRSFL, $\mathsf{Alg}^{\mathrm{I}}$ specifies which facilities are opened in the first stage. After a scenario A is realized, $\mathsf{Alg}^{\mathrm{II}}$ specifies which facilities are opened in the second stage, as well as the assignments of clients in A to facilities that have been opened in either stage.

3.2 A general class of two-stage DRS problems

Abstracting away the key properties of the applications that we consider in Chapter 6, we now define a generic two-stage DRS problem to which our frameworks apply. We remark that these assumptions hold for all the applications we consider in Chapter 6, and for various other two-stage problems considered in the CS literature (see, e.g., [33, 46, 60, 80, 114]).

To get a better handle on the problem, it will be convenient to consider fractional relaxations of the DRS problem obtained by enlarging the first-stage and second-stage decision sets to suitable polytopes. We expand X and Z to polytopes $\mathcal{P} \supseteq X$ and $\mathcal{Z} \supseteq Z$ respectively. We assume that $\mathcal{P}$ is specified either explicitly by a set of $\mathsf{poly}(\mathcal{I})$ linear constraints, or implicitly by a $\mathsf{poly}(\mathcal{I})$-time *separation oracle*, which is an algorithm that, given a point x, either decides that $x \in \mathcal{P}$, or returns a hyperplane separating x from $\mathcal{P}$. In the remainder of the thesis, we refer to elements of X and Z as *integer* first-stage and second-stage decisions respectively; we refer to elements of $\mathcal{P}$ and $\mathcal{Z}$ as *fractional* first-stage and second-stage decisions respectively. (This is simply for the sake of exposition, since in most applications we have $X = \mathcal{P} \cap \mathbb{Z}^m$ and $Z = \mathcal{Z} \cap \mathbb{Z}^n$. Our framework does not require the elements of X and Z to be integer points.) For every scenario $A \in \mathcal{A}$, we enlarge the set of integer feasible solutions $F(A)$ to a polytope $\mathcal{F}(A)$ such that $F(A) = \mathcal{F}(A) \cap (X \times Z)$.

Since we are interested in obtaining two-stage algorithms, it will be convenient to reformulate problem (DRSO) as a problem that seeks only an optimal first-stage decision, and incorporates the second-stage decisions in an implicit manner. We denote by $g(x, A)$

the second-stage cost incurred if we choose the best possible *fractional* second-stage solution for scenario $A \in \mathcal{A}$, given the fractional first-stage decision $x \in \mathcal{P}$; that is, we define

$$g(x, A) := \min \left\{ \text{cost of } z^A : \left(x, z^A\right) \in \mathcal{F}(A) \right\}.$$

To ensure that $g(x, A)$ is well defined, we assume that there is always a fractional second-stage decision z^A such that $\left(x, z^A\right) \in \mathcal{F}(A)$; this is a standard assumption in the study of two-stage optimization problems. Since (fractional or integer) second-stage decisions have nonnegative costs, we have $g(x, A) \geq 0$. Given a first-stage decision $x \in \mathcal{P}$, let $z(\mathring{p}\,; x) := \sup_{p: L(\mathring{p}, p) \leq r} \mathbb{E}_{A \sim p}[g(x, A)]$ be the expected cost incurred in the second stage when the worst possible distribution in the ambiguity set D is realized, and let $h(\mathring{p}\,; x) := c^\intercal x + z(\mathring{p}\,; x)$ denote the total cost incurred if we choose x as a first-stage decision, along with optimal fractional decisions in the second stage. We consider the relaxation of (DRSO) with integer first-stage decisions and (implicit) fractional second-stage decisions,

$$\min_{x \in X} h(\mathring{p}\,; x)\ , \tag{Q($\mathring{p}$)}$$

and its further relaxation wherein first-stage decisions are also fractional,

$$\min_{x \in \mathcal{P}} h(\mathring{p}\,; x)\ . \tag{Q^{fr}($\mathring{p}$)}$$

One benefit of moving from (DRSO) to the relaxations (Q($\mathring{p}$)) and (Q^{fr}($\mathring{p}$)) is that, in the applications we consider in Chapter 6, for every scenario A, the function $x \mapsto g(x, A)$ is convex over $\mathcal{P}$, and we can efficiently compute its value, as well as a subgradient, at any given point. Furthermore, as we discuss in Section 3.4.2, one can convert approximate solutions for the fractional relaxations into approximate solutions for the original (discrete) problem via LP-rounding algorithms.

We now state the assumptions that we make. Assumption (A1) sets a generous limit on the size of the first-stage decision set. Note that without this assumption, we would not be able to represent integer first-stage decisions using $\mathsf{poly}(\mathcal{I})$ bits (since the number of distinct binary strings using at most N bits is $\mathrm{O}\left(2^N\right)$).

(A1) $\log |X| = \mathsf{poly}(\mathcal{I})$.

Assumptions (A2) and (A3) are lifted from Charikar, Chekuri, and Pál [24], who use them to prove an SAA result for two-stage stochastic problems. Informally, (A2) says that

the empty first-stage decision (i.e., $x = 0$) is allowed, and is the first-stage decision that helps the least in the second stage; (A3) says that, if we choose the empty decision in the first stage, then the "regret" we face in the second stage, relative to any other fractional first-stage decision x, is no larger than λ times the cost of x.

(A2) We have $0 \in X$ and $g(0, A) \geq g(x, A)$ for every $x \in \mathcal{P}$ and $A \in \mathcal{A}$.

(A3) We know an *inflation factor* $\lambda \geq 1$ such that $g(0, A) \leq g(x, A) + \lambda c^\intercal x$ for every $x \in \mathcal{P}$ and $A \in \mathcal{A}$.

A common characteristic of all the frameworks we develop in Chapters 4–7 for designing approximation algorithms for two-stage DRS problems is that they involve obtaining approximate solutions for one of the relaxations $\{(\mathrm{Q}(\mathring{p})), (\mathrm{Q}^{\mathrm{fr}}(\mathring{p}))\}$, typically using an ellipsoid-based method (either the classical ellipsoid method for convex optimization, or one of its variants discussed in Section 3.4.1). Determining the number of iterations of these methods requires bounds on the polytope $\mathcal{P}$ in terms of enclosed and enclosing balls; this is captured by (A4), which is directly lifted from Shmoys and Swamy [114]. Note that the vast majority of two-stage problems involve $\{0, 1\}$ decisions, so we have $X = \{0, 1\}^m$ and $\mathcal{P} = [0, 1]^m$, and assumption (A4) is readily satisfied. As in [114], for any given scenario $A \in \mathcal{A}$, we need a value oracle and a subgradient oracle for the function $x \mapsto g(x, A)$, which is a benign requirement since $g(x, A)$ is the optimal value of a polytime-solvable LP in all our applications, and subgradients can be obtained from optimal solutions to the dual of this LP. Whereas Shmoys and Swamy [114] define a syntactic class of two-stage stochastic LPs and show (implicitly) that they satisfy this requirement, we explicitly isolate this requirement in assumption (A5).

(A4) We have positive bounds $R_{\mathrm{small}} \leq 1$ and R_{large} such that:

- $\mathcal{P}$ contains a Euclidean ball of radius R_{small};
- $\mathcal{P}$ is contained in the Euclidean ball of radius R_{large} centered at the origin; and
- $\log \frac{R_{\mathrm{large}}}{R_{\mathrm{small}}} = \mathsf{poly}(\mathcal{I})$.

(A5) For every scenario $A \in \mathcal{A}$, the function $x \mapsto g(x, A)$ is convex over $\mathcal{P}$. Furthermore, given any point $x \in \mathcal{P}$, we can compute in $\mathsf{poly}(\mathcal{I})$ time the value of this function at x *and* a subgradient d at x such that $\|d\| \leq K$, where $\log K = \mathsf{poly}(\mathcal{I})$. By Lemma 1.1, this also implies that the function $x \mapsto g(x, A)$ is K-Lipschitz continuous.

Assumption (A6) is also lifted from Shmoys and Swamy [114], and allows us to express our main results in a more convenient form, with purely multiplicative guarantees. We say a scenario $A \in \mathcal{A}$ is a *null scenario* if we have $g(x, A) = g(0, A)$ for every fractional first-stage decision $x \in \mathcal{P}$. (For example, for DRS facility location, $A = \emptyset$ is a null scenario.)

(A6) For every fractional first-stage decision $x \in \mathcal{P}$ and every non-null scenario $A \in \mathcal{A}$ we have $c^\intercal x + g(x, A) \geq 1$.

In the Wasserstein setting, where L is defined with respect to a scenario metric ℓ, we make the following additional assumption, which relates ℓ to the second-stage costs $g(\cdot, \cdot)$. This is a rather mild assumption and holds for all the applications we consider in Chapter 6 (see the discussion in the introduction of that chapter).

(A7) We have a number $\tau \geq 1$ with $\log \tau = \mathsf{poly}(\mathcal{I})$ such that $g(x, A') - g(x, A) \leq \tau \cdot \ell(A, A')$ for every fractional first-stage decision $x \in \mathcal{P}$ and every pair of scenarios (A, A') with $\ell(A, A') > 0$.

We also suppose that $\ell(A, A) = 0$ for all $A \in \mathcal{A}$, and that we are given an upper bound $\ell_{\max} > 0$ on the pairwise scenario distances $\{\ell(A, A')\}$, with $\log \ell_{\max} = \mathsf{poly}(\mathcal{I})$. We assume without loss of generality that $r \leq \ell_{\max}$ (note that for $r \geq \ell_{\max}$, *every* distribution p satisfies $L_{\mathrm{W}}(\mathring{p}, p) \leq r$).

3.3 Overview of results and techniques

In this section, we give an overview of our results for two-stage DRS optimization under a Wasserstein ball and an L_∞ ball, and of the techniques used to obtain them.

As mentioned before, our frameworks involve relaxations of the discrete problem (DRSO) with fractional second-stage (and possibly first-stage) decisions. We now define two types of algorithms that we utilize to convert fractional solutions into integer ones.

When we work with the relaxation (Q($\mathring{p}$)), we often obtain only an integer first-stage decision $\widehat{x} \in X$, and a bound on the quality of the solution formed by $\widehat{x}$ coupled with optimal *fractional second-stage decisions*. To obtain a solution for (DRSO), we also need *integer* second-stage decisions.

Definition 3.3. A *second-stage α-approximation algorithm* is an algorithm that, given an integer first-stage decision $\widehat{x} \in X$ and a scenario $A \in \mathcal{A}$, computes in $\mathsf{poly}(\mathcal{I})$ time an integer second-stage decision $\widehat{z}^A \in Z$ such that $\left(\widehat{x}, \widehat{z}^A\right)$ is feasible for scenario A and (cost of $\widehat{z}^A$) $\leq \alpha \cdot g(\widehat{x}, A)$.

Typically, the problem of computing suitable second-stage actions for a scenario A boils down to solving an instance of the deterministic version of the underlying problem. For example, for DRSFL, this boils down to (approximately) solving an instance with client set A, where each facility i has opening cost 0 if it has already been opened in the first stage, and f_i^{II} otherwise. So an LP-based α-approximation algorithm for the deterministic problem yields a second-stage α-approximation algorithm.

When we work with the relaxation ($\mathrm{Q}^{\mathrm{fr}}(\mathring{p})$), we obtain a *fractional* first-stage decision $x \in \mathcal{P}$, and a guarantee on the quality of the solution formed by x coupled with optimal fractional second-stage decisions. To obtain a solution for (DRSO), we need to compute both *integer* first-stage and second-stage decisions; we do so using a suitable two-stage algorithm. (Recall from Definition 3.2 that two-stage algorithms compute second-stage decisions in an implicit manner.)

Definition 3.4. A *local ρ-approximation algorithm* is a two-stage algorithm that, given a fractional first-stage decision $x \in \mathcal{P}$, computes in $\mathsf{poly}(\mathcal{I})$ time a feasible (integer) solution $\left(\widehat{x}, \left\{\widehat{z}^A\right\}_{A \in \mathcal{A}}\right)$ for (DRSO) with the following guarantees: $c^\intercal \widehat{x} \leq \rho \cdot c^\intercal x$ and (cost of $\widehat{z}^A$) $\leq \rho \cdot g(x, A)$ for every scenario $A \in \mathcal{A}$.

Local approximation algorithms exist for various two-stage combinatorial-optimization problems such as set cover, edge cover, vertex cover, and facility location with approximation factors that are comparable to the approximation factors known for their deterministic counterparts (see Shmoys and Swamy [114]).

3.3.1 DRS optimization under a Wasserstein ball

In Chapters 4–6 we consider the discrete DRS problem

$$\min_{\substack{x \in X, z \in Z^{\mathcal{A}}: \\ \left(x, z^A\right) \in F(A) \quad \forall A \in \mathcal{A}}} \left\{ c^\intercal x + \sup_{p : L_{\mathrm{W}}(\mathring{p}, p) \leq r} \mathbb{E}_{A \sim p}\left[\text{cost of } z^A\right] \right\}, \tag{DRSO$_{\mathrm{W}}$}$$

where L_W is a Wasserstein metric defined with respect to a scenario metric ℓ. We consider two choices for the scenario collection: the *unrestricted setting* (i.e., $\mathcal{A} = 2^U$) and the *k-bounded setting* (i.e., $\mathcal{A} = \mathcal{A}_{\leq k} := \{A \subseteq U : |A| \leq k\}$), for some ground set U. We consider Wasserstein metrics defined relative to the following scenario metrics: the discrete metric, defined by $\ell^{\mathsf{disc}}(A, A') = 1$ if $A \neq A'$, and 0 otherwise (so L_W is the $\frac{1}{2}L_1$ metric); and the asymmetric metric $\ell_\infty^{\mathsf{asym}}$ with respect to an underlying distance function w over U, defined by $\ell_\infty^{\mathsf{asym}}(A, A') := \max_{j' \in A'} w(j', A)$, where $w(j', A) := \min_{j \in A} w_{j'j}$. Table 3.1 summarizes the results we obtain for the DRS versions of various discrete-optimization problems under a Wasserstein ball.

Problem	ℓ^{disc}		$\ell_\infty^{\mathsf{asym}}$		General $\mathcal{A}$, ℓ
	$\mathcal{A} = 2^U$	$\mathcal{A}_{\leq k}$	$\mathcal{A} = 2^U$	$\mathcal{A}_{\leq k}$	β=approx. for (Π)
Facility location	21.96	196	21.96	196	$O(\beta)$
Vertex cover	16	101.3	–	–	$O(\beta)$
Edge cover	12	36	–	–	$O(\beta)$
Set cover	$O(\log \lvert U \rvert)$	$O\big(\log^2 \lvert U \rvert\big)$	–	–	$O(\beta \log \lvert U \rvert)$
Steiner tree	160	*	160	*	*

Table 3.1: A summary of the approximation factors we obtain for various applications in the Wasserstein setting. We have omitted the $O(\varepsilon)$ terms that appear in the approximation factors. The $\ell_\infty^{\mathsf{asym}}$ setting does not apply to vertex cover, edge cover, and set cover. The approximation factor β for (Π) is the factor $\beta_1\beta_2$ in Theorem 3.6. The * entries are open questions.

We relate the approximability of the discrete DRS problem (DRSO_W) to that of the following *deterministic* problem.

(Π) Given an integer first-stage decision $x \in X, y \geq 0$, and a scenario $A \in \mathcal{A}$, solve

$$g(x, y, A) := \max_{A' \in \mathcal{A}} \{g(x, A') - y \cdot \ell(A, A')\} \ .$$

Note that problem (Π) ties together three distinct sources of complexity in the two-stage DRS problem: the combinatorial complexity of the underlying optimization problem, captured by $g(x, A')$; the complexity of the scenario collection $\mathcal{A}$; and the complexity of the scenario metric ℓ, captured by the $y \cdot \ell(A, A')$ term.

Under the standard notion of approximation, it is impossible to obtain any approximation guarantee for problem (Π) due to its mixed-sign objective (see Theorem 5.9-(b)). To evade this difficulty, we consider the following non-standard notion of approximation.

Note that a $(1,1)$-approximation algorithm for (Π) corresponds to an exact algorithm.

Definition 3.5. Let $\beta_1, \beta_2 \geq 1$. A (β_1, β_2)-*approximation algorithm* for problem (Π) is an algorithm that, given an instance $(x, y, A) \in X \times \mathbb{R}_+ \times \mathcal{A}$, computes in $\mathsf{poly}(\mathcal{I})$ time a scenario $\overline{A} \in \mathcal{A}$ such that

$$g\big(x, \overline{A}\big) - y \cdot \ell\big(A, \overline{A}\big) \geq \max_{A' \in \mathcal{A}} \left\{ \frac{1}{\beta_1} g(x, A') - \beta_2 y \cdot \ell(A, A') \right\} .$$

Our main result for DRS optimization under a Wasserstein-ball ambiguity set is the following. See Definitions 3.2–3.5 for terminology.

Theorem 3.6 (see proof in Section 6.1)**.**
Suppose that assumptions (A1)–(A7) hold, and that we have:

(1) a second-stage α-approximation algorithm;
(2) a (β_1, β_2)-approximation algorithm for problem (Π), with $\log \beta_1 = \mathsf{poly}(\mathcal{I})$; and
(3) a local ρ-approximation algorithm.

Then there exists a two-stage algorithm that, given $\varepsilon > 0$ and $\delta \in (0,1)$, computes a $4\alpha\beta_1\beta_2\rho(1+\varepsilon)$-approximate solution for ($\mathrm{DRSO_W}$) with probability at least $1-\delta$ in time $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log \frac{1}{\delta}\big)$.

Ingredients (1) and (3) can often be obtained using known results for deterministic and two-stage-stochastic optimization respectively; ingredient (2) is the new component we need to supply to instantiate Theorem 3.6 and obtain results for specific two-stage DRS problems. In various settings, we show that an approximation algorithm for (Π) (in the sense of Definition 3.5) can be obtained via an approximation algorithm for the constrained problem $\max_{A' \in \mathcal{A}: \ell(A,A') \leq \mu} g(x, A')$, given $(x, \mu, A) \in X \times \mathbb{R}_+ \times \mathcal{A}$ (see Lemma 6.2). For all the applications that we consider in Chapter 6, this reduces to the simpler max-min problem $\max_{A' \in \mathcal{A}} g(x, A')$, encountered in two-stage robust optimization. Note that this can be seen as the special case of problem (Π) wherein $y = 0$. In the k-bounded setting (i.e., $\mathcal{A} = \mathcal{A}_{\leq k}$), by setting $x = 0$ we recover the k-max-min problem that has been studied in the literature for various applications (see Section 2.1). In particular, this reduction from (Π) to the max-min problem $\max_{A' \in \mathcal{A}} g(x, A')$ problem applies to the $\frac{1}{2}L_1$-metric. Theorem 3.6 thus provides a novel, useful *reduction* from two-stage DRS optimization to deterministic and two-stage {stochastic, robust} optimization. (For instance, Gupta, Nagarajan, and Ravi [59] devise approximations for the max-min problem for various applications, with scenario collection $\mathcal{A}$ defined by matroid-independence and/or knapsack constraints; we

are able to export these guarantees to the corresponding two-stage DRS problem under a $\frac{1}{2}L_1$-ball.)

In Chapter 6, we utilize Theorem 3.6 to obtain approximation algorithms for two-stage DRS versions of various combinatorial-optimization problems. We also discuss variants of Theorem 3.6 that lead to improved approximation factors for specific applications. Our strongest results are for set cover, vertex cover, edge cover, and facility location. For Steiner tree, we are not able to directly apply Theorem 3.6 because we do not have a local approximation algorithm, but we are still able to obtain approximation algorithms in the unrestricted setting ($\mathcal{A} = 2^U$), using a weaker type of rounding algorithm (see Section 6.8).

Technical contributions. The reduction in Theorem 3.6 is obtained by supplementing tools from two-stage {stochastic, robust} optimization with various additional ideas. Its proof consists of two main components, *both of which are of independent interest.*

• **Sample average approximation (SAA) for DRS problems.** In Chapter 4, we prove that a simple and appealing approach in stochastic optimization called the *sample-average-approximation* (SAA) method can be applied to *reduce the relaxed DRS problem* ($\mathrm{Q}(\mathring{p})$) *to the setting where the central distribution $\mathring{p}$ has a moderate-size support and is represented explicitly.* In the SAA method, we draw some N samples to construct an empirical estimate $\widehat{p}$ of $\mathring{p}$, and solve the DRS problem obtained by replacing $\mathring{p}$ with $\widehat{p}$. (See Section 2.2 for a discussion of the use of this method in two-stage stochastic optimization.) Roughly speaking, we show that by taking $N = \mathsf{poly}(\mathcal{I}, \lambda)$ samples, we can ensure that an approximation algorithm for the SAA problem, in conjunction with an approximate value oracle for the SAA objective function, can be used to obtain an approximate solution for the original problem with high probability (see Theorem 4.1). It is well known that $\Omega(\lambda)$ samples are needed even for (standard) two-stage stochastic problems in the black-box model (see Shmoys and Swamy [114]). Our SAA result substantially expands the scope of problems for which the SAA method is known to be effective with $\mathsf{poly}(\mathcal{I}, \lambda)$ sample size. Previously, as discussed in Section 2.2, such results were known for the special case of two-stage stochastic problems, and multi-stage stochastic problems with a constant number of stages, given an *exact* algorithm for the SAA problem and an *exact* evaluation oracle for its objective function.

• **Solving the explicit central-distribution case.** Complementing the above SAA result, we show in Chapter 5 how to approximately solve a two-stage DRS problem with a central distribution $\widehat{p}$ that is represented explicitly. It is natural to move to a fractional relaxation of the problem, by enlarging the first-stage and second-stage decision sets to

suitable polytopes. In *stark contrast* with two-stage {stochastic, robust} optimization, where the fractional relaxation of a problem with an explicit list of scenarios immediately gives a polynomial-size LP and therefore can be solved exactly in polynomial time, it is substantially more challenging to even approximately solve the fractional DRS problem with an explicit central distribution. In fact, this is perhaps the technically more-challenging part of obtaining an approximation algorithm for DRS problems. The crux of the problem is that, while $\widehat{p}$ has moderate-size support, there are (numerous) distributions p in the ambiguity set D that have *exponential-size support*, and one needs to optimize over such distributions. In particular, if we reformulate $\min_{x\in\mathcal{P}} \left\{c^\intercal x + \max_{p:L_{\mathrm{W}}(\widehat{p},p)\le r} \mathbb{E}_{A\sim p}[g(x,A)]\right\}$ as a minimization LP (by taking the dual of the LPs defining $g(x,A)$, and then the dual of the inner maximization problem), we obtain an LP with *an exponential number of both constraints and variables*. (See the discussion in Chapter 5.) Thus, while we started with a central distribution of moderate-size support, we have ended up in a situation similar to that in two-stage stochastic or robust optimization with an exponential number of scenarios. To surmount these obstacles, we work with the problem $\min_{x\in X} h(\widehat{p};x)$, whose fractional relaxation $\min_{x\in\mathcal{P}} h(\widehat{p};x)$ is a *convex program*, and solve this approximately by leveraging the ellipsoid-based algorithm from Theorem 3.14 (see Theorem 5.1). Not surprisingly, this poses various fresh difficulties, which are discussed in detail in Chapter 5.

Remark 3.7. As noted earlier, we would like to be able to handle settings with an exponential number of scenarios, which arise naturally in a variety of discrete-optimization problems; hence our focus is on the black-box model. However, we remark that if (i) the number of scenarios is polynomial (i.e., $|\mathcal{A}|$ is $\mathsf{poly}(\mathcal{I})$) and (ii) the central distribution $\mathring{p}$ is represented explicitly, then it becomes much simpler to solve a fractional relaxation of the DRS problem (where we allow fractional first-stage and second-stage decisions) since one can then leverage LP duality to cast this relaxation as a compact LP. A more general result along these lines is discussed in Section 5.3.

Remark 3.8. We choose to work with the relaxation ($\mathrm{Q}(\mathring{p})$) (with integer first-stage decisions) instead of the relaxation ($\mathrm{Q}^{\mathrm{fr}}(\mathring{p})$) (with fractional first-stage decisions) for two reasons. First, it is slightly easier to prove an SAA result for ($\mathrm{Q}(\mathring{p})$) (see Chapter 4), showing that approximate solutions for the SAA version ($\mathrm{Q}(\widehat{p})$) of the relaxed problem translate into approximate solutions for the original relaxed problem ($\mathrm{Q}(\mathring{p})$). Second, to solve the SAA problem ($\mathrm{Q}(\widehat{p})$) in Chapter 5, we utilize the ellipsoid method with a generalized first-order oracle (see Theorem 3.14). In order to use the (β_1,β_2)-approximation for problem (Π), the generalized first-order oracle needs to round a fractional first-stage decision to an integer point. Thus, the rounding is embedded inside the ellipsoid-based method, and the algorithm directly returns an integer first-stage solution.

3.3.2 DRS optimization under an L_∞ ball

In Chapter 7 we consider the problem

$$\min_{\substack{x\in X, z\in Z^{\mathcal{A}}:\\ (x,z^A)\in F(A)\quad \forall A\in\mathcal{A}}} \left\{ c^\intercal x + \sup_{p: L_\infty(\mathring{p},p)\le r} \mathbb{E}_{A\sim p}\big[\text{cost of } z^A\big] \right\} . \tag{DRSO$_\infty$}$$

Table 3.2 summarizes the results we obtain for the DRS versions of various discrete-optimization problems under an L_∞ ball.

Problem	$\mathcal{A} = 2^U$
Facility location	11
Vertex cover	8
Edge cover	6
Set cover	$O(\log \lvert U\rvert)$

Table 3.2: A summary of the approximation factors we obtain for DRS optimization under an L_∞ ball. We have omitted the $O(\varepsilon)$ terms that appear in the approximation factors.

Recall from Section 3.2 that $g(x, A)$ denotes the optimal second-stage cost of scenario A given x as the first-stage decision, when we allow *fractional* second-stage actions. We assume the following stronger version of assumption (A5).

(A5') For every scenario $A \in \mathcal{A}$, the function $x \mapsto g(x, A)$ is convex over $\mathcal{P}$. Furthermore, given a point $x \in \mathcal{P}$, we can compute in $\mathsf{poly}(\mathcal{I})$ time the value of this function at x *and* a subgradient d at x such that $-\lambda c \le d \le 0$. By Lemma 1.1, this also implies that the function $x \mapsto g(x, A)$ is $\lambda \|c\|$-Lipschitz continuous.

Shmoys and Swamy [114] define a broad class of two-stage problems for which assumption (A5') holds, which includes set cover, facility location, and Steiner tree.

Our main result for DRS optimization under an L_∞-ball ambiguity set is that we can obtain an approximation algorithm for problem (DRSO$_\infty$) via an algorithm for the following problem.

(Υ) Given a fractional first-stage decision $x \in \mathcal{P}$ and $1 \le t \le \min\left\{|\mathcal{A}|, \frac{1}{r}\right\}$, find the t scenarios $A \in \mathcal{A}$ with largest $g(x, A)$ value.

Theorem 3.9 (combination of Theorem 7.1 and Lemma 3.16)**.**
Suppose that assumptions (A1)–(A4), (A5'), and (A6) hold, and that we have:

(1) an algorithm for problem (Υ) *with* $\mathsf{poly}(\mathcal{I}, t)$ *running time; and*
(2) a local ρ*-approximation algorithm.*

Then there exists a two-stage algorithm that, given $\varepsilon > 0$ *and* $\delta \in (0, 1)$*, computes a* $(2+\varepsilon)\rho$*-approximate solution for* (DRSO_∞) *with probability at least* $1 - \delta$ *in time* $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{r}, \frac{1}{\varepsilon}, \log \frac{1}{\delta}\big)$.

As mentioned before, ingredient (2) can often be obtained using known results for two-stage stochastic optimization. Ingredient (1) is the new component we need to supply to instantiate Theorem 3.9 and obtain results for specific two-stage DRS problems. We show that this can be obtained in the unrestricted setting ($\mathcal{A} = 2^U$) when we have $g(x, A) \leq g(x, A')$ for every fractional first-stage decision $x \in \mathcal{P}$ and for every pair of scenarios (A, A') with $A \subseteq A'$, a condition that holds for covering problems, and in particular for the applications we consider in Chapter 6. We utilize Theorem 3.9 to obtain approximation results for two-stage DRS versions of set cover, edge cover, vertex cover, and facility location in the unrestricted setting.

Instead of using an SAA approach to first move to an empirical estimate $\widehat{p}$ of the central distribution $\mathring{p}$, we directly consider the fractional relaxation ($\mathrm{Q}^{\mathrm{fr}}(\mathring{p})$) of ($\mathrm{DRSO}_\infty$) with fractional first-stage decisions and (implicit) fractional second-stage decisions. As in the Wasserstein setting, even for a central distribution with polynomial-size support, solving this relaxation is quite challenging since it again leads to an LP with exponentially many variables and constraints. We move to a proxy objective function that is pointwise close to the true objective, and show that approximate subgradients of the proxy objective function can be computed efficiently at any point (for a certain notion of approximate subgradients introduced by Shmoys and Swamy [114], and presented in Section 3.4.1). This enables to use the algorithm of [114] to find an approximate minimizer of the proxy function; rounding this solution using a local approximation algorithm yields an approximate solution for the discrete DRS problem (DRSO_∞).

3.4 Some preliminary results and definitions

This section contains some preliminary results that are used by our frameworks. In Section 3.4.1, we discuss different approaches for minimizing functions over the set of

{fractional, integer} first-stage decisions. In Section 3.4.2, we discuss how rounding algorithms can be used to translate approximate solutions for the relaxed DRS problems $\{(\mathrm{Q}(\mathring{p})), (\mathrm{Q}^{\mathrm{fr}}(\mathring{p}))\}$ into approximate solutions for the discrete DRS problem (DRSO).

3.4.1 Optimizing over $\mathcal{P}$ or X via ellipsoid-based methods

Throughout this thesis, we encounter multiple times the task of solving a problem of the type

$$\min_{x \in \overline{X}} H(x) \ ,$$

for a certain function $H : \mathcal{P} \to \mathbb{R}$, where $\overline{X}$ is either the polytope $\mathcal{P}$ of fractional first-stage decisions, or the set X of integer first-stage decisions.

Let us recall the assumptions that we stated for the polytope $\mathcal{P}$ in Section 3.2, on which the results stated in this section rely. We assume that we have a $\mathsf{poly}(\mathcal{I})$-time separation oracle for the polytope $\mathcal{P}$. By assumption (A4), we have that $\mathcal{P}$ contains a Euclidean ball of radius R_{small} and is contained in the Euclidean ball of radius R_{large} centered at the origin, where $\log \frac{R_{\mathrm{large}}}{R_{\mathrm{small}}} = \mathsf{poly}(\mathcal{I})$. Furthermore, we have $0 \in X$ (and hence $0 \in \mathcal{P}$) by assumption (A2).

When H is a convex function and $\overline{X} = \mathcal{P}$ (or, more generally, $\overline{X}$ is a convex set), a classical tool that could be used for solving this problem is the *ellipsoid method* introduced by Nemirovski and Yudin [93] and Shor [117]. The algorithm accesses the objective function H via a *first-order oracle*, which, given any point $x \in \mathcal{P}$, computes a tuple $(x, H(x), d)$ where d is a subgradient of H at x.[4]

Theorem 3.10 (Nemirovski and Yudin [93])**.** *Suppose that the function $H : \mathcal{P} \to \mathbb{R}$ is convex and $\widetilde{K}$-Lipschitz continuous, and that we have a first-order oracle for H with running time T_{oracle}. Then, given $\eta > 0$, we can compute in $\mathsf{poly}\left(\mathcal{I}, \log \widetilde{K}, T_{\mathrm{oracle}}, \log \frac{1}{\eta}\right)$ time a solution $\widetilde{x} \in \mathcal{P}$ such that*

$$H(\widetilde{x}) \leq \min_{x \in \mathcal{P}} H(x) + \eta \ .$$

The algorithm from Theorem 3.10 has two phases. In the first phase, starting with an ellipsoid that contains the entire feasible region, at each step, we add a cut (i.e., a linear

[4] We are following the nomenclature used e.g. by Nesterov [94] and Ben-Tal and Nemirovski [7]. Note that in other contexts, the term *first-order oracle* is used for an algorithm that computes *only* a subgradient at any given point (and not the value of the function).

inequality) passing through the center $\widetilde{x}$ of the current ellipsoid to chop off a half-ellipsoid that does not contain points of interest. If $\widetilde{x}$ is infeasible, we use the separation oracle for $\mathcal{P}$ to obtain such a cut. Otherwise, we find a subgradient d of H at $\widetilde{x}$ and use the cut $d^\intercal(x - \widetilde{x}) \leq 0$; the definition of subgradient ensures that any point x discarded by this cut satisfies $H(x) > H(\widetilde{x})$. We then replace the current ellipsoid with a smaller one, and iterate this process until the volume of the current ellipsoid becomes sufficiently small, which happens after a suitably small number of iterations. In the second phase, we select the best solution among the feasible solutions produced in the first phase.

In various two-stage stochastic-optimization settings, as well as distributionally robust optimization, evaluating the objective function, even approximately, is an intractable (#P-hard) problem since the expectation is computed over an exponential number of scenarios (see Dyer and Stougie [37, 38] and Hanasusanto, Kuhn, and Wiesemann [68]). Similarly, computing subgradients exactly is also difficult. To deal with these issues, Shmoys and Swamy [114] give an algorithm for minimizing a convex function over a convex set utilizing the following notion of approximate subgradients.

Definition 3.11. Consider a function $H : \mathcal{P} \to \mathbb{R}$, and let $\omega \geq 0$. A vector $\widehat{d} \in \mathbb{R}^m$ is an *ω-subgradient* of H at a point $x \in \mathcal{P}$ if for every $x' \in \mathcal{P}$ we have

$$H(x') - H(x) \geq \widehat{d} \cdot (x' - x) - \omega H(x) \ .$$

Note that (exact) subgradients correspond to ω-subgradients with $\omega = 0$. Using a variant of the ellipsoid method, Shmoys and Swamy [114] showed the following result.

Theorem 3.12 (see Theorem 4.7 and Lemma 4.14 in Shmoys and Swamy [114])**.** *Suppose that the function $H : \mathcal{P} \to \mathbb{R}$ is convex and $\widetilde{K}$-Lipschitz continuous. Let $\varepsilon > 0$, $\eta > 0$, and $\delta \in (0, 1)$. Define $\widetilde{N} := \left\lceil 2m^2 \ln\left(\frac{16\widetilde{K}R_{\text{large}}^2}{R_{\text{small}} \cdot \eta}\right)\right\rceil$, $\widetilde{n} := \widetilde{N} \cdot \ln\left(\frac{8\widetilde{N}\widetilde{K}R_{\text{large}}}{\eta}\right)$, and $\omega := \frac{\min\{\varepsilon,1\}}{4\widetilde{n}}$. Suppose that given any point $x \in \mathcal{P}$ we can compute a vector that is an ω-subgradient of H at x with probability at least $1 - \delta$ in time $T(\omega, \delta)$.[5] Then we can compute in $\mathsf{poly}\left(\mathcal{I}, \log \widetilde{K}, T\left(\omega, \frac{\delta}{\widetilde{N}+\widetilde{n}}\right), \log \frac{1}{\eta}\right)$ time a solution $\widetilde{x} \in \mathcal{P}$ that with probability at least $1 - \delta$ satisfies*

$$H(\widetilde{x}) \leq (1 + \varepsilon) \cdot \min_{x \in \mathcal{P}} H(x) + \eta \ .$$

Theorem 3.12, in addition to relaxing the requirement of being able to compute exact subgradients, completely dispenses with the requirement of an (even approximate)

[5] We need *not* be able to certify the correctness of the output of this algorithm.

objective-value oracle (the second phase of the ellipsoid method is replaced with an iterative binary-search based on approximate subgradients). We will utilize the algorithm from Theorem 3.12 for DRS optimization under an L_∞ ball (see Chapter 7). However, for DRS optimization under a Wasserstein ball, we will not even be able to compute an ω-subgradient of the objective function for as small an ω as required by Theorem 3.12. To deal with such cases, we develop another variant of the ellipsoid method, which involves the following generalized notion of first-order oracles.

Definition 3.13. Consider a function $H : \mathcal{P} \to \mathbb{R}$, and let $\psi \geq 1$ and $\overline{X} \subseteq \mathcal{P}$. A $\big(\psi, \overline{X}\big)$*-first-order oracle for* H is an algorithm that, given a point $\widetilde{x} \in \mathcal{P}$, computes a tuple $(\overline{x}, f, d) \in \overline{X} \times \mathbb{R} \times \mathbb{R}^m$ such that (i) $H(\overline{x}) \leq f$; and (ii) $H(y) \geq \frac{1}{\psi} f$ for every $y \in \overline{X}$ such that $d^\intercal(y - \widetilde{x}) \geq 0$.

Note that classical first-order oracles correspond to $(1, \mathcal{P})$-first-order oracles that always return $\overline{x} = \widetilde{x}$. The intuition behind this definition is that, given any $\widetilde{x} \in \mathcal{P}$, a $\big(\psi, \overline{X}\big)$-first-order oracle yields an estimate of the objective value at a related point $\overline{x} \in \overline{X}$ and a hyperplane passing through $\widetilde{x}$ such that the points chopped off by this hyperplane are better off by at most a factor ψ compared to $\overline{x}$. This suggests that running a modified ellipsoid method that uses d in lieu of a subgradient at each iteration in the first phase, and uses the estimates $\{f\}$ to select a solution in the second phase, yields an approximate minimizer of H over $\overline{X}$. We now prove that this is indeed the case. Note that the function H is *not* required to be convex.

Theorem 3.14. *Let $H : \mathcal{P} \to \mathbb{R}$ be a $\widetilde{K}$-Lipschitz continuous function, and suppose that we have a $\big(\psi, \overline{X}\big)$-first-order oracle for H with running time T_{oracle}, where $\psi \geq 1$ and $\overline{X} \subseteq \mathcal{P}$. Then, given $\eta > 0$, we can compute in $\mathsf{poly}\Big(\mathcal{I}, \widetilde{K}, T_{\text{oracle}}, \log \frac{1}{\eta}\Big)$ time a solution $\overline{x} \in \overline{X}$ and an estimate f such that*

$$H(\overline{x}) \leq f \leq \psi \cdot \left(\min_{x \in \overline{X}} H(x) + \eta \right) .$$

Proof. We state below the algorithm used to obtain the theorem.

1. Set $k \leftarrow 0$, $\widetilde{x}^0 \leftarrow 0$, $\mu \leftarrow \min\left\{1, \frac{\eta}{2\widetilde{K}R_{\text{large}}}\right\}$, $N \leftarrow \left\lceil 2m^2 \ln \frac{2R_{\text{large}}}{\mu R_{\text{small}}} \right\rceil$, $E_0 \leftarrow \{x \in \mathbb{R}^m : \|x\| \leq R_{\text{large}}\}$, and $\mathcal{P}_0 \leftarrow \mathcal{P}$.
2. For $i \leftarrow 0, \ldots, N-1$ do the following.
   (We maintain that E_i is an ellipsoid centered at $\widetilde{x}^i$ containing $\mathcal{P}_k$.)
   a) If $\widetilde{x}^i \notin \mathcal{P}_k$, let $a^\intercal x \leq b$ be an inequality that is satisfied by all $x \in \mathcal{P}_k$ but violated by $\widetilde{x}^i$. (This is obtained either from a separation oracle for $\mathcal{P}$, or from inequalities added in prior iterations.) Let S be the halfspace $\{x \in \mathbb{R}^m : a^\intercal(x - \widetilde{x}^i) \leq 0\}$.
   b) If $\widetilde{x}^i \in \mathcal{P}_k$, let $\left(\overline{x}^k, f^k, d^k\right)$ be the output of the $\left(\psi, \overline{X}\right)$-first-order oracle, when run with input $\widetilde{x}^i$. If $d^k = 0$, then set $k \leftarrow k+1$ and go to step 3. Otherwise, let S be the halfspace $\left\{x \in \mathbb{R}^m : d^k \cdot (x - \widetilde{x}^i) \leq 0\right\}$; set $\mathcal{P}_{k+1} \leftarrow \mathcal{P}_k \cap S$ and $k \leftarrow k+1$.
   c) Let E_{i+1} be the ellipsoid of minimum volume containing the half-ellipsoid $E_i \cap S$, and let $\widetilde{x}^{i+1}$ be its center.
3. Let $j \leftarrow \operatorname{argmin}_{i \in \{0,1,\ldots,k-1\}} f^i$. Return $\left(\overline{x}^j, f^j\right)$.

First, note that since $\widetilde{x}^0 \in \mathcal{P}_0$ by assumption (A2), we increment k in the first iteration, and hence j is well defined when we reach step 3. Let $(\overline{x}, f)$ denote the output of the algorithm. Note that $H(\overline{x}) \leq f$ follows immediately from the properties of the $\left(\psi, \overline{X}\right)$-first-order oracle. To obtain an upper bound on f, we rework the arguments in the proof of Lemma 4.5 from Shmoys and Swamy [114]. Since $f = \min_{i \in \{0,1,\ldots,k-1\}} f^i$, it suffices to show that for every $x^* \in \overline{X}$, there exists $l \in \{0, 1, \ldots, k-1\}$ such that $f^l \leq \psi \cdot (H(x^*) + \eta)$. We let $x^* \in \overline{X}$ be arbitrary, and work toward showing that such an index l exists. If $d^l \cdot \left(x^* - \widetilde{x}^l\right) \geq 0$ for some l (this includes the case when $d^l = 0$), then by the properties of the $\left(\psi, \overline{X}\right)$-first-order oracle we have $H(x^*) \geq \frac{1}{\psi} f^l$. Along with $\eta > 0$, this implies the upper bound on f^l that we sought.

Now, suppose that there exists no index l such that $d^l \cdot \left(x^* - \widetilde{x}^l\right) \geq 0$, and note that this implies that $x^* \in \mathcal{P}_k$. Let $W \subseteq \mathcal{P}$ be the image of $\mathcal{P}$ under the affine transformation $x \mapsto x^* + \mu(x - x^*)$. So W is a shrunken version of $\mathcal{P}$ (by a factor μ), and we have $x^* \in \mathcal{P}_k \cap W$.

We claim that W is not contained in $\mathcal{P}_k$. Since $W \subseteq \mathcal{P} = \mathcal{P}_0$, and since $\mathcal{P}_k$ is obtained from $\mathcal{P}_0$ by adding the constraints $d^l \cdot \left(x - \widetilde{x}^l\right) \leq 0$ (for $l = 0, \ldots, k-1$), this implies that one of these constraints chops off a portion of W. Therefore there exists a point x' on

the boundary of W such that $d^l \cdot \left(x' - \widetilde{x}^l\right) = 0$ for some index l. By the properties of the $\left(\psi, \overline{X}\right)$-first-order oracle, this implies that $H(x') \geq \frac{1}{\psi} f^l$. Let $x \in \mathcal{P}$ be the point that is mapped to x' by the affine transformation mentioned above. We have

$$\|x' - x^*\| = \mu \|x - x^*\| \leq \frac{\eta}{2\widetilde{K} R_{\text{large}}} (2R_{\text{large}}) = \frac{\eta}{\widetilde{K}} \ ,$$

where the inequality follows from the definition of μ and the fact that $\mathcal{P}$ is contained in a ball of radius R_{large}. Since H is $\widetilde{K}$-Lipschitz continuous, this implies that $H(x') \leq H(x^*)+\eta$, and so we obtain $f^l \leq \psi \cdot H(x') \leq \psi \cdot (H(x^*) + \eta)$.

We now proceed to prove the claim that W is not fully contained in $\mathcal{P}_k$, which we do by showing that the volume of $\mathcal{P}_k$ is smaller than that of W. For $\mathcal{Q} \subseteq \mathbb{R}^m$, let $\mathsf{vol}(\mathcal{Q})$ denote the volume of $\mathcal{Q}$. Let vol_m denote the volume of an m-dimensional Euclidean unit ball. We have

$$\begin{aligned}
\mathsf{vol}(\mathcal{P}_k) &\leq \mathsf{vol}(E_N) \\
&\leq e^{-N/(2m)} \cdot \mathsf{vol}(E_0) \\
&\leq e^{-m \ln \frac{2R_{\text{large}}}{\mu R_{\text{small}}}} \cdot \mathsf{vol}(E_0) \\
&= \left(\frac{\mu R_{\text{small}}}{2R_{\text{large}}}\right)^m \cdot \mathsf{vol}(E_0) \\
&= \left(\frac{\mu R_{\text{small}}}{2}\right)^m \cdot \mathsf{vol}_m \\
&< \mu^m \mathsf{vol}(\mathcal{P}) \\
&= \mathsf{vol}(W) \ .
\end{aligned}$$

The first step follows because $\mathcal{P}_k \subseteq E_N$, which holds due to the invariant maintained by the for loop. The second step follows from the way in which the volume of the ellipsoid decreases from one iteration of the algorithm to the following one: it is well known that $\frac{\mathsf{vol}(E_{i+1})}{\mathsf{vol}(E_i)} \leq e^{-1/2m}$ for every $i \geq 0$ (see, e.g., Grötschel, Lovász, and Schrijver [57]). The third step follows from the definition of N. The fifth step follows because $\mathsf{vol}(E_0) = R_{\text{large}}^m \mathsf{vol}_m$. The sixth step uses $\mathsf{vol}(\mathcal{P}) \geq R_{\text{small}}^m \mathsf{vol}_m$, which holds because $\mathcal{P}$ contains a ball of radius R_{small}. The final step follows because W is obtained by shrinking $\mathcal{P}$ by a factor of μ.

A final point that needs to be addressed is how to implement step 2c). It is well known that there is an explicit formula that can be used to compute the centers $\{\widetilde{x}^i\}$ of the ellipsoids $\{E_i\}$. However the formula involves square roots and hence may lead to irrational numbers. It is well known that this difficulty can be circumvented by carrying

out the computations within a suitably small accuracy and adjusting the algorithm to account for rounding errors (see, e.g., Grötschel, Lovász, and Schrijver [57]). □

3.4.2 Rounding fractional solutions

As previously mentioned, our frameworks work with the fractional relaxations (Q($\mathring{p}$)) and (Q$^{\text{fr}}$($\mathring{p}$)) of the discrete problem (DRSO). In order to obtain integer solutions to (DRSO), we rely on *second-stage approximation algorithms*, which produce an integer second-stage decision for a given scenario, and a given integer first-stage decision; and *local-rounding approximation algorithms*, which round a fractional first-stage decision to obtain integer first-stage and second-stage decisions (see Definitions 3.3 and 3.4). (For DRS Steiner tree, we utilize a third type of algorithm, which we discuss in Section 6.8.) In this section, we show that these two types of algorithms allow converting approximate solutions for the relaxations (Q($\mathring{p}$)) and (Q$^{\text{fr}}$($\mathring{p}$)) into approximate solutions for the discrete DRS problem (DRSO).

Lemma 3.15. *Let $\widehat{x} \in X$ be a ψ-approximate solution for the relaxed DRS problem* (Q($\mathring{p}$)). *For each scenario $A \in \mathcal{A}$, let $\widehat{z}^A$ be the output of a second-stage α-approximation algorithm when given $\widehat{x}$ and A as input. Then $\left(\widehat{x}, \{\widehat{z}^A\}_{A\in\mathcal{A}}\right)$ is an $\alpha\psi$-approximate solution for the discrete DRS problem* (DRSO).

Proof. The solution obtained for problem (DRSO) attains objective value

$$\begin{aligned} c^\intercal \widehat{x} + \sup_{p\in D} \mathbb{E}_{A\sim p}\left[\text{cost of } \widehat{z}^A\right] &\le c^\intercal \widehat{x} + \alpha \cdot \sup_{p\in D} \mathbb{E}_{A\sim p}[g(x,A)] \\ &\le \alpha \cdot \left(c^\intercal \widehat{x} + \sup_{p\in D} \mathbb{E}_{A\sim p}[g(x,A)]\right) \\ &\le \alpha\psi \cdot \text{OPT}(\text{Q}(\mathring{p})) \\ &\le \alpha\psi \cdot \text{OPT}(\text{DRSO}) \ . \end{aligned}$$

The first step uses the guarantees of the second-stage approximation algorithm. The second step follows because $\alpha \ge 1$. The third step uses the fact that $\widehat{x}$ is a ψ-approximate solution for (Q($\mathring{p}$)). The final step follows since (Q($\mathring{p}$)) is a relaxation of (DRSO). □

Lemma 3.16. *Let $x \in \mathcal{P}$ be a ψ-approximate solution for the relaxed DRS problem* (Q$^{\text{fr}}$($\mathring{p}$)). *Let $\left(\widehat{x}, \{\widehat{z}^A\}_{A\in\mathcal{A}}\right)$ be the output of a local ρ-approximation algorithm when given x as input. Then $\left(\widehat{x}, \{\widehat{z}^A\}_{A\in\mathcal{A}}\right)$ is a $\psi\rho$-approximate solution for the discrete DRS problem* (DRSO).

Proof. The solution obtained for problem (DRSO) attains objective value

$$\begin{aligned} c^\intercal \widehat{x} + \sup_{p \in D} \mathbb{E}_{A \sim p}\big[\text{cost of } \widehat{z}^A\big] &\leq \rho \cdot \left(c^\intercal x + \sup_{p \in D} \mathbb{E}_{A \sim p}[g(x, A)] \right) \\ &\leq \psi\rho \cdot \mathrm{OPT}(\mathrm{Q}^{\mathrm{fr}}(\mathring{p})) \\ &\leq \psi\rho \cdot \mathrm{OPT}(\mathrm{DRSO}) \ . \end{aligned}$$

The first step uses the guarantees of the local approximation algorithm. The second one uses the fact that x is a ψ-approximate solution for $(\mathrm{Q}^{\mathrm{fr}}(\mathring{p}))$. The final one follows since $(\mathrm{Q}^{\mathrm{fr}}(\mathring{p}))$ is a relaxation of (DRSO). □

Chapter 4

DRS optimization under a Wasserstein ball: sample average approximation

In this chapter, we consider the discrete DRS problem under a Wasserstein ball

$$\min_{\substack{x\in X, z\in Z^{\mathcal{A}}:\\ \left(x,z^A\right)\in F(A)\quad \forall A\in\mathcal{A}}} \left\{ c^\intercal x + \sup_{p: L_{\mathrm{W}}(\mathring{p},p)\le r} \mathbb{E}_{A\sim p}\big[\text{cost of } z^A\big] \right\}, \tag{DRSO$_{\mathrm{W}}$}$$

where L_{W} is a Wasserstein metric defined relative to a scenario metric ℓ, and the central distribution $\mathring{p}$ is given by a sampling oracle. Recall that for a fractional first-stage decision $x \in \mathcal{P}$, we denote by $z(\mathring{p}\,; x) := \sup_{p: L_{\mathrm{W}}(\mathring{p},p)\le r} \mathbb{E}_{A\sim p}[g(x,A)]$ the expected cost incurred in the second stage, if we choose x in the first stage, and optimal fractional decisions in the second stage (and if the scenario A is drawn according to the worst possible distribution in the ambiguity set). We focus on obtaining an SAA result for the relaxation of (DRSO$_{\mathrm{W}}$) with integer first-stage decisions and (implicit) fractional second-stage decisions,

$$\min_{x\in X} \left\{ h(\mathring{p}\,; \widehat{x}) := c^\intercal x + z(\mathring{p}\,; x) \right\}. \tag{Q($\mathring{p}$)}$$

The sample average approximation (SAA) approach is the following simple, intuitive idea: draw some number N of samples from the central probability distribution $\mathring{p}$ and solve the DRS problem obtained by replacing $\mathring{p}$ with the empirical distribution $\widehat{p}$ induced by those samples. (The *empirical distribution* $\widehat{p}$ induced by samples $A_1, \ldots, A_N \in \mathcal{A}$ is defined by

$\widehat{p}_A := \frac{1}{N}\,|\{i \in [N] : A_i = A\}|$ for every $A \in \mathcal{A}$.) That is, we consider the problem

$$\min_{x \in X} \{h(\widehat{p}\,; x) := c^\intercal x + z(\widehat{p}\,; x)\}\,. \tag{Q($\widehat{p}$)}$$

We refer to (Q($\mathring{p}$)) as the *original problem*, and to (Q($\widehat{p}$)) as the *SAA problem*.

We now state the main result of this chapter, which shows that, using a moderate number of samples, we can translate approximate solutions for a collection of SAA problems into an approximate solution for the original problem, as long as we have an approximate objective-value oracle for the SAA problems.

Theorem 4.1 (see proof in Section 4.4)**.**
Given $\varepsilon > 0$, $\eta > 0$, and $\delta \in (0,1)$, there exist numbers $k = \mathsf{poly}\big(\frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$ and $N = \mathsf{poly}\Big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\eta}, \log\frac{1}{\delta}\Big)$ such that the following holds. Let $\widehat{p}^1, \ldots, \widehat{p}^k$ be empirical estimates of $\mathring{p}$, each constructed using N independent samples. Suppose that for each $i \in [k]$ we have an integer first-stage decision $\widehat{x}^i \in X$ and an estimate f^i such that

$$h\big(\widehat{p}^{\,i}\,; \widehat{x}^i\big) \le f^i \le \psi \cdot \min_{x \in X} h\big(\widehat{p}^{\,i}\,; x\big)\ ,$$

where $\psi \ge 1$. Let $j := \operatorname{argmin}_{i \in [k]} f^i$. Then with probability at least $1 - \delta$ we have

$$h\big(\mathring{p}\,; \widehat{x}^j\big) \le 4\psi(1+\varepsilon) \cdot \min_{x \in X} h(\mathring{p}\,; x) + \psi\eta\ .$$

Organization of this chapter. In Section 4.1, we provide an overview of the techniques used in the proof of Theorem 4.1; we present the proof in detail in Sections 4.2–4.5.

4.1 Overview of the techniques

Our starting point is the work of Charikar, Chekuri, and Pál [24], who proved the following result for a two-stage stochastic problem with bounded inflation factor. Part (i) shows that we can translate optimal solutions of an SAA problem constructed using a certain number of samples into near-optimal solutions of the original problem, whereas part (ii) shows that we can translate *approximate solutions* for a collection of SAA problems into an approximate solution for the original problem, as long as we also have a suitable approximate objective-value oracle for the SAA problems. Note that conditions (1) and (2) in the statement of the theorem below are analogous to assumptions (A2) and (A3).

Theorem 4.2 (see Theorems 1 and 2 in Charikar, Chekuri, and Pál [24][1]). *Consider a two-stage stochastic problem*

$$\min_{x\in\widetilde{X}} \left\{ \widetilde{h}(\widetilde{p}\,;x) := \widetilde{c}^{\intercal}x + \mathbb{E}_{A\sim\widetilde{p}}[\widetilde{g}(x,A)] \right\}, \qquad (\mathrm{Stoc}(\widetilde{p}))$$

with scenario collection $\widetilde{\mathcal{A}}$. *Suppose that: (1)* $\widetilde{X} \subseteq \mathbb{R}^m_+$ *is a finite set with* $0 \in \widetilde{X}$; *and (2) for every* $x \in \widetilde{X}$ *and* $A \in \widetilde{\mathcal{A}}$, *we have* $0 \le \widetilde{g}(0,A) - \widetilde{g}(x,A) \le \Lambda \cdot \widetilde{c}^{\intercal}x$, *for a certain* inflation factor $\Lambda \ge 1$. *Given* $\varepsilon > 0$ *and* $\delta \in (0,1)$, *we have the following two SAA results.*

(i) There exists $N = \mathsf{poly}\left(\log\left|\widetilde{X}\right|, \Lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\right)$ *such that, if we construct an empirical estimate* $\widehat{p}$ *of* $\widetilde{p}$ *using* N *independent samples, then with probability at least* $1-\delta$ *every optimal solution of the SAA problem* $(Stoc(\widehat{p}))$ *is a* $(1+\varepsilon)$*-approximate solution for the original problem* $(\mathrm{Stoc}(\widetilde{p}))$.

(ii) There exist $k = \mathsf{poly}\left(\frac{1}{\varepsilon}, \log\frac{1}{\delta}\right)$ *and* $N = \mathsf{poly}\left(\log\left|\widetilde{X}\right|, \Lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\right)$ *such that the following holds. Let* $\widehat{p}^1, \ldots, \widehat{p}^k$ *be empirical estimates of* $\widetilde{p}$, *each constructed using* N *independent samples. Suppose that for each* $i \in [k]$ *we have a* ψ_1*-approximate solution* $\widetilde{x}^i \in \widetilde{X}$ *for the SAA problem* $(Stoc(\widehat{p}^i))$ *and an estimate* $\widetilde{f}^i$ *such that* $\widetilde{h}(\widehat{p}^i\,;\widetilde{x}^i) \le \widetilde{f}^i \le \psi_2 \cdot \widetilde{h}(\widehat{p}^i\,;\widetilde{x}^i)$. *Let* $j := \operatorname{argmin}_{i\in[k]} \widetilde{f}^i$. *Then with probability at least* $1-\delta$, *we have that* $\widetilde{x}^j$ *is an* $\mathrm{O}((1+\varepsilon)\psi_1\psi_2)$*-approximate solution for the original problem* $(\mathrm{Stoc}(\widetilde{p}))$.

Plugging in the definition of L_{W}, we obtain the following equivalent definition for the expected second-stage cost incurred under a first-stage decision $x \in \mathcal{P}$:

$$z(\mathring{p}\,;x) = \max \quad \sum_{A,A'} \gamma_{A,A'} g(x,A') \qquad (\mathrm{T}(\mathring{p},x))$$

$$\text{s.t.} \quad \sum_{A'} \gamma_{A,A'} \le \mathring{p}_A \quad \forall A \in \mathcal{A} \qquad (4.1)$$

$$\sum_{A,A'} \ell(A,A')\gamma_{A,A'} \le r \qquad (4.2)$$

$$\gamma \ge 0\ . \qquad (4.3)$$

[1]Part (ii) follows from a small modification in the proof of Theorem 2 of Charikar, Chekuri, and Pál [24], which corresponds to the special case where $\psi_2 = 1$. It suffices to modify inequalities (7) and (10) by noting that, instead of $\widehat{f}^i(\overline{x}^i) \le \widehat{f}^j(\overline{x}^j)$, we only have $\widehat{f}^i(\overline{x}^i) \le \psi_2 \widehat{f}^j(\overline{x}^j)$.

Note that although we have not enforced equality in (4.1), in an optimal solution this family of constraints can always be assumed to be tight since increasing $\gamma_{A,A}$ for any $A \in \mathcal{A}$ does not violate constraints (4.2) and (4.3) (recall that $\ell(A, A) = 0$ for every $A \in \mathcal{A}$), and does not decrease the objective value.

The DRS problem (Q($\mathring{p}$)) is *not* a standard two-stage stochastic problem because constraint (4.2) couples the various scenarios, which prevents us from directly applying Theorem 4.2 to (Q($\mathring{p}$)).

The SAA result of Swamy and Shmoys [123] applies to two-stage stochastic problems with fractional first-stage and second-stage decisions, and works whenever the objective functions of the original and the SAA problems satisfy a certain "closeness-in-subgradients" property. A subgradient of $h(\mathring{p}\,;\cdot)$ at a point $x \in \mathcal{P}$ can be obtained from an optimal distribution p to the inner maximization problem in (Q($\mathring{p}$)), which in turn can be obtained via an optimal solution γ for the LP (T($\mathring{p}, x$)). This is however an exponential-size object, and utilizing this to prove closeness-in-subgradients seems quite daunting.

Our first insight, detailed in Section 4.2, is that we can decouple the scenarios by *Lagrangifying* constraint (4.2) using a dual variable $y \geq 0$; we obtain, with some additional work, the following reformulation of (Q($\mathring{p}$)):

$$\min_{x \in X, y \in [0,\tau]} \quad h(\mathring{p}\,; x, y) := c^{\intercal} x + ry + \mathbb{E}_{A \sim \mathring{p}}[g(x, y, A)] \;, \tag{R($\mathring{p}$)}$$

where $g(x, y, A) := \max_{A' \in \mathcal{A}} \{g(x, A') - y \cdot \ell(A, A')\}$; see Lemmas 4.5 and 4.6. Here, τ is the parameter from assumption (A7).

We can view problem (R($\mathring{p}$)) as a classical two-stage stochastic problem as follows: the first-stage action-set is $X \times [0, \tau]$, and the optimal second-stage cost of scenario A under the first-stage decision (x, y) is given by $g(x, y, A)$. However, as we show in Lemma 4.7, it turns out that the inflation parameter Λ for (R($\mathring{p}$)) can be as large as $\ell_{\max}/r$ (recall that $\ell_{\max}$ is an upper bound on $\max_{A,A'} \ell(A, A')$), and so applying the SAA analysis in Charikar, Chekuri, and Pál [24] and Swamy and Shmoys [123] does not yield the sample-complexity bounds that we are aiming for.

A second crucial insight, detailed in Section 4.3, is that instead of considering the true objective function $h(\mathring{p}\,; x)$ (for the original and SAA problems), we can move to a *proxy* objective function $\overline{h}(\mathring{p}\,; x) := c^{\intercal} x + z^{\text{short}}(\mathring{p}\,; x)$, where $z^{\text{short}}(\mathring{p}\,; x)$ restricts the flow γ in (T($\mathring{p}, x$)) to only use (A, A') edges with $\ell(A, A') \leq \lambda r$. We show that: (a) for any central distribution $\widetilde{p}$, we have that $\overline{h}(\widetilde{p}\,; x) + z^{\text{long}}(\widetilde{p}\,; 0)$ is pointwise-close to $h(\widetilde{p}\,; x)$, where $z^{\text{long}}(\widetilde{p}\,; 0)$ is a constant that bounds the contribution to $z(\widetilde{p}\,; x)$ from the remaining "long" edges (see Lemma 4.10); and (b) after Lagrangifying constraint (4.2), the inflation factor

of the resulting two-stage stochastic problem is at most λ (see Lemma 4.12).

The "splitting" of $z(\mathring{p}\,;x)$ into $z^{\text{short}}(\mathring{p}\,;x)$ and $z^{\text{long}}(\mathring{p}\,;0)$ is similar in spirit to the separation into low and high (cost) scenarios in Charikar, Chekuri, and Pál [24], but there are some technical differences, which lead to various complications in our setting that we discuss in the remainder of this section. To prove part (ii) of Theorem 4.2, [24] use the fact that the contribution from high-cost scenarios to the total expected cost is linear in $\mathring{p}$ to argue that with high probability the contribution from high scenarios in one of the SAA problems is at most $(1 + \mathrm{O}(\varepsilon))$ times the optimal value of the original problem. In our case, the contribution $z^{\text{long}}(\mathring{p}\,;0)$ from long edges is not linear in $\mathring{p}$, but we are able to adapt the arguments of [24], exploiting the fact that $z^{\text{long}}(\mathring{p}\,;0)$ is concave in $\mathring{p}$ (see Lemma 4.14).

The main lemma leading to the proof of Theorem 4.1 is Lemma 4.13, which shows that we can translate approximate solutions for $\min_{x\in X} \overline{h}(\widehat{p}\,;x) + C(\widehat{p})$ into approximate solutions for $\min_{x\in X} \overline{h}(\mathring{p}\,;x) + C(\mathring{p})$ for *any* nonnegative concave function $C(\cdot)$. We show in Section 4.4 that Theorem 4.1 follows from Lemma 4.13, and we prove Lemma 4.13 in Section 4.5.

One difficulty in proving the above SAA result is that we do not have much control over the $C(\widehat{p})$ term: even in the specific case of interest to us, where $C(p) = z^{\text{long}}(p\,;0)$, the contribution $C(\widehat{p})$ could be as large as $z(\widehat{p}\,;x)$, and so approximating the true-SAA problem $\min_{x\in X} h(\widehat{p}\,;x)$ need not yield approximate solutions to the proxy-SAA problem $\min_{x\in X} \overline{h}(\widehat{p}\,;x)$. The subtle issue that arises is that we would like to apply the result of [24] to the Lagrangified version of the proxy-SAA problem (since property (b) of the proxy function stated above shows that this has a small inflation factor) and thereby transfer approximation guarantees from the proxy-SAA problem to the original-proxy problem $\min_{x\in X} \overline{h}(\mathring{p}\,;x)$. However, we do not have a starting point to apply this result, since (approximately) solving the true-SAA problem $\min_{x\in X} h(\widehat{p}\,;x)$ does not yield an approximate solution for the proxy-SAA problem $\min_{x\in X} \overline{h}(\widehat{p}\,;x)$. The way around this is to realize that we seek an approximation guarantee for the original problem under the true objective function $h(\mathring{p}\,;x)$ and *not* the proxy objective function $\overline{h}(\mathring{p}\,;x)$. We adapt the arguments of [24] to work toward this end.

A final impediment is that we do *not* have an approximate value oracle for the objective function $h(\widehat{p}\,;x,y)$ of the Lagrangified true-SAA problem (or the objective function $\overline{h}(\widehat{p}\,;x,y)$ of the Lagrangified proxy-SAA problem), as the underlying recourse problem $g(x,y,A)$ turns out to be an inapproximable mixed-sign optimization problem (see Theorem 5.9-(b)). However, we show that an approximate value oracle for $h(\widehat{p}\,;x)$ suffices. (In Chapter 5, we show that such an oracle can be obtained using an algorithm for Problem (Π) with the non-standard type of approximation guarantee introduced in Definition 3.5; see

Lemmas 5.3 and 5.5).

We remark that the proxy function is used *only in the analysis*. One takeaway here is that we derive a *substantially improved sample-complexity bound by taking a slight hit in the approximation ratio* when moving from the SAA problems to the original problem. This is a novel, nuanced result regarding the effectiveness of the SAA method for two-stage DRS problems. We do not know of any other setting where one obtains drastically improved sample complexity by settling for a worse than $(1+\varepsilon)$-factor (but still $O(1)$) loss when moving from the SAA problems to the original problem. In particular, no such result is known for standard two-stage stochastic optimization problems.

Remark 4.3. In this chapter we focus on the relaxation ($\mathrm{Q}(\mathring{p})$) and (approximate) reformulations thereof that we will introduce, all of which deal with *integer* first-stage decisions. However, we state several of the preliminary lemmas (that we eventually utilize to obtain Theorem 4.1) relative to *fractional* first-stage decisions, which is more general than is needed for proving Theorem 4.1. We do so because this greater level of generality is useful in later chapters, and it does not complicate the proofs.

4.2 Reformulating ($\mathrm{Q}(\mathring{p})$) as a two-stage stochastic problem

We first reformulate the DRS problem ($\mathrm{Q}(\mathring{p})$) as a classical two-stage stochastic problem, thus making it more amenable to utilize the SAA machinery developed for two-stage stochastic problems. Recall that for every $(x,y,A) \in X \times \mathbb{R}_+ \times \mathcal{A}$ we have $g(x,y,A) := \max_{A' \in \mathcal{A}} \{g(x,A') - y \cdot \ell(A,A')\}$, and that $\ell_{\max}$ is an upper bound on the pairwise scenario distances $\{\ell(A,A')\}$.

Lemma 4.4 (combination of Lemmas 4.6 and 4.7)**.** *The relaxed DRS problem* ($\mathrm{Q}(\mathring{p})$) *is equivalent to the problem*

$$\min_{x \in X, y \in [0,\tau]} \{h(\mathring{p}\,; x, y) := c^\intercal x + ry + \mathbb{E}_{A \sim \mathring{p}}[g(x,y,A)]\} \ . \tag{R($\mathring{p}$)}$$

Furthermore, for every $(x,y,A) \in X \times [0,\tau] \times \mathcal{A}$ *we have* $g(x,y,A) \geq 0$ *and*

$$0 \leq g(0,0,A) - g(x,y,A) \leq \max\left\{\lambda, \frac{\ell_{\max}}{r}\right\} \cdot (c^\intercal x + ry) \ .$$

We can view problem (R($\mathring{p}$)) as a classical two-stage stochastic problem as follows: the first-stage action-set is $X \times [0, \tau]$, and the optimal second-stage cost of scenario A under the first-stage decision (x, y) is given by $g(x, y, A)$.

In the remainder of this section, we explain how to obtain Lemma 4.4. We start by using LP duality to obtain an alternative way of expressing the objective function of (Q($\mathring{p}$)), thus decoupling the scenarios.

Lemma 4.5. *Let $\widetilde{p}$ be a probability distribution over $\mathcal{A}$, and let $x \in \mathcal{P}$ be a fractional first-stage decision. Then*

$$z(\widetilde{p}\,; x) = \min_{y \geq 0} \left\{ ry + \mathbb{E}_{A \sim \widetilde{p}}[g(x, y, A)] \right\} \ .$$

Proof. Note that the feasible region of the LP (T($\widetilde{p}, x$)) is bounded, since constraints (4.1) and (4.3) imply that in a feasible solution all variables are in the range $[0, 1]$. Furthermore, $\gamma = 0$ is a feasible solution. Therefore this LP has an optimal solution. Taking its dual, using dual variables $\{\mu_A\}_{A \in \mathcal{A}}$ and y for constraints (4.1) and (4.2) respectively, we obtain

$$\begin{aligned} z(\widetilde{p}\,; x) = \min \quad & ry + \sum_A \widetilde{p}_A \mu_A \\ \text{s.t.} \quad & \mu_A \geq g(x, A') - y \cdot \ell(A, A') \qquad \forall A, A' \in \mathcal{A} \\ & y \geq 0 \ . \end{aligned}$$

Note that for a fixed value of y, the best choice of μ is obtained by setting

$$\mu_A = \max_{A' \in \mathcal{A}} \{ g(x, A') - y \cdot \ell(A, A') \} = g(x, y, A)$$

for every scenario $A \in \mathcal{A}$. It follows that

$$z(\widetilde{p}\,; x) = \min_{y \geq 0} \left\{ ry + \sum_{A \in \mathcal{A}} \widetilde{p}_A g(x, y, A) \right\} = \min_{y \geq 0} \left\{ ry + \mathbb{E}_{A \sim \widetilde{p}}[g(x, y, A)] \right\} \ . \qquad \square$$

Given Lemma 4.5, we can reformulate (Q($\mathring{p}$)) as

$$\min_{x \in X, y \geq 0} \left\{ h(\mathring{p}\,; x, y) := c^\intercal x + ry + \mathbb{E}_{A \sim \mathring{p}}[g(x, y, A)] \right\} \ .$$

We can exploit assumption (A7) to show that we may limit y to the range $[0, \tau]$, thus obtaining the formulation (R($\mathring{p}$)).

Lemma 4.6. *Let $\widetilde{p}$ be a probability distribution over $\mathcal{A}$, and let $x \in \mathcal{P}$ be a fractional first-stage decision. Then*

$$h(\widetilde{p}\,; x) = \min_{y \in [0,\tau]} h(\widetilde{p}\,; x, y) \ .$$

Proof. Let $y \geq \tau$, and consider a scenario $A \in \mathcal{A}$. For every scenario $A' \in \mathcal{A}$ such that $\ell(A, A') > 0$, we have

$$\begin{aligned} g(x, A) - y \cdot \ell(A, A) &= g(x, A) \\ &\geq g(x, A') - \tau \cdot \ell(A, A') \\ &\geq g(x, A') - y \cdot \ell(A, A') \ , \end{aligned}$$

where the first inequality follows from assumption (A7). This implies that there is a maximizer A' of $g(x, A') - y \cdot \ell(A, A')$ with $\ell(A, A') = 0$, and so $g(x, y, A) = \max_{A' \in \mathcal{A} : \ell(A,A')=0} g(x, A')$. Since this holds for every $A \in \mathcal{A}$, we obtain $\mathbb{E}_{A \sim \hat{p}}[g(x, y, A)] = \mathbb{E}_{A \sim \hat{p}}[g(x, \tau, A)]$. Since $ry \geq r\tau$, we obtain $h(\widetilde{p}\,; x, y) \geq h(\widetilde{p}\,; x, \tau)$. The result then follows from Lemma 4.5. □

It is easy to see that the second-stage cost $g(x, y, A)$ is nonnegative for every $(x, y, A) \in \mathcal{P} \times \mathbb{R}_+ \times \mathcal{A}$, since $g(x, y, A) \geq g(x, A) - y \cdot \ell(A, A) = g(x, A) \geq 0$. To conclude, we bound the inflation factor of the second-stage costs.

Lemma 4.7. *Let $x \in \mathcal{P}$, $y \geq 0$, and $A \in \mathcal{A}$. Then*

$$0 \leq g(0, 0, A) - g(x, y, A) \leq \max\left\{\lambda, \frac{\ell_{\max}}{r}\right\} \cdot (c^\intercal x + ry) \ .$$

Proof. The first inequality follows because for every scenario $A' \in \mathcal{A}$ we have

$$g(x, A') - y \cdot \ell(A, A') \leq g(0, A') - 0 \cdot \ell(A, A') \ ,$$

since $g(x, A') \leq g(0, A')$ by assumption (A2) and since $\ell(A, A') \geq 0$.

We now prove the second inequality. Let $\overline{A} \in \mathcal{A}$ such that $g(0, 0, A) = g\big(0, \overline{A}\big)$ (equivalently, let $\overline{A} := \operatorname{argmax}_{A' \in \mathcal{A}} g(0, A')$). Then we have

$$\begin{aligned} g(0, 0, A) - g(x, y, A) &\leq g\big(0, \overline{A}\big) - \big(g\big(x, \overline{A}\big) - y \cdot \ell\big(A, \overline{A}\big)\big) \\ &\leq \lambda \cdot c^\intercal x + y \cdot \ell_{\max} \\ &\leq \max\left\{\lambda, \tfrac{\ell_{\max}}{r}\right\} \cdot (c^\intercal x + ry) \ , \end{aligned}$$

where the second step follows from assumption (A3) and the fact that $\ell\big(A, \overline{A}\big) \leq \ell_{\max}$. □

4.3 Reducing the inflation factor

Given Lemma 4.4, and by suitably discretizing the y-range $[0,\tau]$, one can use Theorem 4.2 (setting the parameter Λ to $\max\{\lambda, \frac{\ell_{\max}}{r}\}$) to show that: if we construct SAA problems $\min_{x\in X} h(\widehat{p}\,;x) \equiv \min_{x\in X, y\in[0,\tau]} h(\widehat{p}\,;x,y)$ using $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{\ell_{\max}}{r}, \frac{1}{\varepsilon}\big)$ samples, and can approximately evaluate the SAA objective value $h(\widehat{p}\,;x,y)$ at any given point, then we can translate ψ-approximate solutions for the SAA problems into an $O(\psi+\varepsilon)$-approximate solution for (Q($\mathring{p}$)), with high probability.

But there are various issues due to which this result does not quite suit our purposes. First, $\frac{\ell_{\max}}{r}$ could be rather large, and is not $\mathsf{poly}(\mathcal{I},\lambda)$.[2] Second, it seems difficult to compute the SAA objective value $h(\widehat{p}\,;x,y)$ at any given point (x,y), or even approximate it. This difficulty arises because problem (Π) (of computing $g(x,y,A)$) encompasses the k-max-min problem in two-stage robust optimization, which is computationally for various underlying combinatorial-optimization problems (see the discussion in Section 2.1). Moreover, the mixed-sign objective in (Π) makes it hard to even approximate it (see Theorem 5.9-(b)).

We need various ideas to circumvent these issues. The main result in this section is an approximate reformulation of (Q($\mathring{p}$)) as a classical two-stage stochastic optimization problem with inflation factor bounded by λ. This shows that we can eliminate the dependence on $\frac{\ell_{\max}}{r}$ altogether, at the expense of a slight deterioration in the approximation ratio when moving from the SAA problems to the original problem.

Examining the proof of Lemma 4.7, we notice that the $\frac{\ell_{\max}}{r}$ term in the inflation factor of (R($\mathring{p}$)) arises because when considering the problem $\max_{A'\in\mathcal{A}}\{g(x,A') - y\cdot\ell(A,A')\}$ for a given first-stage decision $(x,y)\in\mathcal{P}\times[0,\tau]$ and a given scenario $A\in\mathcal{A}$, we may encounter a scenario A' such that $\ell(A,A')$ is as large as $\ell_{\max}$. To eliminate this possibility and reduce the sample complexity to $\mathsf{poly}(\mathcal{I},\lambda)$, we define a distance threshold $M := \lambda r$, and work toward suitably modifying the objective function $h(\mathring{p}\,;x,y)$ of (R($\mathring{p}$)) to enforce that we never encounter pairs of scenarios (A,A') with $\ell(A,A') > M$. We call such pairs *long* edges, and the pairs with $\ell(A,A')\le M$ *short* edges. For every $(x,y,A)\in X\times\mathbb{R}_+\times\mathcal{A}$, let

$$\overline{g}(x,y,A) := \max_{A'\in\mathcal{A}:\ell(A,A')\le M}\{g(x,A') - y\cdot\ell(A,A')\}\ .$$

Let $z^{\text{short}}(\mathring{p}\,;x)$ be obtained from $z(\mathring{p}\,;x)$ by restricting γ to only send flow on pairs (A,A') with $\ell(A,A')\le M$ (see the precise definition in (4.4)).

[2] The problem persists even if we utilize the closeness-in-subgradients machinery by Swamy and Shmoys [123] to the further relaxation of (R($\mathring{p}$)) with fractional first-stage decisions. Computing sufficiently accurate subgradients would involve estimating $\mathbb{E}_{A\sim\mathring{p}}[\ell(A,\pi(x,y,A))]$ to within an εr term, where $\pi(x,y,A) := \operatorname{argmax}_{A'\in\mathcal{A}}\{g(x,A') - y\cdot\ell(A,A')\}$, which requires $\Omega\big(\frac{\ell_{\max}}{\varepsilon r}\big)$ samples.

Lemma 4.8 (combination of Lemmas 4.10 and 4.12)**.** *Consider the* proxy problem

$$\min_{x \in X} \left\{ \overline{h}(\mathring{p}\,; x) := c^\intercal x + z^{\text{short}}(\mathring{p}\,; x) \right\} . \tag{$\overline{\text{Q}}(\mathring{p})$}$$

Its objective function serves as a proxy for the objective function of ($\text{Q}(\mathring{p})$)*: there exists a constant* $C(\mathring{p})$ *such that for every* $x \in X$ *we have*

$$h(\mathring{p}\,; x) \le \overline{h}(\mathring{p}\,; x) + C(\mathring{p}) \le 2h(\mathring{p}\,; x) .$$

Furthermore, ($\overline{\text{Q}}(\mathring{p})$) *is equivalent to the two-stage stochastic problem*

$$\min_{x \in X, y \in [0,\tau]} \left\{ \overline{h}(\mathring{p}\,; x, y) := c^\intercal x + ry + \mathbb{E}_{A \sim \mathring{p}}[\overline{g}(x, y, A)] \right\} . \tag{$\overline{\text{R}}(\mathring{p})$}$$

For every $(x, y, A) \in X \times [0, \tau] \times \mathcal{A}$ *we have* $\overline{g}(x, y, A) \ge 0$ *and*

$$0 \le \overline{g}(0, 0, A) - \overline{g}(x, y, A) \le \lambda \cdot (c^\intercal x + ry) .$$

We now discuss how to obtain Lemma 4.8. The following lemma gives a bound on the amount of flow that can be sent on long edges by a flow from the central distribution to another distribution in the ambiguity set.

Lemma 4.9. *Let* $\widetilde{p}$ *be a probability distribution over* $\mathcal{A}$*,* $x \in \mathcal{P}$ *be a fractional first-stage decision, and* γ *be a feasible solution for the LP* $(\text{T}(\widetilde{p}, x))$*. Then we have*

$$\sum_{A, A' : \ell(A, A') > M} \gamma_{A,A'} \le \frac{r}{M} = \frac{1}{\lambda} .$$

Proof. By constraint (4.2), we have

$$r \ge \sum_{A, A'} \gamma_{A,A'} \ell(A, A') \ge M \cdot \sum_{A, A' : \ell(A, A') > M} \gamma_{A,A'} .$$

Plugging in the definition of M yields the result. □

Motivated by Lemma 4.9, we "decompose" $z(\mathring{p}\,; x)$ into $z^{\text{short}}(\mathring{p}\,; x)$ and $z^{\text{long}}(\mathring{p}\,; x)$, which are upper bounds on the contributions to $z(\mathring{p}\,; x)$ from the short and long edges respectively.

We define $z^{\text{short}}(\mathring{p}\,;x)$ and $z^{\text{long}}(\mathring{p}\,;x)$ as follows:

$$z^{\text{short}}(\mathring{p}\,;x) := \max\left\{ \sum_{A,A'} \gamma_{A,A'} g(x,A') \;\middle|\; \begin{array}{l} \gamma \text{ is feasible for } (\mathrm{T}(\mathring{p},x)), \\ \gamma_{A,A'} = 0 \;\text{ if }\; \ell(A,A') > M \end{array} \right\} , \tag{4.4}$$

$$z^{\text{long}}(\mathring{p}\,;x) := \max\left\{ \sum_{A,A'} \gamma_{A,A'} g(x,A') \;\middle|\; \gamma \text{ is feasible for } (\mathrm{T}(\mathring{p},x)), \sum_{A,A'} \gamma_{A,A'} \le \frac{1}{\lambda} \right\} .$$

We show that decomposing $z(\mathring{p}\,;x)$ into the maximum contributions from short and long edges as discussed above *and* replacing the contribution from the long edges with $z^{\text{long}}(\mathring{p}\,;0)$ yields a function that approximates the objective function of $(\mathrm{Q}(\mathring{p}))$ to within a factor of 2.

Lemma 4.10. *Let $\widetilde{p}$ be a probability distribution over $\mathcal{A}$, and let $x \in \mathcal{P}$ be a fractional first-stage decision. Then*

$$h(\widetilde{p}\,;x) \le c^{\intercal}x + z^{\text{short}}(\widetilde{p}\,;x) + z^{\text{long}}(\widetilde{p}\,;0) \le 2h(\widetilde{p}\,;x) .$$

Proof. We claim that

$$z(\widetilde{p}\,;x) \le z^{\text{short}}(\widetilde{p}\,;x) + z^{\text{long}}(\widetilde{p}\,;x) \le 2z(\widetilde{p}\,;x) \tag{4.5}$$

and

$$z^{\text{long}}(\widetilde{p}\,;x) \le z^{\text{long}}(\widetilde{p}\,;0) \le z^{\text{long}}(\widetilde{p}\,;x) + c^{\intercal}x . \tag{4.6}$$

Recalling that $h(\widetilde{p}\,;x) = c^{\intercal}x + z(\widetilde{p}\,;x)$, and using the first inequalities in (4.5) and (4.6), we obtain

$$h(\widetilde{p}\,;x) \le c^{\intercal}x + z^{\text{short}}(\widetilde{p}\,;x) + z^{\text{long}}(\widetilde{p}\,;x) \le c^{\intercal}x + z^{\text{short}}(\widetilde{p}\,;x) + z^{\text{long}}(\widetilde{p}\,;0) ,$$

proving the first part of the lemma. Using the second inequalities in (4.6) and (4.5), we obtain

$$c^{\intercal}x + z^{\text{short}}(\widetilde{p}\,;x) + z^{\text{long}}(\widetilde{p}\,;0) \le 2c^{\intercal}x + z^{\text{short}}(\mathring{p}\,;x) + z^{\text{long}}(\mathring{p}\,;x) \le 2h(\widetilde{p}\,;x) ,$$

proving the second part of the lemma.

It remains to prove the two claims. We start by proving (4.5). Note that the LP used in the definition of $z(\widetilde{p}\,;x)$ is a relaxation of the LPs used in the definitions of $z^{\text{short}}(\widetilde{p}\,;x)$ and $z^{\text{long}}(\widetilde{p}\,;x)$. It follows that $z^{\text{short}}(\widetilde{p}\,;x) \le z(\widetilde{p}\,;x)$ and $z^{\text{long}}(\widetilde{p}\,;x) \le z(\widetilde{p}\,;x)$; adding these two inequalities yields the second inequality in (4.5). To prove the first inequality in (4.5), let γ^* be an optimal solution for the LP $(\mathrm{T}(\widetilde{p},x))$. We decompose γ^* into a flow supported on short edges and one supported on long edges. That is, we write $\gamma^* = \gamma^{\text{short}} + \gamma^{\text{long}}$, where γ^{short} only sends flow on short edges, and γ^{long} only sends flow on long edges. Note that γ^{short} is feasible for the LP defining $z^{\text{short}}(\widetilde{p}\,;x)$, and γ^{long} is feasible for the LP defining $z^{\text{long}}(\widetilde{p}\,;x)$ (by Lemma 4.9). It follows that $z^{\text{short}}(\widetilde{p}\,;x) \ge \sum_{A,A':\ell(A,A')\le M} \gamma^*_{A,A'} g(x,A')$ and $z^{\text{long}}(\widetilde{p}\,;x) \ge \sum_{A,A':\ell(A,A')>M} \gamma^*_{A,A'} g(x,A')$. Summing these two inequalities yields

$$z^{\text{short}}(\widetilde{p}\,;x) + z^{\text{long}}(\widetilde{p}\,;x) \ge \sum_{A,A'} \gamma^*_{A,A'} g(x,A') = z(\widetilde{p}\,;x)\ .$$

Finally, we prove claim (4.6). Note that $z^{\text{long}}(\widetilde{p}\,;0)$ and $z^{\text{long}}(\widetilde{p}\,;x)$ are defined as the optimal values of two LPs with the same feasible region. Let γ^0 and γ^x be optimal solutions for these two LPs respectively. We have

$$z^{\text{long}}(\widetilde{p}\,;0) \ge \sum_{A,A'} \gamma^x_{A,A'} g(0,A') \ge \sum_{A,A'} \gamma^x_{A,A'} g(x,A') = z^{\text{long}}(\widetilde{p}\,;x)\ ,$$

where the second inequality follows from Assumption (A2). This proves the first part of (4.6). Furthermore, we have

$$z^{\text{long}}(\widetilde{p}\,;x) \ge \sum_{A,A'} \gamma^0_{A,A'} g(x,A') \ge \sum_{A,A'} \gamma^0_{A,A'} (g(0,A') - \lambda c^\intercal x) \ge z^{\text{long}}(\widetilde{p}\,;0) - c^\intercal x\ ,$$

where the second inequality follows from assumption (A3), and the third one follows from the fact that $\sum_{A,A'} \gamma^0_{A,A'} \le \frac{1}{\lambda}$. This yields the second part of (4.6). □

Remark 4.11. By increasing the threshold M that demarcates short and long edges to $\lambda r/\varepsilon$, one could strengthen (4.6) to $z^{\text{long}}(\widetilde{p}\,;x) \le z^{\text{long}}(\widetilde{p}\,;0) \le z^{\text{long}}(\widetilde{p}\,;x) + \varepsilon c^\intercal x$, and so the contribution $z^{\text{long}}(\widetilde{p}\,;x)$ from long edges would be within an $\varepsilon c^\intercal x$ term of the constant $z^{\text{long}}(\widetilde{p}\,;0)$. This is analogous to what happens with the decomposition of the scenario collection into "low-cost" and "high-cost" scenarios in the proof of the SAA result of Charikar, Chekuri, and Pál [24]. Note however that this change in the threshold M does not avoid the factor-2 loss in Lemma 4.10.

Given Lemma 4.10, we focus on the *proxy problem*

$$\min_{x\in X} \left\{\overline{h}(\mathring{p}\,;x) := c^\intercal x + z^{\text{short}}(\mathring{p}\,;x)\right\} \ . \tag{$\overline{\text{Q}}(\mathring{p})$}$$

Recall that $\overline{g}(x,y,A) := \max_{A'\in\mathcal{A}:\ell(A,A')\le M}\{g(x,A') - y\cdot\ell(A,A')\}$. Using LP duality, following the same arguments as in Lemmas 4.5 and 4.6, we have

$$\overline{h}(\mathring{p}\,;x) = c^\intercal x + \min_{y\in[0,\tau]}\{ry + \mathbb{E}_{A\sim\mathring{p}}[\overline{g}(x,y,A)]\} \tag{4.7}$$

for every fractional first-stage decision $x\in\mathcal{P}$, and so we can reformulate ($\overline{\text{Q}}(\mathring{p})$) as follows:

$$\min_{x\in X, y\in[0,\tau]} \left\{\overline{h}(\mathring{p}\,;x,y) := c^\intercal x + ry + \mathbb{E}_{A\sim\mathring{p}}[\overline{g}(x,y,A)]\right\} \ . \tag{$\overline{\text{R}}(\mathring{p})$}$$

Note that, as was the case for ($\text{R}(\mathring{p})$), we can view ($\overline{\text{R}}(\mathring{p})$) as a two-stage stochastic problem, with first-stage decision set $X\times[0,\tau]$ and optimal second-stage costs $\overline{g}(x,y,A)$ for every $(x,y,A)\in X\times[0,\tau]\times\mathcal{A}$. It is easy to see that the second-stage cost $\overline{g}(x,y,A)$ is nonnegative for every $(x,y,A)\in\mathcal{P}\times\mathbb{R}_+\times\mathcal{A}$, since $\overline{g}(x,y,A)\ge g(x,A) - y\cdot\ell(A,A) = g(x,A)\ge 0$.

The chief advantage of, and reason for, moving from ($\text{R}(\mathring{p})$) to ($\overline{\text{R}}(\mathring{p})$) is that, as we now show, we reduce the inflation factor to λ, which is the inflation factor of the discrete DRS problem we started with.

Lemma 4.12. *Let $x\in\mathcal{P}$, $y\ge 0$, and $A\in\mathcal{A}$. Then*

$$0\le \overline{g}(0,0,A) - \overline{g}(x,y,A) \le \lambda\cdot(c^\intercal x + ry) \ .$$

Proof. We mimic the proof of Lemma 4.7. The first inequality follows because for every scenario $A'\in\mathcal{A}$ we have

$$g(x,A') - y\cdot\ell(A,A') \le g(0,A') - 0\cdot\ell(A,A') \ ,$$

since $g(x,A')\le g(0,A')$ by assumption (A2) and since $\ell(A,A')\ge 0$.

We now prove the second inequality. Let $\overline{A} := \operatorname{argmax}_{A'\in\mathcal{A}:\ell(A,A')\le M} g(0,A')$. Then

$$\begin{aligned}\overline{g}(0,0,A) - \overline{g}(x,y,A) &\le \overline{g}\big(0,\overline{A}\big) - \big(g\big(x,\overline{A}\big) - y\cdot\ell\big(A,\overline{A}\big)\big)\\ &\le \lambda\cdot c^\intercal x + y\cdot M\\ &= \lambda\cdot(c^\intercal x + ry) \ ,\end{aligned}$$

where the second inequality follows from (A3) and the fact that $\ell\big(A,\overline{A}\big)\le M$. □

4.4 Main lemma and proof of the SAA theorem

After suitably discretizing the y-range $[0, \tau]$ and applying the SAA result from Charikar, Chekuri, and Pál [24] (Theorem 4.2), we can show that approximate solutions for the SAA problem $(\overline{\mathrm{R}}(\widehat{p}))$ can be translated into approximate solutions for $(\overline{\mathrm{R}}(\mathring{p}))$, with an improved $\mathsf{poly}(\mathcal{I}, \lambda)$ sample complexity. Given the equivalence between $(\overline{\mathrm{R}}(\widehat{p}))$ and $(\overline{\mathrm{Q}}(\widehat{p}))$ (and between $(\overline{\mathrm{R}}(\mathring{p}))$ and $(\overline{\mathrm{Q}}(\mathring{p}))$) shown by (4.7), we can use this to translate approximate solutions for $(\overline{\mathrm{Q}}(\widehat{p}))$ into approximate solutions for $(\overline{\mathrm{Q}}(\mathring{p}))$. Since for every distribution $\widetilde{p}$ over $\mathcal{A}$ we have that $h(\widetilde{p}\,; x)$ (which is the objective function of $(\mathrm{Q}(\widetilde{p}))$) and $\overline{h}(\widetilde{p}\,; x) + z^{\mathrm{long}}(\widetilde{p}\,; 0)$ (which equals the objective function of $(\overline{\mathrm{Q}}(\widetilde{p}))$ plus a constant) are pointwise close by Lemma 4.10, this seems to indicate that we can also translate approximate solutions for $(\mathrm{Q}(\widehat{p}))$ into approximate solutions for $(\mathrm{Q}(\mathring{p}))$.

However, two sources of difficulty remain. First, the fact that $h(\widehat{p}\,; x)$ and $\overline{h}(\widehat{p}\,; x) + z^{\mathrm{long}}(\widehat{p}\,; 0)$ are pointwise close does not mean that approximate solutions for $(\mathrm{Q}(\widehat{p}))$ translate into approximate solutions for $(\overline{\mathrm{Q}}(\widehat{p}))$, since the term $z^{\mathrm{long}}(\widehat{p}\,; 0)$ could be significant compared to $h(\widehat{p}\,; x)$, as indicated by the factor-2 loss in Lemma 4.10.

Second, note that the SAA result for $(\overline{\mathrm{R}}(\mathring{p}))$ obtained via Theorem 4.2 involves using estimates for the objective function $\overline{h}(\widehat{p}\,; x, y)$ of $(\overline{\mathrm{R}}(\widehat{p}))$, which we do not have. However, we will show in Chapter 5 that if we have an approximation algorithm for problem (Π) that gives the type of guarantee stated in Definition 3.5, then given an integer first-stage decision $x \in X$ one can obtain an approximate solution for $(\mathrm{T}(\widehat{p}, x))$ (see Lemma 5.5), which can be used to estimate $h(\widehat{p}\,; x)$. While this is not the same as a value oracle for $\overline{h}(\widehat{p}\,; x, y)$, we show that this nevertheless suffices.

We now state the main lemma that we use to prove Theorem 4.1.

Lemma 4.13 (see proof in Section 4.5)**.** *Let $\varepsilon > 0$, $\eta > 0$, and $\delta \in (0, 1)$. Let $C(p)$ be a nonnegative concave function defined over the set of probability distributions over $\mathcal{A}$. There exist $k = \mathsf{poly}\big(\frac{1}{\varepsilon}, \log \frac{1}{\delta}\big)$ and $N = \mathsf{poly}\Big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log \frac{1}{\eta}, \log \frac{1}{\delta}\Big)$ such that the following holds. Let $\widehat{p}^1, \ldots, \widehat{p}^k$ be empirical estimates of $\mathring{p}$, each constructed using N independent samples. Suppose that for each $i \in [k]$ we have an integer first-stage decision $\widehat{x}^i \in X$ and an estimate f^i such that*

$$\overline{h}\big(\widehat{p}^{\,i}\,; \widehat{x}^i\big) + C\big(\widehat{p}^{\,i}\big) \le f^i \le \psi \cdot \min_{x \in X} \big\{\overline{h}\big(\widehat{p}^{\,i}\,; x\big) + C\big(\widehat{p}^{\,i}\big)\big\} \ ,$$

where $\psi \ge 1$. Let $j := \operatorname{argmin}_{i \in [k]} f^i$. Then with probability at least $1 - \delta$ we have

$$\overline{h}\big(\mathring{p}\,; \widehat{x}^j\big) + C(\mathring{p}) \le \psi(1 + \varepsilon) \cdot \min_{x \in X} \big\{\overline{h}(\mathring{p}\,; x) + C(\mathring{p})\big\} + \psi\eta \ .$$

We proceed to show that Theorem 4.1 follows from Lemma 4.13, but first we need the following preliminary lemma.

Lemma 4.14. *The function $p \mapsto z^{\text{long}}(p\,;0)$, defined over the probability distributions over $\mathcal{A}$, is nonnegative and concave.*

Proof. Let p be a probability distribution over $\mathcal{A}$. Note that all feasible solutions γ for the LP defining $z^{\text{long}}(p\,;0)$ are nonnegative by constraint (4.3), and that the coefficients of the objective function are nonnegative. It follows that $z^{\text{long}}(p\,;0)$ is nonnegative.

To prove that the function is concave, consider any two distributions p and q, and let $\widetilde{p} = \theta \cdot p + (1-\theta) \cdot q$, where $\theta \in [0,1]$. Let γ^p and γ^q be optimal solutions for the LPs defining $z^{\text{long}}(p\,;0)$ and $z^{\text{long}}(q\,;0)$ respectively. Then $\gamma^{\widetilde{p}} := \theta \cdot \gamma^p + (1-\theta) \cdot \gamma^q$ is a feasible solution for the LP defining $z^{\text{long}}(\widetilde{p}\,;0)$, which implies that

$$\begin{aligned} z^{\text{long}}(\widetilde{p}\,;0) &\geq \sum_{A,A'} \gamma^{\widetilde{p}}_{A,A'} g(0,A') \\ &= \theta \cdot \sum_{A,A'} \gamma^{p}_{A,A'} g(0,A') + (1-\theta) \cdot \sum_{A,A'} \gamma^{q}_{A,A'} g(0,A') \\ &= \theta \cdot z^{\text{long}}(p\,;0) + (1-\theta) \cdot z^{\text{long}}(q\,;0)\ . \end{aligned}$$

□

Proof of Theorem 4.1. Let the number k of SAA problems and the number N of samples for each such problem be given by Lemma 4.13, with parameters $\left(\varepsilon, \frac{\eta}{2}, \delta\right)$. We show that for every $i \in [k]$, we have that $(\widehat{x}^i, 2f^i)$ satisfies the conditions of Lemma 4.13 taking the parameter ψ in the lemma statement to be 2ψ, and setting $C(p) := z^{\text{long}}(p\,;0)$ (which is a nonnegative concave function by Lemma 4.14). That is, we show that

$$\overline{h}\big(\widehat{p}^i\,;\widehat{x}^i\big) + z^{\text{long}}(\widehat{p}^i\,;0) \leq 2f^i \leq 2\psi \cdot \min_{x \in X} \left\{\overline{h}\big(\widehat{p}^i\,;x\big) + z^{\text{long}}(\widehat{p}^i\,;0)\right\}\ .$$

To see this, note that

$$\begin{aligned} \overline{h}\big(\widehat{p}^i\,;\widehat{x}^i\big) + z^{\text{long}}(\widehat{p}^i\,;0) &\leq 2h\big(\widehat{p}^i\,;\widehat{x}^i\big) \\ &\leq 2f^i \\ &\leq 2\psi \cdot \min_{x \in X} h\big(\widehat{p}^i\,;x\big) \\ &\leq 2\psi \cdot \min_{x \in X} \left\{\overline{h}\big(\widehat{p}^i\,;x\big) + z^{\text{long}}(\widehat{p}^i\,;0)\right\}\ . \end{aligned}$$

The first and the last inequalities follow from Lemma 4.10. The second and the third inequalities use the guarantee on $(\widehat{x}^i, f^i)$ given in the theorem statement.

Furthermore, since $j \in \operatorname{argmin}_{i\in[k]} f^i$, we also have $j \in \operatorname{argmin}_{i\in[k]} \{2f^i\}$. Applying Lemma 4.13, we have that with probability $1-\delta$,

$$\overline{h}\big(\mathring{p}\,;\widehat{x}^j\big) + z^{\mathrm{long}}(\mathring{p}\,;0) \le 2\psi(1+\varepsilon)\cdot \min_{x\in X}\big\{\overline{h}(\mathring{p}\,;x) + z^{\mathrm{long}}(\mathring{p}\,;0)\big\} + \psi\eta\ .$$

Using Lemma 4.10 again, we obtain that the left side is greater than or equal to $h(\mathring{p}\,;\widehat{x}^j)$, and the right side is less than or equal to $4\psi(1+\varepsilon)\cdot\min_{x\in X} h(\mathring{p}\,;x) + \psi\eta$, and the theorem follows. □

Remark 4.15. Note that by applying Lemma 4.13 with $C(p) := 0$ for every p, we can also convert approximate solutions for $\min_{x\in X}\overline{h}(\widehat{p}\,;x)$ (for $\widehat{p} = \widehat{p}^1,\ldots,\widehat{p}^k$) into an approximate solution for $\min_{x\in X}\overline{h}(\mathring{p}\,;x)$. This can be used, along with Lemma 4.10, to obtain a variant of Theorem 4.1 showing that ψ-approximate solutions for $\min_{x\in X}\overline{h}(\widehat{p}\,;x)$ translate into a solution for $\min_{x\in X} h(\mathring{p}\,;x)$ that is approximately optimal, within a $2\psi(1+\varepsilon)$-factor and a $\psi\eta$ additive term. With some additional work, this can be used to halve the approximation factor in our main result for DRS optimization under a Wasserstein ball (Theorem 3.6), if we require an approximation algorithm for computing $\overline{g}(x,y,A)$ (rather than for the problem (Π) of computing $g(x,y,A)$).

4.5 Proof of Lemma 4.13

4.5.1 Overview

We discretize the y-range $[0,\tau]$ suitably to obtain a set $Y \subseteq [0,\tau]$ such that for every integer first-stage decision $x \in X$ and any distribution $\widetilde{p}$, there exists $y \in Y$ such that $\overline{h}(\widetilde{p}\,;x,y)$ is close to $\overline{h}(\widetilde{p}\,;x)$ (see Lemma 4.16). This allows translating approximate solutions for $\min_{x\in X}\overline{h}(\widetilde{p}\,;x)$ into approximate solutions for $\min_{x\in X,y\in Y}\overline{h}(\widetilde{p}\,;x,y)$ and vice versa.

Let $\widehat{p}$ denote a generic empirical estimate of $\mathring{p}$ (which could be any of $\widehat{p}^1,\ldots,\widehat{p}^k$). The arguments in Charikar, Chekuri, and Pál [24] show that approximate solutions for $\min_{x\in X,y\in Y}\overline{h}(\widehat{p}\,;x,y)$ (for $\widehat{p} = \widehat{p}^1,\ldots,\widehat{p}^k$) can be used to obtain an approximate solution for $\min_{x\in X,y\in Y}\overline{h}(\mathring{p}\,;x,y)$ (given a suitable value oracle for $\overline{h}(\widehat{p}\,;x,y)$). Recall that $\overline{h}(p\,;x,y) := c^\intercal x + ry + \mathbb{E}_{A\sim p}[\overline{g}(x,y,A)]$. The proof in [24] proceeds by decomposing $\mathbb{E}_{A\sim p}[\overline{g}(x,y,A)]$ into two terms, $\mathbb{E}^{\mathrm{low}}_{A\sim p}\big[\overline{g}(x,y,A)\big]$ and $\mathbb{E}^{\mathrm{high}}_{A\sim p}\big[\overline{g}(x,y,A)\big]$, which are the contributions from "low-cost" and "high-cost" scenarios respectively. For the low scenarios, Hoeffding's inequality (Theorem 2.3) implies that $\mathbb{E}^{\mathrm{low}}_{A\sim\widehat{p}}\big[\overline{g}(\cdot,\cdot,A)\big]$ and $\mathbb{E}^{\mathrm{low}}_{A\sim\mathring{p}}\big[\overline{g}(\cdot,\cdot,A)\big]$ are pointwise close (see Lemma 4.17).

The contribution to $\overline{h}(p\,;x,y)$ from high scenarios, however, could be quite different in the SAA and original problems, although in both problems, this contribution is essentially independent of (x,y) since the definition of high scenarios ensures that they occur with small probability (see Lemmas 4.18 and 4.19).

Since $\mathbb{E}^{\text{high}}_{A\sim p}\big[\overline{g}(0,0,A)\big]$ is *linear in* p, the expected value of $\mathbb{E}^{\text{high}}_{A\sim \widehat{p}}\big[\overline{g}(0,0,A)\big]$ (over the random selection of the scenarios used to construct $\widehat{p}$), is precisely $\mathbb{E}^{\text{high}}_{A\sim \mathring{p}}[\overline{g}(0,0,A)]$. Thus, we can use Markov's inequality (Theorem 2.1) to show that for at least one of our multiple SAA problems (say, the one with empirical distribution $\widehat{p}^{\,t}$), we have that $\mathbb{E}^{\text{high}}_{A\sim \widehat{p}^{\,t}}[\overline{g}(0,0,A)]$ is not much larger than $\mathbb{E}^{\text{high}}_{A\sim \mathring{p}}[\overline{g}(0,0,A)]$. This can be used to show that a ψ-approximate solution for this SAA problem is also a $\psi(1+\mathrm{O}(\varepsilon))$-approximate solution for the original problem. But we do not a priori know this index t, and evaluating or estimating $\mathbb{E}_{A\sim \mathring{p}}[\overline{g}(x,y,A)]$ (and hence, $\overline{h}(\mathring{p}\,;x,y)$) is challenging because (other than the difficulty of evaluating $\overline{g}(x,y,A)$ for a specific scenario A) $\mathring{p}$ can have exponential support; in fact, this is often #P-hard even for standard two-stage stochastic problems. In [24], it is shown that if one can estimate the objective value $\overline{h}(\widehat{p}\,;x,y)$ for the SAA problem (which seems easier since $\widehat{p}$ has small support), then choosing the solution corresponding to the SAA problem with best SAA objective-value estimate works.

In our case, we actually want to evaluate the objective value $\overline{h}(\mathring{p}\,;x,y)+C(\mathring{p})$ for the solution returned by the SAA problem. While we can once again decompose $\mathbb{E}_{A\sim p}[\overline{g}(x,y,A)]$ into $\mathbb{E}^{\text{low}}_{A\sim p}\big[\overline{g}(x,y,A)\big]$ and $\mathbb{E}^{\text{high}}_{A\sim p}\big[\overline{g}(x,y,A)\big]$, the term $C(p)$ could have very different contributions in the SAA and original problems (as is the case for the term $\mathbb{E}^{\text{high}}_{A\sim p}\big[\overline{g}(x,y,A)\big]$), and we need to reason about this separately. Moreover, a complicating factor is that this term is *not* linear in p. But since it is *concave* in p, we are still able to use Markov's inequality as above. In the proof below, we consider the combined term $\mathbb{E}^{\text{high}}_{A\sim \widehat{p}}\big[\overline{g}(0,0,A)\big]+C(\widehat{p})$, and apply Markov's inequality to show that there is an index $t\in[k]$ for which this term for $\widehat{p}=\widehat{p}^{\,t}$ is not much larger than $\mathbb{E}^{\text{high}}_{A\sim \mathring{p}}[\overline{g}(0,0,A)]+C(\mathring{p})$ (see Lemma 4.20).

Finally, we show that, although we do not know this index t, and we do not know how to evaluate $h(\widehat{p}\,;x,y)$ or $\overline{h}(\widehat{p}\,;x,y)$, the index j corresponding to the best estimate f^i works as well as t; this is captured by inequality (4.19).

4.5.2 Some preliminary lemmas

Throughout the proof, we assume that $\varepsilon\le 7/3$. Note that this can be done without loss of generality: if this does not hold, we can set $\varepsilon:=\min\{\varepsilon,7/3\}$ and then reason as below. Let $\varepsilon':=\varepsilon/14\le 1/6$. We set the number of SAA problems to $k:=\left\lceil \frac{2}{\varepsilon'}\ln\frac{3}{\delta}\right\rceil$ (note that $k\ge 1$).

We introduce a discretization Y of the y-range $[0,\tau]$.[3] Let $\eta' := \frac{\eta}{4+4\varepsilon'}$, and define

$$Y := \{0,\tau\} \cup \left\{\text{integer multiples of } \frac{\eta'}{\lambda r} \text{ in } (0,\tau)\right\} .$$

We now bound the error introduced by replacing $y \in [0,\tau]$ with $y \in Y$ in (4.7).

Lemma 4.16. *Let $\widetilde{p}$ be a probability distribution over $\mathcal{A}$ and $x \in X$ be an integer first-stage decision. Then there exists $y \in Y$ such that*

$$0 \le \overline{h}(\widetilde{p}\,; x, y) - \overline{h}(\widetilde{p}\,; x) \le \eta' .$$

Proof. The first inequality holds for every $y \in Y$ by (4.7), so we focus on showing that the second part holds for some choice of y. By (4.7), there exists $y' \in [0,\tau]$ such that $\overline{h}(\widetilde{p}\,; x) = \overline{h}(\widetilde{p}\,; x, y')$. Let y be the largest number in Y that is no larger than y'. Because of the way in which we defined the discretization Y, we have $y' - \frac{\eta'}{\lambda r} \le y \le y'$.

Note that

$$\begin{aligned} \overline{h}(\widetilde{p}\,; x, y) - \overline{h}(\widetilde{p}\,; x) &= \overline{h}(\widetilde{p}\,; x, y) - \overline{h}(\widetilde{p}\,; x, y') \\ &= (c^\intercal x + ry + \mathbb{E}_{A\sim\widetilde{p}}[\overline{g}(x, y, A)]) \\ &\quad - (c^\intercal x + ry' + \mathbb{E}_{A\sim\widetilde{p}}[\overline{g}(x, y', A)]) \\ &\le \mathbb{E}_{A\sim\widetilde{p}}[\overline{g}(x, y, A) - \overline{g}(x, y', A)] . \end{aligned} \tag{4.8}$$

Note that for every scenario $A \in \mathcal{A}$, we have

$$\overline{g}(x, y, A) - \overline{g}(x, y', A) \le (y' - y) \cdot M . \tag{4.9}$$

This follows because, if we let $\overline{A} := \operatorname{argmax}_{A' \in \mathcal{A}: \ell(A,A') \le M} \{g(x, A') - y \cdot \ell(A, A')\}$, then we have

$$\begin{aligned} \overline{g}(x, y', A) &\ge g\big(x, \overline{A}\big) - y' \cdot \ell\big(A, \overline{A}\big) \\ &= \big(g\big(x, \overline{A}\big) - y \cdot \ell\big(A, \overline{A}\big)\big) + (y - y') \cdot \ell\big(A, \overline{A}\big) \\ &= \overline{g}(x, y, A) + (y - y') \cdot \ell\big(A, \overline{A}\big) \\ &\ge \overline{g}(x, y, A) + (y - y') \cdot M , \end{aligned}$$

[3] The discretization considered in [24] is incorrect: it assumes implicitly that the search region of the SAA problem is (or may be) restricted to points whose first-stage cost is within some factor of the optimum of the original problem, but this need not hold. It also assumes that the grid points lie in the feasible region, which again need not hold.

where the last inequality follows because $y - y' \le 0$ and $\ell\big(A, \overline{A}\big) \le M$. Combining (4.8) and (4.9), we obtain

$$\overline{h}(\widetilde{p}\,; x, y) - \overline{h}(\widetilde{p}\,; x) \le (y' - y) \cdot M \le \frac{\eta'}{\lambda r} M = \eta' \ . \qquad \square$$

We now introduce the classification of the scenario collection $\mathcal{A}$ into low-cost and high-cost scenarios, and prove the key properties of this classification (Lemmas 4.17, 4.19, and 4.20). We classify the scenarios according to the values $\overline{g}(0, 0, A)$, using the threshold $H := \frac{\lambda}{\varepsilon'} \cdot \widetilde{O}$, where $\widetilde{O} := \min_{x \in X} \overline{h}(\mathring{p}\,; x) + C(\mathring{p})$. We define the collections of low and high scenarios as $\mathcal{A}^{\text{low}} := \{A \in \mathcal{A} : \overline{g}(0, 0, A) \le H\}$ and $\mathcal{A}^{\text{high}} := \{A \in \mathcal{A} : \overline{g}(0, 0, A) > H\}$ respectively.

Before proving the key properties of the classification of the scenarios, we define some notation for convenience. Let p be an arbitrary distribution. It will be cumbersome to carry around the $C(p)$ term, so we define $\widetilde{h}(p\,; x) := \overline{h}(p\,; x) + C(p)$ and $\widetilde{h}(p\,; x, y) := \overline{h}(p\,; x, y) + C(p)$. We use $\mathbb{E}^{\text{low}}_{A \sim p}[\cdot]$ and $\mathbb{E}^{\text{high}}_{A \sim p}[\cdot]$ to denote the contribution to the expectation $\mathbb{E}_{A \sim p}[\cdot]$ from the low and the high scenarios respectively (so we have $\mathbb{E}_{A \sim p}[\cdot] = \mathbb{E}^{\text{low}}_{A \sim p}[\cdot] + \mathbb{E}^{\text{high}}_{A \sim p}[\cdot]$).

Let x^* be an optimal solution for $\min_{x \in X} \overline{h}(\mathring{p}\,; x)$ (which is also an optimal solution for $\min_{x \in X} \widetilde{h}(\mathring{p}\,; x)$). By (4.7), there exists $y^* \in [0, \tau]$ such that $\overline{h}(\mathring{p}\,; x^*) = \overline{h}(\mathring{p}\,; x^*, y^*)$. Recall that our goal is to bound the quality of the solution $\widehat{x}^j$ with respect to $\widetilde{O}$. Note that since $C(\mathring{p}) \ge 0$, we have $\widetilde{O} = \widetilde{h}(\mathring{p}\,; x^*, y^*) \ge \overline{h}(\mathring{p}\,; x^*, y^*)$.

Lemma 4.17. *There exists* $N_1 = \mathsf{poly}\Big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log \frac{1}{\eta}, \log \frac{1}{\delta}\Big)$ *such that the following holds with probability at least* $1 - \frac{\delta}{3}$ *as long as* $N \ge N_1$*:*

$$\Big|\mathbb{E}^{\text{low}}_{A \sim \widehat{p}^i}[\overline{g}(x, y, A)] - \mathbb{E}^{\text{low}}_{A \sim \mathring{p}}[\overline{g}(x, y, A)]\Big| \le \varepsilon' \widetilde{O} \qquad \forall i \in [k],\ \forall (x, y) \in X \times Y \ .$$

Proof. Let us fix $i \in [k]$ and $(x, y) \in X \times Y$. Consider the random variable

$$W := \begin{cases} \overline{g}(x, y, A) & \text{, if } A \in \mathcal{A}^{\text{low}} \ ; \\ 0 & \text{, otherwise,} \end{cases}$$

where A is a scenario sampled according to the central distribution $\mathring{p}$. Note that W is in the range $[0, H] = \Big[0, \frac{\lambda \widetilde{O}}{\varepsilon'}\Big]$. Moreover, note that $\mathbb{E}[W] = \mathbb{E}^{\text{low}}_{A \sim \mathring{p}}[\overline{g}(x, y, A)]$, and that $\mathbb{E}^{\text{low}}_{A \sim \widehat{p}^i}[\overline{g}(x, y, A)]$ can be seen as an empirical estimate of W computed using N samples.

By Hoeffding's inequality (Corollary 2.4), there exists

$$N_1 = \mathsf{poly}\left(\frac{\frac{\lambda\widetilde{O}}{\varepsilon'}}{\varepsilon'\widetilde{O}}, \log\frac{1}{\frac{\delta}{3k|X||Y|}}\right) = \mathsf{poly}\left(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log|Y|, \log\frac{1}{\delta}\right)$$

such that, as long as $N \geq N_1$, we have

$$\mathbf{Pr}\left[\left|\mathbb{E}^{\mathrm{low}}_{A\sim\widehat{p}^i}[\overline{g}(x,y,A)] - \mathbb{E}^{\mathrm{low}}_{A\sim\mathring{p}}[\overline{g}(x,y,A)]\right| > \varepsilon'\widetilde{O}\right] \leq \frac{\delta}{3k\,|X|\,|Y|}\ .$$

Taking the union bound over all tuples (i,x,y), we get that the inequality in the lemma statement holds for all of them with probability at least $1 - k\,|X|\,|Y| \cdot \frac{\delta}{3k|X||Y|} = 1 - \frac{\delta}{3}$.

To conclude, note that $\log|Y| = \mathrm{O}\left(\log\left(\frac{\tau}{\frac{\eta'}{\lambda r}}\right)\right) = \mathsf{poly}\left(\mathcal{I}, \log\lambda, \log\frac{1}{\eta}\right)$. □

Lemma 4.18. *We have* $\sum_{A\in\mathcal{A}^{\mathrm{high}}} \mathring{p}_A \leq \frac{2\varepsilon'}{\lambda}$.

Proof. First, consider the case where $\widetilde{O} = 0$. We show that the support of $\mathring{p}$ only contains low scenarios, which implies the result. Let $A \in \mathsf{supp}(\mathring{p})$. Since we have $\overline{h}(\mathring{p}\,; x^*, y^*) \leq \widetilde{O} = 0$, we obtain $c^\intercal x^* + ry^* = 0$ and $\overline{g}(x^*, y^*, A) = 0$. Lemma 4.12 then yields $\overline{g}(0,0,A) \leq \overline{g}(x^*, y^*, A) + \lambda\cdot(c^\intercal x^* + ry^*) = 0 = H$, which implies that $A \in \mathcal{A}^{\mathrm{low}}$.

We now turn to the case where $\widetilde{O} > 0$. We have

$$\begin{aligned}
\widetilde{O} &\geq \overline{h}(\mathring{p}\,; x^*, y^*)\\
&\geq \mathbb{E}^{\mathrm{high}}_{A\sim\mathring{p}}[\overline{g}(x^*, y^*, A)]\\
&\geq \sum_{A\in\mathcal{A}^{\mathrm{high}}} \mathring{p}_A \cdot (\overline{g}(0,0,A) - \lambda\cdot(c^\intercal x^* + ry^*))\\
&\geq \sum_{A\in\mathcal{A}^{\mathrm{high}}} \mathring{p}_A \cdot \left(\frac{\lambda(1-\varepsilon')}{\varepsilon'}\right)\widetilde{O}\ ,
\end{aligned}$$

where the third inequality follows from Lemma 4.12, and the final inequality follows because $\overline{g}(0,0,A) > H = \frac{\lambda}{\varepsilon'}\widetilde{O}$ for every scenario $A \in \mathcal{A}^{\mathrm{high}}$ and $\widetilde{O} \geq \overline{h}(\mathring{p}\,; x^*, y^*) \geq c^\intercal x^* + ry^*$. Solving for $\sum_{A\in\mathcal{A}^{\mathrm{high}}} \mathring{p}_A$, and using the fact that $\varepsilon' \leq \frac{1}{6}$, we obtain

$$\sum_{A\in\mathcal{A}^{\mathrm{high}}} \mathring{p}_A \leq \frac{\varepsilon'}{\lambda(1-\varepsilon')} \leq \frac{2\varepsilon'}{\lambda}\ .$$

□

Lemma 4.19. *There exists* $N_2 = \mathsf{poly}\big(\lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$ *such that the following holds with probability at least* $1 - \frac{\delta}{3}$ *as long as* $N \geq N_2$*:*

$$\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A)] - \mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(x,y,A)] \leq 2\varepsilon'(c^\intercal x + ry) \quad \forall (x,y) \in X \times Y \ ; \tag{4.10}$$

$$\mathbb{E}^{\text{high}}_{A\sim\widehat{p}^i}[\overline{g}(0,0,A)] - \mathbb{E}^{\text{high}}_{A\sim\widehat{p}^i}[\overline{g}(x,y,A)] \leq 3\varepsilon'(c^\intercal x + ry) \ \forall i \in [k], \forall (x,y) \in X \times Y \ . \tag{4.11}$$

Proof. We start by showing that (4.10) holds with probability 1 (for every $N \geq 1$). For every $(x,y) \in X \times Y$, we have

$$\begin{aligned}\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A)] - \mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(x,y,A)] &= \mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A) - \overline{g}(x,y,A)] \\ &\leq \mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\lambda \cdot (c^\intercal x + ry)] \leq 2\varepsilon'(c^\intercal x + ry) \ .\end{aligned}$$

The first inequality follows from Lemma 4.12, and the second inequality follows because $\sum_{A\in\mathcal{A}^{\text{high}}} \mathring{p}_A \leq \frac{2\varepsilon'}{\lambda}$ by Lemma 4.18.

Now we focus on proving (4.11). Let A be a scenario sampled according to the central distribution $\mathring{p}$, and consider the indicator random variable

$$W := \begin{cases} 1 & \text{, if } A \in \mathcal{A}^{\text{high}} \ ; \\ 0 & \text{, otherwise.} \end{cases}$$

Note that $\mathbb{E}[W] = \sum_{A\in\mathcal{A}^{\text{high}}} \mathring{p}_A$, and that $\sum_{A\in\mathcal{A}^{\text{high}}} \widehat{p}^i_A$ can be seen as an empirical estimate of W computed using N samples, for every $i \in [k]$. By Hoeffding's inequality (Corollary 2.4), there exists

$$N_2 = \mathsf{poly}\left(\frac{1}{\frac{\varepsilon'}{\lambda}}, \log\frac{1}{\frac{\delta}{3k}}\right) = \mathsf{poly}\left(\lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\right)$$

such that, as long as $N \geq N_2$, we have

$$\mathbf{Pr}\left[\left|\sum_{A\in\mathcal{A}^{\text{high}}} \widehat{p}^i_A - \sum_{A\in\mathcal{A}^{\text{high}}} \mathring{p}_A\right| > \frac{\varepsilon'}{\lambda}\right] \leq \frac{\delta}{3k}$$

for every $i \in [k]$. By the union bound, with probability at least $1 - k\frac{\delta}{3k} = 1 - \frac{\delta}{3}$ we have that for every $i \in [k]$, the inequality $\left|\sum_{A\in\mathcal{A}^{\text{high}}} \widehat{p}^i_A - \sum_{A\in\mathcal{A}^{\text{high}}} \mathring{p}_A\right| \leq \frac{\varepsilon'}{\lambda}$ holds, and so $\sum_{A\in\mathcal{A}^{\text{high}}} \widehat{p}^i_A \leq \sum_{A\in\mathcal{A}^{\text{high}}} \mathring{p}_A + \frac{\varepsilon'}{\lambda} \leq \frac{3\varepsilon'}{\lambda}$, where the last inequality follows from Lemma 4.18. We can then prove that (4.11) holds by following the reasoning we used to prove (4.10):

for every $i \in [k]$ and $(x, y) \in X \times Y$, we have

$$\begin{aligned}\mathbb{E}^{\text{high}}_{A\sim\widehat{p}^i}[\overline{g}(0,0,A)] - \mathbb{E}^{\text{high}}_{A\sim\widehat{p}^i}[\overline{g}(x,y,A)] &= \mathbb{E}^{\text{high}}_{A\sim\widehat{p}^i}[\overline{g}(0,0,A) - \overline{g}(x,y,A)] \\ &\le \mathbb{E}^{\text{high}}_{A\sim\widehat{p}^i}[\lambda \cdot (c^\intercal x + ry)] \le 3\varepsilon'(c^\intercal x + ry) \ ,\end{aligned}$$

where the last inequality follows because $\sum_{A\in\mathcal{A}^{\text{high}}} \widehat{p}^i_A \le 3\frac{\varepsilon'}{\lambda}$. □

Lemma 4.20. *With probability at least $1 - \frac{\delta}{3}$, there exists $t \in [k]$ such that*

$$\left(\mathbb{E}^{\text{high}}_{A\sim\widehat{p}^t}[\overline{g}(0,0,A)] + C\big(\widehat{p}^t\big)\right) - \left(\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A)] + C(\mathring{p})\right) \le \varepsilon'\widetilde{O} \ . \tag{4.12}$$

Proof. Let $i \in [k]$. The expected value (over the random selection of the scenarios used to construct $\widehat{p}^i$) of $\mathbb{E}^{\text{high}}_{A\sim\widehat{p}^i}[\overline{g}(0,0,A)]$ is $\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A)]$. Since $C(\cdot)$ is a concave function by assumption, and since the expected value of $\widehat{p}^i$ is $\mathring{p}$, by Jensen's inequality (Theorem 2.2) the expected value of $C(\widehat{p}^i)$ is at most $C(\mathring{p})$. Using Markov's inequality (Theorem 2.1), we obtain that

$$\mathbb{E}^{\text{high}}_{A\sim\widehat{p}^i}[\overline{g}(0,0,A)] + C\big(\widehat{p}^i\big) > (1+\varepsilon')\left(\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A)] + C(\mathring{p})\right) \tag{4.13}$$

holds with probability at most $\frac{1}{1+\varepsilon'}$. Since the samples used to construct the empirical distributions are independent, the probability that this holds for every $i \in [k]$ is at most

$$\left(\frac{1}{1+\varepsilon'}\right)^k \le \left(1 - \frac{\varepsilon'}{2}\right)^{\frac{2}{\varepsilon'}\ln\frac{3}{\delta}} \le \left(e^{-\frac{\varepsilon'}{2}}\right)^{\frac{2}{\varepsilon'}\ln\frac{3}{\delta}} = \frac{\delta}{3} \ ,$$

where the first inequality follows from $\frac{1}{1+\varepsilon'} \le 1 - \frac{\varepsilon'}{2}$, which holds for $\varepsilon' \in [0,1]$, and from the definition of k.

Therefore, with probability at least $1 - \frac{\delta}{3}$, inequality (4.13) is violated for some index $i = t$, and so we have

$$\begin{aligned}&\left(\mathbb{E}^{\text{high}}_{A\sim\widehat{p}^t}[\overline{g}(0,0,A)] + C\big(\widehat{p}^t\big)\right) - \left(\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A)] + C(\mathring{p})\right) \\ &\le \varepsilon' \cdot \left(\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A)] + C(\mathring{p})\right) \\ &\le \varepsilon' \cdot \left(\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(x^*,y^*,A) + \lambda(c^\intercal x^* + ry^*)] + C(\mathring{p})\right) \\ &\le \varepsilon' \cdot \left(c^\intercal x^* + ry^* + \mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(x^*,y^*,A)] + C(\mathring{p})\right) \\ &\le \varepsilon'\widetilde{O} \ .\end{aligned}$$

The second inequality follows from Lemma 4.12. The third inequality uses Lemma 4.18 and the fact that $\varepsilon' \le \frac{1}{6}$: we have $\sum_{A\in\mathcal{A}^{\text{high}}} \mathring{p}_A \le \frac{2\varepsilon'}{\lambda} \le \frac{1}{\lambda}$. The final inequality follows because $\widetilde{O} = \widetilde{h}(\mathring{p}\,; x^*, y^*) \ge c^\intercal x^* + r y^* + \mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(x^*, y^*, A)] + C(\mathring{p})$. □

4.5.3 Details of the proof

We set $N := \max\{N_1, N_2\}$, where $N_1 = \mathsf{poly}\Big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\eta}, \log\frac{1}{\delta}\Big)$ and $N_2 = \mathsf{poly}\big(\lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$ are given by Lemmas 4.17 and 4.19 respectively. Note that $N = \mathsf{poly}\Big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\eta}, \log\frac{1}{\delta}\Big)$.

By Lemmas 4.17, 4.19, and 4.20, and using the union bound, we have all of the following with probability at least $1-\delta$:

$$\big|\mathbb{E}^{\text{low}}_{A\sim\widehat{p}^i}[\overline{g}(x,y,A)] - \mathbb{E}^{\text{low}}_{A\sim\mathring{p}}[\overline{g}(x,y,A)]\big| \le \varepsilon'\widetilde{O} \qquad \forall i\in[k],\ \forall (x,y)\in X\times Y\ ; \tag{4.14}$$

$$\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A)] - \mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(x,y,A)] \le 2\varepsilon'(c^\intercal x + ry) \ \ \forall (x,y)\in X\times Y\ ; \tag{4.15}$$

$$\mathbb{E}^{\text{high}}_{A\sim\widehat{p}^i}[\overline{g}(0,0,A)] - \mathbb{E}^{\text{high}}_{A\sim\widehat{p}^i}[\overline{g}(x,y,A)] \le 3\varepsilon'(c^\intercal x + ry)\ \forall i\in[k], \forall (x,y)\in X\times Y\ ; \tag{4.16}$$

$$\Big(\mathbb{E}^{\text{high}}_{A\sim\widehat{p}^t}[\overline{g}(0,0,A)] + C\big(\widehat{p}^t\big)\Big) - \Big(\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A)] + C(\mathring{p})\Big) \le \varepsilon'\widetilde{O}, \text{ for some } t\in[k]\ . \tag{4.17}$$

In the sequel, we suppose that (4.14)–(4.17) hold. By Lemma 4.16, there exist $\overline{y}$ and $y^j \in Y$ such that $\overline{h}(\widetilde{p}\,; x^*, \overline{y}) \le \overline{h}(\widetilde{p}\,; x^*, y^*) + \eta' \le \widetilde{O} + \eta'$ and $\overline{h}(\widehat{p}^j\,; \widehat{x}^j, y^j) \le \overline{h}(\widehat{p}^j\,; \widehat{x}^j) + \eta'$. We first use the properties of the estimates $\{f^i\}$ and the choice of the index j to relate the quality of $(\widehat{x}^j, y^j)$ under the j-th SAA problem $\min_{x\in X, y\in[0,\tau]} \widetilde{h}(\widehat{p}^j\,; x, y)$ to the quality of $(x^*, \overline{y})$ under any of the SAA problems $\min_{x\in X, y\in[0,\tau]} \widetilde{h}(\widehat{p}^i\,; x, y)$ (where $i\in[k]$). We have

$$\begin{aligned}
\widetilde{h}\big(\widehat{p}^j\,; \widehat{x}^j, y^j\big) &\le \widetilde{h}\big(\widehat{p}^j\,; \widehat{x}^j\big) + \eta' \\
&\le f^j + \eta' \\
&\le f^i + \eta' \\
&\le \psi \cdot \min_{x\in X} \widetilde{h}\big(\widehat{p}^i\,; x\big) + \eta' \\
&= \psi \cdot \min_{x\in X, y\in[0,\tau]} \widetilde{h}\big(\widehat{p}^i\,; x, y\big) + \eta' \\
&\le \psi \cdot \widetilde{h}\big(\widehat{p}^i\,; x^*, \overline{y}\big) + \eta'\ .
\end{aligned} \tag{4.18}$$

The first step follows from the definition of y^j. The second step follows from the definition of f^j in the lemma statement. The third step follows because we chose j as the index corresponding to the smallest estimate. The fourth step follows from the definition of f^i in the lemma statement. The fifth step follows from (4.7). The final step follows because $(x^*, \overline{y})$ is a feasible solution for the i-th SAA problem.

Applying (4.18) to $i = j$ and $i = t$, we obtain

$$\widetilde{h}\big(\widehat{p}^j\,; \widehat{x}^j, y^j\big) \le \psi \cdot \widetilde{h}\big(\widehat{p}^j\,; x^*, \overline{y}\big) + \eta' \ ,$$
$$\widetilde{h}\big(\widehat{p}^j\,; \widehat{x}^j, y^j\big) \le \psi \cdot \widetilde{h}\big(\widehat{p}^t\,; x^*, \overline{y}\big) + \eta' \ .$$

Taking a convex combination of the two inequalities above with coefficients $\frac{1}{\psi}$ and $1 - \frac{1}{\psi}$ respectively, we obtain

$$\widetilde{h}\big(\widehat{p}^j\,; \widehat{x}^j, y^j\big) \le \widetilde{h}\big(\widehat{p}^j\,; x^*, \overline{y}\big) + (\psi - 1) \cdot \widetilde{h}\big(\widehat{p}^t\,; x^*, \overline{y}\big) + \eta' \ . \tag{4.19}$$

Next, we evaluate the quality of $(\widehat{x}^j, y^j)$ in the original problem. We have

$$\widetilde{h}\big(\mathring{p}\,; \widehat{x}^j, y^j\big) = c^\intercal \widehat{x}^j + r y^j + \mathbb{E}^{\text{low}}_{A\sim\mathring{p}}\big[\overline{g}\big(\widehat{x}^j, y^j, A\big)\big] + \mathbb{E}^{\text{high}}_{A\sim\mathring{p}}\big[\overline{g}\big(\widehat{x}^j, y^j, A\big)\big] + C(\mathring{p}) \ . \tag{4.20}$$

To bound the contribution from low scenarios, we use (4.14), obtaining

$$\mathbb{E}^{\text{low}}_{A\sim\mathring{p}}\big[\overline{g}\big(\widehat{x}^j, y^j, A\big)\big] \le \mathbb{E}^{\text{low}}_{A\sim\widehat{p}^j}\big[\overline{g}\big(\widehat{x}^j, y^j, A\big)\big] + \varepsilon' \widetilde{O} \ . \tag{4.21}$$

Next we bound the contribution from high scenarios. For any $i \in [k]$, let

$$\Delta^i := \Big(\mathbb{E}^{\text{high}}_{A\sim\widehat{p}^i}[\overline{g}(0,0,A)] + C\big(\widehat{p}^i\big)\Big) - \Big(\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A)] + C(\mathring{p})\Big) \ .$$

We have

$$\begin{aligned}
\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}\big[\overline{g}\big(\widehat{x}^j, y^j, A\big)\big] &\le \mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A)] \\
&= \mathbb{E}^{\text{high}}_{A\sim\widehat{p}^j}[\overline{g}(0,0,A)] - \Delta^j + C\big(\widehat{p}^j\big) - C(\mathring{p}) \\
&\le \mathbb{E}^{\text{high}}_{A\sim\widehat{p}^j}\big[\overline{g}\big(\widehat{x}^j, y^j, A\big)\big] + 3\varepsilon'\big(c^\intercal \widehat{x}^j + r y^j\big) - \Delta^j + C\big(\widehat{p}^j\big) - C(\mathring{p}) \, .
\end{aligned} \tag{4.22}$$

The first inequality follows from Lemma 4.12. The equality follows from the definition of Δ^j. The final inequality follows from (4.16).

Substituting (4.21) and (4.22) in (4.20), and grouping terms by recalling that

$$\widetilde{h}\big(\widehat{p}^{\,j}\,;\widehat{x}^{j},y^{j}\big)=c^{\intercal}\widehat{x}^{j}+ry^{j}+\mathbb{E}^{\text{low}}_{A\sim\widehat{p}^{\,j}}\big[\overline{g}\big(\widehat{x}^{j},y^{j},A\big)\big]+\mathbb{E}^{\text{high}}_{A\sim\widehat{p}^{\,j}}\big[\overline{g}\big(\widehat{x}^{j},y^{j},A\big)\big]+C\big(\widehat{p}^{\,j}\big)\ ,$$

we obtain

$$\begin{aligned}\widetilde{h}\big(\mathring{p}\,;\widehat{x}^{j},y^{j}\big)&\leq\widetilde{h}\big(\widehat{p}^{\,j}\,;\widehat{x}^{j},y^{j}\big)+\varepsilon'\widetilde{O}+3\varepsilon'\big(c^{\intercal}\widehat{x}^{j}+ry^{j}\big)-\Delta^{j}\\&\leq\Big[\widetilde{h}\big(\widehat{p}^{\,j}\,;x^{*},\overline{y}\big)-\Delta^{j}\Big]+(\psi-1)\cdot\widetilde{h}\big(\widehat{p}^{\,t}\,;x^{*},\overline{y}\big)\\&\quad+\varepsilon'\widetilde{O}+3\varepsilon'\big(c^{\intercal}\widehat{x}^{j}+ry^{j}\big)+\eta'\ ,\end{aligned}\tag{4.23}$$

where the second inequality follows from (4.19).

We now proceed to bound $\widetilde{h}(\widehat{p}^{\,j}\,;x^{*},\overline{y})-\Delta^{j}$ and $\widetilde{h}(\widehat{p}^{\,t}\,;x^{*},\overline{y})$. We have

$$\begin{aligned}\widetilde{h}\big(\widehat{p}^{\,j}\,;x^{*},\overline{y}\big)-\Delta^{j}&=c^{\intercal}x^{*}+r\overline{y}+\mathbb{E}^{\text{low}}_{A\sim\widehat{p}^{\,j}}[\overline{g}(x^{*},\overline{y},A)]+\mathbb{E}^{\text{high}}_{A\sim\widehat{p}^{\,j}}[\overline{g}(x^{*},\overline{y},A)]+C\big(\widehat{p}^{\,j}\big)\\&\quad-\Big(\mathbb{E}^{\text{high}}_{A\sim\widehat{p}^{\,j}}[\overline{g}(0,0,A)]+C\big(\widehat{p}^{\,j}\big)\Big)+\Big(\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A)]+C(\mathring{p})\Big)\\&\leq c^{\intercal}x^{*}+r\overline{y}+\mathbb{E}^{\text{low}}_{A\sim\widehat{p}^{\,j}}[\overline{g}(x^{*},\overline{y},A)]+\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(0,0,A)]+C(\mathring{p})\\&\leq(1+2\varepsilon')(c^{\intercal}x^{*}+r\overline{y})+\mathbb{E}^{\text{low}}_{A\sim\mathring{p}}[\overline{g}(x^{*},\overline{y},A)]\\&\quad+\mathbb{E}^{\text{high}}_{A\sim\mathring{p}}[\overline{g}(x^{*},\overline{y},A)]+C(\mathring{p})+\varepsilon'\widetilde{O}\\&\leq(1+2\varepsilon')\widetilde{h}(\mathring{p}\,;x^{*},\overline{y})+\varepsilon'\widetilde{O}\\&\leq(1+3\varepsilon')\widetilde{O}+(1+2\varepsilon')\eta'\ .\end{aligned}\tag{4.24}$$

The first step only expands the definitions of $\widetilde{h}(\widehat{p}^{\,j}\,;x^{*},\overline{y})$ and Δ^{j}. The second step follows because $\overline{g}(x^{*},\overline{y},A)\leq\overline{g}(0,0,A)$ for every $A\in\mathcal{A}$ by Lemma 4.12. The third step uses (4.14) to bound the term involving low scenarios and (4.15) to bound the term involving high scenarios. The fourth step follows because $\widetilde{h}(\mathring{p}\,;x^{*},\overline{y})=c^{\intercal}x^{*}+r\overline{y}+\mathbb{E}_{A\sim\mathring{p}}[\overline{g}(x^{*},\overline{y},A)]+C(\mathring{p})$. The final step follows because $\widetilde{h}(\mathring{p}\,;x^{*},\overline{y})\leq\widetilde{O}+\eta'$.

Similarly, we have

$$\widetilde{h}\big(\widehat{p}^{\,t}\,;x^{*},\overline{y}\big)\leq(1+3\varepsilon')\widetilde{O}+(1+2\varepsilon')\eta'+\Delta^{t}\leq(1+4\varepsilon')\widetilde{O}+(1+2\varepsilon')\eta'\ .\tag{4.25}$$

The first inequality is obtained by following the same steps used to derive (4.24) (but using t instead of j). The second inequality follows from (4.17).

Substituting (4.24) and (4.25) in (4.23), we obtain

$$\begin{aligned}\widetilde{h}\big(\mathring{p}\,;\widehat{x}^j,y^j\big) &\le \Big[(1+3\varepsilon')\widetilde{O}+(1+2\varepsilon')\eta'\Big]+(\psi-1)\Big[(1+4\varepsilon')\widetilde{O}+(1+2\varepsilon')\eta'\Big]\\ &\quad+\varepsilon'\widetilde{O}+3\varepsilon'\big(c^{\intercal}\widehat{x}^j+ry^j\big)+\eta'\\ &=\psi(1+4\varepsilon')\widetilde{O}+(1+\psi(1+2\varepsilon'))\eta'+3\varepsilon'\big(c^{\intercal}\widehat{x}^j+ry^j\big)\\ &\le \psi(1+4\varepsilon')\widetilde{O}+\psi(2+2\varepsilon')\eta'+3\varepsilon'\widetilde{h}\big(\mathring{p}\,;\widehat{x}^j,y^j\big)\ ,\end{aligned} \tag{4.26}$$

where the last inequality follows because $\psi \ge 1$ and $c^{\intercal}\widehat{x}^j + ry^j \le \widetilde{h}(\mathring{p}\,;\widehat{x}^j,y^j)$.

Therefore we obtain

$$\begin{aligned}\widetilde{h}\big(\mathring{p}\,;\widehat{x}^j\big) &\le \widetilde{h}\big(\mathring{p}\,;\widehat{x}^j,y^j\big)\\ &\le \psi\frac{1+4\varepsilon'}{1-3\varepsilon'}\widetilde{O}+\psi\frac{2+2\varepsilon'}{1-3\varepsilon'}\eta'\\ &\le \psi(1+\varepsilon)\widetilde{O}+\psi\eta\ .\end{aligned}$$

The first inequality follows from (4.7). The second one follows from (4.26). The final inequality follows because $\frac{1+4\varepsilon'}{1-3\varepsilon'} = 1+\frac{7\varepsilon'}{1-3\varepsilon'} \le 1+14\varepsilon' = 1+\varepsilon$ and $\frac{2+2\varepsilon'}{1-3\varepsilon'}\eta' \le (4+4\varepsilon')\eta' = \eta$ (since $1-3\varepsilon' \ge \frac{1}{2}$). □

Chapter 5

DRS optimization under a Wasserstein ball: polynomial-size central distribution

Recall that in Chapter 4, we proved an SAA result for a discrete DRS optimization problem under a Wasserstein ball,

$$\min_{\substack{x\in X, z\in Z^{\mathcal{A}}:\\ (x,z^A)\in F(A)\quad \forall A\in\mathcal{A}}} \left\{ c^\intercal x + \sup_{p: L_{\mathrm{W}}(\mathring{p},p)\le r} \mathbb{E}_{A\sim p}\big[\text{cost of } z^A\big] \right\}, \tag{DRSO$_{\mathrm{W}}$}$$

where L_{W} is a Wasserstein metric defined relative to a scenario metric ℓ, and the central distribution $\mathring{p}$ is given by a sampling oracle.

We moved from (DRSO$_{\mathrm{W}}$) to its relaxation with integer first-stage decisions and (implicit) fractional second-stage decisions,

$$\min_{x\in X} \{h(\mathring{p}\,; x) := c^\intercal x + z(\mathring{p}\,; x)\}. \tag{Q($\mathring{p}$)}$$

Informally, we gave a reduction from (Q($\mathring{p}$)) to a collection of DRS problems of the type

$$\min_{x\in X} \{h(\widehat{p}\,; x) := c^\intercal x + z(\widehat{p}\,; x)\}, \tag{Q($\widehat{p}$)}$$

where the central distribution $\widehat{p}$ has a polynomially-bounded support size (see Theorem 4.1 for the precise statement). Our goal in this chapter is to design a framework for computing

an approximate solution for ($\mathrm{Q}(\widehat{p})$), as well as an estimate of its objective value. Combined with the SAA result (Theorem 4.1), this allows us to obtain an approximation algorithm for the relaxed DRS problem ($\mathrm{Q}(\mathring{p})$) with a black-box central distribution, which can then be converted into an approximate solution for the discrete DRS problem ($\mathrm{DRSO}_{\mathrm{W}}$) via a second-stage approximation algorithm.

The central distribution $\widehat{p}$ is represented explicitly, that is, we have a collection of pairs $\{(A, \widehat{p}_A)\}_{A \in \mathcal{A}^{\sup}}$ specifying the probabilities of the scenarios in the support of $\widehat{p}$, which we denote by $\mathcal{A}^{\sup}$. The scenarios that do not appear in this collection of pairs have probability zero under the distribution $\widehat{p}$. The input size of the SAA problem ($\mathrm{Q}(\widehat{p})$) is denoted by $\widehat{\mathcal{I}}$, and is defined as the sum of the input size $\mathcal{I}$ of the original problem ($\mathrm{DRSO}_{\mathrm{W}}$) and the encoding size of the pairs $\{(A, \widehat{p}_A)\}_{A \in \mathcal{A}^{\sup}}$ used to specify the central distribution. Note that $|\mathcal{A}^{\sup}| = \mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ by definition (and when $\widehat{p}$ is obtained via the SAA result from Theorem 4.1 with parameters $(\varepsilon, \eta, \delta)$, we have $|\mathcal{A}^{\sup}| = \mathsf{poly}\left(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\eta}, \log\frac{1}{\delta}\right)$).

We now state the main result in this chapter, which gives a reduction from ($\mathrm{Q}(\widehat{p})$) to the following problem.

(Π) Given an integer first-stage decision $x \in X, y \geq 0$, and a scenario $A \in \mathcal{A}$, solve
$$g(x, y, A) := \max_{A' \in \mathcal{A}} \{g(x, A') - y \cdot \ell(A, A')\} \ .$$

Recall from Definition 3.5 that a (β_1, β_2)-approximation algorithm for (Π) returns a scenario $\overline{A} \in \mathcal{A}$ such that
$$g\big(x, \overline{A}\big) - y \cdot \ell\big(A, \overline{A}\big) \geq \max_{A' \in \mathcal{A}} \left\{ \frac{1}{\beta_1} g(x, A') - \beta_2 y \cdot \ell(A, A') \right\} \ .$$

Theorem 5.1 (combination of Theorem 3.14 and Lemmas 5.2 and 5.6)**.** *Suppose that we have (i) a (β_1, β_2)-approximation algorithm for problem (Π); and (ii) a local ρ-approximation algorithm. Then, given $\eta > 0$, we can compute in $\mathsf{poly}\left(\widehat{\mathcal{I}}, \log\frac{1}{\eta}\right)$ time an integer first-stage decision $\widehat{x} \in X$ and an estimate f such that*
$$h(\widehat{p}; \widehat{x}) \leq f \leq \beta_1 \beta_2 \rho \left(\min_{x \in X} h(\widehat{p}; x) + \eta \right) \ .$$

Organization of this chapter. In Section 5.1, we provide an overview of the techniques used in the proof of Theorem 5.1; we present the proof in detail in Section 5.2. In Sec-

tion 5.3, we show that for certain choices of the scenario collection $\mathcal{A}$ and the scenario metric ℓ, we can solve the fractional SAA problem ($\mathrm{Q}^{\mathrm{fr}}(\widehat{p})$) (with *fractional* first-stage and second-stage decisions) exactly, by reformulating it as a compact LP. In Section 5.4, we present hardness results for some problems related to solving the DRS problem ($\mathrm{Q}(\widehat{p})$).

5.1 Overview of the techniques

Perhaps the most natural approach for obtaining an approximation algorithm for the DRS problem ($\mathrm{Q}(\widehat{p})$) would be to move to its further relaxation with fractional first-stage decisions,

$$\min_{x \in \mathcal{P}} \{h(\widehat{p}\,; x) := c^{\intercal} x + z(\widehat{p}\,; x)\} \ , \tag{$\mathrm{Q}^{\mathrm{fr}}(\widehat{p})$}$$

compute a (fractional) approximate solution for it, then round it via a local approximation algorithm. Unlike the case with two-stage {stochastic, robust} optimization, where the fractional relaxation of a problem with an explicit polynomial-size list of scenarios gives a compact LP and can therefore be solved in polynomial time, it is substantially more challenging to even approximately solve ($\mathrm{Q}^{\mathrm{fr}}(\widehat{p})$) with a polynomial-size central distribution. As we now explain, reformulating ($\mathrm{Q}^{\mathrm{fr}}(\widehat{p})$) as an LP leads to an LP with exponentially many variables *and* constraints. Recall from Section 4.1 that we can express $z(\widehat{p}\,; x)$ as the optimal value of the following LP:

$$z(\widehat{p}\,; x) = \max \quad \sum_{A, A'} \gamma_{A,A'} g(x, A') \tag{$\mathrm{T}(\widehat{p}, x)$}$$

$$\text{s.t.} \quad \sum_{A'} \gamma_{A,A'} \leq \widehat{p}_A \quad \forall A \in \mathcal{A} \tag{5.1}$$

$$\sum_{A, A'} \ell(A, A') \gamma_{A,A'} \leq r \tag{5.2}$$

$$\gamma \geq 0 \ . \tag{5.3}$$

Note that, because of constraints (5.1) and (5.3), we can simplify the above LP so that we only have variables $\gamma_{A,A'}$ for $A \in \mathcal{A}^{\mathrm{sup}}$, and constraints (5.1) are only present for $A \in \mathcal{A}^{\mathrm{sup}}$. Taking the dual of this simplified LP, we obtain a reformulation of $z(\widehat{p}\,; x)$ as a

minimization LP; this yields the following reformulation of ($\mathrm{Q}^{\mathrm{fr}}(\widehat{p})$):

$$\begin{aligned} \min \quad & c^\intercal x + ry + \sum_{A \in \mathcal{A}^{\sup}} \widehat{p}_A \mu_A \\ \text{s.t.} \quad & \mu_A \geq g(x, A') - y \cdot \ell(A, A') \qquad \forall A \in \mathcal{A}^{\sup}, A' \in \mathcal{A} \\ & x \in \mathcal{P},\ \mu \in \mathbb{R}_+^{\mathcal{A}^{\sup}},\ y \geq 0\ . \end{aligned}$$

Suppose that under each scenario A, the cost of a fractional second-stage decision z^A is given by $s^A \cdot z^A$, where $s^A \in \mathbb{R}^n_+$. Substituting $g(x, A') = \min\left\{s^{A'} \cdot z^{A'} : \left(x, z^{A'}\right) \in \mathcal{F}(A')\right\}$ in the reformulation above, we obtain the following reformulation of ($\mathrm{Q}^{\mathrm{fr}}(\widehat{p})$) as an LP:

$$\begin{aligned} \min \quad & c^\intercal x + ry + \sum_{A \in \mathcal{A}^{\sup}} \widehat{p}_A \mu_A \\ \text{s.t.} \quad & \mu_A \geq s^{A'} \cdot z^{A'} - y \cdot \ell(A, A') \qquad \forall A \in \mathcal{A}^{\sup}, A' \in \mathcal{A} \\ & (x, z^{A'}) \in \mathcal{F}(A') \qquad \forall A' \in \mathcal{A} \\ & x \in \mathcal{P}, \mu \in \mathbb{R}_+^{\mathcal{A}^{\sup}},\ y \geq 0\ . \end{aligned}$$

The numbers of variables and constraints are both $\Omega(|\mathcal{A}|)$. Thus, even though the central distribution $\widehat{p}$ has polynomial-size support, the LP-reformulation of ($\mathrm{Q}^{\mathrm{fr}}(\widehat{p})$) has *exponentially many variables and constraints*, and we are unable to solve it efficiently.[1]

It seems therefore preferable to work directly with ($\mathrm{Q}^{\mathrm{fr}}(\widehat{p})$) as a *convex program*. One can show that solving this convex program reduces to the following problem.

(*) Given a *fractional* first-stage decision $x \in \mathcal{P}, y \geq 0$, and a scenario $A \in \mathcal{A}$, solve

$$g(x, y, A) := \max_{A' \in \mathcal{A}} \{g(x, A') - y \cdot \ell(A, A')\}\ .$$

(Note that problem (*) is more general than (Π), which is the problem for which we assume the existence of a non-standard type of approximation algorithm.) Indeed, if we had an efficient algorithm for (*), then given any point $x \in \mathcal{P}$ we would be able to efficiently separate the (exponentially many) constraints of the dual of the simplified version of the LP ($\mathrm{T}(\widehat{p}, x)$). We would therefore be able to compute an optimal solution for this LP,

[1] An exception to this is the unrestricted setting (i.e., $\mathcal{A} = 2^U$ for some ground set U) with the discrete scenario metric ℓ^{disc} (so L_{W} is the $\frac{1}{2}L_1$ metric), under the assumption that $g(x, A) \leq g(x, A')$ for all $x \in \mathcal{P}$, $A \subseteq A'$, which holds for covering problems. Here, we can reformulate $z(\widehat{p}; x)$ as an LP with polynomially many variables $\{\gamma_{A,A'}\}$ and hence, obtain a reformulation of ($\mathrm{Q}^{\mathrm{fr}}(\widehat{p})$) as a compact LP. Theorem 5.8 shows a more general result along these lines.

which could be used to compute both the objective value $h(\widehat{p}\,;x)$ and a subgradient of $h(\widehat{p}\,;\cdot)$ at x. We would therefore be able to approximately solve ($\mathrm{Q}^{\mathrm{fr}}(\widehat{p})$) via the classical ellipsoid method (Theorem 3.10).

It turns out that (*) is generally a complicated problem that, as we show in Theorem 5.9-(b), can capture the k-max-min problem $\max_{A\subseteq U:|A|\leq k} g(0,A)$ encountered in two-stage robust optimization, which is NP-hard in various settings (see Section 2.1). Moreover, due to its mixed-sign objective, (*) is often inapproximable under the standard notion of approximation (see Theorem 5.9-(b)).

We therefore consider the non-standard notion of approximation given by Definition 3.5. If we have a (β_1,β_2)-approximation algorithm for (*), then we can compute a β-approximate solution γ for ($\mathrm{T}(\widehat{p},x)$), given any fractional first-stage decision $x\in\mathcal{P}$, where $\beta=\beta_1\beta_2$ (see Lemma 5.5). As shown in Lemma 5.3, this solution γ could then be used to compute both (i) a β-approximate estimate of $h(\widehat{p}\,;x)$ and (ii) a $(1-1/\beta)$-subgradient of $h(\widehat{p}\,;\cdot)$ at x. This suggests that we try to utilize ellipsoid-based algorithms based on approximate subgradients. We discussed two such algorithms in Section 3.4.1, namely, the algorithm by Shmoys and Swamy [114] based on ω-subgradients (see Theorem 3.12) and the alternate algorithm based on $\big(\psi,\overline{X}\big)$-first-order oracles (see Theorem 3.14).

The algorithm by Shmoys and Swamy [114] requires the ability to compute ω-subgradients for quite small values of ω, which amounts to obtaining a (β_1,β_2)-approximation algorithm for (*) with $\beta_1\beta_2=1+\varepsilon$. This is impossible for various problems, because, as mentioned before, (*) can be used to encode the k-max-min problem $\max_{A\subseteq U:|A|\leq k} g(0,A)$, which is APX-hard in various settings.

Therefore, we move to the ellipsoid-based algorithm from Theorem 3.14. One can show that an approximate evaluation oracle for $h(\widehat{p}\,;\cdot)$ can be combined with an algorithm for computing ω-subgradients to obtain a generalized first-order oracle in the sense of Definition 3.13; this holds for *any* value of ω. However, one final difficulty remains: even in cases where we know how to (approximately) solve the k-max-min problem, which as noted earlier is a special case of (*), we only have an algorithm that works with *integer* first-stage decisions. (This is also the reason why in Theorem 5.1, we only assume the approximability of (Π), which involves computing $g(x,y,A)$ for *integer* first-stage decisions x.) However, in (*), x could be fractional. To remedy this, we utilize the flexibility of the alternate first-order oracle from Definition 3.13, where we are allowed to move to a different point. Here, when faced with a fractional point $x\in\mathcal{P}$, we will first round x using a local ρ-approximation algorithm to an integer point $\widehat{x}\in X$, and then find a β-approximate solution for ($\mathrm{T}(\widehat{p},\widehat{x})$), using the (β_1,β_2)-approximation algorithm for (Π) (see Lemma 5.5). We show in Lemma 5.4 that this indeed yields a $(\beta\rho,X)$-first-order oracle, which we then

use in the ellipsoid-based algorithm from Theorem 3.14.

5.2 Proof of Theorem 5.1

We show in Lemma 5.2 that we can utilize a (β_1, β_2)-approximation algorithm for (Π), in conjunction with a local ρ-approximation algorithm, to obtain a $(\beta_1\beta_2\rho, X)$-first-order oracle for $h(\widehat{p}\,;\cdot)$. Recall from Definition 3.13 that this oracle is an algorithm that, given any fractional first-stage decision $x \in \mathcal{P}$, computes a tuple $(\widehat{x}, f, d) \in X \times \mathbb{R} \times \mathbb{R}^m$ such that (i) $h(\widehat{p}\,;\widehat{x}) \leq f$; and (ii) $h(\widehat{p}\,;x') \geq \frac{1}{\psi} f$ for every $x' \in X$ such that $d^\intercal(x' - x) \geq 0$. Equipped with this generalized first-order oracle and the bound on the Lipschitz constant of $h(\widehat{p}\,;\cdot)$ given by Lemma 5.6, we can then immediately obtain Theorem 5.1 by utilizing the ellipsoid-based method from Theorem 3.14.

Lemma 5.2 (combination of Lemmas 5.4 and 5.5)**.** *Suppose that we have (i) a (β_1, β_2)-approximation algorithm for problem (Π); and (ii) a local ρ-approximation algorithm. Then we can obtain a $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$-time $(\beta_1\beta_2\rho, X)$-first-order oracle for $h(\widehat{p}\,;\cdot)$.*

We now discuss how to obtain Lemma 5.2. First, we show in Lemma 5.3 that, given any fractional first-stage decision $x \in \mathcal{P}$, an approximate solution for ($\mathrm{T}(\widehat{p}, x)$) can be utilized to compute an approximate subgradient of $h(\widehat{p}\,;\cdot)$ at x. Then, we show in Lemma 5.4 that given any $x \in \mathcal{P}$, if we round it to an integer first-stage decision $\widehat{x} \in X$ using a local approximation algorithm, then an approximate solution for $(\mathrm{T}(\widehat{p}, \widehat{x}))$ can be utilized to compute a valid output for a generalized first-order oracle, when x is given as the input. Finally, we show in Lemma 5.5 that one can compute an approximate solution for $(\mathrm{T}(\widehat{p}, \widehat{x}))$ utilizing the approximation algorithm for problem (Π).

Recall from assumption (A5) that for any scenario $A \in \mathcal{A}$, the function $x \mapsto g(x, A)$ is convex over $\mathcal{P}$, and at every $x \in \mathcal{P}$, we can efficiently compute its value, and a subgradient $d^{x,A}$.

Lemma 5.3. *Let $\widetilde{p}$ be a probability distribution over $\mathcal{A}$. Let $x \in \mathcal{P}$ be a fractional first-stage decision, and let γ be a β-approximate solution for $(\mathrm{T}(\widetilde{p}, x))$. Define $f := c^\intercal x + \sum_{A,A'} \gamma_{A,A'} g(x, A')$ and $d := c + \sum_{A,A'} \gamma_{A,A'} d^{x,A'}$. Then we have (i) $f \leq h(\widetilde{p}\,;x) \leq \beta f$ and (ii) d is a $(1 - 1/\beta)$-subgradient of $h(\widetilde{p}\,;\cdot)$ at x.*

Proof. We start by proving (i). Since γ is feasible for $(\mathrm{T}(\widetilde{p}, x))$, we have

$$h(\widetilde{p}\,;x) \geq c^\intercal x + \sum_{A,A'} \gamma_{A,A'} g(x, A') = f \ ,$$

which proves the first part of (i). To prove the second part of (i), note that

$$h(\widetilde{p}\,;x) = c^\intercal x + z(\widetilde{p}\,;x) \le c^\intercal x + \beta \sum_{A,A'} \gamma_{A,A'} g(x,A') \le \beta f \ ,$$

where the first inequality holds because γ is a β-approximate solution for $(\mathrm{T}(\widetilde{p},x))$.

Now we prove (ii). Let $x' \in \mathcal{P}$. Consider the function $\zeta : \overline{x} \mapsto c^\intercal \overline{x} + \sum_{A,A'} \gamma_{A,A'} g(\overline{x},A')$. We claim that d is a subgradient of $\zeta(\cdot)$ at x. Assuming this, we obtain

$$h(\widetilde{p}\,;x') - \zeta(x) \ge \zeta(x') - \zeta(x) \ge d^\intercal(x'-x) \ ,$$

where the first inequality follows because γ is a feasible solution for $(\mathrm{T}(\widetilde{p},x'))$, and the second inequality follows because d is a subgradient of $\zeta(\cdot)$ at x. It follows that

$$h(\widetilde{p}\,;x') - h(\widetilde{p}\,;x) \ge d^\intercal(x'-x) + \zeta(x) - h(\widetilde{p}\,;x) \ge d^\intercal(x'-x) - \left(1-\frac{1}{\beta}\right) h(\widetilde{p}\,;x) \ ,$$

where the final inequality follows from part (i): we have $\zeta(x) = f \ge \frac{1}{\beta} h(\widetilde{p}\,;x)$. Since this holds for every $x' \in \mathcal{P}$, it follows that d is a $(1-1/\beta)$-subgradient of $h(\widetilde{p}\,;\cdot)$ at x.

It remains to prove the claim. For any $x' \in \mathcal{P}$, we have

$$\begin{aligned}
\zeta(x') - \zeta(x) &= c^\intercal(x'-x) + \sum_{A,A'} \gamma_{A,A'} (g(x',A') - g(x,A')) \\
&\ge c^\intercal(x'-x) + \sum_{A,A'} \gamma_{A,A'} d^{x,A'} \cdot (x'-x) \\
&= d^\intercal(x'-x) \ ,
\end{aligned}$$

where the second inequality follows because $d^{x,A'}$ is a subgradient of $g(\cdot,A')$ at x. □

Lemma 5.4. *Let $x \in \mathcal{P}$ be a fractional first-stage decision, and let $\widehat{x}$ be obtained by rounding x via a local ρ-approximation algorithm. Let γ be a β-approximate solution for $(\mathrm{T}(\widehat{p},\widehat{x}))$. Define $f := \beta \cdot \left(c^\intercal \widehat{x} + \sum_{A,A'} \gamma_{A,A'} g(\widehat{x},A')\right)$ and $d := c + \sum_{A,A'} \gamma_{A,A'} d^{x,A}$. Then we have (i) $h(\widehat{p}\,;\widehat{x}) \le f$ and (ii) $h(\widehat{p}\,;x') \ge \frac{1}{\beta\rho} f$ for every $x' \in X$ such that $d^\intercal(x'-x) \ge 0$.*

Proof. Part (i) follows immediately from part (i) of Lemma 5.3, setting $(\widetilde{p},x)$ to $(\widehat{p},\widehat{x})$.

Now we prove (ii). Let $x' \in X$ such that $d^\intercal(x'-x) \ge 0$. Consider the function $\zeta : \overline{x} \mapsto c^\intercal \overline{x} + \sum_{A,A'} \gamma_{A,A'} g(\overline{x},A')$. By repeating the arguments used in the proof of

Lemma 5.3, we obtain that $h(\widehat{p}; x') - \zeta(x) \geq d^{\intercal}(x' - x) \geq 0$. Since $\widehat{x}$ is obtained by rounding x via a local ρ-approximation algorithm, we have $\zeta(\widehat{x}) \leq \rho\zeta(x)$, so we obtain $h(\widehat{p}; x') \geq \zeta(x) \geq \frac{1}{\rho}\zeta(\widehat{x}) = \frac{1}{\beta\rho}f$. □

Lemma 5.5. *Suppose that we have a (β_1, β_2)-approximation algorithm for problem (Π). Then, given an integer first-stage decision $\widehat{x} \in X$, we can compute in $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ time a $\beta_1\beta_2$-approximate solution for $(\mathrm{T}(\widehat{p}, \widehat{x}))$.*

Proof. Recall that, as noted in Section 5.1, we can simplify the LP $(\mathrm{T}(\widehat{p}, \widehat{x}))$ so that we only have variables $\gamma_{A,A'}$ for $A \in \mathcal{A}^{\sup}$, and constraints (5.1) are only present for $A \in \mathcal{A}^{\sup}$. The simplified LP, which has a polynomial number of constraints, is stated below.

$$\begin{aligned} \max \quad & \sum_{A \in \mathcal{A}^{\sup}, A' \in \mathcal{A}} \gamma_{A,A'} g(\widehat{x}, A') && \text{(P)} \\ \text{s.t.} \quad & \sum_{A' \in \mathcal{A}} \gamma_{A,A'} \leq \widehat{p}_A \quad \forall A \in \mathcal{A}^{\sup} \\ & \sum_{A \in \mathcal{A}^{\sup}, A' \in \mathcal{A}} \ell(A, A')\gamma_{A,A'} \leq r \\ & \gamma \in \mathbb{R}_+^{\mathcal{A}^{\sup} \times \mathcal{A}} \, . \end{aligned}$$

The dual of the LP above is

$$\begin{aligned} \min \quad & ry + \sum_{A \in \mathcal{A}^{\sup}} \widehat{p}_A \mu_A && \text{(D)} \\ \text{s.t.} \quad & \mu_A \geq g(\widehat{x}, A') - y \cdot \ell(A, A') \qquad \forall A \in \mathcal{A}^{\sup}, A' \in \mathcal{A} && (5.4) \\ & \mu \in \mathbb{R}_+^{\mathcal{A}^{\sup}}, \; y \geq 0 \, . && (5.5) \end{aligned}$$

Notice that (D) is an LP (since $\widehat{x}$ is fixed) with only $\mathrm{O}(|\mathcal{A}^{\sup}|) = \mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ variables, but $\Theta(|\mathcal{A}^{\sup}|\,|\mathcal{A}|)$ constraints, which may be exponential in the input size. It is evident that a (β_1, β_2)-approximation algorithm Alg for problem (Π) yields some type of approximate separation oracle for (D). Using a standard technique in approximation algorithms, we prove that (D), and the primal (P), can be solved approximately (see, e.g., [23, 47, 49, 70, 75, 76, 78]).

Let $\mathcal{Q}(\nu)$ denote the set of feasible solutions of (D) with objective value at most ν, that is, $\mathcal{Q}(\nu) := \left\{(\mu, y) : (5.4),\ (5.5),\ ry + \sum_{A \in \mathcal{A}^{\sup}} \widehat{p}_A \mu_A \leq \nu\right\}$. Note that OPT(D) is the smallest value of ν such that $\mathcal{Q}(\nu) \neq \emptyset$. We use Alg to obtain a $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$-time approx-

imate separation oracle in the following sense. Given ν and (μ, y), we either show that $(\beta_1\mu, \beta_1\beta_2 y) \in \mathcal{Q}(\beta_1\beta_2\nu)$, or we exhibit a hyperplane separating (μ, y) from $\mathcal{Q}(\nu)$. Thus, given ν, by running the ellipsoid method with this approximate separation oracle, we can in $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ time either certify that $\mathcal{Q}(\nu) = \emptyset$, or obtain (μ, y) with $(\beta_1\mu, \beta_1\beta_2 y) \in \mathcal{Q}(\beta_1\beta_2\nu)$.

We now describe the separation oracle. Given ν and (μ, y), we first check if (5.5) and $ry + \sum_{A\in\mathcal{A}^{\sup}} \widehat{p}_A\mu_A \le \nu$ hold, and if not, we use the appropriate inequality as the separating hyperplane. Next, for every scenario $A \in \mathcal{A}^{\sup}$, we run Alg for the input $(\widehat{x}, y, A)$, thus obtaining a scenario $\overline{A} \in \mathcal{A}$. Then we verify if $\mu_A \ge g\big(\widehat{x}, \overline{A}\big) - y \cdot \ell\big(A, \overline{A}\big)$. If this constraint is violated, then we use it as the separating hyperplane. Note that the running time of this separation oracle is $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$.

If no violated constraint is found by this approximate separation oracle, then for every $A \in \mathcal{A}^{\sup}$ and $A' \in \mathcal{A}$, letting $\overline{A}$ denote the output of Alg when given input $(\widehat{x}, y, A)$, we have

$$\mu_A \ge g\big(\widehat{x}, \overline{A}\big) - y \cdot \ell\big(A, \overline{A}\big) \ge \frac{1}{\beta_1} g(\widehat{x}, A') - \beta_2 y \cdot \ell(A, A') \ .$$

This implies that $(\beta_1\mu, \beta_1\beta_2 y) \in \mathcal{Q}(\beta_1\beta_2\nu)$.

We claim that by running the ellipsoid method as described above for different values of ν, using binary search, we can compute an approximate solution for (D). To define the range of ν on which to do binary search, we need to find lower and upper bounds on OPT(D). First, note that OPT(D) ≥ 0, since every feasible solution (μ, y) of (D) is nonnegative by (5.5). To obtain an upper bound, we run the ellipsoid method as described above for $\nu = 2^0, 2^1, 2^2, 2^3, \ldots$, until we find $\nu = \bar{\nu}$ such that the algorithm returns a solution in $\mathcal{Q}(\beta_1\beta_2\bar{\nu})$. Then $\mathsf{UB} := \beta_1\beta_2\bar{\nu}$ is an upper bound on OPT(D). Note that this takes $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ time, since we consider $O(\log \text{OPT(D)})$ different values of ν, and OPT(D) = OPT(P) $\le \max_{A'\in\mathcal{A}} g(\widehat{x}, A')$, and we have $\log(\max_{A'\in\mathcal{A}} g(\widehat{x}, A')) = \mathsf{poly}(\mathcal{I})$ by assumption (A5).

Let $\varepsilon > 0$. Since OPT(D) $\in [0, \mathsf{UB}]$, we have that $\mathcal{Q}(-\beta_1\beta_2\varepsilon) = \emptyset$ and $\mathcal{Q}(\mathsf{UB}) \ne \emptyset$, and so we can perform binary search on the interval $[-\varepsilon, \mathsf{UB}]$ to find a value ν^* such that the ellipsoid method, when run for $\nu = \nu^*$ (with the approximate separation oracle introduced above), returns a solution (μ^*, y^*) with $(\beta_1\mu^*, \beta_1\beta_2 y^*) \in \mathcal{Q}(\beta_1\beta_2\nu^*)$, and when run with $\nu = \nu^* - \varepsilon$, certifies that $\mathcal{Q}(\nu^* - \varepsilon) = \emptyset$.[2] Since $\mathcal{Q}(\beta_1\beta_2\nu^*) \ne \emptyset$, we have $\beta_1\beta_2\nu^* \ge$ OPT(D) = OPT(T$(\widehat{p}, \widehat{x})$). The ellipsoid method with $\nu = \nu^* - \varepsilon$ generates a

[2]Note that there does not necessarily exist a threshold ν separating the two different types of outcomes for the ellipsoid method, namely: (i) certifying that $\mathcal{Q}(\nu) = \emptyset$; or (ii) producing a solution (μ, y) with $(\beta_1\mu, \beta_1\beta_2 y) \in \mathcal{Q}(\beta_1\beta_2\nu)$. Indeed, while we can say that we have outcome (i) when $\nu < \frac{\text{OPT(D)}}{\beta_1\beta_2}$ and

subset of $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ constraints from the family of constraints (5.4). Consider the LP ($\widetilde{\mathrm{D}}$), obtained from (D) by restricting the family of constraints (5.4) to this subset. Since the constraints generated by the ellipsoid method certify that $\mathcal{Q}(\nu^* - \varepsilon) = \emptyset$, it follows that $\mathrm{OPT}(\widetilde{\mathrm{D}}) > \nu^* - \varepsilon$. We can choose ε small enough with $\log \frac{1}{\varepsilon} = \mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ so that this also implies $\mathrm{OPT}(\widetilde{\mathrm{D}}) \geq \nu^*$. Note that this choice also guarantees that $\log \frac{\mathsf{UB}}{\varepsilon} = \mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ and hence the binary search takes $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ time.

Now, consider the dual ($\widetilde{\mathrm{P}}$) of ($\widetilde{\mathrm{D}}$). By strong duality, we obtain $\mathrm{OPT}(\widetilde{\mathrm{P}}) = \mathrm{OPT}(\widetilde{\mathrm{D}}) \geq \nu^*$. Note that ($\widetilde{\mathrm{P}}$) corresponds to the LP obtained from (P) by restricting γ to use only the variables $\gamma_{A,A'}$ corresponding to constraints from the family (5.4) that were retained in ($\widetilde{\mathrm{D}}$). Therefore, by computing an optimal solution for ($\widetilde{\mathrm{P}}$) (which we can do in $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ time, as it has $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ variables and constraints), we obtain a feasible solution for (P) with objective value

$$\mathrm{OPT}(\widetilde{\mathrm{P}}) \geq \nu^* \geq \frac{1}{\beta_1 \beta_2} \cdot \mathrm{OPT}(\mathrm{T}(\widehat{p}, \widehat{x})) \ . \qquad \square$$

We now bound the Lipschitz constant of the objective function of ($\mathrm{Q}^{\mathrm{fr}}(\widehat{p})$). Note that $\log(\|c\| + K) = \mathsf{poly}(\mathcal{I})$ by assumption (A5).

Lemma 5.6. *Let $\widetilde{p}$ be a probability distribution over $\mathcal{A}$. Then the function $h(\widetilde{p}\,;\cdot)$ is $(\|c\| + K)$-Lipschitz continuous over $\mathcal{P}$.*

Proof. We show that for every fractional first-stage decision $x \in \mathcal{P}$, there exists a subgradient of $h(\widetilde{p}\,;\cdot)$ at x with Euclidean norm at most $\|c\| + K$. The result then follows from Lemma 1.1. Indeed, let $x \in \mathcal{P}$, and let γ be an optimal solution for ($\mathrm{T}(\widetilde{p}, x)$). Then by Lemma 5.3 (setting $\beta = 1$), we have that $d := c + \sum_{A,A'} \gamma_{A,A'} d^{x,A'}$ is a subgradient of $h(\widetilde{p}\,;\cdot)$ at x. We have

$$\|d\| = \left\| c + \sum_{A,A'} \gamma_{A,A'} d^{x,A'} \right\| \leq \|c\| + \sum_{A,A'} \gamma_{A,A'} \left\| d^{x,A'} \right\| \leq \|c\| + K \ ,$$

where the first inequality follows from the triangle inequality, and the second one follows because $\left\|d^{x,A'}\right\| \leq K$ for every $A' \in \mathcal{A}$ by assumption (A5) and $\sum_{A,A'} \gamma_{A,A'} \leq 1$. $\square$

outcome (ii) when $\nu \geq \mathrm{OPT}(\mathrm{D})$, both outcomes are possible for $\nu \in \left[\frac{\mathrm{OPT}(\mathrm{D})}{\beta_1 \beta_2}, \mathrm{OPT}(\mathrm{D})\right)$.

5.3 Solving $(\mathrm{Q}^{\mathrm{fr}}(\widehat{p}))$ exactly in certain settings

In this section we show that, for certain choices of the scenario collection $\mathcal{A}$ and the scenario metric ℓ, we can reformulate the fractional SAA problem $(\mathrm{Q}^{\mathrm{fr}}(\widehat{p}))$ as a compact LP, and hence solve it exactly.

Definition 5.7. We say that the scenario collection $\mathcal{A}$ is *collapsible under the scenario metric* ℓ if given any scenario $A \in \mathcal{A}$, we can compute in $\mathsf{poly}(\mathcal{I})$ time a collection of scenarios $\phi(A) \subseteq \mathcal{A}$ such that for every fractional first-stage decision $x \in \mathcal{P}$ and every $y \geq 0$, we have

$$g(x, y, A) = \max_{A' \in \phi(A)} \{g(x, A') - y \cdot \ell(A, A')\} \ .$$

Note that if we only have a polynomial number number of scenarios (i.e., if $|\mathcal{A}|$ is $\mathsf{poly}(\mathcal{I})$), then clearly $\mathcal{A}$ is collapsible under *any* scenario metric ℓ, since we can simply define $\phi(A) := \mathcal{A}$ for every scenario A. Therefore the discussion in the sequel implies that in this case, we can solve $(\mathrm{Q}^{\mathrm{fr}}(\widehat{p}))$ efficiently by reformulating it as a compact LP.

Theorem 5.8. *Suppose that the scenario collection $\mathcal{A}$ is collapsible under the scenario metric ℓ, and that the second-stage costs $\{g(x, A')\}$ are given by compact LPs, say, $g(x, A') = \min\left\{s^{A'} \cdot z^{A'} : \left(x, z^{A'}\right) \in \mathcal{F}(A')\right\}$. Then we can compute an optimal solution for $(\mathrm{Q}^{\mathrm{fr}}(\widehat{p}))$, and its objective value, in $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ time.*

Proof. Recall from the discussion in Section 5.1 that we can reformulate $(\mathrm{Q}^{\mathrm{fr}}(\widehat{p}))$ as the following convex program:

$$\begin{aligned} \min \quad & c^{\intercal}x + ry + \sum_{A \in \mathcal{A}^{\mathrm{sup}}} \widehat{p}_A \mu_A \\ \text{s.t.} \quad & \mu_A \geq g(x, A') - y \cdot \ell(A, A') \qquad \forall A \in \mathcal{A}^{\mathrm{sup}}, A' \in \mathcal{A} \\ & x \in \mathcal{P}, \mu \in \mathbb{R}_+^{\mathcal{A}^{\mathrm{sup}}}, \ y \geq 0 \ . \end{aligned} \tag{5.6}$$

If $\mathcal{A}$ is collapsible under ℓ, then the exponentially many constraints in (5.6) are equivalent to

$$\mu_A \geq g(x, A') - y \cdot \ell(A, A') \qquad \forall A \in \mathcal{A}^{\mathrm{sup}}, A' \in \phi(A) \ . \tag{5.7}$$

Substituting constraints (5.6) with (5.7) and replacing $g(x, A')$ with its LP formulation,

we obtain the compact LP

$$\begin{aligned} \min \quad & c^\intercal x + ry + \sum_{A \in \mathcal{A}^{\sup}} \widehat{p}_A \mu_A \\ \text{s.t.} \quad & \mu_A \geq s^{A'} \cdot z^{A'} - y \cdot \ell(A, A') && \forall A \in \mathcal{A}^{\sup}, A' \in \phi(A) \\ & (x, z^{A'}) \in \mathcal{F}(A') && \forall A' \in \cup_{A \in \mathcal{A}^{\sup}} \phi(A) \\ & x \in \mathcal{P}, \mu \in \mathbb{R}_+^{\mathcal{A}^{\sup}}, \ y \geq 0 \ . \end{aligned}$$

Solving this LP yields an optimal solution for $(\mathrm{Q}^{\mathrm{fr}}(\widehat{p}))$. □

5.4 Some hardness results

Recall the k-max-min problem

$$\max_{A \in \mathcal{A}_{\leq k}} g(0, A) \ ,$$

where $\mathcal{A}_{\leq k} := \{A \subseteq U : |A| \leq k\}$ for some ground set U. For several underlying combinatorial-optimization problems, this problem is APX-hard (see the discussion in Section 2.1). In this section, we provide reductions that relate the difficulty of some of the tasks that we encounter when developing our framework for DRS optimization under a Wasserstein ball to that of the k-max-min problem. We focus on the explicit central-distribution setting, wherein we have a DRS problem of the type

$$\min_{x \in X} \{h(\widehat{p}; x) := c^\intercal x + z(\widehat{p}; x)\} \ , \tag{$\mathrm{Q}(\widehat{p})$}$$

and the central distribution $\widehat{p}$ is specified by the collection of pairs $\{(A, \widehat{p}_A)\}_{A \in \mathsf{supp}(\widehat{p})}$.

We show that under some conditions, given this two-stage DRS problem, the following tasks can capture the k-max-min problem (and hence are computationally hard if the latter problem is hard): (a) evaluating the objective function $h(\widehat{p}; x)$ of the DRS problem $(\mathrm{Q}(\widehat{p}))$; (b) solving the problem (Π) of computing $g(x, y, A) := \max_{A' \in \mathcal{A}} \{g(x, A') - y \cdot \ell(A, A')\}$, given $(x, y, A) \in X \times \mathbb{R}_+ \times \mathcal{A}$; and (c) evaluating the objective function $h(\widehat{p}; x, y) := c^\intercal x + ry + \mathbb{E}_{A \sim \widehat{p}}[g(x, y, A)]$ of the reformulation $(\mathrm{R}(\widehat{p}))$ of $(\mathrm{Q}(\widehat{p}))$ (obtained via Lemma 4.4).

Note that (a) and (c) are in contrast with classical two-stage stochastic optimization, where (using fractional second-stage decisions) evaluating the objective function of the problem is computationally hard when the probability distribution is given by a sampling oracle, but straightforward when it is given explicitly. For (b) and (c), we also show

impossibility of obtaining *any multiplicative approximation guarantee*. All the reductions in this section hold for two-stage DRS problems both in the k-bounded setting and in the unrestricted setting. Furthermore, the results for (a) and (c) hold even if the central distribution $\widehat{p}$ is restricted to having constant support size, which shows that, as alluded to above, these hardness results do not stem from the complexity of the central distribution $\widehat{p}$. (Note that (b) does not involve the distribution $\widehat{p}$.)

Theorem 5.9.
Consider the two-stage DRS problem under a Wasserstein ball $\min_{x\in X} h(\widehat{p}\,; x)$*, where the central distribution* $\widehat{p}$ *is given explicitly. Consider the following two settings:*

(B1) the k-bounded setting (i.e., $\mathcal{A} = \mathcal{A}_{\leq k}$) with ℓ as the discrete scenario metric ℓ^{disc}; and

(B2) the unrestricted setting (i.e., $\mathcal{A} = 2^U$) with ℓ given by: $\ell(A, A) = 0$ for all $A \in \mathcal{A}$; for $A \neq A' \in \mathcal{A}$, we have $\ell(A, A') = 1$ if $|A|, |A'| \leq k$, and $2Z$ otherwise, where $Z \geq \max_{\overline{A}\in\mathcal{A}} g\big(0, \overline{A}\big)$.

*Assume that $g(0, \emptyset) = 0$, and that the k-*max-min *problem*

$$\max_{A\in\mathcal{A}_{\leq k}} g(0, A) \tag{Ξ}$$

is NP*-hard and has optimal value at least* 1*. We have the following hardness results in both settings, assuming* P≠NP*.*

(a) One can choose the central distribution $\widehat{p}$ and the radius r so that the problem of computing $z(\widehat{p}\,; 0)$ is equivalent to (Ξ) (and hence admits no polytime algorithm).

(b) No polytime multiplicative approximation is possible for computing $g(0, y, \emptyset)$, given $y \geq 0$ in the input.

(c) By choosing the central distribution $\widehat{p}$ suitably, the hardness result in (b) carries over to the problem of computing $\mathbb{E}_{A\sim\widehat{p}}[g(0, y, A)]$, given $y \geq 0$ in the input.

Proof. Let $A^* \in \mathcal{A}_{\leq k}$ be an optimal solution for (Ξ), and $\mathrm{OPT}(\Xi) := g(0, A^*)$ be its objective value.

Part (a). For the setting (B1), it suffices to set $r := \ell_{\max}$ (and $\widehat{p}$ may be arbitrary). Then the ambiguity set $D = \{p : L(\widehat{p}, p) \le r\}$ encompasses *all* the distributions over $\mathcal{A}_{\le k}$. It follows that

$$z(\widehat{p}\,; 0) = \max_{p: L(\widehat{p},p) \le r} \mathbb{E}_{A\sim p}[g(0, A)] = \max_{A \in \mathcal{A}_{\le k}} g(0, A) \ .$$

For the setting (B2), we set $r := 1$, and take $\widehat{p}$ to be the distribution that puts weight of 1 on $\emptyset$ (and 0 on the remaining scenarios). We claim that $z(\widehat{p}\,; 0)$ is again equivalent to the problem (Ξ). Recall that $z(\widehat{p}\,; 0)$ is the optimal value of the LP

$$\begin{aligned} \max \quad & \sum_{A,A'} \gamma_{A,A'} g(0, A') && (\mathrm{T}(\widehat{p}, 0)) \\ \text{s.t.} \quad & \sum_{A'} \gamma_{A,A'} \le \widehat{p}_A \quad \forall A \in \mathcal{A} \\ & \sum_{A,A'} \ell(A, A') \gamma_{A,A'} \le r \\ & \gamma \ge 0 \ . \end{aligned}$$

Setting $\gamma_{\emptyset,A^*} = 1$, and setting the remaining variables to zero, yields a feasible solution for ($\mathrm{T}(\widehat{p}, 0)$) with objective value $g(0, A^*) = \mathrm{OPT}(\Xi)$. We now work toward showing that no feasible solution γ attains a higher objective value. Let α denote the amount of flow sent on the edges $(\emptyset, A')$ with $|A'| > k$. Note that this flow contributes $\alpha \cdot 2Z$ to the ℓ-cost of γ, and at most $\alpha \cdot Z$ to its objective value. Let θ be the amount of flow on the edge $(\emptyset, \emptyset)$. This does not contribute to the ℓ-cost of γ, or to its objective value. The flow on the remaining edges has volume $1 - \alpha - \theta$; it contributes $(1 - \alpha - \theta) \cdot 1$ to the ℓ-cost of γ and at most $(1 - \alpha - \theta) \cdot \mathrm{OPT}(\Xi)$ to its objective value. Therefore, the ℓ-cost of γ is $\alpha \cdot 2Z + (1 - \alpha - \theta) \cdot 1 = 2\alpha Z + (1 - \alpha - \theta)$; since γ is feasible for ($\mathrm{T}(\widehat{p}, 0)$), we obtain $2\alpha Z + (1 - \alpha - \theta) \le 1$. We can therefore bound the objective value of γ as follows:

$$\begin{aligned} \sum_{A,A'} \gamma_{A,A'} g(0, A') &\le \alpha \cdot Z + (1 - \alpha - \theta) \cdot \mathrm{OPT}(\Xi) \\ &\le \frac{\alpha + \theta}{2} + (1 - \alpha - \theta) \cdot \mathrm{OPT}(\Xi) \\ &\le \left(1 - \frac{\alpha + \theta}{2}\right) \cdot \mathrm{OPT}(\Xi) \\ &\le \mathrm{OPT}(\Xi) \ , \end{aligned}$$

where the third inequality follows because $\mathrm{OPT}(\Xi) \ge 1$ by assumption.

Part (b). We consider the setting (B1) first. First, note that

$$g(0,0,\emptyset) = \max_{A' \in \mathcal{A}_{\leq k}} \{g(0,A') - 0 \cdot \ell(\emptyset, A')\} = \max_{A' \in \mathcal{A}_{\leq k}} g(0,A) \ ,$$

and so for $y = 0$, computing $g(0,y,\emptyset)$ is equivalent to (Ξ).

By exploiting the mixed-sign objective, we can argue that any multiplicative approximation for $g(0,y,\emptyset)$ would allow us to solve the decision version of (Ξ): given $\nu \geq 0$, decide whether OPT(Ξ) $> \nu$. Suppose we have such an approximation algorithm, and run it with input $y = \nu$. If OPT(Ξ) $> \nu$, then we have

$$g(0,y,\emptyset) \geq g(0,A^*) - y \cdot \ell(\emptyset, A^*) > \nu - \nu \cdot 1 = 0 \ ,$$

and so the approximation algorithm would return a solution with a positive objective value. If on the contrary we have OPT(Ξ) $\leq \nu$, then for every scenario $A' \in \mathcal{A}_{\leq k}$ with $A' \neq \emptyset$ we have

$$g(0,A') - y \cdot \ell(\emptyset, A') = g(0,A') - \nu \leq 0 \ .$$

Since

$$g(0,\emptyset) - y \cdot \ell(\emptyset,\emptyset) = 0 - \nu \cdot 0 = 0 \ ,$$

we conclude that $g(0,y,\emptyset) = 0$, and so the approximation algorithm must return a solution with objective value 0. So we can distinguish between OPT(Ξ) $> \nu$ and OPT(Ξ) $\leq \nu$.

Now consider the setting (B2). Again, we suppose that we have an approximation algorithm for computing $g(0,y,\emptyset)$, and we show that given any $\nu \geq 0$, we can decide whether OPT(Ξ) $> \nu$. Since by assumption OPT(Ξ) ≥ 1, we may assume that $\nu \geq 1$, as otherwise the answer is clearly yes. Suppose we run the algorithm with input $y = \nu$. If OPT(Ξ) $> \nu$, then

$$g(0,y,\emptyset) \geq g(0,A^*) - y \cdot \ell(\emptyset, A^*) > \nu - \nu \cdot 1 = 0 \ ,$$

and so the approximation algorithm must return a solution with positive objective value. If on the contrary we have OPT(Ξ) $\leq \nu$, then we claim that $g(0,y,\emptyset) = 0$, and so the approximation algorithm must return a solution with objective value 0. Thus, we can distinguish between OPT(Ξ) $> \nu$ and OPT(Ξ) $\leq \nu$. To prove the claim, we consider three types of scenarios separately. For $A' = \emptyset$, we have

$$g(0,A') - y \cdot \ell(\emptyset, A') = 0 - \nu \cdot 0 = 0 \ ;$$

for $A' \neq \emptyset$ with $|A'| \leq k$, we have

$$g(0, A') - y \cdot \ell(\emptyset, A') \leq \nu - \nu \cdot 1 = 0 \ ;$$

finally, for scenarios A' with $|A'| > k$, we have

$$g(0, A') - y \cdot \ell(\emptyset, A') \leq Z - \nu \cdot 2Z = (1 - 2\nu)Z \leq 0 \ ,$$

where the last inequality follows because $\nu \geq 1$.

Part (c). This follows from part (b) by simply taking $\widehat{p}$ to be the distribution that puts a weight of 1 on the scenario $\emptyset$ (and 0 on the remaining scenarios); then for every $y \geq 0$ we have $\mathbb{E}_{A \sim \widehat{p}}[g(0, y, A)] = g(0, y, \emptyset)$, so the hardness result in part (b) carries over. $\square$

Chapter 6

DRS optimization under a Wasserstein ball: applications

In this chapter we demonstrate the versatility of our framework—i.e., Theorem 3.6 and the two main components used in its proof, namely Theorems 4.1 and 5.1—for handling general two-stage DRS problems under a Wasserstein ball by applying it to obtain the *first* approximation guarantees for the DRS versions of various combinatorial-optimization problems, namely set cover, vertex cover, edge cover, facility location, and Steiner tree. Except for set cover, our approximation factors are within constant factors of the guarantees known for the deterministic counterparts of these problems.

For convenience, we restate Theorem 3.6 below.

Theorem 3.6 (see proof in Section 6.1)**.**
Suppose that assumptions (A1)–(A7) hold, and that we have:

(1) a second-stage α-approximation algorithm;
(2) a (β_1, β_2)-approximation algorithm for problem (Π), with $\log \beta_1 = \mathsf{poly}(\mathcal{I})$; and
(3) a local ρ-approximation algorithm.

Then there exists a two-stage algorithm that, given $\varepsilon > 0$ and $\delta \in (0, 1)$, computes a $4\alpha\beta_1\beta_2\rho(1+\varepsilon)$-approximate solution for ($\mathrm{DRSO_W}$) with probability at least $1 - \delta$ in time $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log \frac{1}{\delta}\big)$.

We also restate assumptions (A1)–(A7) below.

(A1) $\log |X| = \mathsf{poly}(\mathcal{I})$.

(A2) We have $0 \in X$ and $g(0, A) \geq g(x, A)$ for every $x \in \mathcal{P}$ and $A \in \mathcal{A}$.

(A3) We know an *inflation factor* $\lambda \geq 1$ such that $g(0, A) \leq g(x, A) + \lambda c^\intercal x$ for every $x \in \mathcal{P}$ and $A \in \mathcal{A}$.

(A4) We have positive bounds $R_{\text{small}} \leq 1$ and R_{large} such that:

- $\mathcal{P}$ contains a Euclidean ball of radius R_{small};
- $\mathcal{P}$ is contained in the Euclidean ball of radius R_{large} centered at the origin; and
- $\log \frac{R_{\text{large}}}{R_{\text{small}}} = \mathsf{poly}(\mathcal{I})$.

(A5) For every scenario $A \in \mathcal{A}$, the function $x \mapsto g(x, A)$ is convex over $\mathcal{P}$. Furthermore, given any point $x \in \mathcal{P}$, we can compute in $\mathsf{poly}(\mathcal{I})$ time the value of this function at x *and* a subgradient d at x such that $\|d\| \leq K$, where $\log K = \mathsf{poly}(\mathcal{I})$. By Lemma 1.1, this also implies that the function $x \mapsto g(x, A)$ is K-Lipschitz continuous.

(A6) For every fractional first-stage decision $x \in \mathcal{P}$ and every non-null scenario $A \in \mathcal{A}$ we have $c^\intercal x + g(x, A) \geq 1$.

(A7) We have a number $\tau \geq 1$ with $\log \tau = \mathsf{poly}(\mathcal{I})$ such that $g(x, A') - g(x, A) \leq \tau \cdot \ell(A, A')$ for every fractional first-stage decision $x \in \mathcal{P}$ and every pair of scenarios (A, A') with $\ell(A, A') > 0$.

We defer the proof of Theorem 3.6 to Section 6.1, where we show that it can be obtained by combining Theorems 4.1 and 5.1. Here, we discuss what is needed in order to apply Theorem 3.6 and obtain results for our applications.

1. Verify that assumptions (A1)–(A7) hold. For the applications we consider, we have $X = \{0, 1\}^m$ and $\mathcal{P} = [0, 1]^m$, so assumptions (A1) and (A4) are readily satisfied by taking $R_{\text{small}} = \frac{1}{2}$ and $R_{\text{large}} = \sqrt{m}$. Assumptions (A2) and (A3) follow because all our applications are covering problems, and every first-stage action is associated with a corresponding second-stage action that is more expensive by a bounded factor. Hence, it is always possible to not take any first-stage action; taking more first-stage actions can never hurt the recourse cost, and (A3) holds by taking λ to be the maximum factor by which the cost of a first-stage action increases in the second stage. Assumption (A5) follows from prior work of Shmoys and Swamy [114], as the underlying two-stage problem falls into the class of two-stage LPs considered therein. Assumption (A6) typically holds if first-stage and second-stage decisions have integer costs. Assumption (A7) can

usually be satisfied by setting $\tau := \max\{1, \mathsf{UB}/\Delta\}$, where UB is a suitable upper bound on $\max_{A\in\mathcal{A}} g(0, A)$ and Δ is a lower bound on $\min_{A,A':\ell(A,A')>0} \ell(A, A')$; for any tuple (x, A, A') fulfilling the conditions of (A7) we obtain $g(x, A') - g(x, A) \le g(x, A') \le g(0, A') \le \mathsf{UB} \le \tau \cdot \ell(A, A')$.

2. Furnish the following algorithms.

   (a) A second-stage α-approximation algorithm: this typically reduces to obtaining an LP-based approximation algorithm for the deterministic version of the problem, and we can simply plug in known approximation results.

   (b) A local ρ-approximation algorithm: we have $\rho = 2\alpha$ for set cover, vertex cover, and edge cover, and $\rho = \mathrm{O}(1)$ for facility location (see Shmoys and Swamy [114]). We do not have such an algorithm for Steiner tree, but we have a weaker type of rounding algorithm for a *monotone* variant of the problem, which is sufficient in some settings—see Section 6.8.

   (c) A (β_1, β_2)-approximation algorithm for problem (Π), of computing $g(x, y, A)$, given $(x, y, A) \in X \times \mathbb{R}_+ \times \mathcal{A}$. This is a new component that we need to devise, whose design will depend on the scenario collection $\mathcal{A}$ and the scenario metric ℓ (and of course the underlying combinatorial-optimization problem). For various problems, we show how to obtain such an algorithm in the k-bounded setting by building upon results known for k-max-min problems.

Theorem 3.6 then shows that, for any $\varepsilon > 0$, we can obtain a $4\alpha\rho\beta_1\beta_2(1+\varepsilon)$-approximate solution for the *discrete* two-stage DRS problem (i.e., with integer first-stage and second-stage decisions) in $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}\big)$ time (and hence, sample complexity).

We consider two choices for the scenario collection $\mathcal{A}$: the k-bounded setting ($\mathcal{A} = \mathcal{A}_{\le k} := \{A \subseteq U : |\mathcal{A}| \le k\}$), and the unrestricted setting ($\mathcal{A} = 2^U$). We consider two choices for the underlying scenario metric ℓ. For all applications, we consider the *discrete metric* ℓ^{disc}, defined by $\ell^{\mathsf{disc}}(A, A') = 0$ if $A = A'$, and $\ell^{\mathsf{disc}}(A, A') = 1$ if $A \ne A'$. For facility location and Steiner tree, we also consider the *asymmetric metric* $\ell_\infty^{\mathsf{asym}}$. Recall that for a distance function w over the ground set U, this asymmetric metric is defined by $\ell_\infty^{\mathsf{asym}}(A, A') := \max_{j'\in A'} w(j', A)$, where $w(j', A) := \min_{j\in A} w_{j'j}$.

Table 6.1 summarizes the approximation factors that we obtain for DRS versions of a variety of applications. For all the nonempty fields in the table, except for Steiner tree, we are able to obtain approximation guarantees via our general framework (Theorem 3.6). However, some of the guarantees stated in Table 6.1 are obtained by adapting the general framework to obtain improved results by exploiting properties that hold in specific settings.

Also, for Steiner tree, we need to modify our approach somewhat because we do not have a local approximation algorithm. Figure 6.1 shows the different approaches we utilize to obtain our results.

Problem	ℓ^{disc}		$\ell_\infty^{\mathsf{asym}}$		General $\mathcal{A}$, ℓ
	$\mathcal{A} = 2^U$	$\mathcal{A}_{\leq k}$	$\mathcal{A} = 2^U$	$\mathcal{A}_{\leq k}$	β=approx. for (Π)
Facility location	21.96	196	21.96	196	$\mathrm{O}(\beta)$
Vertex cover	16	101.3	–	–	$\mathrm{O}(\beta)$
Edge cover	12	36	–	–	$\mathrm{O}(\beta)$
Set cover	$\mathrm{O}(\log \lvert U \rvert)$	$\mathrm{O}\big(\log^2 \lvert U \rvert\big)$	–	–	$\mathrm{O}(\beta \log \lvert U \rvert)$
Steiner tree	160	*	160	*	*

Table 6.1: A summary of the approximation factors we obtain for various applications in the Wasserstein setting. We have omitted the $\mathrm{O}(\varepsilon)$ terms that appear in the approximation factors. The $\ell_\infty^{\mathsf{asym}}$ setting does not apply to vertex cover, edge cover, and set cover. The approximation factor β for (Π) is the factor $\beta_1\beta_2$ in Theorem 3.6. The * entries are open questions.

Organization of this chapter. In Section 6.1, we prove our main result for DRS optimization under a Wasserstein ball (Theorem 3.6). In Section 6.2, we discuss how to obtain an approximation algorithm for problem (Π), which is one of the ingredients required to apply our general framework. We show that this reduces to obtaining an approximation algorithm for the constrained problem $\max_{A' \in \mathcal{A}: \ell(A,A') \leq \mu} g(x, A')$, given $(x, \mu, A) \in X \times \mathbb{R}_+ \times \mathcal{A}$ (see Lemma 6.2). In Section 6.3, we present an alternative approach for solving DRS problems under a Wasserstein ball in the unrestricted setting, which yields improved approximation factors. Then, each of the Sections 6.4–6.8 delves into a specific application.

6.1 Proof of Theorem 3.6

In this section, we show how Theorems 4.1 and 5.1 can be utilized to prove Theorem 3.6. We need to do a little more work than directly combining Theorems 4.1 and 5.1, since the latter theorems involve *multiplicative + additive* approximations, whereas Theorem 3.6 aims for a purely multiplicative approximation. To overcome this difficulty, we exploit assumption (A6) to obtain a lower bound on the optimal value of $\min_{x \in \mathcal{P}} h(\widetilde{p}; x)$ for every distribution $\widetilde{p}$, and then use this lower bound to convert the additive errors into multiplicative errors.

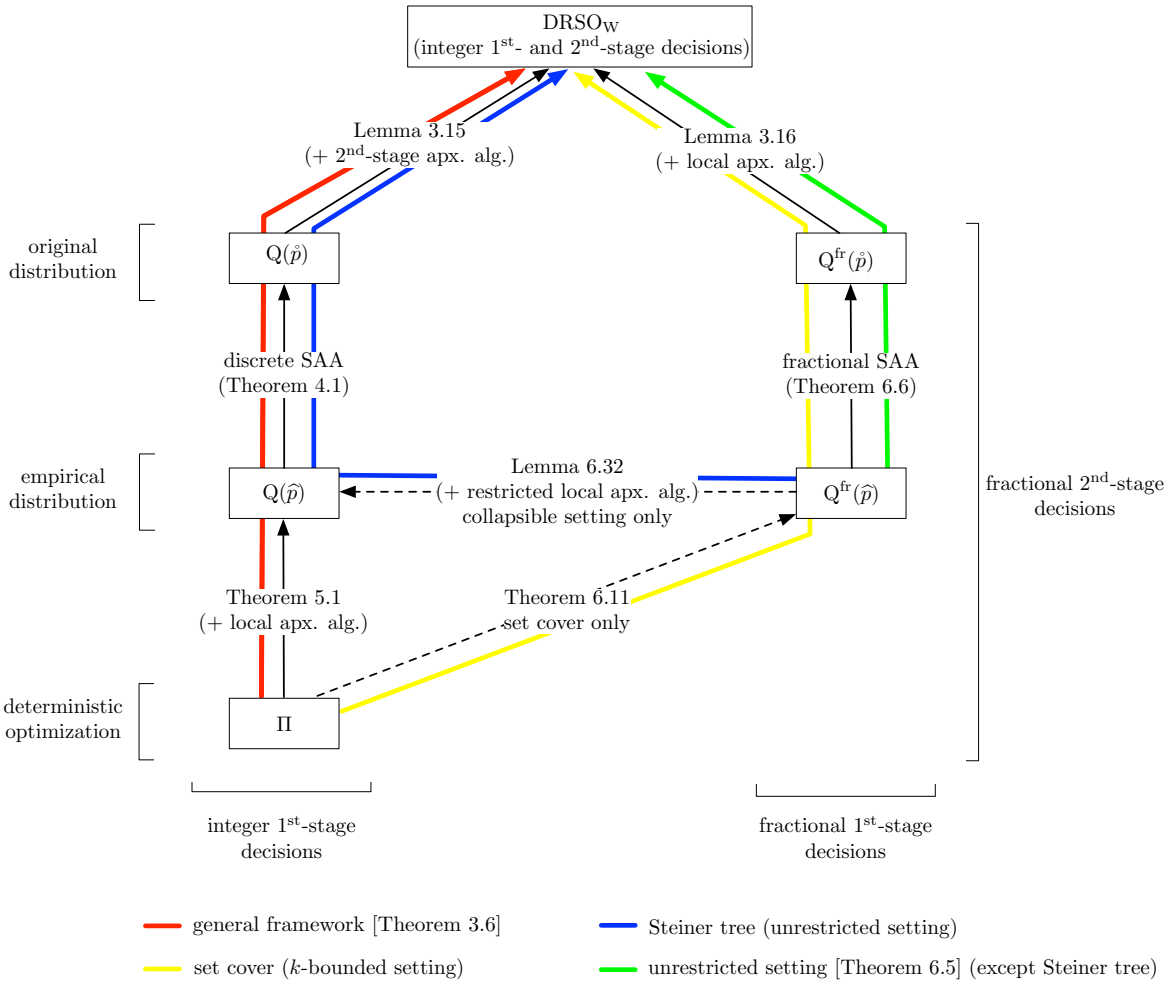

Figure 6.1: The collection of problem reductions utilized by our frameworks for DRS optimization under a Wasserstein ball. A black arrow from a problem to another indicates that an approximation algorithm for the first one can be used to obtain an approximation algorithm for the second one; the label indicates the relevant result and any additional ingredients required by the reduction. Solid black arrows indicate general reductions; dashed black arrows indicate reductions that apply only to special cases. The red arrow indicates the sequence of reductions utilized by our general framework; the remaining (green, blue, and yellow) arrows indicate alternative paths used in some applications.

Lemma 6.1. *Suppose that we have a (β_1, β_2)-approximation algorithm for problem (Π). Then in $\mathsf{poly}(\mathcal{I})$ time we can either (i) determine that $x = 0$ is an optimal solution for $\min_{x\in\mathcal{P}} h(\widetilde{p}\,; x)$ for every distribution $\widetilde{p}$; or (ii) obtain a lower bound LB on $\min_{x\in\mathcal{P}} h(\widetilde{p}\,; x)$ for every distribution $\widetilde{p}$, where $\log \frac{1}{\mathsf{LB}} = \mathsf{poly}(\mathcal{I})$.*

Proof. We first show that if $\mathcal{A}$ contains any non-null scenario, then $\min_{x\in\mathcal{P}} h(\widetilde{p}\,; x) \geq \frac{r}{\ell_{\max}}$ for every distribution $\widetilde{p}$; otherwise $x = 0$ is an optimal solution for $\min_{x\in\mathcal{P}} h(\widetilde{p}\,; x)$ for every $\widetilde{p}$. Recall that a null scenario is a scenario A such that $g(x, A) = g(0, A)$ for every fractional first-stage decision $x \in \mathcal{P}$.

Suppose that $A^* \in \mathcal{A}$ is a non-null scenario. For any fractional first-stage decision $x \in \mathcal{P}$ and any distribution $\widetilde{p}$, there is a feasible solution γ for $(\mathrm{T}(\widetilde{p}, x))$ that sends at least $\frac{r}{\ell_{\max}}$ flow to A^*, i.e., $\sum_{A\in\mathcal{A}} \gamma_{A,A^*} \geq \frac{r}{\ell_{\max}}$. It follows that $z(\widetilde{p}\,; x) \geq \frac{r}{\ell_{\max}} \cdot g(x, A^*)$, and so

$$h(\widetilde{p}\,; x) \geq c^\intercal x + \frac{r}{\ell_{\max}} g(x, A^*) \geq \frac{r}{\ell_{\max}} \cdot (c^\intercal x + g(x, A^*)) \geq \frac{r}{\ell_{\max}} \ .$$

The second inequality follows because $r \leq \ell_{\max}$ by assumption, and the third inequality follows from assumption (A6) as A^* is a non-null scenario. This shows that $\min_{x\in\mathcal{P}} h(\widetilde{p}\,; x) \geq \frac{r}{\ell_{\max}}$ for every distribution $\widetilde{p}$.

Now, suppose that all scenarios in $\mathcal{A}$ are null scenarios. Then we have $g(x, A) = g(0, A)$ for every $x \in \mathcal{P}, A \in \mathcal{A}$, and so $z(\widetilde{p}\,; x) = z(\widetilde{p}\,; 0)$ for every $x \in \mathcal{P}$ and for every distribution $\widetilde{p}$. It follows that $x = 0$ is an optimal solution for $\min_{x\in\mathcal{P}} h(\widetilde{p}\,; x)$ for every distribution $\widetilde{p}$.

While we will not quite be able to determine if $\mathcal{A}$ contains a non-null scenario, we can do the following. We run the (β_1, β_2)-approximation algorithm for problem (Π) with input $(x, y, A) = (0, 0, A)$, for an arbitrary scenario $A \in \mathcal{A}$. Let $\overline{A}$ be the output obtained. If $g\big(0, \overline{A}\big) < \frac{1}{\beta_1}$, then we claim that (i) holds; otherwise, we claim that (ii) holds for $\mathsf{LB} = \frac{r}{\beta_1 \ell_{\max}}$.

We now argue that the algorithm is correct. First, suppose that $g\big(0, \overline{A}\big) < \frac{1}{\beta_1}$. Because of the guarantee of the approximation algorithm for (Π), we have

$$g\big(0, \overline{A}\big) - 0 \cdot \ell\big(A, \overline{A}\big) \geq \max_{A'\in\mathcal{A}} \left\{ \frac{1}{\beta_1} g(0, A') - \beta_2 \cdot 0 \cdot \ell(A, A') \right\} \ ,$$

which reduces to $g\big(0, \overline{A}\big) \geq \frac{1}{\beta_1} \max_{A'\in\mathcal{A}} g(0, A')$. It follows that $g(0, A') \leq \beta_1 \cdot g\big(0, \overline{A}\big) < 1$ for every scenario $A' \in \mathcal{A}$. This implies that all the scenarios are null, since a non-null scenario A must satisfy $c^\intercal 0 + g(0, A') \geq 1$. As we have shown above, when all the scenarios are null scenarios, (i) holds.

Now, consider the case where $g\big(0, \overline{A}\big) \geq \frac{1}{\beta_1}$. We have two subcases to consider. First, suppose that there exists a non-null scenario. Then we have shown that $\frac{r}{\ell_{\max}}$ is a lower bound on $\min_{x \in \mathcal{P}} h(\widetilde{p}\,; x)$ for every distribution $\widetilde{p}$. Since $\frac{r}{\ell_{\max}} \geq \frac{r}{\beta_1 \ell_{\max}}$ as $\beta_1 \geq 1$, it follows that (ii) holds as claimed. The second subcase is when all the scenarios are null. As we have shown above, this implies that $x = 0$ is an optimal solution for $\min_{x \in \mathcal{P}} h(\widetilde{p}\,; x)$ for every distribution $\widetilde{p}$. Note that there exists a feasible solution for $(\mathrm{T}(\widetilde{p}, 0))$ that sends at least $\frac{r}{\ell_{\max}}$ flow to $\overline{A}$, and so

$$\min_{x \in \mathcal{P}} h(\widetilde{p}\,; x) = h(\widetilde{p}\,; 0) = z(\widetilde{p}\,; 0) \geq \frac{r}{\ell_{\max}} g\big(0, \overline{A}\big) \geq \frac{r}{\beta_1 \ell_{\max}} \ ,$$

which yields (ii). □

We are now ready to prove Theorem 3.6.

Proof of Theorem 3.6. We first run the algorithm from Lemma 6.1 to either (i) determine that $x = 0$ is an optimal solution for $\min_{x \in \mathcal{P}} h(\widetilde{p}\,; x)$ for every distribution $\widetilde{p}$; or (ii) obtain a lower bound LB on $\min_{x \in \mathcal{P}} h(\widetilde{p}\,; x)$ for every distribution $\widetilde{p}$, where $\log \frac{1}{\mathsf{LB}} = \mathsf{poly}(\mathcal{I})$.

We claim that in either case we can obtain a $4\beta_1\beta_2\rho(1+\varepsilon)$-approximate solution for $(\mathrm{Q}(\mathring{p}))$ with probability at least $1-\delta$. We then use the second-stage α-approximation algorithm to obtain integer second-stage decisions, and the theorem follows from Lemma 3.15.

In case (i), we have that $x = 0$ is an optimal solution of $\min_{x \in \mathcal{P}} h(\mathring{p}\,; x)$; since this is a relaxation of $(\mathrm{Q}(\mathring{p}))$ (and since $0 \in X$), it follows that $\widehat{x} = 0$ is an optimal solution for $(\mathrm{Q}(\mathring{p}))$.

Now, suppose that we are in case (ii). Let $\varepsilon' := \varepsilon/5$, and assume without loss of generality that $\varepsilon' \leq 1$. We let k and N be given by Theorem 4.1, with parameters $(\varepsilon', \eta := \varepsilon' \cdot \mathsf{LB}, \delta)$, and construct empirical distributions $\widehat{p}^1, \ldots, \widehat{p}^k$, each using N independent samples. For each empirical distribution $\widehat{p}^i$, we run the algorithm from Theorem 5.1 to obtain an integer first-stage decision $\widehat{x}^i \in X$ and an estimate f^i such that

$$h\big(\widehat{p}\,; \widehat{x}^i\big) \leq f^i \leq \beta_1\beta_2\rho\left(\min_{x \in X} h\big(\widehat{p}^{\,i}\,; x\big) + \eta\right) \leq \beta_1\beta_2\rho(1+\varepsilon') \cdot \min_{x \in X} h\big(\widehat{p}^{\,i}\,; x\big) \ ,$$

where the final inequality follows because

$$\eta = \varepsilon' \cdot \mathsf{LB} \leq \varepsilon' \cdot \min_{x \in \mathcal{P}} h\big(\widehat{p}^{\,i}\,; x\big) \leq \varepsilon' \cdot \min_{x \in X} h\big(\widehat{p}^{\,i}\,; x\big) \ . \tag{6.1}$$

Let $j := \operatorname{argmin}_{i\in[k]} f^i$. By Theorem 4.1 (setting $\psi = \beta_1\beta_2\rho(1+\varepsilon')$), it follows that with probability at least $1-\delta$ we have

$$\begin{aligned} h\big(\mathring{p}\,;\widehat{x}^j\big) &\le 4\beta_1\beta_2\rho(1+\varepsilon')^2 \cdot \min_{x\in X} h(\mathring{p}\,;x) + \beta_1\beta_2\rho(1+\varepsilon')\eta \\ &\le 4\beta_1\beta_2\rho\left((1+\varepsilon')\left(1+\frac{5}{4}\varepsilon'\right)\right) \cdot \min_{x\in X} h(\mathring{p}\,;x) \\ &\le 4\beta_1\beta_2\rho(1+\varepsilon) \cdot \min_{x\in X} h(\mathring{p}\,;x)\ , \end{aligned}$$

which proves the claim. The second step follows from (6.1), and the final step uses the fact that $\varepsilon' = \frac{\varepsilon}{5} \le 1$.

Note that the running time is $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$ as claimed because we solve $k = \mathsf{poly}\big(\frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$ SAA problems, each with input size $\widehat{\mathcal{I}} = \mathsf{poly}(\mathcal{I}, N) = \mathsf{poly}\Big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\eta}, \log\frac{1}{\delta}\Big)$. Each call to the algorithm from Theorem 5.1 takes $\mathsf{poly}\Big(\widehat{\mathcal{I}}, \log\frac{1}{\eta}\Big) = \mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$ time, since $\log\frac{1}{\eta} = \mathsf{poly}\big(\mathcal{I}, \log\frac{1}{\varepsilon}\big)$. □

6.2 Obtaining an approximation algorithm for (Π)

Recall that in order to apply Theorem 3.6 to a DRS problem, we need to furnish an approximation algorithm for the following problem.

(Π) Given an integer first-stage decision $x \in X, y \ge 0$, and a scenario $A \in \mathcal{A}$, solve

$$g(x,y,A) := \max_{A'\in\mathcal{A}} \{g(x,A') - y\cdot\ell(A,A')\}\ .$$

In this section, we show that this can be obtained via an approximation algorithm for the constrained problem

$$\max_{A'\in\mathcal{A}:\ell(A,A')\le\mu} g(x,A')\ , \tag{Φ}$$

given $(x,\mu,A) \in X\times\mathbb{R}_+\times\mathcal{A}$.

Lemma 6.2. *Let $\mathcal{L}\subseteq\mathbb{R}_+$ be a finite set containing all the pairwise scenario distances $\{\ell(A,A')\}_{A,A'\in\mathcal{A}}$. Let $(x,y,A)\in\mathcal{P}\times\mathbb{R}_+\times A$. For each $\mu\in\mathcal{L}$, let A_μ be a β-approximate solution for the instance (x,μ,A) of problem (Φ), and let*

$$\mu^* := \operatorname{argmax}_{\mu\in\mathcal{L}} \{g(x,A_\mu) - y\cdot\ell(A,A_\mu)\}\ .$$

Then we have

$$g(x, A_{\mu^*}) - y \cdot \ell(A, A_{\mu^*}) \geq \max_{A' \in \mathcal{A}} \left\{ \frac{1}{\beta} g(x, A') - y \cdot \ell(A, A') \right\} .$$

Proof. For every scenario $A' \in \mathcal{A}$, letting $\mu' := \ell(A, A')$, we have

$$g(x, A_{\mu^*}) - y \cdot \ell(A, A_{\mu^*}) \geq g(x, A_{\mu'}) - y \cdot \ell(A, A_{\mu'}) \geq \frac{1}{\beta} g(x, A') - y \cdot \ell(A, A') .$$

The first inequality follows from the definition of μ^*, and the second follows from the definition of $A_{\mu'}$. □

For all applications and choices of scenario metrics ℓ that we consider, we can construct a set $\mathcal{L}$ as in Lemma 6.2 with $|\mathcal{L}| = \mathsf{poly}(\mathcal{I})$. Lemma 6.2 shows that we can utilize a β-approximation algorithm for (Φ) to obtain a $(\beta, 1)$-approximation algorithm for problem (Π). We remark that this reduction can be generalized to yield a $(\beta, 1+\varepsilon)$-approximation algorithm for problem (Π) using $\mathrm{O}\left(\log_{1+\varepsilon} \frac{\ell_{\max}}{\ell_{\min}}\right)$ calls to an approximation algorithm for problem (Φ), where $\ell_{\max}$ is an upper bound on $\max_{A,A'} \ell(A, A')$, and $\ell_{\min}$ is a positive lower bound on $\min_{A,A':\ell(A,A')>0} \ell(A, A')$.

6.3 Improved results in the unrestricted setting: a reduction from ($\mathrm{DRSO_W}$) to the fractional SAA problem ($\mathrm{Q^{fr}}(\widehat{p})$)

In this section, we show that for covering problems in the unrestricted setting (i.e., $\mathcal{A} = 2^U$), we can improve upon our general framework (i.e., Theorem 3.6), and obtain a $4\rho(1+\varepsilon)$-approximation using a local ρ-approximation algorithm for the underlying combinatorial-optimization problem.

The improvement comes from two sources. First, in the unrestricted setting and for the scenario metrics of interest, we are able to solve the fractional SAA problem ($\mathrm{Q^{fr}}(\widehat{p})$) *exactly*; this will follow from Theorem 5.8, since we show that $\mathcal{A}$ is collapsible for these scenario metrics (see Lemma 6.4). Second, we give a better and more direct reduction from ($\mathrm{DRSO_W}$) to the fractional SAA problem (see Theorem 6.3). Given optimal solutions $\{\overline{x}^i\}$ for ($\mathrm{Q^{fr}}(\widehat{p})$) (for $\widehat{p} = \widehat{p}^1, \widehat{p}^2, \dots$), instead of rounding these solutions to integer

first-stage decisions using a local approximation algorithm, then utilizing the SAA result of Theorem 4.1, and finally obtaining integer second-stage actions using a second-stage rounding algorithm, we proceed as follows. We prove an analogue of Theorem 4.1 for the *fractional* SAA problem showing that we can use the solutions $\{\overline{x}^i\}$ to obtain an approximate solution for ($\mathrm{Q}^{\mathrm{fr}}(\mathring{p})$) (see Theorem 6.6), and then use a local approximation algorithm to round this solution and obtain integer first-stage and second-stage decisions.

Theorem 6.3. *Suppose that we have:*

(1) a local ρ-approximation algorithm; and

(2) a $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$-time algorithm that computes a fractional first-stage decision $\overline{x} \in \mathcal{P}$ and an estimate f such that

$$h(\widehat{p}\,;\overline{x}) \leq f \leq \psi \cdot \min_{x \in \mathcal{P}} h(\widehat{p}\,;x) \ .$$

Then there exists a two-stage algorithm that, given $\varepsilon > 0$ and $\delta \in (0,1)$, computes a $4\rho\psi(1+\varepsilon)$-approximate solution for ($\mathrm{DRSO}_{\mathrm{W}}$) with probability at least $1-\delta$ in time $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$.

Lemma 6.4. *Suppose that we are in the unrestricted setting (i.e., $\mathcal{A} = 2^U$), and that for every fractional first-stage decision $x \in \mathcal{P}$ and every pair of scenarios $A, A' \in \mathcal{A}$ with $A \subseteq A'$ we have $g(x,A) \leq g(x,A')$. Then the collection of scenarios $\mathcal{A}$ is collapsible under the discrete metric ℓ^{disc} and the asymmetric metric $\ell_\infty^{\mathsf{asym}}$.*

Proof. Let $A \in \mathcal{A}$ be an arbitrary scenario. If ℓ is the discrete metric ℓ^{disc}, then we set $\phi(A) := \{A, U\}$. If ℓ is the asymmetric metric $\ell_\infty^{\mathsf{asym}}$, defined with respect to the metric w over the ground set U, then we set

$$\phi(A) := \{\{j \in U : w(j,A) \leq \mu\} : \mu \in \mathcal{L}\}\,,$$

where $\mathcal{L} := \{w_{jj'} : j, j' \in U\}$ is the set of all the pairwise distances over the ground set. Note that in both settings, if we choose an arbitrary pair $(x,\mu) \in \mathcal{P} \times \mathbb{R}_+$, the collection of scenarios $\phi(A)$ contains the (unique) maximal solution A' for the instance (x,μ,A) of the constrained problem (Φ). By the monotonicity property of the second-stage costs $g(\cdot,\cdot)$ imposed in the lemma statement, A' is optimal for this instance. By Lemma 6.2, it follows that $\phi(A)$ contains an optimal solution for $\max_{A' \in \mathcal{A}} \{g(x,A') - y \cdot \ell(A,A')\}$ for every $(x,y) \in \mathcal{P} \times \mathbb{R}_+$. It follows that $\mathcal{A}$ is collapsible under ℓ. □

We now show that combining Theorem 6.3 with Theorem 5.8 and Lemma 6.4, we obtain an improved $4\rho(1+\varepsilon)$ approximation factor for the discrete DRS problem ($\mathrm{DRSO}_{\mathrm{W}}$) in the unrestricted setting, for certain choices of the scenario metric ℓ.

Theorem 6.5. *Suppose that we are in the unrestricted setting (i.e., $\mathcal{A} = 2^U$), that the scenario metric ℓ is either the discrete metric ℓ^{disc} or the asymmetric metric $\ell_\infty^{\mathsf{asym}}$, and that the second-stage costs $\{g(x, A')\}$ are given by compact LPs, say, $g(x, A') = \min\{s^{A'} \cdot z^{A'} : (x, z^{A'}) \in \mathcal{F}(A')\}$. Moreover, suppose that we have a local ρ-approximation algorithm. Then there exists a two-stage algorithm that, given $\varepsilon > 0$ and $\delta \in (0,1)$, computes a $4\rho(1+\varepsilon)$-approximate solution for (DRSO$_{\mathrm{W}}$) with probability at least $1-\delta$ in time $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$.*

Proof. By Lemma 6.4, we have that $\mathcal{A}$ is collapsible under ℓ. Theorem 5.8 then implies that we can compute an optimal solution for the fractional SAA problem (Q$^{\mathrm{fr}}(\widehat{p})$), as well as its objective value. The result then follows from Theorem 6.3, setting $\psi = 1$. □

In the remainder of this section, we focus on proving Theorem 6.3. We use the following variant of Theorem 4.1, which allows translating approximate solutions for SAA problems $\min_{x\in\mathcal{P}} h(\widehat{p}^{\,i}\,; x)$ into an approximate solution for $\min_{x\in\mathcal{P}} h(\mathring{p}\,; x)$.

Theorem 6.6. *Let $\varepsilon > 0$ and $\delta \in (0,1)$, and let $\mathsf{LB} > 0$ be a lower bound on $\min_{x\in\mathcal{P}} h(\widetilde{p}\,; x)$ for every distribution $\widetilde{p}$, with $\log\frac{1}{\mathsf{LB}} = \mathsf{poly}(\mathcal{I})$. There exist numbers $k = \mathsf{poly}\big(\frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$ and $N = \mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$ such that the following holds. Let $\widehat{p}^{\,1}, \ldots, \widehat{p}^{\,k}$ be empirical estimates of $\mathring{p}$, each constructed using N independent samples. Suppose that for each $i \in [k]$ we have a fractional first-stage decision $\overline{x}^i \in \mathcal{P}$ and an estimate $\overline{f}^{\,i}$ such that*

$$h\big(\widehat{p}^{\,i}\,; \overline{x}^i\big) \le \overline{f}^{\,i} \le \overline{\psi} \cdot \min_{x\in\mathcal{P}} h\big(\widehat{p}^{\,i}\,; x\big) \ ,$$

where $\overline{\psi} \ge 1$. Let $j := \operatorname{argmin}_{i\in[k]} \overline{f}^{\,i}$. Then with probability at least $1-\delta$ we have

$$h\big(\mathring{p}\,; \overline{x}^j\big) \le 4\overline{\psi}(1+\varepsilon) \cdot \min_{x\in\mathcal{P}} h(\mathring{p}\,; x) \ .$$

Proof. The result follows by applying Theorem 4.1 to a suitable discretization of the polytope $\mathcal{P}$. Let $\varepsilon' := \varepsilon/9$, and assume without loss of generality that $\varepsilon' \le 1$. By Lemma 5.6, we have that $h(\widetilde{p}\,; \cdot)$ is $\widetilde{K}$-Lipschitz continuous for every distribution $\widetilde{p}$, where $\widetilde{K} := \|c\| + K$. Recall that by assumption (A4), $\mathcal{P}$ contains a ball of radius $R_{\mathrm{small}} \le 1$ and is contained in a ball of radius R_{large} centered at the origin, with $\log\frac{R_{\mathrm{large}}}{R_{\mathrm{small}}} = \mathsf{poly}(\mathcal{I})$. We discretize $\mathcal{P}$ as in Swamy and Shmoys [123]. Let $\Delta = \frac{\varepsilon' \cdot \mathsf{LB} \cdot R_{\mathrm{small}}}{8\widetilde{K} R_{\mathrm{large}}\sqrt{m}}$, and consider the grid

$$\mathcal{G} := \{x \in \mathcal{P} : x_i = n_i\Delta, \quad n_i \in \mathbb{Z}_+ \text{ for all } i \in [m]\} \ .$$

(Note that R_{small} needs to be a part of the specification of the grid step size; otherwise, a "flat" polytope $\mathcal{P}$ could evade the grid across arbitrarily large distances.) As shown in [123], we have: (i) $|\mathcal{G}| \leq \left(\frac{2R_{\text{large}}}{\Delta}\right)^m$, and so $\log|\mathcal{G}| = \mathsf{poly}\big(\mathcal{I}, \log\frac{1}{\varepsilon}\big)$ (since $m, \log\widetilde{K}, \log\frac{1}{\mathsf{LB}}$, and $\log\frac{R_{\text{large}}}{R_{\text{small}}}$ are all $\mathsf{poly}(\mathcal{I})$); and (ii) for every $x \in \mathcal{P}$, letting $\phi(x)$ denote the point in $\mathcal{G}$ closest to x in Euclidean distance, we have $\|x - \phi(x)\| \leq \frac{\varepsilon'\mathsf{LB}}{\widetilde{K}}$, and hence, $|h(\widetilde{p}\,; x) - h(\widetilde{p}\,; \phi(x))| \leq \varepsilon'\mathsf{LB}$ for every distribution $\widetilde{p}$.

Consider the two-stage DRS problem $\min_{x\in\mathcal{G}} h(\mathring{p}\,; x)$, and note that it satisfies properties (A1)–(A7). Properties (A2)–(A7) are directly inherited from the DRS problem $\min_{x\in X} h(\mathring{p}\,; x)$, so it suffices to show property (A1). Let $\mathcal{I}'$ denote the size of the input of the DRS problem over the grid $\mathcal{G}$, which consists of the input of the original DRS problem over X along with the parameters ε' and LB (which are used for determining the grid step size Δ). Since $\log|\mathcal{G}| = \mathsf{poly}\big(\mathcal{I}, \log\frac{1}{\varepsilon}\big)$, we have $\log|\mathcal{G}| = \mathsf{poly}(\mathcal{I}')$.

Let the number k of SAA problems and the number N of samples used to construct each empirical estimate $\widehat{p}^{\,i}$ be as given by Theorem 4.1, when we apply it to the two-stage DRS problem $\min_{x\in\mathcal{G}} h(\mathring{p}\,; x)$ (i.e., we take X in the theorem statement to be the grid $\mathcal{G}$), with parameters $(\varepsilon', \eta := \varepsilon'\mathsf{LB}, \delta)$. To invoke the theorem, we need to supply the required points $\{\widehat{x}^{\,i}\}$ and estimates $\{f^i\}$. We set $\widehat{x}^{\,i} := \phi\big(\overline{x}^{\,i}\big)$ and $f^i := \overline{f}^{\,i} + \varepsilon'\mathsf{LB}$ for every $i \in [k]$. We claim that these satisfy the precondition in the statement of Theorem 4.1, with $\psi = \overline{\psi} + \varepsilon'$. To see this, consider any $i \in [k]$. On the one hand, we have

$$h\big(\widehat{p}^{\,i}\,; \widehat{x}^{\,i}\big) \leq h\big(\widehat{p}^{\,i}\,; \overline{x}^{\,i}\big) + \varepsilon'\mathsf{LB} \leq \overline{f}^{\,i} + \varepsilon'\mathsf{LB} = f^i \ .$$

On the other hand, we have

$$f^i = \overline{f}^{\,i} + \varepsilon'\mathsf{LB} \leq \overline{\psi} \cdot \min_{x\in\mathcal{P}} h\big(\widehat{p}^{\,i}\,; x\big) + \varepsilon'\mathsf{LB} \leq \big(\overline{\psi} + \varepsilon'\big) \cdot \min_{x\in\mathcal{P}} h\big(\widehat{p}^{\,i}\,; x\big) \leq \psi \cdot \min_{x\in\mathcal{G}} h\big(\widehat{p}^{\,i}\,; x\big) \ .$$

Moreover, the index j, which is a minimizer for the estimates $\left\{\overline{f}^{\,i}\right\}_{i\in[k]}$, is also a minimizer for the new estimates $\{f^i\}_{i\in[k]}$. So applying Theorem 4.1, we obtain that with probability at least $1 - \delta$,

$$\begin{aligned} h\big(\mathring{p}\,; \widehat{x}^{\,j}\big) &\leq 4\psi(1+\varepsilon') \cdot \min_{x\in\mathcal{G}} h(\mathring{p}\,; x) + \psi\eta \\ &= 4\psi(1+\varepsilon') \cdot \min_{x\in\mathcal{G}} h(\mathring{p}\,; x) + \psi\varepsilon'\mathsf{LB} \ . \end{aligned}$$

Note that

$$\min_{x\in\mathcal{G}} h(\mathring{p}\,;x) \le \min_{x\in\mathcal{P}} h(\mathring{p}\,;\phi(x)) \le \min_{x\in\mathcal{P}} h(\mathring{p}\,;x) + \varepsilon'\mathsf{LB}\ .$$

Therefore, we obtain

$$\begin{aligned} h\big(\mathring{p}\,;\widehat{x}^j\big) &\le 4\psi(1+\varepsilon')\cdot\min_{x\in\mathcal{G}} h(\mathring{p}\,;x) + \psi\varepsilon'\mathsf{LB} \\ &\le 4\psi(1+\varepsilon')\cdot\min_{x\in\mathcal{P}} h(\mathring{p}\,;x) + (\psi\varepsilon' + 4\psi(1+\varepsilon')\varepsilon')\mathsf{LB} \\ &\le 4\overline{\psi}(1+\varepsilon)\cdot\min_{x\in\mathcal{P}} h(\mathring{p}\,;x)\ . \end{aligned}$$

The final step follows because $\mathsf{LB} \le \min_{x\in\mathcal{P}} h(\mathring{p}\,;x)$, and also uses $\psi = \overline{\psi} + \varepsilon' \le \overline{\psi}(1+\varepsilon')$ and $\varepsilon' = \frac{\varepsilon}{9} \le 1$. □

Proof of Theorem 6.3. The proof follows very closely the structure of the proof of Theorem 3.6, so we omit some details. Using Lemma 6.1, we can either find an optimal solution $\overline{x}$ for $\min_{x\in\mathcal{P}} h(\mathring{p}\,;x)$, or obtain a lower bound LB on $\min_{x\in\mathcal{P}} h(\widetilde{p}\,;x)$ for every distribution $\widetilde{p}$. In the latter case, given that we have a ψ-approximation algorithm for the fractional SAA problem $\min_{x\in\mathcal{P}} h(\widehat{p}\,;x)$, we invoke the SAA result from Theorem 6.6 (setting $\overline{\psi} = \psi$) to obtain a $4\psi(1+\varepsilon)$-approximate solution $\overline{x}$ for $\min_{x\in\mathcal{P}} h(\mathring{p}\,;x)$. Using the local approximation algorithm to round $\overline{x}$, we obtain a $4\rho\psi(1+\varepsilon)$-approximate solution for ($\mathrm{DRSO}_{\mathrm{W}}$) by Lemma 3.16. □

6.4 DRS set cover

In two-stage DRS set cover (DRSSC), we have a collection $\mathcal{S}$ of subsets of a ground set U. A scenario is a subset of U, and specifies the set of elements to be covered in that scenario. We may buy a set $S \in \mathcal{S}$ in either stage, incurring costs of $c^{\mathrm{I}}_S \in \mathbb{Z}_+$ and $c^{\mathrm{II}}_S \in \mathbb{Z}_+$ in the first and in the second stage respectively. For each scenario A, the sets chosen in the first stage and in scenario A (in the second stage) must together cover A. The goal is to choose some first-stage sets $\mathcal{S}^{\mathrm{I}} \subseteq \mathcal{S}$ and sets $\mathcal{S}^A \subseteq \mathcal{S}$ in each scenario A so as to minimize

$$\sum_{S\in\mathcal{S}^{\mathrm{I}}} c^{\mathrm{I}}_S + \sup_{p:L(\mathring{p},p)\le r} \mathbb{E}_{A\sim p}\left[\sum_{S\in\mathcal{S}^A} c^{\mathrm{II}}_S\right].$$

The input size $\mathcal{I}$ is the encoding size of $\big(U,\mathcal{S},c^{\mathrm{I}},c^{\mathrm{II}},r\big)$. We may assume that $\bigcup_{S\in\mathcal{S}} S = U$, as otherwise the problem is infeasible.

It is easy to see that DRSSC can be cast as an instance of (DRSO): the first-stage decision set X and the second-stage decision set Z are $X = Z = \{0,1\}^U$; the corresponding sets of fractional first-stage and second-stage decisions are given by the polytopes $\mathcal{P} = \mathcal{Z} = [0,1]^U$. The polytope specifying the feasibility conditions for a scenario A is

$$\mathcal{F}(A) := \left\{ \left(x, z^A\right) \in \mathcal{P} \times \mathcal{Z} : \sum_{S \in \mathcal{S}: e \in S} \left(x_S + z_S^A\right) \geq 1 \ \forall e \in A \right\}.$$

The inflation factor λ is defined as $\max\left\{1, \max_{S \in \mathcal{S}} \frac{c_S^{\mathrm{II}}}{c_S^{\mathrm{I}}}\right\}$. Note that we may assume that $c_S^{\mathrm{II}} = 0$ if $c_S^{\mathrm{I}} = 0$ since we can always buy S (for free) in the first stage; in the above expression for λ, we adopt the convention that $0/0 = 0$.

Different scenarios could be quite unrelated, so there does not seem to be a natural choice for a (non-discrete) scenario metric ℓ; we therefore consider the discrete scenario metric ℓ^{disc} (and so L is the $\frac{1}{2}L_1$ metric). We can take $\ell_{\max} := 1$.

Lemma 6.7. *Assumptions (A1)–(A6) hold for* DRSSC. *Moreover, when ℓ is the discrete metric ℓ^{disc}, assumption (A7) holds for $\tau = \max\left\{1, \sum_{S \in \mathcal{S}} c_S^{\mathrm{II}}\right\}$.*

Proof. Properties (A1)–(A5) follow from the discussion in the introduction of this chapter.

We now prove that property (A6) holds. Note that if $A \in \mathcal{A}$ is a non-null scenario, then A contains an element e such that every set $S \in \mathcal{S}$ containing e satisfies $c_S^{\mathrm{II}} \geq 1$, and hence $c_S^{\mathrm{I}} \geq 1$. Therefore, any fractional feasible solution $\left(x, z^A\right)$ for scenario A has cost

$$\sum_{S \in \mathcal{S}} \left(c_S^{\mathrm{I}} x_S + c_S^{\mathrm{II}} z_S^A\right) \geq \sum_{S \in \mathcal{S}: e \in S} \left(c_S^{\mathrm{I}} x_S + c_S^{\mathrm{II}} z_S^A\right) \geq \sum_{S \in \mathcal{S}: e \in S} \left(x_S + z_S^A\right) \geq 1 \ .$$

To see that property (A7) holds, note that $\min_{A, A' : \ell(A,A') > 0} \ell(A, A') \geq 1$, and that $\max_{A \in \mathcal{A}} g(0, A) \leq \sum_{S \in \mathcal{S}} c_S^{\mathrm{II}}$. The latter inequality follows because, given any scenario $A \in \mathcal{A}$, setting $z_S^A = 1$ for every set $S \in \mathcal{S}$ yields a feasible solution for scenario A under the first-stage decision $x = 0$, with cost $\sum_{S \in \mathcal{S}} c_S^{\mathrm{II}}$. The result then follows from the discussion in the introduction of this chapter. □

Shmoys and Swamy [114] give a local ρ-approximation algorithm for DRSSC with $\rho = \mathrm{O}(\log |U|)$. Applying Theorem 6.5 immediately yields the following result in the unrestricted setting.

Theorem 6.8. *Consider* DRSSC *under a Wasserstein ball in the unrestricted setting, with ℓ being the discrete metric ℓ^{disc}. There exists a two-stage algorithm that, given $\varepsilon > 0$ and $\delta \in (0,1)$, computes an $\mathrm{O}((1+\varepsilon)\log|U|)$-approximate solution with probability at least $1-\delta$ in $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$ time.*

We now consider the k-bounded setting. To apply Theorem 3.6, we need to furnish a second-stage α-approximation algorithm; a (β_1, β_2)-approximation for problem (Π); and a local ρ-approximation algorithm. We can set $\alpha = \mathrm{O}(\log|U|)$ using the well-known LP-based $\mathrm{O}(\log|U|)$-approximation algorithm for (deterministic) set cover (see Chvátal [27]), and Shmoys and Swamy [114] show that we can set $\rho = 2\alpha = \mathrm{O}(\log|U|)$. We now show that we can set $(\beta_1, \beta_2) = (\mathrm{O}(\log|U|), 1)$.

Lemma 6.9. *For* DRSSC *in the k-bounded setting with ℓ being the discrete metric ℓ^{disc}, we can obtain an $(\mathrm{O}(\log|U|), 1)$-approximation algorithm for problem (Π).*

Proof. By Lemma 6.2, it suffices to obtain $\mathrm{O}(\log|U|)$-approximation algorithm for the constrained problem (Φ). Consider an instance $(x, \mu, A) \in X \times \mathbb{R}_+ \times \mathcal{A}$ of (Φ). If $\mu < 1$, then the only feasible scenario is A itself, so we can solve this instance exactly. Otherwise, all the scenarios are feasible, so the problem reduces to $\max_{A \in \mathcal{A}_{\leq k}} g(x, A)$. Note that this is equivalent to a k-max-min fractional set cover problem (i.e., the problem of finding a scenario $A \in \mathcal{A}_{\leq k}$ so as to maximize the cost of an optimal fractional set cover of A), where the cost of buying a set $S \in \mathcal{S}$ is set to c^{Π}_S if $x_S = 0$, and to 0 if $x_S = 1$. Gupta, Nagarajan, and Ravi [60] give an $\mathrm{O}(\log|U|)$-approximation algorithm for k-max-min *integer* set cover, wherein the goal is to choose a scenario $A \in \mathcal{A}_{\leq k}$ so as to maximize the cost of an optimal *integral* set cover for A. It is implicit in their analysis that this also yields an $\mathrm{O}(\log|U|)$-approximation for k-max-min fractional set cover.[1] We therefore obtain a $\mathrm{O}(\log|U|)$-approximation algorithm for (Φ). □

We can therefore use Theorem 3.6 to obtain an approximation factor $\mathrm{O}(\alpha\beta_1\beta_2\rho) = \mathrm{O}\big(\log^3|U|\big)$ for DRSSC under a $\frac{1}{2}L_1$-ball in the k-bounded setting. By incorporating a decoupling idea of Shmoys and Swamy [114] in our ellipsoid-based algorithm (in a manner similar to Feige, Jain, Mahdian, and Mirrokni [46] in their work on two-stage robust set cover), we can avoid the use of a local approximation algorithm inside the ellipsoid method, thus obtaining an improved approximation ratio.

[1] See Theorem 4.2 and Claim 4.3 in [60]; Theorem 4.2 proves that the optimal fractional cost of the set-cover instance $(S, \mathcal{F})$ is at most $c(\Phi^*) + 12T^*$.

Theorem 6.10. *Consider* DRSSC *in the k-bounded setting, with ℓ being the discrete metric ℓ^{disc}. There exists a two-stage algorithm that, given $\varepsilon > 0$ and $\delta \in (0,1)$, computes an $\mathrm{O}\big((1+\varepsilon)\log^2|U|\big)$-approximate solution with probability at least $1-\delta$ in $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$ time.*

The key ingredient for proving Theorem 6.10 is the following result, which can be seen as a variant of Theorem 5.1. Whereas Theorem 5.1 shows that in general one can obtain an approximation algorithm for an SAA problem $\min_{x\in X} h(\widehat{p}\,;x)$ by using an approximation algorithm for problem (Π) and a local approximation rounding algorithm, the result below shows that, for set cover, we can obtain an approximation algorithm for the SAA problem $\min_{x\in\mathcal{P}} h(\widehat{p}\,;x)$ (with *fractional* first-stage decisions) using *only* an approximation algorithm for problem (Π).

Theorem 6.11. *Consider a* DRSSC *problem $\min_{x\in\mathcal{P}} h(\widehat{p}\,;x)$, where the distribution $\widehat{p}$ is given explicitly. Let $\widehat{\mathcal{I}}$ denote the input size of this problem. Suppose that we have a (β_1,β_2)-approximation algorithm for problem (Π). Then, given $\eta > 0$, we can compute in $\mathsf{poly}\Big(\widehat{\mathcal{I}}, \log\frac{1}{\eta}\Big)$ time a fractional first-stage decision $\overline{x} \in \mathcal{P}$ and an estimate f such that*

$$h(\widehat{p}\,;\overline{x}) \le f \le 2\beta_1\beta_2\left(\min_{x\in\mathcal{P}} h(\widehat{p}\,;x) + \eta\right).$$

We now show how Theorem 6.11 can be used to obtain Theorem 6.10.

Proof of Theorem 6.10. The proof follows very closely the structure of the proof of Theorem 3.6, so we omit some details. Using Lemma 6.1, we can either find an optimal solution $\overline{x}$ for $\min_{x\in\mathcal{P}} h(\mathring{p}\,;x)$, or obtain a lower bound LB on $\min_{x\in\mathcal{P}} h(\widetilde{p}\,;x)$ for every distribution $\widetilde{p}$. In the latter case, Theorem 6.11 (setting $\eta = \varepsilon\mathsf{LB}$ and $(\beta_1,\beta_2) = (\mathrm{O}(\log|U|),1)$, using Lemma 6.9) yields an $\mathrm{O}((1+\varepsilon)\log|U|)$-approximation algorithm for the fractional SAA problem $\min_{x\in\mathcal{P}} h(\widehat{p}\,;x)$. We can then invoke the SAA result from Theorem 6.6 to obtain an $\mathrm{O}((1+\varepsilon)\log|U|)$-approximate solution $\overline{x}$ for $\min_{x\in\mathcal{P}} h(\mathring{p}\,;x)$. Using the local $\mathrm{O}(\log|U|)$-approximation algorithm from Shmoys and Swamy [114] to round $\overline{x}$, we obtain an $\mathrm{O}\big((1+\varepsilon)\log^2|U|\big)$-approximate solution for ($\mathrm{DRSO_W}$) by Lemma 3.16. □

In the remainder of this section, we explain how to obtain Theorem 6.11. Let us consider an instance $\big(U, \mathcal{S}, c^{\mathrm{I}}, c^{\mathrm{II}}, r\big)$ of DRSSC. The improvement comes from a better way of generating a cut passing through the center $\overline{x} \in \mathcal{P}$ of the current ellipsoid. Instead of using a local ρ-approximation algorithm to round $\overline{x}$ to an integer first-stage decision

$\widehat{x} \in X$, then using an approximation algorithm for (Π) at $\widehat{x}$ to generate a suitable cut at $\overline{x}$ (as we did in Lemma 5.2 for the general framework), we do the following. Let $S_{\overline{x}} := \left\{e \in U : \sum_{S \in \mathcal{S}: e \in S} \overline{x}_S \geq 1/2\right\}$ be the set of elements covered to an extent of at least $1/2$ by the sets (fractionally) bought in the first stage under the decision $\overline{x}$. Since elements in $S_{\overline{x}}$ are mostly covered by $\overline{x}$, and the remaining elements are mostly uncovered, intuitively only these remaining elements should matter. Indeed, we argue that approximate solutions to

$$\max_{A' \in \mathcal{A}} \{g(0, A' \setminus S_{\overline{x}}) - y \cdot \ell(A, A')\}$$

(for different values of y and A) can be used to obtain a suitable cut at $\overline{x}$. Note that this problem can be cast as $g(0, y, A)$ for a modified instance where we add $S_{\overline{x}}$ to our set system $\mathcal{S}$, with costs $c^{\mathrm{I}}_{S_{\overline{x}}} = c^{\mathrm{II}}_{S_{\overline{x}}} = 0$. Thus, we avoid the ρ-factor loss that was incurred in the ellipsoid-based method due to the use of the local approximation algorithm.

Consider the following LP:

$$\begin{aligned}
\max \quad & \sum_{A, A'} \gamma_{A,A'} g(0, A' \setminus S_{\overline{x}}) && && (\mathrm{W}_{\overline{x}}) \\
\text{s.t.} \quad & \sum_{A'} \gamma_{A,A'} \leq \widehat{p}_A && \forall A \in \mathcal{A}^{\sup} \\
& \sum_{A, A'} \ell(A, A') \gamma_{A,A'} \leq r \\
& \gamma \geq 0 \ .
\end{aligned}$$

We prove an analogue of Lemma 5.5 showing that one can compute an approximate solution γ for ($\mathrm{W}_{\overline{x}}$) using an approximation algorithm for (Π) (see Lemma 6.12). Then we prove an analogue of Lemma 5.4, showing that we can use γ to both approximate $h(\widehat{p}; \widetilde{x})$ for a related point $\widetilde{x}$ (Lemma 6.13 (i)), and obtain a suitable cut passing through $\overline{x}$ (Lemma 6.13 (ii)).

Lemma 6.12. *Suppose that we have a (β_1, β_2)-approximation algorithm for problem (Π). Then, given any $\overline{x} \in \mathcal{P}$, we can compute a $\beta_1\beta_2$-approximate solution for ($\mathrm{W}_{\overline{x}}$) in* $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ *time.*

Proof. Consider the instance of DRSSC obtained from the original instance $\left(U, \mathcal{S}, c^{\mathrm{I}}, c^{\mathrm{II}}, r\right)$ by adding the set $S_{\overline{x}}$ to $\mathcal{S}$, with costs $c^{\mathrm{I}}_{S_{\overline{x}}} = c^{\mathrm{II}}_{S_{\overline{x}}} = 0$. Let $\{g^{\mathrm{new}}(x, A)\}_{x \in \mathcal{P}, A \in \mathcal{A}}$ denote the second-stage costs for this new instance of DRSSC. Note that, for every scenario $A \in \mathcal{A}$, we have $g^{\mathrm{new}}(0, A) = g(0, A \setminus S_{\overline{x}})$. Therefore, if we were to write the LP ($\mathrm{T}(\widehat{p}, 0)$) for this modified instance of DRSSC (i.e., ($\mathrm{T}(\widehat{p}, 0)$) with g substituted by g^{new}), we would

obtain ($W_{\overline{x}}$). This means that we can obtain a $\beta_1\beta_2$-approximate solution γ to ($W_{\overline{x}}$) by applying Lemma 5.5 to the modified instance (using the (β_1, β_2)-approximation algorithm for (Π) given to us, also applied to the modified instance). □

Lemma 6.13. *Let $\overline{x} \in \mathcal{P}$ and $\widetilde{x} := (\min\{2\overline{x}_S, 1\})_{S\in\mathcal{S}}$. Let γ be a β-approximate solution for ($W_{\overline{x}}$). Define $f := \beta\cdot\Big(2c^\intercal\overline{x} + \sum_{A,A'}\gamma_{A,A'}g(0, A'\setminus S_{\overline{x}})\Big)$ and $d := c+\sum_{A,A'}\gamma_{A,A'}d^{\overline{x},A'\setminus S_{\overline{x}}}$. Then we have (i) $h(\widehat{p};\widetilde{x}) \le f$ and (ii) $h(\widehat{p};x) \ge \frac{1}{2\beta}f$ for every $x \in \mathcal{P}$ such that $d^\intercal(x-\overline{x}) \ge 0$.*

Proof.

Part (i). Let γ^* be an optimal solution for ($T(\widehat{p},\widetilde{x})$). We claim that for every scenario $A' \in \mathcal{A}$, we have $g(\widetilde{x}, A') \le g(0, A'\setminus S_{\overline{x}})$. Assuming this, we obtain

$$\begin{aligned} h(\widehat{p};\widetilde{x}) &= c^\intercal\widetilde{x} + \sum_{A,A'}\gamma^*_{A,A'}g(\widetilde{x}, A') \\ &\le 2c^\intercal\overline{x} + \sum_{A,A'}\gamma^*_{A,A'}g(0, A'\setminus S_{\overline{x}}) \\ &\le 2c^\intercal\overline{x} + \beta\sum_{A,A'}\gamma_{A,A'}g(0, A'\setminus S_{\overline{x}}) \\ &\le \beta\cdot f\ . \end{aligned}$$

The first inequality follows from $\widetilde{x} \le 2\overline{x}$ and from the claim. The second inequality follows because γ is a β-approximate solution for ($W_{\overline{x}}$). The final inequality uses the fact that $\beta \ge 1$.

It remains to prove the claim. Consider a scenario $A' \in \mathcal{A}$, and let z^* be an optimal fractional second-stage decision for scenario $A'\setminus S_{\overline{x}}$ given the first-stage decision $x = 0$. Since z^* fully covers all elements of $A'\setminus S_{\overline{x}}$ and $\widetilde{x}$ fully covers all the elements of $S_{\overline{x}}$, we have that $\widetilde{x} + z^*$ fully covers A', and so z^* is feasible for scenario A' given the first-stage decision $\widetilde{x}$; this implies that $g(\widetilde{x}, A') \le (\text{cost of } z^*) = g(0, A'\setminus S_{\overline{x}})$.

Part (ii). Consider the function $\zeta : x \mapsto c^\intercal x + \sum_{A,A'}\gamma_{A,A'}g(x, A'\setminus S_{\overline{x}})$. We claim that d is a subgradient of $\zeta(\cdot)$ at $\overline{x}$. Assuming this, let $x \in \mathcal{P}$ such that $d^\intercal(x-\overline{x}) \ge 0$. We obtain

$$h(\widehat{p};x) - \zeta(\overline{x}) \ge \zeta(x) - \zeta(\overline{x}) \ge d^\intercal(x-\overline{x}) \ge 0\ .$$

The first inequality follows because γ is a feasible solution for $(\mathrm{T}(\widehat{p}, x))$. The second inequality follows because d is a subgradient of $\zeta(\cdot)$ at $\overline{x}$. The final inequality follows from the definition of x. It follows that $h(\widehat{p}\,; x) \geq \zeta(\overline{x})$.

Now, note that for every scenario $A' \in \mathcal{A}$, we have $g(\overline{x}, A' \setminus S_{\overline{x}}) \geq \frac{1}{2} g(0, A' \setminus S_{\overline{x}})$. To see this, let z^* be an optimal fractional second-stage solution for scenario $A' \setminus S_{\overline{x}}$ given $\overline{x}$ as the first-stage decision. Then z^* covers elements of $A' \setminus S_{\overline{x}}$ to an extent of at least $\frac{1}{2}$, and so $(\min\{2z^*_S, 1\})_{S \in \mathcal{S}}$ is a feasible second-stage solution for scenario $A' \setminus S_{\overline{x}}$ given 0 as the first-stage decision. This implies that

$$g(0, A' \setminus S_{\overline{x}}) \leq (\text{cost of } (\min\{2z^*_S, 1\})_{S \in \mathcal{S}}) \leq 2(\text{cost of } z^*) = 2g(\overline{x}, A' \setminus S_{\overline{x}}) \ .$$

So we obtain

$$\begin{aligned} h(\widehat{p}\,; x) &\geq \zeta(\overline{x}) \\ &= c^\intercal \overline{x} + \sum_{A, A'} \gamma_{A,A'} g(\overline{x}, A' \setminus S_{\overline{x}}) \\ &\geq c^\intercal \overline{x} + \frac{1}{2} \sum_{A, A'} \gamma_{A,A'} g(0, A' \setminus S_{\overline{x}}) \\ &= \frac{1}{2\beta} f \ . \end{aligned}$$

It remains to prove the claim. For any $x \in \mathcal{P}$, we have

$$\begin{aligned} \zeta(x) - \zeta(\overline{x}) &= c^\intercal (x - \overline{x}) + \sum_{A, A'} \gamma_{A,A'} (g(x, A' \setminus S_{\overline{x}}) - g(\overline{x}, A' \setminus S_{\overline{x}})) \\ &\geq c^\intercal (x - \overline{x}) + \sum_{A, A'} \gamma_{A,A'} d^{\overline{x}, A' \setminus S_{\overline{x}}} \cdot (x - \overline{x}) \\ &= d^\intercal (x - \overline{x}) \ , \end{aligned}$$

where the second inequality follows because $d^{\overline{x}, A' \setminus S_{\overline{x}}}$ is a subgradient of $g(\cdot, A' \setminus S_{\overline{x}})$ at $\overline{x}$. □

We are now ready to prove Theorem 6.11.

Proof of Theorem 6.11. Using Lemmas 6.12 and 6.13, along with the given approximation algorithm for problem (Π), we obtain a $(2\beta_1\beta_2, \mathcal{P})$-first-order oracle for $h(\widehat{p}\,; \cdot)$ with running

time $T_{\text{oracle}} = \mathsf{poly}\left(\widehat{\mathcal{I}}\right)$. Recall that $h(\widehat{p}\,;\cdot)$ is $(\|c\| + K)$-Lipschitz continuous by Lemma 5.6. We can therefore obtain $(\overline{x}, f)$ with the sought guarantees via the ellipsoid-based algorithm from Theorem 3.14. □

6.5 DRS vertex cover

In two-stage DRS vertex cover (DRSVC), we are given a graph $G = (V, E)$. A scenario is a collection of edges $A \subseteq E$. We may buy a vertex $v \in V$ in either stage, incurring costs $c_v^{\mathrm{I}} \in \mathbb{Z}_+$ and $c_v^{\mathrm{II}} \in \mathbb{Z}_+$ in the first and in the second stage respectively. For each scenario A, the vertices bought in the first stage and in scenario A (in the second stage) must together cover A (a collection of vertices is said to cover A if it contains at least one endpoint of each edge in A). The goal is to choose some first-stage vertices $V^{\mathrm{I}} \subseteq V$ and vertices $V^A \subseteq V$ in each scenario A so as to minimize

$$\sum_{v \in V^{\mathrm{I}}} c_v^{\mathrm{I}} + \sup_{p : L(\mathring{p}, p) \leq r} \mathbb{E}_{A \sim p}\left[\sum_{v \in V^A} c_v^{\mathrm{II}}\right].$$

The input size $\mathcal{I}$ is defined as the encoding size of $\left(V, E, c^{\mathrm{I}}, c^{\mathrm{II}}, r\right)$.

Note that DRSVC can be seen as a special case of DRSSC: an instance $\left(V, E, c^{\mathrm{I}}, c^{\mathrm{II}}, r\right)$ of DRSVC is equivalent to an instance of DRSSC with ground set $U := E$, and with a collection of sets $\mathcal{S} := \{S_v : v \in V\}$, where for each vertex $v \in V$ the set S_v consists of the edges incident with v, and has costs c_v^{I} and c_v^{II} in the first and in the second stage respectively. It is therefore clear that we can use (DRSO) to model DRSVC, while satisfying assumptions (A1)–(A7) (see Lemma 6.7). We again consider ℓ to be the discrete scenario metric ℓ^{disc} (and so L is the $\frac{1}{2}L_1$ metric), and take $\ell_{\max} := 1$.

The following lemma gives the ingredients we use to obtain our results in the unrestricted and in the k-bounded setting.

Lemma 6.14. *We have the following algorithms for* DRSVC*:*

(i) a second-stage 2-approximation algorithm;

(ii) a $\left(\frac{2e}{e-1}, 1\right)$-approximation for problem (Π) in the k-bounded setting, with ℓ being the discrete scenario metric ℓ^{disc}; and

(iii) a local 4-approximation algorithm.

Proof.

Part (i). Consider an integer first-stage decision $\widehat{x} \in X$ and a scenario $A \subseteq E$. Note that $g(\widehat{x}, A)$ is the minimum cost of a fractional vertex cover of the graph G' obtained from (V, A) by deleting the edges that are covered by vertices v with $\widehat{x}_{S_v} = 1$ (where the cost of a vertex v is c_v^{II}). We can therefore compute an integer vertex cover for G' with cost at most $2g(\widehat{x}, A)$ (see, e.g., Section 1.3 of Williamson and Shmoys [129]), which induces a suitable integer second-stage decision $\widehat{z}^A$.

Part (ii). We show how to obtain a $\frac{2e}{e-1}$-approximation algorithm for problem (Φ). The result then follows from Lemma 6.2.

Consider an instance $(\widehat{x}, \mu, A) \in X \times \mathbb{R}_+ \times \mathcal{A}$ of (Φ). If $\mu < 1$, then A is the unique feasible solution for (Φ), so we are done. Otherwise, (Φ) reduces to $\max_{A \in \mathcal{A}_{\leq k}} g(\widehat{x}, A)$, a problem for which Feige, Jain, Mahdian, and Mirrokni [46] give a $\frac{2e}{e-1}$-approximation algorithm.

Part (iii). This follows from part (i) and Theorem 2.1 in Shmoys and Swamy [114], which shows that one can convert a second-stage α-approximation algorithm into a local 2α-approximation algorithm for set cover (and in particular for vertex cover). □

We obtain the following results for DRSVC.

Theorem 6.15. *Consider* DRSVC *under a Wasserstein ball, with ℓ being the discrete metric ℓ^{disc}. There exists a two-stage algorithm that, given $\varepsilon > 0$ and $\delta \in (0, 1)$, computes a ψ-approximate solution with probability at least $1 - \delta$ in $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log \frac{1}{\delta}\big)$ time, where $\psi = 16 + \mathrm{O}(\varepsilon)$ in the unrestricted setting and $\psi = 101.3 + \mathrm{O}(\varepsilon)$ in the k-bounded setting.*

Proof. The result for the unrestricted setting follows from Theorem 6.5, setting $\rho = 4$ (using the local approximation algorithm from Lemma 6.14-(iii)). The result for the k-bounded setting follows from Theorem 3.6, setting $\alpha = 2, (\beta_1, \beta_2) = \big(\frac{2e}{e-1}, 1\big)$, and $\rho = 4$ (using the algorithms from Lemma 6.14-(i), Lemma 6.14-(ii), and Lemma 6.14-(iii) respectively). □

6.6 DRS edge cover

In two-stage DRS edge cover (DRSEC), we are given a graph $G = (V, E)$. A scenario is a collection of vertices $A \subseteq V$. We may buy an edge $e \in E$ in either stage, incurring costs

$c_e^{\mathrm{I}} \in \mathbb{Z}_+$ and $c_e^{\mathrm{II}} \in \mathbb{Z}_+$ in the first and in the second stage respectively. For each scenario A, the edges bought in the first stage and in scenario A (in the second stage) must together cover A (a collection of edges is said to cover A if it contains at least one edge incident with each vertex in A). The goal is to choose some first-stage edges $E^{\mathrm{I}} \subseteq E$ and edges $E^A \subseteq E$ in each scenario A so as to minimize

$$\sum_{e \in E^{\mathrm{I}}} c_e^{\mathrm{I}} + \sup_{p : L(\mathring{p}, p) \leq r} \mathbb{E}_{A \sim p} \left[\sum_{e \in E^A} c_e^{\mathrm{II}} \right] .$$

The input size $\mathcal{I}$ is defined as the encoding size of $\left(V, E, c^{\mathrm{I}}, c^{\mathrm{II}}, r\right)$.

Note that DRSEC can be seen as a special case of DRSSC: an instance $\left(V, E, c^{\mathrm{I}}, c^{\mathrm{II}}, r\right)$ of DRSEC is equivalent to an instance of DRSSC with ground set $U := V$, and with a collection of sets $\mathcal{S} := \{S_e := \{u, v\}\}_{e=uv \in E}$; for each edge $e \in E$, the first-stage and second-stage cost of the corresponding set are defined as $c^{\mathrm{I}}_{S_e} := c^{\mathrm{I}}_e$ and $c^{\mathrm{II}}_{S_e} := c^{\mathrm{II}}_e$ respectively. It is therefore clear that we can use (DRSO) to model DRSEC, while satisfying assumptions (A1)–(A7) (see Lemma 6.7). We again consider ℓ to be the discrete scenario metric ℓ^{disc} (and so L is the $\frac{1}{2}L_1$ metric), and take $\ell_{\max} := 1$.

The following lemma gives the ingredients we use to obtain our results in the unrestricted and in the k-bounded setting.

Lemma 6.16. *We have the following algorithms for* DRSEC*:*

(i) a second-stage $\frac{3}{2}$-approximation algorithm;

(ii) a $(2, 1)$-approximation for problem (Π) in the k-bounded setting, with ℓ being the discrete scenario metric ℓ^{disc}; and

(iii) a local 3-approximation algorithm.

Proof.

Part (i). Consider an integer first-stage decision $\widehat{x}$ and a scenario $A \subseteq V$. Note that $g(\widehat{x}, A)$ is the minimum cost of a fractional edge cover of A if we set the cost of each edge $e \in E$ to c_e^{II} if $\widehat{x}_e = 0$, and to 0 if $\widehat{x}_e = 1$. It is well known that we can compute an edge cover of A of cost at most $\frac{3}{2} g(\widehat{x}, A)$, which induces a suitable integer second-stage decision $\widehat{z}^A$ (more generally, for an instance of set cover where each set contains at most f elements, one can compute an integer set cover whose cost is at most $1 + \frac{1}{2} + \cdots + \frac{1}{f}$ times the minimum cost of a fractional set cover—see Chvátal [27]).

Part (ii). We show how to obtain a 2-approximation algorithm for the constrained problem (Φ). The result then follows from Lemma 6.2.

Consider an instance $(\widehat{x}, \mu, A) \in X \times \mathbb{R}_+ \times \mathcal{A}$ of (Φ). If $\mu < 1$, then A is the unique feasible solution for (Φ), so we are done. Otherwise, (Φ) reduces to $\max_{A \in \mathcal{A}_{\leq k}} g(\widehat{x}, A)$, a problem for which Feige, Jain, Mahdian, and Mirrokni [46] give a 2-approximation algorithm.

Part (iii). This follows from part (i) and Theorem 2.1 in Shmoys and Swamy [114], which shows that one can convert a second-stage α-approximation algorithm into a local 2α-approximation algorithm for set cover (and in particular for edge cover). □

We obtain the following results for DRSEC.

Theorem 6.17. *Consider* DRSEC *under a Wasserstein ball, with ℓ being the discrete metric ℓ^{disc}. There exists a two-stage algorithm that, given $\varepsilon > 0$ and $\delta \in (0,1)$, computes a ψ-approximate solution with probability at least $1-\delta$ in $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log \frac{1}{\delta}\big)$ time, where $\psi = 12 + \mathrm{O}(\varepsilon)$ in the unrestricted setting and $\psi = 36 + \mathrm{O}(\varepsilon)$ in the k-bounded setting.*

Proof. The result for the unrestricted setting follows from Theorem 6.5, setting $\rho = 3$ (using the local approximation algorithm from Lemma 6.16-(iii)). The result for the k-bounded setting follows from Theorem 3.6, setting $\alpha = 3, (\beta_1, \beta_2) = (2,1)$, and $\rho = 3/2$ (using the algorithms from Lemma 6.16-(i), Lemma 6.16-(ii), and Lemma 6.16-(iii) respectively). □

6.7 DRS facility location

In two-stage DRS facility location (DRSFL), we have a metric space $\Big(\mathcal{F} \cup \mathcal{C}, \{w_{ij}\}_{i,j \in \mathcal{F} \cup \mathcal{C}}\Big)$, where $\mathcal{F}$ is a set of facilities, and $\mathcal{C}$ is a set of clients. We assume that $w_{ij} \in \mathbb{Z}_+$ for every $i, j \in \mathcal{F} \cup \mathcal{C}$. A scenario is a subset of $\mathcal{C}$ indicating the set of clients that need to be served in that scenario. (We can model integer demands by creating colocated clients.)

We may open a facility $i \in \mathcal{F}$ in either stage, incurring costs of $f_i^{\mathrm{I}} \in \mathbb{Z}_+$ and $f_i^{\mathrm{II}} \in \mathbb{Z}_+$ respectively. In scenario A, we need to assign every client $j \in A$ to a facility $i^A(j)$ opened in the first stage or in scenario A (in the second stage). The goal is to minimize

$$\sum_{i \text{ opened in stage I}} f_i^{\mathrm{I}} + \max_{p : L(\hat{p}, p) \leq r} \mathbb{E}_{A \sim p} \left[\sum_{i \text{ opened in scenario } A} f_i^{\mathrm{II}} + \sum_{j \in A} w_{i^A(j) j} \right].$$

The input size $\mathcal{I}$ is defined as the encoding size of $\left(\mathcal{F}, \mathcal{C}, w, f^{\mathrm{I}}, f^{\mathrm{II}}, r\right)$.

It is easy to see that DRSFL can be cast as an instance of (DRSO): the first-stage decision set X and the second-stage decision set Z are $X = \{0,1\}^{\mathcal{F}}$ and $Z = \{0,1\}^{\mathcal{F} \cup (\mathcal{F} \times \mathcal{C})}$; the corresponding sets of fractional first-stage and second-stage decisions are given by the polytopes $\mathcal{P} = [0,1]^{\mathcal{F}}$ and $\mathcal{Z} = [0,1]^{\mathcal{F} \cup (\mathcal{F} \times \mathcal{C})}$. The polytope specifying the feasibility conditions for a scenario A is

$$\mathcal{F}(A) := \left\{ \left(x, z^A\right) \in \mathcal{P} \times \mathcal{Z} : \sum_{i \in \mathcal{F}} z_{ij}^A \geq 1 \ \forall j \in A, \quad z_{ij}^A \leq x_i + z_i^A \ \ \forall i \in \mathcal{F}, j \in A \right\} .$$

The inflation factor λ is defined as $\max\left\{1, \max_{i \in \mathcal{F}} \frac{f_i^{\mathrm{II}}}{f_i^{\mathrm{I}}}\right\}$. Note that we may assume that $f_i^{\mathrm{I}} = 0$ if $f_i^{\mathrm{II}} = 0$ since we can always open facility i (for free) in the first stage; in the above expression for λ, we adopt the convention that $0/0 = 0$.

We consider two choices for the scenario metric ℓ: the discrete metric ℓ^{disc} and the asymmetric metric $\ell_\infty^{\mathsf{asym}}$ (defined with respect to the underlying metric w). We set $\ell_{\max} := 1$ and $\ell_{\max} := \max_{i,i' \in \mathcal{C}} w_{i,i'}$ in these two settings respectively.

Lemma 6.18. *Assumptions (A1)–(A6) hold for* DRSFL*. Moreover, assumption (A7) holds for* $\tau = \max\left\{1, \sum_{i \in \mathcal{F}} f_i^{\mathrm{II}} + \sum_{i \in \mathcal{F}, j \in \mathcal{C}} w_{ij}\right\}$ *when* ℓ *is the discrete metric* ℓ^{disc} *or the asymmetric metric* $\ell_\infty^{\mathsf{asym}}$*,*

Proof. Properties (A1)–(A5) follow from the discussion in the introduction of this chapter.

We now prove that property (A6) holds. Note that if $A \in \mathcal{A}$ is a non-null scenario, then A contains a client $j' \in \mathcal{C}$ such that $f_i^{\mathrm{II}} + w_{ij'} \geq 1$ for every facility $i \in \mathcal{F}$. Therefore, for every facility i we have $f_i^{\mathrm{II}} \geq 1$ (and hence $f_i^{\mathrm{I}} \geq 1$) or $w_{ij'} \geq 1$, and so $\min\left\{f_i^{\mathrm{I}}, f_i^{\mathrm{II}}\right\} + w_{ij'} \geq 1$. Therefore, any fractional feasible solution $\left(x, z^A\right)$ for scenario A has cost

$$\begin{aligned}
\sum_{i \in \mathcal{F}} f_i^{\mathrm{I}} x_i + \sum_{i \in \mathcal{F}} f_i^{\mathrm{II}} z_i^A + \sum_{i \in F, j \in \mathcal{C}} w_{ij} z_{ij}^A &\geq \sum_{i \in \mathcal{F}} \left(f_i^{\mathrm{I}} x_i + f_i^{\mathrm{II}} z_i^A + w_{ij'} z_{ij'}^A\right) \\
&\geq \sum_{i \in \mathcal{F}} \left(\min\left\{f_i^{\mathrm{I}}, f_i^{\mathrm{II}}\right\} \left(x_i + z_i^A\right) + w_{ij'} z_{ij'}^A\right) \\
&\geq \sum_{i \in \mathcal{F}} \left(\min\left\{f_i^{\mathrm{I}}, f_i^{\mathrm{II}}\right\} + w_{ij'}\right) z_{ij'}^A \\
&\geq \sum_{i \in \mathcal{F}} z_{ij'}^A \geq 1 \ ,
\end{aligned}$$

where the third and the fifth steps follow from the assumption that $\left(x, z^A\right)$ is feasible for A.

To see that property (A7) holds, note that $\min_{A,A':\ell(A,A')>0} \ell(A, A') \geq 1$, and that $\max_{A\in\mathcal{A}} g(0, A) \leq \sum_{i\in\mathcal{F}} f_i^{\mathrm{II}} + \sum_{i\in\mathcal{F},j\in\mathcal{C}} w_{ij}$. The latter inequality follows because, given any scenario $A \in \mathcal{A}$, by opening an arbitrary facility $i' \in \mathcal{F}$ in the second stage and connecting all clients of A to it, we obtain a feasible second-stage solution for A, given $x = 0$ as a first-stage decision, with cost $f_{i'}^{\mathrm{II}} + \sum_{j\in A} w_{i'j} \leq \sum_{i\in\mathcal{F}} f_i^{\mathrm{II}} + \sum_{i\in\mathcal{F},j\in\mathcal{C}} w_{ij}$. The result then follows from the discussion in the introduction of this chapter. □

Lemma 6.19 and Theorem 6.20 give the main ingredients we use to obtain our results for DRSFL in the unrestricted and in the k-bounded setting.

Lemma 6.19. *We have the following algorithms for* DRSFL*:*

(i) a second-stage 1.488*-approximation algorithm; and*

(ii) a local 5.488*-approximation algorithm.*

Proof.

Part (i). Consider an integer first-stage decision $\widehat{x}$ and a scenario $A \subseteq \mathcal{C}$. Note that $g(\widehat{x}, A)$ is the minimum cost of a fractional solution for the (deterministic) facility-location instance defined as follows: the set of facilities is $\mathcal{F}$, the set of clients is A, the distances are $\{w_{ij}\}_{i\in\mathcal{F},j\in A}$, and the opening cost of each facility $i \in \mathcal{F}$ is 0 if $\widehat{x}_i = 1$, and f_i^{II} if $\widehat{x}_i = 0$. We can compute an integer solution for this instance of cost at most $1.488 \cdot g(\widehat{x}, A)$ using the algorithm by Li [83]; such a solution induces a suitable integer second-stage decision $\widehat{z}^A$.

Part (ii). Shmoys and Swamy [114] showed that an LP-relative ϱ-approximation for deterministic facility location having a certain "demand-obliviousness" property, combined with a second-stage α-approximation algorithm, can be turned into a $(\varrho + \alpha)$-approximation algorithm for two-stage stochastic facility location. If the ϱ-approximation algorithm has the property that it returns a solution where every cost component of the rounded solution—i.e., the facility cost, and *each* client's assignment cost—is at most ϱ times the corresponding cost component of the fractional solution, then the resulting algorithm is a local approximation algorithm. Using the deterministic 4-approximation algorithm of Shmoys, Tardos, and Aardal [115] and the algorithm from part (i) gives a local ρ-approximation with $\rho = 5.488$. □

Theorem 6.20 (see proof in Section 6.7.1)**.** *Consider* DRSFL *in the k-bounded setting (i.e., $\mathcal{A} = \mathcal{A}_{\leq k}$), and suppose that ℓ is either the discrete scenario metric ℓ^{disc} or the asymmetric metric $\ell_\infty^{\mathsf{asym}}$. Then we have a $(6,1)$-approximation for problem* (II).

We obtain the following results for DRSFL.

Theorem 6.21. *Consider* DRSFL *under a Wasserstein ball, where the underlying scenario metric ℓ is either the discrete metric ℓ^{disc} or the asymmetric metric $\ell_\infty^{\mathsf{asym}}$. There exists a two-stage algorithm that, given $\varepsilon > 0$ and $\delta \in (0,1)$, computes a $(\psi + \mathrm{O}(\varepsilon))$-approximate solution with probability at least $1-\delta$ in $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$ time, where $\psi = 21.96$ in the unrestricted setting, and $\psi = 196$ in the k-bounded setting.*

Proof. The result for the unrestricted setting follows from Theorem 6.5, setting $\rho = 5.488$ (using the local approximation algorithm from Lemma 6.19-(ii)).

The result for the k-bounded setting follows from Theorem 3.6, setting $\alpha = 1.488$, $(\beta_1, \beta_2) = (6,1)$, and $\rho = 5.488$ (see Lemma 6.19-(i), Theorem 6.20, and Lemma 6.19-(ii) respectively). □

6.7.1 Proof of Theorem 6.20

In this section, we prove Theorem 6.20. We give a 6-approximation algorithm for the constrained problem (Φ). The result then follows immediately from Lemma 6.2. Consider an instance $(\widehat{x}, \mu, A) \in X \times \mathbb{R}_+ \times \mathcal{A}$ of (Φ). If ℓ is the discrete metric and $\mu < 1$, then we return A, which is the only feasible solution. In every other case, we claim that we can compute a set of clients $\widetilde{\mathcal{C}} \subseteq \mathcal{C}$ such that a scenario $A' \in \mathcal{A}$ is feasible if and only if $A' \subseteq \widetilde{\mathcal{C}}$. If ℓ is the discrete metric and $\mu \geq 1$, then we set $\widetilde{\mathcal{C}} := \mathcal{C}$. If ℓ is the asymmetric metric $\ell_\infty^{\mathsf{asym}}$, then we set $\widetilde{\mathcal{C}} := \{j \in \mathcal{C} : w(j, A) \leq \mu\}$. It follows that approximately solving this instance of (Φ) amounts to approximately solving the k-max-min fractional facility location problem for an underlying facility-location instance $\Big(\mathcal{F}, \widetilde{\mathcal{C}}, \{w_{ij}\}_{i,j \in \mathcal{F} \cup \widetilde{\mathcal{C}}}, \big\{\widetilde{f}_i\big\}_{i \in \mathcal{F}}\Big)$, where $\widetilde{f}_i := 0$ if $\widehat{x}_i = 1$, and $\widetilde{f}_i := f_i^{\mathrm{II}}$ otherwise. We now give a 6-approximation algorithm for this problem.

Theorem 6.22. *There exists a 6-approximation algorithm for the k-*max-min *facility location problem: given an instance of* DRSFL*, compute $\max_{A \in \mathcal{A}_{\leq k}} g(0, A)$.*

Khandekar, Kortsarz, Mirrokni, and Salavatipour [80] give a 10-approximation for the version of integral k-max-min facility location where a scenario may place an *arbitrary*

number of colocated clients at a location in $\mathcal{C}$ (and the global number of clients is constrained to be at most k).[2] However, in our setting, we may place at most one client at any location in $\mathcal{C}$, so the algorithm in [80] does not work for our purposes. (Clearly, our setting is more general, since we can encode the scenario setting of [80] by creating k colocated copies at every $j \in \mathcal{C}$.) As noted earlier, we can model more general settings, where clients have (integer) demands, by creating a fixed number of colocated clients at locations in $\mathcal{C}$; but, here again, we have a constraint that *limits* the number of colocated clients at any $j \in \mathcal{C}$.

We therefore need to develop new techniques to devise an approximation algorithm for fractional k-max-min facility location. The key tool that we exploit here is that of *cost-sharing schemes*. We uncover a novel connection between cost-sharing schemes and k-max-min problems by demonstrating that one can exploit a cost-sharing scheme for facility location having certain properties to obtain an approximation algorithm for k-max-min {integral, fractional} facility location. Previously, cost-sharing schemes have been exploited in the boosted-sampling approach of Gupta, Pál, Ravi, and Sinha [62] for two-stage stochastic optimization, however to our knowledge, they have not been used previously to tackle k-max-min problems. Our result also improves the approximation factor for integral k-max-min facility location from 10 to 6 (see Remark 6.24).

A cost-sharing method is a function $\xi : 2^{\mathcal{C}} \times \mathcal{C} \to \mathbb{R}_+$, where $\xi(S, j)$ for $j \in S$ intuitively gives the contribution of j toward the cost incurred in satisfying the client set S (i.e., the cost of opening facilities and assigning clients in S to these open facilities). For sets $S, T \subseteq \mathcal{C}$, define $\xi(S, T) := \sum_{j \in T} \xi(S, j)$. Pál and Tardos [95] devised a cost-sharing method ξ satisfying the following properties.

- $\xi(S, j) = 0$ if $j \notin S$.
- (Competitiveness) For every $S \subseteq \mathcal{C}$, we have $\xi(S, S) \leq g(0, S)$.
- (Cost-recovery) For every $S \subseteq \mathcal{C}$, we have $\xi(S, S) \geq g(0, S)/3$.
- (Cross-monotonicity) For all $S_1 \subseteq S_2 \subseteq \mathcal{C}$ and every client $j \in \mathcal{C}$, we have $\xi(S_2, j) \leq \xi(S_1, j)$.

We will prove an additional useful property about ξ, for which we very briefly describe how ξ is computed. For every $S \subseteq \mathcal{C}$ and $i \in \mathcal{F}$, we compute a certain *time* $t(S, i) \geq 0$.

[2] Since the gap between the integral and fractional optimal values for facility location is at most $\alpha = 1.488$ (see Li [83]), a β-approximation for the integral (respectively fractional) version implies an $\alpha\beta$-approximation for k-max-min fractional (respectively integral) facility location.

The cost-share of a client $j \in S$ is then defined as $\xi(S, j) := \min_{i \in \mathcal{F}} \max\{t(S, i), w_{ij}\}$. The function $t(\cdot, \cdot)$ satisfies the following property: for every set $S \subseteq \mathcal{C}$, every client $j \notin S$, and every facility $i \in \mathcal{F}$, we have $t(S + j, i) \leq t(S, i)$. Furthermore, if this inequality is strict, then $t(S + j, i) \geq w_{ij}$.

Lemma 6.23. *Consider $S \subseteq \mathcal{C}$ and two clients $j_1 \in S$ and $j_2 \notin J$. Then $\xi(S + j_2, j_1) \geq \min\{\xi(S, j_1), \xi(S + j_2, j_2)\}$.*

Proof. By cross-monotonicity, we have $\xi(S + j_2, j_1) \leq \xi(S, j_1)$. If this holds at equality, then the result follows immediately. So assume otherwise. By the way in which the cost-shares are defined, $\xi(S + j_2, j_1) < \xi(S, j_1)$ implies that $\xi(S + j_2, j_1) = t(S + j_2, i)$ for some facility i and $t(S + j_2, i) < t(S, i)$. This implies that $t(S + j_2, i) \geq w_{ij_2}$, and it follows that $\xi(S + j_2, j_2) \leq \max\{t(S + j_2, i), w_{ij_2}\} = t(S + j_2, i) = \xi(S + j_2, j_1)$. □

Proof of Theorem 6.22. We may assume that $k \leq |\mathcal{C}|$ (otherwise, we simply set $k = |\mathcal{C}|$). Consider the following simple greedy algorithm. Initialize $t \leftarrow 0$ and $S_0 \leftarrow \emptyset$. For $t = 1, \ldots, k$, we find $\bar{j} \leftarrow \operatorname{argmax}_{j \in \mathcal{C} \setminus S_{t-1}} \xi(S_{t-1} + j, j)$, and set $S_t \leftarrow S_{t-1} \cup \{\bar{j}\}$.

Let O^* be an optimal solution for the k-max-min problem $\max_{A \in \mathcal{A}_{\leq k}} g(0, A)$. We claim that $\xi(S_k, S_k) \geq \xi(S_k \cup O^*, S_k \cup O^*)/2$. This will complete the proof since this implies that

$$g(0, S_k) \geq \xi(S_k, S_k) \geq \frac{\xi(S_k \cup O^*, S_k \cup O^*)}{2} \geq \frac{g(0, S_k \cup O^*)}{6} \geq \frac{g(0, O^*)}{6} , \tag{6.2}$$

where the first inequality uses the competitiveness property, and the third inequality uses the cost-recovery property.

We now prove the above claim. Suppose that $S_k \neq O^*$ (otherwise, the claim immediately follows). For any $t \in [k]$, we show that $\xi(S_t, j) \geq M_t := \max_{j' \in \mathcal{C} \setminus S_{t-1}} \xi(S_{t-1} + j', j')$ for all $j \in S_t$. We prove this by induction on t. Note that $M_t \geq M_{t+1}$ due to cross-monotonicity, and since $\mathcal{C} \setminus S_{t-1} \supseteq \mathcal{C} \setminus S_t$. The statement is clearly true for $t = 1$. Suppose that it is true for index t, and consider index $t + 1$. Let $\bar{j}$ be the element added to S_t in iteration $t + 1$. By definition of $\bar{j}$, we have $\xi\big(S_{t+1}, \bar{j}\big) = M_{t+1}$. For every $j \in S_t$, we have

$$\xi(S_{t+1}, j) \geq \min\big\{\xi(S_t, j), \xi\big(S_{t+1}, \bar{j}\big)\big\} \geq \min\{M_t, M_{t+1}\} = M_{t+1} .$$

The first inequality follows from Lemma 6.23. The second inequality follows from the induction hypothesis and the fact that $\xi\big(S_{t+1}, \bar{j}\big) = M_{t+1}$. Thus, for every $j \in S_{t+1}$, we have $\xi(S_{t+1}, j) \geq M_{t+1}$. This completes the induction step.

Therefore, we obtain

$$\begin{aligned}
\xi(S_k, S_k) &\geq k \cdot M_k \\
&\geq k \cdot \max_{j \in O^* \setminus S_k} \xi(S_{k-1} + j, j) \\
&\geq k \cdot \max_{j \in O^* \setminus S_k} \xi(S_k \cup O^*, j) \\
&\geq k \cdot \frac{\xi(S_k \cup O^*, O^* \setminus S_k)}{|O^* \setminus S_k|} \\
&\geq \xi(S_k \cup O^*, O^* \setminus S_k) \\
&= \xi(S_k \cup O^*, S_k \cup O^*) - \xi(S_k \cup O^*, S_k) \\
&\geq \xi(S_k \cup O^*, S_k \cup O^*) - \xi(S_k, S_k) \ .
\end{aligned}$$

The first inequality follows from the statement proved in the previous paragraph. The second one is simply because we restricted $\mathcal{C} \setminus S_{k-1}$ to $O^* \setminus S_k$. The third one follows from cross-monotonicity. The fourth one is because we replaced max by an average and all cost shares are nonnegative. The fifth one is because $|O^*| \leq k$. The last inequality is again due to cross-monotonicity. We obtain $\xi(S_k, S_k) \geq \xi(S_k \cup O^*, S_k \cup O^*)/2$ as claimed. □

Remark 6.24. In fact Pál and Tardos [95] show a stronger form of cost recovery, namely, that for every scenario $S \subseteq \mathcal{C}$ there exists an integer solution $\widehat{z}^S$ feasible for scenario S such that $\xi(S, S) \geq \big(\text{cost of } \widehat{z}^S\big)/3$. Modifying the proof of Theorem 6.22 by using this stronger property in the chain of inequalities (6.2), one can also obtain a 6-approximation algorithms for *integral* k-max-min facility location.

6.8 DRS Steiner tree

In two-stage DRS Steiner tree (DRSST), we are given a complete graph $G = (V, E)$ with metric edge costs $\{c_e\}_{e \in E}$, a root vertex $s \in V$, and an inflation factor $\lambda \geq 1$. A scenario is a set of vertices $A \subseteq V$ (called *terminals*) specifying the nodes that need to be connected to the root s. We may buy an edge $e \in E$ in either stage, incurring costs $c_e^{\mathrm{I}} \in \mathbb{Z}_+$ or $c_e^{\mathrm{II}} = \lambda c_e$ in the first and in the second stage respectively.[3] For each scenario A, the union

[3] With non-uniform inflation factors for different edges, even two-stage stochastic Steiner tree becomes at least as hard to approximate as group Steiner tree (see Ravi and Sinha [98]), which is known not to admit an $\mathrm{O}\big(\log^{2-\varepsilon}(\text{number of groups})\big)$-approximation unless $\mathsf{NP} \subseteq \mathsf{ZTIME}\big(n^{\mathsf{polylog}(n)}\big)$ (see Halperin and Krauthgamer [66]).

of the edges $F \subseteq E$ bought in the first stage and $F^A \subseteq E$ bought in scenario A (in the second stage) must connect all nodes in A to s, and we want to minimize

$$\sum_{e \in F} c_e^{\mathrm{I}} + \max_{p : L(\mathring{p}, p) \leq r} \mathbb{E}_{A \sim p} \left[\sum_{e \in F^A} c_e^{\mathrm{II}} \right] .$$

An impediment for obtaining an approximation for DRSST utilizing the results we discussed so far is that we do not have a local approximation algorithm for DRSST. There is however a weaker type of rounding algorithm for a *monotone* version of DRSST (which we refer to as MDRSST), wherein we require that for every scenario A, the set of edges $F \cup F^A$ contain a path from each node $v \in A$ to the root s consisting of a segment starting at v comprising edges from F^A, followed by a segment ending at s comprising edges from F. A path from v to s with this property is said to be *monotone*. This monotonicity property was stipulated by Dhamdhere, Goyal, Ravi, and Singh [33] and Gupta, Ravi, and Sinha [64] in the context of two-stage {stochastic, robust} Steiner tree respectively, where they show that imposing this condition only incurs a factor-2 loss. We now state a result used by Dhamdhere, Goyal, Ravi, and Singh [33] to show that approximation guarantees for two-stage robust Steiner tree can be obtained via its monotone counterpart, then we use it to show that the same holds in the DRS setting.

Lemma 6.25 (see Lemma 4.1 in Dhamdhere, Goyal, Ravi, and Singh [33])**.** *Let $\overline{\mathcal{A}} \subseteq 2^V$. Consider edge sets $\left(F, \left\{F^A\right\}_{A \in \overline{\mathcal{A}}}\right)$ such that $F \cup F^A$ contains a path from v to s for every $A \in \overline{\mathcal{A}}$ and every $v \in A$. Then there exist edge sets $\left(\widetilde{F}, \left\{\widetilde{F}^A\right\}_{A \in \overline{\mathcal{A}}}\right)$ such that:*

(i) $\widetilde{F} \cup \widetilde{F}^A$ contains a monotone path from v to s for every $A \in \overline{\mathcal{A}}$ and every $v \in A$;

(ii) $\sum_{e \in \widetilde{F}} c_e^{\mathrm{I}} \leq 2 \cdot \sum_{e \in F} c_e^{\mathrm{I}}$; and

(iii) $\sum_{e \in \widetilde{F}^A} c_e^{\mathrm{II}} \leq 2 \cdot \sum_{e \in F^A} c_e^{\mathrm{II}}$ for every set $A \in \overline{\mathcal{A}}$.

Lemma 6.26. *Consider an instance I of DRSST, and let I^{mon} be the corresponding instance of MDRSST. Then every ψ-approximate solution for I^{mon} is a 2ψ-approximate solution for I.*

Proof. By applying Lemma 6.25 to an optimal solution for I (and setting $\overline{\mathcal{A}} := \mathcal{A}$), we infer that $\mathrm{OPT}(I^{\mathrm{mon}}) \leq 2 \cdot \mathrm{OPT}(I)$. Therefore, given any ψ-approximate solution for I^{mon}, it is also feasible for I (since I is a relaxation of I^{mon}) and attains objective value at most $\psi \cdot \mathrm{OPT}(I^{\mathrm{mon}}) \leq 2\psi \cdot \mathrm{OPT}(I)$. □

Given Lemma 6.26 (and the absence of a local approximation algorithm for DRSST, as mentioned above), we focus on obtaining an approximation algorithm for MDRSST. The input size $\mathcal{I}$ is defined as the encoding size of $\left(G, c^{\mathrm{I}}, s, \lambda, r\right)$. We now show that MDRSST can be cast as an instance of (DRSO). The set of integer first-stage decisions is set to $X := \{0,1\}^E$, and the polytope of fractional first-stage decisions is set to $\mathcal{P} := [0,1]^E$.

The representation of second-stage decisions is based on the IP formulation used by Gupta, Ravi, and Sinha [64]. We use a vector of binary variables $q^A \in \{0,1\}^E$ to indicate the edges bought in scenario A. For notational simplicity, we assume that $s \notin A$; clearly, this can always be ensured without changing the problem. To encode the requirement that there is a monotone v-s path for every $v \in A$, we bidirect the edges to obtain the set of arcs $\overleftrightarrow{E}$, and use two vectors of binary flow variables $f^{\mathrm{I},A,v}, f^{\mathrm{II},A,v} \in \{0,1\}^{\overleftrightarrow{E}}$ to specify the segments of v's path comprising first-stage and second-stage edges respectively. For a vertex $v \in V$, let $\delta^{\mathrm{in}}(v)$ (respectively $\delta^{\mathrm{out}}(v)$) denote the arcs of $\overleftrightarrow{E}$ entering (respectively leaving) v. For an arc $e \in \overleftrightarrow{E}$, we abuse notation and use x_e to denote the component of x corresponding to the undirected version of e. We can then express the cost incurred in scenario A if we make optimal integer second-stage decisions, given the integer first-stage decision $x \in X$, as the optimal value of the following IP.

$$\min \quad \sum_{e \in E} c_e^{\mathrm{II}} q_e^A$$

$$\text{s.t.} \quad \sum_{e \in \delta^{\mathrm{out}}(v)} (f_e^{\mathrm{I},A,v} + f_e^{\mathrm{II},A,v}) - \sum_{e \in \delta^{\mathrm{in}}(v)} (f_e^{\mathrm{I},A,v} + f_e^{\mathrm{II},A,v}) \geq 1 \qquad \forall v \in A \tag{6.3}$$

$$\sum_{e \in \delta^{\mathrm{out}}(u)} (f_e^{\mathrm{I},A,v} + f_e^{\mathrm{II},A,v}) = \sum_{e \in \delta^{\mathrm{in}}(u)} (f_e^{\mathrm{I},A,v} + f_e^{\mathrm{II},A,v}) \qquad \forall v \in A, u \in V \setminus \{s,v\} \tag{6.4}$$

$$f_e^{\mathrm{I},A,v} \leq x_e, \quad f_e^{\mathrm{II},A,v} \leq q_e^A \qquad \forall v \in A, \forall e \in \overleftrightarrow{E} \tag{6.5}$$

$$\sum_{e \in \delta^{\mathrm{in}}(u)} f_e^{\mathrm{I},A,v} \leq \sum_{e \in \delta^{\mathrm{out}}(u)} f_e^{\mathrm{I},A,v} \qquad \forall v \in A, u \in V \setminus \{s,v\} \tag{6.6}$$

$$q_e^A \in \{0,1\} \qquad \forall e \in E \tag{6.7}$$

$$f_e^{\mathrm{I},A,v}, f_e^{\mathrm{II},A,v} \in \{0,1\} \qquad \forall v \in A, e \in \overleftrightarrow{E}. \tag{6.8}$$

Constraints (6.3) and (6.4) enforce that $f^{\mathrm{I},A,v} + f^{\mathrm{II},A,v}$ sends one unit of flow from v to s for every terminal $v \in A$ (so it dominates a directed $v \rightsquigarrow s$ path),[4] and (6.5) enforces that this flow is supported on edges bought in the first and second stages. Constraints (6.6) encode the monotonicity requirement on the v-s path.

To incorporate the IP formulation above into (DRSO), we set the polytope specifying the fractional second-stage decisions to $\mathcal{Z} := [0,1]^{E \times \left(\overleftrightarrow{E} \times V\right) \times \left(\overleftrightarrow{E} \times V\right)}$. For a scenario $A \subseteq V$, the polytope $\mathcal{F}(A)$ is composed of the tuples $\left(x, \left(q^A, \left\{f_{\cdot}^{\mathrm{I},A,\cdot}\right\}, \left\{f_{\cdot}^{\mathrm{II},A,\cdot}\right\}\right)\right) \in \mathcal{P} \times \mathcal{Z}$ that are feasible for the LP relaxation of the IP above (wherein we replace constraints (6.7) and (6.8) with nonnegativity constraints).

We obtain results in the unrestricted setting, and leave the k-bounded setting for future work. We consider two choices for the scenario metric ℓ: the discrete metric ℓ^{disc} and the asymmetric metric $\ell_\infty^{\mathsf{asym}}$ (defined with respect to the underlying metric c^{I}). We set $\ell_{\max} := 1$ and $\ell_{\max} := \max_{u,v \in V} c^{\mathrm{I}}_{u,v}$ in these two settings respectively.

Lemma 6.27. *Assumptions (A1)–(A6) hold for* MDRSST*. Moreover, when ℓ is the discrete metric ℓ^{disc} or the asymmetric metric $\ell_\infty^{\mathsf{asym}}$, assumption (A7) holds with $\tau = \max\left\{1, \sum_{e \in E} c_e^{\mathrm{II}}\right\}$.*

Proof. Properties (A1)–(A5) follow from the discussion in the introduction of this chapter.

We now prove that property (A6) holds. Note that if $A \in \mathcal{A}$ is a non-null scenario, then A contains a vertex v such that every $v \rightsquigarrow s$ path has c-cost greater than or equal to 1. Consider any fractional feasible solution $\left(x, \left(q^A, \left\{f_{\cdot}^{\mathrm{I},A,\cdot}\right\}, \left\{f_{\cdot}^{\mathrm{II},A,\cdot}\right\}\right)\right)$ for scenario A. Then $f^{\mathrm{I},A,v} + f^{\mathrm{II},A,v}$ sends one unit of flow from v to s, and so $x + q^A$ dominates a convex combination of $v \rightsquigarrow s$ paths, which implies that the total cost of this solution for scenario A is greater than or equal to 1.

To see that property (A7) holds, note that $\min_{A,A' : \ell(A,A') > 0} \ell(A, A') \geq 1$, and that $\max_{A \in \mathcal{A}} g(0, A) \leq \sum_{e \in E} c_e^{\mathrm{II}}$. The latter inequality follows because, given any scenario $A \in \mathcal{A}$, by buying every edge $e \in E$ in the second stage, we can obtain a feasible second-stage solution for A, given $x = 0$ as a first-stage decision, of cost $\sum_{e \in E} c_e^{\mathrm{II}}$. The result then follows from the discussion in the introduction of this chapter. □

While we do not have a local approximation algorithm for MDRSST (or DRSST), we have a weaker type of rounding algorithm that resembles a local approximation algorithm.

[4] Constraint (6.3) is slightly modified with respect to the formulation in [62], which instead of enforcing that the *net outgoing flow* from any terminal is at least 1, only enforces that the *outgoing flow* is at least 1.

Definition 6.28. A *restricted local ρ-approximation algorithm* is an algorithm that takes as input a fractional first-stage decision $x \in \mathcal{P}$ and a collection of scenarios $\overline{\mathcal{A}}$, and computes in $\mathsf{poly}\big(\mathcal{I}, |\overline{\mathcal{A}}|\big)$ time an integer first-stage decision $\widehat{x} \in X$ and integer second-stage decisions $\big\{\widehat{z}^A\big\}_{A \in \overline{\mathcal{A}}}$ such that:

(i) $\big(\widehat{x}, \widehat{z}^A\big)$ is feasible for scenario A, for every $A \in \overline{\mathcal{A}}$;

(ii) $c^\intercal \widehat{x} \le \rho \cdot c^\intercal x$; and

(iii) (cost of $\widehat{z}^A$) $\le \rho \cdot g(x, A)$ for every scenario $A \in \overline{\mathcal{A}}$.

To see that this is indeed a weaker type of rounding algorithm, note that a local ρ-approximation algorithm $(\mathsf{Alg}^{\mathrm{I}}, \mathsf{Alg}^{\mathrm{II}})$ immediately yields a restricted ρ-approximation algorithm: given an instance $\big(x, \overline{\mathcal{A}}\big) \in \mathcal{P} \times 2^U$, it suffices to run $\mathsf{Alg}^{\mathrm{I}}$ with input x, then run $\mathsf{Alg}^{\mathrm{II}}$ for each of the scenarios in $\overline{\mathcal{A}}$. Gupta, Ravi, and Sinha [64] presented a restricted local 20-approximation algorithm for MDRSST, and Gupta [65] improved the approximation factor to 10.

Recall from Definition 5.7 that we say that the scenario collection $\mathcal{A}$ is *collapsible* under a scenario metric ℓ if given any scenario $A \in \mathcal{A}$, we can compute in $\mathsf{poly}(\mathcal{I})$ time a collection of scenarios $\phi(A) \subseteq \mathcal{A}$ such that for every fractional first-stage decision $x \in \mathcal{P}$ and every $y \ge 0$, we have

$$g(x, y, A) = \max_{A' \in \phi(A)} \{g(x, A') - y \cdot \ell(A, A')\} \ .$$

We show that in the collapsible setting, a *restricted* local approximation algorithm and a second-stage approximation suffice to obtain an approximation algorithm for the DRS problem ($\mathrm{DRSO}_{\mathrm{W}}$) (see Theorem 6.29). Instantiating this result for Steiner tree yields approximation algorithms for DRSST in the unrestricted setting, when ℓ is the discrete metric ℓ^{disc} or the asymmetric $\ell_\infty^{\mathsf{asym}}$ (see Theorem 6.31).

Theorem 6.29 (see proof in Section 6.8.1)**.** *Consider a generic DRS problem under a Wasserstein ball* ($\mathrm{DRSO}_{\mathrm{W}}$) *satisfying assumptions (A1)–(A7). Suppose that the scenario collection $\mathcal{A}$ is collapsible under the scenario metric ℓ, and that the second-stage costs $\{g(x, A')\}$ are given by compact LPs, say, $g(x, A') = \min\big\{s^{A'} \cdot z^{A'} : \big(x, z^{A'}\big) \in \mathcal{F}(A')\big\}$. Moreover, suppose that we have:*

(1) a second-stage α-approximation algorithm; and
(2) a restricted local ρ-approximation algorithm.

Then there exists a two-stage algorithm that, given $\varepsilon > 0$ *and* $\delta \in (0,1)$*, computes a* $4\alpha\rho(1+\varepsilon)$*-approximate solution for* ($\mathrm{DRSO_W}$) *with probability at least* $1-\delta$ *in time* $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$.

Before exploiting the theorem above to obtain our main result for DRSST, we show that we have a second-stage approximation algorithm for MDRSST.

Lemma 6.30. *We have a second-stage* 2*-approximation algorithm for* MDRSST.

Proof. Consider an integer first-stage decision $\widehat{x} \in X$ and a scenario A. Our goal is to show that we can compute an integer second-stage decision that is feasible for scenario A (under the first-stage decision $\widehat{x}$) with cost at most $2g(\widehat{x}, A)$.

Let T denote the vertex set of the connected component of $(V, \{e \in E : \widehat{x}_e = 1\})$ that contains the root s, and let $\overline{G}$ be the graph obtained from G by contracting T to a single vertex $\overline{s}$. One can show that a vector q^A coming from an optimal solution for the LP defining $g(\widehat{x}, A)$ yields a feasible fractional solution for an instance of Steiner tree on $\overline{G}$, with root $\overline{s}$, terminals $A \setminus T$, and edge costs given by c^{II}. Using the primal-dual algorithm by Agrawal, Klein, and Ravi [1], we can obtain a solution for this Steiner tree problem of cost at most $2\sum_{e\in E} c^{\mathrm{II}}_e q^A_e = 2g(\widehat{x}, A)$; this solution induces a suitable integer second-stage decision. □

We are now ready to prove our results for DRSST.

Theorem 6.31. *Consider* DRSST *under a Wasserstein ball in the unrestricted setting, where the underlying scenario metric* ℓ *is either the discrete metric* ℓ^{disc} *or the asymmetric metric* $\ell^{\mathsf{asym}}_\infty$*, defined relative to the underlying metric* c *on the vertex set* V*. There exists a two-stage algorithm that, given* $\varepsilon > 0$ *and* $\delta \in (0,1)$*, computes a* $(160 + \mathrm{O}(\varepsilon))$*-approximate solution with probability at least* $1-\delta$ *in time* $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\varepsilon}, \log\frac{1}{\delta}\big)$.

Proof. Note that the scenario collection $\mathcal{A}$ is collapsible under ℓ by Lemma 6.4. Given an instance I of DRSST, we apply Theorem 6.29 to the corresponding instance I^{mon} of MDRSST, with $\alpha = 2$ (using the second-stage approximation algorithm from Lemma 6.30) and $\rho = 10$ (using the restricted local approximation algorithm from Gupta [65]). This yields an $(80 + \mathrm{O}(\varepsilon))$-approximate solution for I^{mon}. By Lemma 6.26, this is a $(160 + \mathrm{O}(\varepsilon))$-approximate solution for I. □

6.8.1 Proof of Theorem 6.29

The following preliminary lemma shows that for a distribution $\widehat{p}$ represented explicitly, one can utilize a restricted local approximation algorithm to efficiently convert an approximate solution for $\min_{x\in\mathcal{P}} h(\widehat{p};x)$ into an approximate solution for $\min_{x\in X} h(\widehat{p};x)$. This can be seen as a variant of Lemma 3.16; the difference is that the conversion algorithm from Lemma 6.32 requires a weaker type of rounding algorithm (i.e., a *restricted* local approximation instead of a local approximation), and can only be performed efficiently in the explicit-distribution setting.

Lemma 6.32. *Consider a two-stage DRS problem under a Wasserstein ball* (DRSO_{W}) *in the explicit central-distribution setting, with input size* $\widehat{\mathcal{I}}$*. Suppose that (i) the scenario collection* $\mathcal{A}$ *is collapsible under the scenario metric* ℓ*; and (ii) we have a restricted local* ρ*-approximation algorithm. Then, given a* ψ*-approximate solution for* $\min_{x\in\mathcal{P}} h(\widehat{p};x)$*, we can compute in* $\mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ *time a* $\rho\psi$*-approximate solution for* $\min_{x\in X} h(\widehat{p};x)$*.*

Proof. Let $\mathcal{A}^{\text{sup}}$ denote the support of $\widehat{p}$, and let $\overline{\mathcal{A}} := \cup_{A\in\mathcal{A}^{\text{sup}}}\phi(A)$, where $\{\phi(A)\}$ are the scenario collections given by Definition 5.7. Note that $\left|\overline{\mathcal{A}}\right| = \mathsf{poly}(\mathcal{I}, |\mathcal{A}^{\text{sup}}|) = \mathsf{poly}\left(\widehat{\mathcal{I}}\right)$. Let $\overline{x} \in \mathcal{P}$ be a ψ-approximate solution for $\min_{x\in\mathcal{P}} h(\widehat{p};x)$, and let $\widehat{x} \in X$ be obtained by running the restricted local ρ-approximation algorithm, giving as input the fractional first-stage decision $\overline{x}$ and the scenario collection $\overline{\mathcal{A}}$; this can be computed in $\mathsf{poly}\left(\mathcal{I}, \left|\overline{\mathcal{A}}\right|\right) = \mathsf{poly}\left(\widehat{\mathcal{I}}\right)$ time.[5] We claim that $h(\widehat{p};\widehat{x}) \le \rho \cdot h(\widehat{p};\overline{x})$; this implies that

$$h(\widehat{p};\widehat{x}) \le \rho\psi \cdot \min_{x\in\mathcal{P}} h(\widehat{p};x) \le \rho\psi \cdot \min_{x\in X} h(\widehat{p};x)$$

as desired. By Lemma 4.5, and by the definition of the scenario collections $\{\phi(A)\}$, we have that

$$h(\widehat{p};x) = c^{\intercal}x + \min_{y\ge 0}\left\{ry + \sum_{A\in\mathcal{A}^{\text{sup}}} \widehat{p}_A \cdot \max_{A'\in\phi(A)}\{g(x,A') - y\cdot\ell(A,A')\}\right\}$$

[5] Although the restricted local approximation algorithm also returns integer second-stage decisions $\left\{\widehat{z}^{A}\right\}_{A\in\overline{\mathcal{A}}}$, we will not use them. We use however the fact that $g(\widehat{x},A) \le \rho\cdot g(\overline{x},A)$ for every scenario $A \in \overline{\mathcal{A}}$, which is implied by these decisions.

for every fractional first-stage decision $x \in \mathcal{P}$. Using this, we obtain

$$\begin{aligned} h(\widehat{p}\,;\widehat{x}) &= c^{\intercal}\widehat{x} + \min_{y \geq 0} \left\{ ry + \sum_{A \in \mathcal{A}^{\sup}} \widehat{p}_A \cdot \max_{A' \in \phi(A)} \{g(\widehat{x}, A') - y \cdot \ell(A, A')\} \right\} \\ &\leq \rho \cdot c^{\intercal}\overline{x} + \min_{y \geq 0} \left\{ ry + \sum_{A \in \mathcal{A}^{\sup}} \widehat{p}_A \cdot \max_{A' \in \phi(A)} \{\rho \cdot g(\overline{x}, A') - y \cdot \ell(A, A')\} \right\} \\ &\leq \rho \left(c^{\intercal}\overline{x} + \min_{y \geq 0} \left\{ ry + \sum_{A \in \mathcal{A}^{\sup}} \widehat{p}_A \cdot \max_{A' \in \phi(A)} \{g(\overline{x}, A') - y \cdot \ell(A, A')\} \right\} \right) \\ &= \rho \cdot h(\widehat{p}\,;\overline{x}) \ , \end{aligned}$$

where the first inequality follows from the guarantees of the restricted local approximation algorithm (since by definition $\phi(A) \subseteq \overline{\mathcal{A}}$ for every $A \in \mathcal{A}^{\sup}$). □

Proof of Theorem 6.29. The proof follows very closely the structure of the proof of Theorem 3.6, so we omit some details. Using Lemma 6.1, we either obtain that $\overline{x} = 0$ is an optimal solution for $\min_{x \in \mathcal{P}} h(\mathring{p}\,;x)$ (and hence for $\min_{x \in X} h(\mathring{p}\,;x)$), or obtain a lower bound on $\min_{x \in \mathcal{P}} h(\widetilde{p}\,;x)$ for every distribution $\widetilde{p}$. In the latter case, Theorem 5.8 gives an exact algorithm for the SAA problem $\min_{x \in \mathcal{P}} h(\widehat{p}\,;x)$. Combining this with Lemma 6.32, we obtain a ρ-approximation algorithm for the SAA problem $\min_{x \in X} h(\widehat{p}\,;x)$. We can then invoke the SAA result from Theorem 4.1 to compute a $4\rho(1+\varepsilon)$-approximate solution $\overline{x}$ for $\min_{x \in X} h(\mathring{p}\,;x)$. (Note that we have an exact value oracle for the SAA objective function $h(\widehat{p}\,;\cdot)$; this follows because we can write $z(\widehat{p}\,;x)$ as a compact LP, mimicking the proof of Theorem 5.8.) We then use the second-stage approximation algorithm to obtain second-stage decisions; by Lemma 3.15, this yields an $4\alpha\rho(1+\varepsilon)$-approximate solution for (DRSO_{W}). □

Chapter 7

DRS optimization under an L_∞ ball

In this chapter, we consider the discrete DRS problem under an L_∞ ball

$$\min_{\substack{x\in X, z\in Z^{\mathcal{A}}:\\ (x,z^A)\in F(A)\quad \forall A\in\mathcal{A}}} \left\{ c^\intercal x + \sup_{p: L_\infty(\mathring{p},p)\le r} \mathbb{E}_{A\sim p}\big[\text{cost of } z^A\big] \right\} , \tag{DRSO$_\infty$}$$

where the central distribution $\mathring{p}$ is given by a sampling oracle. Throughout this chapter, we assume without loss of generality that $r \le 1$ (since $L_\infty(\mathring{p}, p) \le 1$ for every distribution p).

Recall that $z(\mathring{p}\,; x) := \max_{p\in D} \mathbb{E}_{A\sim p}[g(x, A)]$ denotes the expected cost incurred in the second stage if we take the fractional first-stage decision $x \in \mathcal{P}$, *and* if we allow fractional second-stage decisions. We work with the relaxation with fractional first-stage and second-stage decisions

$$\min_{x\in\mathcal{P}} \{h(\mathring{p}\,; x) := c^\intercal x + z(\mathring{p}\,; x)\} \ . \tag{Q$^{\text{fr}}$}$$

Our main result in this chapter is that we can obtain an approximation algorithm for problem (Q$^{\text{fr}}$) via an algorithm for the following problem:

(Υ) Given a fractional first-stage decision $x \in \mathcal{P}$ and $1 \le t \le \min\left\{ |\mathcal{A}|, \frac{1}{r} \right\}$, find the t scenarios $A \in \mathcal{A}$ with largest $g(x, A)$ value.

Theorem 7.1 (see proof in Section 7.6)**.**
Suppose that we have a $\mathsf{poly}(\mathcal{I}, t)$*-time algorithm for problem* (Υ)*. Given* $\varepsilon > 0$ *and* $\delta \in (0, 1)$*, we can compute a fractional first-stage decision* $\overline{x} \in \mathcal{P}$ *such that*

$$h(\mathring{p}\,; \overline{x}) \leq (2 + \varepsilon) \cdot \min_{x \in \mathcal{P}} h(\mathring{p}\,; x)$$

with probability at least $1 - \delta$ *in* $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{r}, \frac{1}{\varepsilon}, \log \frac{1}{\delta}\big)$ *time.*

Combining this with a local ρ-approximation algorithm yields an approximation factor $(2 + \varepsilon)\rho$ for the discrete DRS problem (DRSO_∞) (see Theorem 3.9).

Note that when $|\mathcal{A}|$ is $\mathsf{poly}(\mathcal{I})$, then problem ($\Upsilon$) can be trivially solved in $\mathsf{poly}(\mathcal{I})$ time. Indeed, as is the case for DRS optimization under a Wasserstein ball (see Remark 3.7 and Section 5.3), if (i) $|\mathcal{A}|$ is $\mathsf{poly}(\mathcal{I})$ and (ii) the central distribution $\mathring{p}$ is represented explicitly, then the fractional relaxation of the DRS problem, $\min_{x \in \mathcal{P}} h(\mathring{p}\,; x)$, can again be cast as a compact LP, and hence solved easily in polynomial time.

Organization of this chapter. In Section 7.1 we give an overview of the techniques used to prove the theorem above; the proof is presented in Sections 7.2–7.6. In Section 7.7 we utilize Theorem 3.9 to obtain approximation algorithms for various applications.

7.1 Overview of the techniques

At a high level, our approach is as follows. We first show how to obtain a suitable convex proxy function $h^{\mathrm{pr}}(\mathring{p}\,; x)$ that is pointwise close to the objective function $h(\mathring{p}\,; x)$ (see Lemma 7.2). Instead of utilizing an SAA approach to move to an SAA version of $h^{\mathrm{pr}}(\mathring{p}\,; x)$ with a central distribution of moderate support size, show that a near-optimal solution for the SAA problem translates to a near-optimal solution for the original problem, and finally show how to approximately solve the SAA problem (which is again challenging and requires the ellipsoid method, since this does not reduce to a polynomial-size LP), it is simpler to directly solve the proxy problem, $\min_{x \in \mathcal{P}} h^{\mathrm{pr}}(\mathring{p}\,; x)$, using the approximate-subgradient based machinery in Shmoys and Swamy [114]. We show that, if we have an algorithm for the problem (Υ) of computing the t worst scenarios for a given fractional first-stage decision $x \in \mathcal{P}$, then we can compute an ω-subgradient of $h^{\mathrm{pr}}(\mathring{p}\,; \cdot)$ at x in $\mathsf{poly}\big(\mathcal{I}, \frac{1}{r}, \lambda, \frac{1}{\omega}\big)$ time (see Lemma 7.7), and hence can directly use the ellipsoid-based approach in [114] to obtain a

solution $\overline{x} \in \mathcal{P}$ such that

$$h^{\mathrm{pr}}(\mathring{p}\,; \overline{x}) \leq (1 + \mathrm{O}(\varepsilon)) \min_{x \in \mathcal{P}} h^{\mathrm{pr}}(\mathring{p}\,; x) + \eta \ .$$

This in turn implies that

$$h(\mathring{p}\,; \overline{x}) \leq (2 + \mathrm{O}(\varepsilon)) \min_{x \in \mathcal{P}} h(\mathring{p}\,; x) + \eta \ .$$

The algorithm from [114] also requires a bound on the Lipschitz constant of the proxy function; we show how this can be obtained in Section 7.5. We can fold the additive error into a multiplicative error by computing a lower bound on $\min_{x \in \mathcal{P}} h^{\mathrm{pr}}(\mathring{p}\,; x)$ (see Lemma 7.12).

7.2 A proxy function for $h(\mathring{p}\,; x)$

In this section, we introduce a proxy function for the objective function $h(\mathring{p}\,; x)$ of problem (Q^{fr}), and show that these two functions are pointwise close. Let p be any distribution in the ambiguity set D (that is, p satisfies $L_\infty(\mathring{p}, p) \leq r$), and consider a scenario $A \in \mathcal{A}$. Note that we must have $p_A \geq \max\{\mathring{p}_A - r, 0\}$. We refer to the right side of this inequality as the *blocked mass* in scenario A. The remainder of the probability mass $\mathring{p}_A$ (i.e., the difference between $\mathring{p}_A$ and the blocked massed) may be moved to other scenarios, and hence we call it the *free mass* in scenario $\mathcal{A}$.

Separating the free mass and the blocked mass of all the scenarios, we obtain a decomposition $\mathring{p} = \mathring{p}^{\mathrm{free}} + \mathring{p}^{\mathrm{blocked}}$, where $\mathring{p}_A^{\mathrm{blocked}} := \max\{\mathring{p}_A - r, 0\}$ and $\mathring{p}_A^{\mathrm{free}} := \mathring{p}_A - \mathring{p}_A^{\mathrm{blocked}} = \min\{\mathring{p}_A, r\}$ for every scenario $A \in \mathcal{A}$. The definition of our proxy function is based on the following (informal) interpretation of the ambiguity set D. A distribution p belongs to D if it can be obtained from $\mathring{p}$ by redistributing the free mass $\mathring{p}^{\mathrm{free}}$ among the different scenarios (while keeping the blocked mass $\mathring{p}^{\mathrm{blocked}}$ steady), and ensuring that this does not increase the total mass of any scenario by more than r. It is not hard to see that, if $p \in D$, then we can perform this redistribution in such a way that each scenario $A \in \mathcal{A}$ only receives free mass from other scenarios, or sends free mass to other scenarios, *but not both*. With this restriction, imposing that the total mass on each scenario increases by at most r is equivalent to imposing that it receives a mass of at most r from other scenarios. Letting $\mathring{P}^{\mathrm{free}} := \sum_{A \in \mathcal{A}} \mathring{p}_A^{\mathrm{free}}$ denote the total free mass in $\mathring{p}$, this motivates "decomposing" $z(\mathring{p}\,; x)$ as

$$\mathbb{E}_{A \sim \mathring{p}}[g(x, A)] + \max_{q \in \mathcal{R}} \sum_{A \in \mathcal{A}} q_A g(x, A) \ , \tag{7.1}$$

where

$$\mathcal{R} := \left\{ q \in \mathbb{R}_+^{\mathcal{A}} : \sum_{A \in \mathcal{A}} q_A \leq \mathring{P}^{\text{free}}, \quad q_A \leq r \ \forall A \in \mathcal{A} \right\} .$$

The first term in (7.1) can be seen as the maximum contribution to $z(\mathring{p}\,; x)$ from mass that remains in its original scenario, whereas the second term can be seen as the maximum contribution from mass that is redistributed.

Note however that we cannot compute $\mathring{P}^{\text{free}}$ exactly, since we only have access to $\mathring{p}$ via a sampling oracle. Our main result in this section is that if we have a suitable estimate $\widehat{P}^{\text{free}}$ of $\mathring{P}^{\text{free}}$, then we can use a decomposition of $z(\mathring{p}\,; x)$ as described above (but using $\widehat{P}^{\text{free}}$ instead of $\mathring{P}^{\text{free}}$) to obtain a proxy function that is pointwise close to $h(\mathring{p}\,; x)$.

Lemma 7.2. *Let $\varepsilon > 0$, and let $\widehat{P}^{\text{free}}$ be an estimate of $\mathring{P}^{\text{free}}$ such that*

$$\mathring{P}^{\text{free}} \leq \widehat{P}^{\text{free}} \leq \min\left\{ (1+\varepsilon)\mathring{P}^{\text{free}}, 1 \right\} .$$

Consider the proxy function

$$h^{\text{pr}}(\mathring{p}\,; x) := c^\intercal x + \mathbb{E}_{A \sim \mathring{p}}[g(x, A)] + \underbrace{\max_{q \in \widehat{\mathcal{R}}} \sum_{A \in \mathcal{A}} q_A g(x, A)}_{(\mathrm{W}_x)} ,$$

where $\widehat{\mathcal{R}} := \left\{ q \in \mathbb{R}_+^{\mathcal{A}} : \sum_{A \in \mathcal{A}} q_A \leq \widehat{P}^{\text{free}}, \quad q_A \leq r \ \forall A \in \mathcal{A} \right\}$. For every fractional first-stage decision $x \in \mathcal{P}$, we have

$$h(\mathring{p}\,; x) \leq h^{\text{pr}}(\mathring{p}\,; x) \leq (2+\varepsilon) \cdot h(\mathring{p}\,; x) .$$

Before proving Lemma 7.2, we need the following preliminary lemma.

Lemma 7.3. *For every fractional first-stage decision $x \in \mathcal{P}$ we have*

$$\mathrm{OPT}(\mathrm{W}_x) \leq (1+\varepsilon) \cdot z(\mathring{p}\,; x) .$$

Proof. Let q^* be an optimal solution for (W_x). We prove that there exists a distribution $\widetilde{q} \in D$ such that $\widetilde{q} \geq \frac{1}{1+\varepsilon} q^*$. Assuming this, the result follows, since we obtain

$$z(\mathring{p}\,; x) \geq \mathbb{E}_{A \sim \widetilde{q}}[g(x, A)] \geq \frac{1}{1+\varepsilon} \sum_{A \in \mathcal{A}} q_A^* g(x, A) = \frac{1}{1+\varepsilon} \mathrm{OPT}(\mathrm{W}_x) .$$

We give a constructive proof of the existence of $\widetilde{q}$, via an iterative algorithm. We start by setting $\widetilde{q} := \mathring{p}^{\text{blocked}} + \frac{1}{1+\varepsilon} q^*$. We claim that

(i) $\widetilde{q} \geq \frac{1}{1+\varepsilon} q^*$;

(ii) $\widetilde{q}_A \geq \max\{\mathring{p}_A - r, 0\}$ for every scenario $A \in \mathcal{A}$; and

(iii) $\widetilde{q}_A \leq \min\{\mathring{p}_A + r, 1\}$ for every scenario $A \in \mathcal{A}$.

Note that (i) follows immediately from the definition of $\widetilde{q}$, as does (ii) (since $\widetilde{q} \geq \mathring{p}^{\text{blocked}}$). To show (iii), note that for every scenario $A \in \mathcal{A}$ we have

$$\widetilde{q}_A = \max\{\mathring{p}_A - r, 0\} + \frac{1}{1+\varepsilon} q^*_A \leq \max\{\mathring{p}_A, r\} \leq \min\{\mathring{p}_A + r, 1\} \ .$$

The first inequality follows because $q^* \in \widehat{\mathcal{R}}$ implies $\frac{1}{1+\varepsilon} q^*_A \leq \frac{1}{1+\varepsilon} r \leq r$.

From now on, we iteratively modify $\widetilde{q}$ to obtain $\sum_{A \in \mathcal{A}} \widetilde{q}_A = 1$, while preserving the invariants (i)–(iii). Note that if we achieve this, then we are done: (ii), (iii), and $\sum_{A \in \mathcal{A}} \widetilde{q}_A = 1$ combined imply that $\widetilde{q}$ is a probability distribution with $L_\infty(\mathring{p}, \widetilde{q}) \leq r$ and hence $\widetilde{q} \in D$.

Note that

$$\sum_{A \in \mathcal{A}} \widetilde{q}_A = \sum_{A \in \mathcal{A}} \left(\mathring{p}^{\text{blocked}}_A + \frac{1}{1+\varepsilon} q^*_A \right) \leq \sum_{A \in \mathcal{A}} \mathring{p}^{\text{blocked}}_A + \mathring{P}^{\text{free}} = 1 \ ,$$

where the inequality follows because $q^* \in \widehat{\mathcal{R}}$ implies $\sum_{A \in \mathcal{A}} q^*_A \leq \widehat{P}^{\text{free}} \leq (1+\varepsilon)\mathring{P}^{\text{free}}$. If this inequality is tight, then we are done. Otherwise, since $\sum_{A \in \mathcal{A}} \mathring{p}_A = 1$, there must exist a scenario $A \in \mathcal{A}$ such that $\widetilde{q}_A < \mathring{p}_A$. We increase the component $\widetilde{q}_A$ until one of the following stopping conditions is reached (whichever happens first): $\sum_{A \in \mathcal{A}} \widetilde{q}_A = 1$ or $\widetilde{q}_A = \mathring{p}_A + r$. If we still have $\sum_{A \in \mathcal{A}} \widetilde{q}_A < 1$, then we repeat the same step with a different scenario. As each step (except possibly the final one) decreases the number of scenarios A such that $\widetilde{q}_A < \mathring{p}_A$, this process eventually stops, and so at this moment we have $\sum_{A \in \mathcal{A}} \widetilde{q}_A = 1$. Note that these operations preserve invariants (i) and (ii), since we are only increasing components of $\widetilde{q}$. Invariant (iii) is also preserved due to the second stopping condition used at each iteration. □

We are now ready to prove Lemma 7.2.

Proof of Lemma 7.2. Let $x \in \mathcal{P}$ be fixed throughout this proof. We start by showing the first inequality. Let $q^* := \operatorname{argmax}_{q: L_\infty(\mathring{p}, q) \le r} \mathbb{E}_{A \sim q}[g(x, A)]$, so that $h(\mathring{p}\,; x) = c^\intercal x + \mathbb{E}_{A \sim q^*}[g(x, A)]$. We decompose q^* into two vectors as follows: we write $q^* = q^1 + q^2$, where $q^1_A := \min\{q^*_A, \mathring{p}_A\}$ and $q^2_A := q^*_A - q^1_A$ for every scenario $A \in \mathcal{A}$. Next we upper bound the contribution of each of these two vectors to the objective value $h(\mathring{p}\,; x)$. Since $q^1 \le \mathring{p}$, we have $\sum_{A \in \mathcal{A}} q^1_A g(x, A) \le \mathbb{E}_{A \sim \mathring{p}}[g(x, A)]$. We claim that $q^2 \in \widehat{\mathcal{R}}$. Assuming this, we obtain

$$\begin{aligned} h(\mathring{p}\,; x) &= c^\intercal x + \sum_{A \in \mathcal{A}} q^1_A g(x, A) + \sum_{A \in \mathcal{A}} q^2_A g(x, A) \\ &\le c^\intercal x + \mathbb{E}_{A \sim \mathring{p}}[g(x, A)] + \max_{q \in \widehat{\mathcal{R}}} \sum_{A \in \mathcal{A}} q_A g(x, A) \\ &= h^{\mathrm{pr}}(\mathring{p}\,; x)\ . \end{aligned}$$

We now prove the claim that $q^2 \in \widehat{\mathcal{R}}$. First, note that for every scenario $A \in \mathcal{A}$ we have $0 \le q^2_A \le r$: if $q^*_A \le \mathring{p}_A$, we have $q^2_A = 0$; otherwise we have $q^2_A = q^*_A - \mathring{p}_A$, and hence $0 \le q^2_A \le r$ since $L_\infty(\mathring{p}, q^*) \le r$. Furthermore, we have

$$\begin{aligned} \sum_{A \in \mathcal{A}} q^2_A &= \sum_{A \in \mathcal{A}} (q^*_A - \min\{q^*_A, \mathring{p}_A\}) \\ &= \sum_{A \in \mathcal{A}: q^*_A > \mathring{p}_A} (q^*_A - \mathring{p}_A) \\ &= \sum_{A \in \mathcal{A}: q^*_A < \mathring{p}_A} (\mathring{p}_A - q^*_A) \\ &\le \sum_{A \in \mathcal{A}: q^*_A < \mathring{p}_A} (\mathring{p}_A - \max\{\mathring{p}_A - r, 0\}) \\ &= \sum_{A \in \mathcal{A}: q^*_A < \mathring{p}_A} \mathring{p}^{\mathrm{free}}_A \\ &\le \mathring{P}^{\mathrm{free}} \\ &\le \widehat{P}^{\mathrm{free}}\ . \end{aligned}$$

The third equality holds because

$$\sum_{A \in \mathcal{A}: q^*_A > \mathring{p}_A} (q^*_A - \mathring{p}_A) - \sum_{A \in \mathcal{A}: q^*_A < \mathring{p}_A} (\mathring{p}_A - q^*_A) = \sum_{A \in \mathcal{A}} (q^*_A - \mathring{p}_A) = 0\ ,$$

where the last step uses the fact that q^* and $\mathring{p}$ are probability distributions. The first inequality follows because $L_\infty(\mathring{p}, q^*) \le r$. This concludes the proof of the claim.

Now we proceed to prove the second inequality in the lemma statement. Since $\mathring{p} \in D$, we have $z(\mathring{p}\,; x) \ge \mathbb{E}_{A\sim\mathring{p}}[g(x, A)]$. By Lemma 7.3, we have $\mathrm{OPT}(\mathrm{W}_x) \le (1+\varepsilon) \cdot z(\mathring{p}\,; x)$. Using these two inequalities, we obtain

$$\begin{aligned} h^{\mathrm{pr}}(\mathring{p}\,; x) &= c^\intercal x + \mathbb{E}_{A\sim\mathring{p}}[g(x, A)] + \mathrm{OPT}(\mathrm{W}_x) \\ &\le c^\intercal x + (2+\varepsilon) z(\mathring{p}\,; x) \\ &\le (2+\varepsilon) \cdot h(\mathring{p}\,; x)\ . \end{aligned}$$

□

7.3 Estimating $\mathring{P}^{\mathrm{free}}$

Recall that the proxy function $h^{\mathrm{pr}}(\mathring{p}\,; \cdot)$ introduced in Section 7.2 requires a suitable estimate of the total free mass $\mathring{P}^{\mathrm{free}} := \sum_{A\in\mathcal{A}} \mathring{p}_A^{\mathrm{free}}$. In this section, we present an algorithm for computing such an estimate.

Lemma 7.4. *Given $\varepsilon > 0$ and $\delta \in (0,1)$, we can compute in $\mathsf{poly}\big(\mathcal{I}, \frac{1}{\varepsilon}, \frac{1}{r}, \log\frac{1}{\delta}\big)$ time an estimate $\widehat{P}^{\mathrm{free}}$ of $\mathring{P}^{\mathrm{free}}$ such that with probability at least $1-\delta$ we have*

$$\mathring{P}^{\mathrm{free}} \le \widehat{P}^{\mathrm{free}} \le \min\Big\{(1+\varepsilon)\mathring{P}^{\mathrm{free}}, 1\Big\}\ .$$

Before proving Lemma 7.4, we need two preliminary lemmas. We start by obtaining a lower bound on $\mathring{P}^{\mathrm{free}}$.

Lemma 7.5. *We have $\mathring{P}^{\mathrm{free}} \ge r$.*

Proof. If there exists a scenario $A \in \mathcal{A}$ with $\mathring{p}_A^{\mathrm{free}} \ge r$, then we have $\mathring{P}^{\mathrm{free}} \ge \mathring{p}_A^{\mathrm{free}} \ge r$. Otherwise, we have $\mathring{P}^{\mathrm{free}} = \sum_{A\in\mathcal{A}} \mathring{p}_A^{\mathrm{free}} = \sum_{A\in\mathcal{A}} \mathring{p}_A = 1 \ge r$. □

We partition the scenario collection $\mathcal{A}$ into a collection of *frequent scenarios* $\mathcal{A}^{\mathrm{freq}} := \{A \in \mathcal{A} : \mathring{p}_A \ge r\}$ and a collection of *rare* scenarios $\mathcal{A}^{\mathrm{rare}} := \{A \in \mathcal{A} : \mathring{p}_A < r\}$. The lemma below shows that, in order to obtain a suitable estimate of $\mathring{P}^{\mathrm{free}}$, it suffices to utilize an empirical estimate of $\widehat{p}$ of $\mathring{p}$ that is accurate enough over a superset of the frequent scenarios.

Lemma 7.6. *Consider a partition $\mathcal{A} = \widehat{\mathcal{A}}^{\mathrm{freq}} \cup \widehat{\mathcal{A}}^{\mathrm{rare}}$ of the scenario collection, with $\mathcal{A}^{\mathrm{freq}} \subseteq \widehat{\mathcal{A}}^{\mathrm{freq}}$ (and hence $\widehat{\mathcal{A}}^{\mathrm{rare}} \subseteq \mathcal{A}^{\mathrm{rare}}$). Let $\widehat{p}$ be a probability distribution such that*

$\sum_{A\in\widehat{\mathcal{A}}^{\text{freq}}} |\widehat{p}_A - \mathring{p}_A| \le \frac{1}{4}\varepsilon r$. *Let* $Q^{\text{free}} := \sum_{A\in\widehat{\mathcal{A}}^{\text{freq}}} \min\{\widehat{p}_A, r\} + \sum_{A\in\widehat{\mathcal{A}}^{\text{rare}}} \widehat{p}_A$ *and* $\widehat{P}^{\text{free}} := \min\left\{Q^{\text{free}} + \frac{1}{2}\varepsilon r, 1\right\}$. *Then we have*

$$\mathring{P}^{\text{free}} \le \widehat{P}^{\text{free}} \le \min\left\{(1+\varepsilon)\mathring{P}^{\text{free}}, 1\right\} .$$

Proof. We first show that the first sum in the definition of Q^{free} is a good estimate of the amount of free mass in $\widehat{\mathcal{A}}^{\text{freq}}$. We have

$$\begin{aligned}
\left|\sum_{A\in\widehat{\mathcal{A}}^{\text{freq}}} \min\{\widehat{p}_A, r\} - \sum_{A\in\widehat{\mathcal{A}}^{\text{freq}}} \mathring{p}_A^{\text{free}}\right| &\le \sum_{A\in\widehat{\mathcal{A}}^{\text{freq}}} \left|\min\{\widehat{p}_A, r\} - \mathring{p}_A^{\text{free}}\right| \\
&= \sum_{A\in\widehat{\mathcal{A}}^{\text{freq}}} |\min\{\widehat{p}_A, r\} - \min\{\mathring{p}_A, r\}| \\
&\le \sum_{A\in\widehat{\mathcal{A}}^{\text{freq}}} |\widehat{p}_A - \mathring{p}_A| \\
&\le \frac{1}{4}\varepsilon r ,
\end{aligned} \tag{7.2}$$

where the first step uses the triangle inequality, and the final step is by assumption.

Next we show that the second sum in the definition of Q^{free} is a good estimate of the amount of free mass in $\widehat{\mathcal{A}}^{\text{rare}}$. We have

$$\begin{aligned}
\left|\sum_{A\in\widehat{\mathcal{A}}^{\text{rare}}} \widehat{p}_A - \sum_{A\in\widehat{\mathcal{A}}^{\text{rare}}} \mathring{p}_A^{\text{free}}\right| &= \left|\sum_{A\in\widehat{\mathcal{A}}^{\text{rare}}} \widehat{p}_A - \sum_{A\in\widehat{\mathcal{A}}^{\text{rare}}} \mathring{p}_A\right| \\
&= \left|\sum_{A\in\widehat{\mathcal{A}}^{\text{freq}}} \widehat{p}_A - \sum_{A\in\widehat{\mathcal{A}}^{\text{freq}}} \mathring{p}_A\right| \\
&\le \sum_{A\in\widehat{\mathcal{A}}^{\text{freq}}} |\widehat{p}_A - \mathring{p}_A| \\
&\le \frac{1}{4}\varepsilon r ,
\end{aligned} \tag{7.3}$$

The first step follows because $\widehat{\mathcal{A}}^{\text{rare}} \subseteq \mathcal{A}^{\text{rare}}$ and $\mathring{p}_A^{\text{free}} = \mathring{p}_A$ for all $A \in \mathcal{A}^{\text{rare}}$. The second step follows because $\mathring{p}$ and $\widehat{p}$ are probability distributions and $\left\{\widehat{\mathcal{A}}^{\text{freq}}, \widehat{\mathcal{A}}^{\text{rare}}\right\}$ is a partition of $\mathcal{A}$. The third step follows from the triangle inequality. The final step holds by assumption.

Combining (7.2) and (7.3) yields $\left|Q^{\text{free}} - \mathring{P}^{\text{free}}\right| \leq \frac{1}{2}\varepsilon r$. We now show that this implies that the estimate $\widehat{P}^{\text{free}}$ has the claimed guarantees. If $\widehat{P}^{\text{free}} = 1$, then we clearly have $\widehat{P}^{\text{free}} \geq \mathring{P}^{\text{free}}$; otherwise we have $\widehat{P}^{\text{free}} = Q^{\text{free}} + \frac{1}{2}\varepsilon r \geq \mathring{P}^{\text{free}}$. On the other hand, we have

$$\widehat{P}^{\text{free}} = \min\left\{Q^{\text{free}} + \frac{1}{2}\varepsilon r, 1\right\} \leq \min\left\{\mathring{P}^{\text{free}} + \varepsilon r, 1\right\} \leq \left\{(1+\varepsilon)\mathring{P}^{\text{free}}, 1\right\} ,$$

where last step follows from Lemma 7.5. □

We are now ready to prove Lemma 7.4.

Proof of Lemma 7.4. We first work toward computing a superset $\widehat{\mathcal{A}}^{\text{freq}}$ of $\mathcal{A}^{\text{freq}}$. Note that $\left|\mathcal{A}^{\text{freq}}\right| \leq \frac{1}{r}$, since

$$1 = \sum_{A \in \mathcal{A}} \mathring{p}_A \geq \sum_{A \in \mathcal{A}^{\text{freq}}} \mathring{p}_A \geq \left|\mathcal{A}^{\text{freq}}\right| \cdot r .$$

Let $\widehat{p}$ be an empirical estimate of $\mathring{p}$ constructed N_1 samples (we will determine the value of N_1 later). For any frequent scenario $A \in \mathcal{A}^{\text{freq}}$, we can see $\widehat{p}_A$ as an empirical estimate (computed using N_1 samples) of the indicator random variable

$$W := \begin{cases} 1 & \text{, if } A' = A \text{ ;} \\ 0 & \text{, otherwise,} \end{cases}$$

where A' is sampled according to the distribution $\mathring{p}$. Note that $\mathbb{E}[W] = \mathring{p}_A$. By Hoeffding's inequality (Corollary 2.4), we can choose $N_1 = \mathsf{poly}\left(\frac{1}{\frac{r}{2}}, \log\frac{1}{\frac{\delta r}{2}}\right) = \mathsf{poly}\left(\frac{1}{r}, \log\frac{1}{\delta}\right)$ such that for every $A \in \mathcal{A}^{\text{freq}}$ we have $\mathbf{Pr}\left[\left|\widehat{p}_A - \mathring{p}_A\right| > \frac{r}{2}\right] \leq \frac{\delta r}{2}$. By the union bound, we have with probability at least $1 - \left|\mathcal{A}^{\text{freq}}\right|\frac{\delta r}{2} \geq 1 - \frac{\delta}{2}$ that for every $A \in \mathcal{A}^{\text{freq}}$ the inequality $\left|\widehat{p}_A - \mathring{p}_A\right| \leq \frac{r}{2}$ holds, and so $\widehat{p}_A \geq \mathring{p}_A - \frac{r}{2} \geq \frac{r}{2}$. We set $\widehat{\mathcal{A}}^{\text{freq}} := \left\{A : \widehat{p}_A \geq \frac{r}{2}\right\}$, so we have $\mathcal{A}^{\text{freq}} \subseteq \widehat{\mathcal{A}}^{\text{freq}}$ with probability at least $1 - \frac{\delta}{2}$.

Now that we have $\widehat{\mathcal{A}}^{\text{freq}}$, we work toward obtaining another empirical estimate $\widehat{p}$ of $\mathring{p}$ using N_2 independent samples (where N_2 will be defined later) that is sufficiently accurate for the scenarios in $\widehat{\mathcal{A}}^{\text{freq}}$, so that we can compute an estimate of $\mathring{P}^{\text{free}}$ as in Lemma 7.6. First, note that $\left|\widehat{\mathcal{A}}^{\text{freq}}\right| \leq \frac{2}{r}$, since

$$1 = \sum_{A \in \mathcal{A}} \widehat{p}_A \geq \sum_{A \in \widehat{\mathcal{A}}^{\text{freq}}} \widehat{p}_A \geq \left|\widehat{\mathcal{A}}^{\text{freq}}\right| \cdot \frac{r}{2} .$$

Using Hoeffding's inequality (Corollary 2.4) and the union bound as above, we can choose

$$N_2 = \mathsf{poly}\left(\frac{1}{\frac{\varepsilon r}{4\left|\widehat{\mathcal{A}}^{\text{freq}}\right|}}, \log \frac{1}{\frac{\delta}{2\left|\widehat{\mathcal{A}}^{\text{freq}}\right|}}\right) = \mathsf{poly}\left(\frac{1}{\varepsilon}, \frac{1}{r}, \log \frac{1}{\delta}\right)$$

such that with probability at least $1 - \left|\widehat{\mathcal{A}}^{\text{freq}}\right| \frac{\delta}{2\left|\widehat{\mathcal{A}}^{\text{freq}}\right|} = 1 - \frac{\delta}{2}$ the inequality $|\widehat{p}_A - \mathring{p}_A| \leq \frac{\varepsilon r}{4\left|\widehat{\mathcal{A}}^{\text{freq}}\right|}$ holds for every $A \in \widehat{\mathcal{A}}^{\text{freq}}$. Adding this inequality over all scenarios $A \in \widehat{\mathcal{A}}^{\text{freq}}$ yields $\sum_{A \in \widehat{\mathcal{A}}^{\text{freq}}} |\widehat{p}_A - \mathring{p}_A| \leq \frac{1}{4}\varepsilon r$. We can then invoke Lemma 7.6 to compute the estimate $\widehat{P}^{\text{free}}$. The success probability is at least $\left(1 - \frac{\delta}{2}\right)^2 \geq 1 - \delta$. □

7.4 Computing approximate subgradients of the proxy function

In this section, we discuss how to compute ω-subgradients of the proxy function $h^{\text{pr}}(\mathring{p}\,;\cdot)$, utilizing an algorithm for problem (Υ).

Lemma 7.7. *Let $\omega > 0$ and $\delta > 0$. Suppose that we have an algorithm for problem (Υ) with running time $\mathsf{poly}(\mathcal{I}, t)$. Given any fractional first-stage decision $x \in \mathcal{P}$, we can compute a vector $\widehat{d}$ that is an ω-subgradient of the proxy function $h^{\text{pr}}(\mathring{p}\,;\cdot)$ at x with probability at least $1 - \delta$ in $\mathsf{poly}\big(\mathcal{I}, \frac{1}{r}, \lambda, \frac{1}{\omega}, \log \frac{1}{\delta}\big)$ time.*

Part (i) of Lemma 7.8 gives an exact expression for a subgradient of $h^{\text{pr}}(\mathring{p}\,;\cdot)$ at an arbitrary point $x \in \mathcal{P}$. Part (ii) shows that we can compute an ω-subgradient at x if we have (a) a vector that is componentwise close to $\mathbb{E}_{A\sim\mathring{p}}\big[d^{x,A}\big]$; and (b) an optimal solution for (W_x). (Recall that $d^{x,A}$ denotes a subgradient of $g(\cdot, A)$ at x with $-\lambda c \leq d^{x,A} \leq 0$, which we can compute efficiently by assumption (A5').) Complementing this, Lemma 7.9 shows that we can obtain (a) by using an empirical estimate of $\mathring{p}$ constructed with a suitable number of samples, and Lemma 7.10 shows that we can obtain (b) using the algorithm for (Υ).

Lemma 7.8. *The function $h^{\text{pr}}(\mathring{p}\,;\cdot)$ is convex. Furthermore, for any fractional first stage-decision $x \in \mathcal{P}$, letting q^* denote an optimal solution to (W_x), we have:*

(i) the vector $d := c + \mathbb{E}_{A\sim\mathring{p}}\big[d^{x,A}\big] + \sum_{A\in\mathcal{A}} q^*_A d^{x,A}$ is a subgradient of $h^{\text{pr}}(\mathring{p}\,;\cdot)$ at x; and

(ii) if $d^{\text{est}} \in \mathbb{R}^m$ *is a vector such that* $-\omega c \le d^{\text{est}} - \mathbb{E}_{A\sim\mathring{p}}\big[d^{x,A}\big] \le 0$*, then* $\widehat{d} := c + d^{\text{est}} + \sum_{A\in\mathcal{A}} q^*_A d^{x,A}$ *is an* ω*-subgradient of* $h^{\text{pr}}(\mathring{p}\,;\cdot)$ *at* x*.*

Proof. Convexity of $h^{\text{pr}}(\mathring{p}\,;\cdot)$ will follow from the fact that we have a subgradient of $h^{\text{pr}}(\mathring{p}\,;\cdot)$ at every point $x \in \mathcal{P}$ (which follows from part (i)). Note that part (i) is a special case of part (ii) with $\omega = 0$, so we focus on proving part (ii). For any $x' \in \mathcal{P}$, we have

$$\begin{aligned}
h^{\text{pr}}(\mathring{p}\,;x') - h^{\text{pr}}(\mathring{p}\,;x) &\ge c^\intercal(x'-x) + \mathbb{E}_{A\sim\mathring{p}}[g(x',A) - g(x,A)] + \sum_{A\in\mathcal{A}} q^*_A(g(x',A) - g(x,A)) \\
&\ge c^\intercal(x'-x) + \mathbb{E}_{A\sim\mathring{p}}\big[d^{x,A}\cdot(x'-x)\big] + \sum_{A\in\mathcal{A}} q^*_A d^{x,A}\cdot(x'-x) \\
&\ge c^\intercal(x'-x) + d^{\text{est}}\cdot(x'-x) \\
&\quad + \sum_{A\in\mathcal{A}} q^*_A d^{x,A}\cdot(x'-x) + \sum_{e:x'_e<x_e} (x'_e - x_e)\omega c_e \\
&\ge \widehat{d}^\intercal(x'-x) - \omega\cdot c^\intercal x \\
&\ge \widehat{d}^\intercal(x'-x) - \omega\cdot h^{\text{pr}}(\mathring{p}\,;x)\ .
\end{aligned}$$

The first inequality follows since q^* is an optimal solution for (W_x) and a feasible solution for $(\mathrm{W}_{x'})$. The second inequality follows since $d^{x,A}$ is a subgradient of $g(\cdot,A)$ at x for every $A \in \mathcal{A}$. The third inequality follows from the componentwise closeness of d^{est} and $\mathbb{E}_{A\sim\mathring{p}}\big[d^{x,A}\big]$. The fourth inequality follows because

$$\sum_{e:x'_e<x_e} (x'_e - x_e)\omega c_e \ge -\omega \sum_{e:x'_e<x_e} c_e x_e \ge -\omega\cdot c^\intercal x\ .$$

The last inequality holds because $h^{\text{pr}}(\mathring{p}\,;x) \ge c^\intercal x$. □

Lemma 7.9. *Let* $\omega > 0$ *and* $\delta \in (0,1)$*. Given any* $x \in \mathcal{P}$*, we can compute in* $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\omega}, \log\frac{1}{\delta}\big)$ *time a vector* $d^{\text{est}} \in \mathbb{R}^m$ *that with probability at least* $1-\delta$ *satisfies*

$$-\omega c \le d^{\text{est}} - \mathbb{E}_{A\sim\mathring{p}}\big[d^{x,A}\big] \le 0\ .$$

Proof. We sample scenarios $A_1, \ldots, A_N$ from the central distribution $\mathring{p}$, and set $\widehat{d} := \frac{1}{N}\sum_{i\in[N]} d^{x,A_i}$, where N will be determined later. Note that for each $e \in [m]$, we can view $\widehat{d}_e$ as an empirical estimate of the random variable $W_e := d^{x,A}_e$, where A is sampled according to $\mathring{p}$. Note that W_e lies in the range $[-\lambda c_e, 0]$, and that $\mathbb{E}[W_e] = \mathbb{E}_{A\sim\mathring{p}}\big[d^{x,A}_e\big]$. By

Hoeffding's inequality (Corollary 2.4), we can choose

$$N = \mathsf{poly}\left(\frac{\lambda c_e}{\frac{\omega c_e}{2}}, \log \frac{1}{\frac{\delta}{m}}\right) = \mathsf{poly}\left(\mathcal{I}, \lambda, \frac{1}{\omega}, \log \frac{1}{\delta}\right)$$

such that

$$\mathbf{Pr}\left[\left|\widehat{d}_e - \mathbb{E}_{A\sim\mathring{p}}\left[d_e^{x,A}\right]\right| > \frac{\omega c_e}{2}\right] \leq \frac{\delta}{m}$$

for every $e \in [m]$. By the union bound, it follows that with probability at least $1 - m\frac{\delta}{m} = 1 - \delta$ we have

$$-\frac{1}{2}\omega c \leq \widehat{d} - \mathbb{E}_{A\sim\mathring{p}}\left[d^{x,A}\right] \leq \frac{1}{2}\omega c\ .$$

We can therefore take $d^{\text{est}} := \widehat{d} - \frac{1}{2}\omega c$. □

Lemma 7.10. *Suppose that we have a* $\mathsf{poly}(\mathcal{I}, t)$*-time algorithm for problem* (Υ)*. Given any fractional first-stage decision* $x \in \mathcal{P}$*, we can compute an optimal solution for* (W_x) *in* $\mathsf{poly}\left(\mathcal{I}, \frac{1}{r}\right)$ *time.*

Proof. We start by setting $t := \min\left\{\left\lceil\frac{\widehat{P}^{\text{free}}}{r}\right\rceil, |\mathcal{A}|\right\}$ and using the algorithm for problem (Υ) to compute the t scenarios A with largest $g(x, A)$ value. (Throughout this proof, suppose for simplicity that there are no ties. If there are ties, they may be broken arbitrarily.) Note that this takes $\mathsf{poly}(\mathcal{I}, t) = \mathsf{poly}\left(\mathcal{I}, \frac{1}{r}\right)$ time. We define q^* as follows: we set $q^*_A := r$ for all the scenarios returned, except for the one with smallest $g(x, A)$ value; for this one, we set $q^*_A := \min\left\{r, \widehat{P}^{\text{free}} - (t-1)r\right\}$. For the remaining scenarios A, we set $q^*_A := 0$.

We now argue that q^* is an optimal solution for (W_x). Consider the polytope $\frac{1}{r}\widehat{\mathcal{R}} := \left\{\frac{1}{r}q : q \in \widehat{\mathcal{R}}\right\}$. Note that ($\mathrm{W}_x$) is equivalent to the problem $\max_{q\in\frac{1}{r}\widehat{\mathcal{R}}}\left\{\sum_{A\in\mathcal{A}} q_A g(x, A)\right\}$ (up to scaling of the solutions), which can be seen as a fractional knapsack problem: we have an item of value $g(x, A)$ and weight 1 for each scenario $A \in \mathcal{A}$; the capacity of the knapsack is set to $\frac{\widehat{P}^{\text{free}}}{r}$. The result then follows by using the well-known fact that one can compute an optimal solution for a fractional knapsack problem in a greedy fashion, by repeatedly picking among the available items the one with the highest value/weight ratio, until the knapsack is full or we run out of items (see, e.g., Dantzig [28]). Computing this solution for the fractional knapsack problem, then scaling it back by an r factor, we obtain q^*. □

Lemma 7.7 now follows easily by combining the preliminary lemmas.

Proof of Lemma 7.7. We start by using Lemma 7.9 to compute a vector d^{est} such that $-\omega c \leq d^{\text{est}} - \mathbb{E}_{A\sim\mathring{p}}\left[d^{x,A}\right] \leq 0$ with probability at least $1-\delta$. Then, we use Lemma 7.10 to compute an optimal solution q^* for (W_x). Finally, we compute and return $c + d^{\text{est}} + \sum_{A\in\mathcal{A}} q^*_A d^{x,A}$. By Lemma 7.8, this is an ω-subgradient of $h^{\text{pr}}(\mathring{p}\,;\cdot)$ at x, as long as the call to the algorithm from Lemma 7.9 to compute d^{est} was successful. □

7.5 Lipschitz-continuity of the proxy function

In this section, we show how to obtain a suitable upper bound on the Lipschitz constant of the proxy function $h^{\text{pr}}(\mathring{p}\,;\cdot)$, which will be necessary when utilizing the ellipsoid-based method of Shmoys and Swamy [114] (Theorem 3.12) to find an approximate solution for $\min_{x\in\mathcal{P}} h^{\text{pr}}(\mathring{p}\,;x)$.

Lemma 7.11. *The proxy function $h^{\text{pr}}(\mathring{p}\,;\cdot)$ is $(2\lambda+1)\,\|c\|$-Lipschitz continuous.*

Proof. By Lemma 1.1, it suffices to show that $h^{\text{pr}}(\mathring{p}\,;\cdot)$ has a subgradient of Euclidean norm at most $(2\lambda+1)\,\|c\|$ at every point. Let $x\in\mathcal{P}$, and let q^* be an optimal solution for (W_x). By Lemma 7.8, we have that $d := c + \mathbb{E}_{A\sim\mathring{p}}\left[d^{x,A}\right] + \sum_{A\in\mathcal{A}} q^*_A d^{x,A}$ is a subgradient of $h^{\text{pr}}(\mathring{p}\,;\cdot)$ at x. We have

$$\|d\| = \left\| c + \mathbb{E}_{A\sim\mathring{p}}\left[d^{x,A}\right] + \sum_{A\in\mathcal{A}} q^*_A d^{x,A} \right\| \leq \|c\| + \sum_{A\in\mathcal{A}} \mathring{p}_A \left\|d^{x,A}\right\| + \sum_{A\in\mathcal{A}} q^*_A \left\|d^{x,A}\right\| \leq (2\lambda+1)\,\|c\|\,.$$

The second step follows from the triangle inequality, and the final step follows because $\left\|d^{x,A}\right\| \leq \lambda\,\|c\|$ for every $A\in\mathcal{A}$ by assumption (A5'), $\sum_{A\in\mathcal{A}} \mathring{p}_A = 1$, and $\sum_{A\in\mathcal{A}} q^*_A \leq \widehat{P}^{\text{free}} \leq 1$. □

7.6 Proof of Theorem 7.1

We now combine our results from Sections 7.2–7.5 to prove Theorem 7.1. First, we exploit assumption (A6) to obtain a lower bound on $\min_{x\in\mathcal{P}} h^{\text{pr}}(\mathring{p}\,;x)$; this allows us to fold the additive error incurred when applying the ellipsoid-based method by Shmoys and Swamy [114] (Theorem 3.12).

Lemma 7.12. *Suppose that* $\mathring{P}^{\text{free}} \leq \widehat{P}^{\text{free}} \leq \min\left\{(1+\varepsilon)\mathring{P}^{\text{free}}, 1\right\}$*, and that we have a* $\mathsf{poly}(\mathcal{I}, t)$*-time algorithm for problem* (Υ)*. Then we can determine in* $\mathsf{poly}(\mathcal{I})$ *time that either (i)* $\min_{x\in\mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x) \geq r$*; or (ii)* $x = 0$ *is an optimal solution for* $\min_{x\in\mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x)$*.*

Proof. We first show that if $\mathcal{A}$ only contains null scenarios, then $x = 0$ is an optimal solution for $\min_{x\in\mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x)$; otherwise, we have $\min_{x\in\mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x) \geq r$. First, suppose that $\mathcal{A}$ only contains null scenarios. Then the value of $\mathbb{E}_{A\sim\mathring{p}}[g(x, A)] + \max_{q\in\widehat{\mathcal{R}}} \sum_{A\in\mathcal{A}} q_A g(x, A)$ is independent of x, and so $x = 0$ is a minimizer of $h^{\text{pr}}(\mathring{p}\,; \cdot)$. Now, suppose that there exists a non-null scenario $A^* \in \mathcal{A}$. Let $q \in \mathbb{R}^{\mathcal{A}}$ be defined by $q_{A^*} := r$ and $q_A = 0$ for every other scenario A. Note that $q \in \widehat{\mathcal{R}}$, since by Lemma 7.5 we have $\widehat{P}^{\text{free}} \geq \mathring{P}^{\text{free}} \geq r$. It follows that for every fractional first-stage decision $x \in \mathcal{P}$ we have

$$h^{\text{pr}}(\mathring{p}\,; x) \geq c^\intercal x + r g(x, A^*) \geq r(c^\intercal x + g(x, A^*)) \geq r \ ,$$

where the second inequality follows because $r \leq 1$ by assumption, and the final inequality follows from assumption (A6) since A^* is a non-null scenario. We conclude that $\min_{x\in\mathcal{P}} h(\mathring{p}\,; x) \geq r$.

While we will not quite be able to determine if $\mathcal{A}$ contains a non-null scenario, we can do the following. First we run the algorithm for problem (Υ) with input $(x, t) = (0, 1)$ to compute $\overline{A} := \operatorname{argmax}_{A\in\mathcal{A}} g(0, A)$. If $g\big(0, \overline{A}\big) \geq 1$, then we claim that (i) holds; otherwise, we claim that (ii) holds.

We now show that the algorithm is correct. First, suppose that $g\big(0, \overline{A}\big) < 1$. Then for every scenario $A \in \mathcal{A}$ we have $c^\intercal 0 + g(0, A) \leq c^\intercal 0 + g\big(0, \overline{A}\big) < 1$. This cannot hold for a non-null scenario A by assumption (A6), and so we conclude that $\mathcal{A}$ only has null scenarios. As shown above, it follows that (ii) holds.

Now, suppose that $g\big(0, \overline{A}\big) \geq 1$. We have already shown that (i) holds if there exists a non-null scenario. Now, suppose that $\mathcal{A}$ only has null scenarios. As explained above, this implies that $x = 0$ is an optimal solution for $\min_{x\in\mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x)$. Let $\overline{q} \in \mathbb{R}^{\mathcal{A}}$ be defined by $\overline{q}_{\overline{A}} := r$ and $\overline{q}_A = 0$ for every other scenario A. Note that $\overline{q} \in \widehat{\mathcal{R}}$, since by Lemma 7.5 we have $\widehat{P}^{\text{free}} \geq \mathring{P}^{\text{free}} \geq r$. It follows that (i) holds, since we obtain

$$\min_{x\in\mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x) = h^{\text{pr}}(\mathring{p}\,; 0) \geq r g\big(0, \overline{A}\big) \geq r \ . \qquad \square$$

Proof of Theorem 7.1. We show how to obtain a solution $\overline{x} \in \mathcal{P}$ such that

$$h(\mathring{p}\,; \overline{x}) \leq (2 + 7\varepsilon) \cdot \min_{x\in\mathcal{P}} h(\mathring{p}\,; x)$$

with probability at least $1 - \delta$ (the theorem then follows by applying this weaker result with parameter $\frac{\varepsilon}{7}$ instead of ε). Let us assume without loss of generality that $\varepsilon \leq 1$.

We start by using the algorithm from Lemma 7.4 to compute an estimate $\widehat{P}^{\text{free}}$ of $\mathring{P}^{\text{free}}$, setting the failure parameter to $\frac{\delta}{2}$. Let us assume in the sequel that $\mathring{P}^{\text{free}} \leq \widehat{P}^{\text{free}} \leq \min\left\{(1+\varepsilon)\mathring{P}^{\text{free}}, 1\right\}$; this happens with probability at least $1 - \frac{\delta}{2}$.

Next, we run the algorithm from Lemma 7.12 to determine that either (i) $\min_{x \in \mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x) \geq r$; or (ii) $x = 0$ is an optimal solution for $\min_{x \in \mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x)$. In case (ii), we set $\overline{x} := 0$. Now, suppose we are in case (i). We run the variant of the ellipsoid method by Shmoys and Swamy [114] (Theorem 3.12) with parameters $\left(\varepsilon, \eta := \varepsilon r, \frac{\delta}{2}\right)$ to find an approximate solution for the problem $\min_{x \in \mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x)$, using the algorithm from Lemma 7.7 to compute approximate subgradients; the Lipschitz constant is set to $\widetilde{K} := (2\lambda + 1)\,\|c\|$ (see Lemma 7.5). Let $\overline{x}$ be the solution returned. Then with probability at least $1 - \frac{\delta}{2}$ we have

$$h^{\text{pr}}(\mathring{p}\,; \overline{x}) \leq (1+\varepsilon) \cdot \min_{x \in \mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x) + \varepsilon r \leq (1+2\varepsilon) \cdot \min_{x \in \mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x)\ ,$$

where the last inequality follows because $\min_{x \in \mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x) \geq r$. Assuming that $h^{\text{pr}}(\mathring{p}\,; \overline{x}) \leq (1+2\varepsilon) \cdot \min_{x \in \mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x)$ (which holds with probability at least $1 - \frac{\delta}{2}$ in case (i), and with probability 1 in case (ii)), we obtain

$$\begin{aligned} h(\mathring{p}\,; \overline{x}) &\leq h^{\text{pr}}(\mathring{p}\,; \overline{x}) \\ &\leq (1+2\varepsilon) \cdot \min_{x \in \mathcal{P}} h^{\text{pr}}(\mathring{p}\,; x) \\ &\leq (1+2\varepsilon)(2+\varepsilon) \cdot \min_{x \in \mathcal{P}} h(\mathring{p}\,; x) \\ &\leq (2+7\varepsilon) \cdot \min_{x \in \mathcal{P}} h(\mathring{p}\,; x)\ . \end{aligned}$$

The first and the third inequality follow from Lemma 7.2; the final inequality follows because $(1+2\varepsilon)(2+\varepsilon) \leq 2 + 7\varepsilon$ holds for every $\varepsilon \leq 1$.

The success probability is at least $\left(1 - \frac{\delta}{2}\right)^2 \geq 1 - \delta$. It remains to bound the running time. The call to the algorithm from Lemma 7.12 takes $\mathsf{poly}(\mathcal{I})$ time. The call to the algorithm from Theorem 3.12 takes

$$\mathsf{poly}\left(\mathcal{I}, \log \widetilde{K}, \mathsf{poly}\left(\mathcal{I}, \frac{1}{r}, \lambda, \frac{1}{\omega}, \log\left(\frac{2\left(\widetilde{N} + \widetilde{n}\right)}{\delta}\right)\right), \log \frac{1}{\eta}\right)$$

time, where $\widetilde{N} := \left\lceil 2m^2 \ln\left(\frac{16\widetilde{K}R_{\text{large}}^2}{R_{\text{small}}\cdot\eta}\right)\right\rceil$, $\widetilde{n} := \widetilde{N}\cdot\ln\left(\frac{8\widetilde{N}\widetilde{K}R_{\text{large}}}{\eta}\right)$, and $\omega := \frac{\min\{\varepsilon,1\}}{4\widetilde{n}}$. (This is obtained by plugging in the running time of the algorithm for computing approximate subgradients from Lemma 7.7 in the bound given by Theorem 3.12.) We now give (loose) bounds on the terms appearing above: we have

1. $\log\widetilde{K} = \log((2\lambda+1)\|c\|) = \mathsf{poly}(\mathcal{I},\lambda)$;
2. $\log\frac{1}{\eta} = \log\frac{1}{\varepsilon r} = \mathsf{poly}\left(\mathcal{I},\frac{1}{\varepsilon}\right)$;
3. $\widetilde{N} = \mathsf{poly}\left(\mathcal{I},\lambda,\frac{1}{\varepsilon}\right)$ (using 1 and 2, along with assumption (A4));
4. $\widetilde{n} = \mathsf{poly}\left(\mathcal{I},\lambda,\frac{1}{\varepsilon}\right)$ (using 1–3, along with assumption (A4));
5. $\frac{1}{\omega} = \mathsf{poly}\left(\mathcal{I},\lambda,\frac{1}{\varepsilon}\right)$ (using 4); and
6. $\log\left(\frac{2(\widetilde{N}+\widetilde{n})}{\delta}\right) = \mathsf{poly}\left(\mathcal{I},\lambda,\frac{1}{\varepsilon},\log\frac{1}{\delta}\right)$ (using 3 and 4).

Combining all these bounds, we conclude that the call to the ellipsoid-based method takes $\mathsf{poly}\left(\mathcal{I},\lambda,\frac{1}{r},\frac{1}{\varepsilon},\log\frac{1}{\delta}\right)$ time. □

7.7 Applications

Recall that the problem (Υ) asks for the t scenarios A with highest $g(x,A)$ value, given a fractional first-stage decision $x \in \mathcal{P}$. In this section we show that, if the underlying problem is a covering problem, then we can solve (Υ) efficiently in the unrestricted setting ($\mathcal{A} = 2^U$); see Lemma 7.13. This, combined with Theorem 3.9, leads to approximation algorithms for DRS optimization under an L_∞ ball in the unrestricted setting for various applications. See Table 7.1 for a summary of the results, and Theorem 7.14 for the precise statement.

Lemma 7.13. *Consider the unrestricted setting (i.e., $\mathcal{A} = 2^U$ for some ground set U). Suppose that $g(x,A) \le g(x,A')$ for every fractional first-stage decision $x \in \mathcal{P}$ and for every pair of scenarios (A,A') with $A \subseteq A'$. Then there is an algorithm for problem (Υ) with running time $\mathsf{poly}(\mathcal{I},t)$.*

Problem	$\mathcal{A} = 2^U$
Facility location	11
Vertex cover	8
Edge cover	6
Set cover	$O(\log \lvert U \rvert)$

Table 7.1: A summary of the approximation factors we obtain for DRS optimization under an L_∞ ball. We have omitted the $O(\varepsilon)$ terms that appear in the approximation factors.

Proof. Consider an instance (x, t) of problem (Υ). Our goal is to construct a sequence of scenarios $A_1, \ldots, A_t$, where A_i is the i-th scenario with highest $g(x, A)$ value.

By the monotonicity assumption of $g(\cdot, \cdot)$, the costliest scenario is U, so we start by setting $A_1 := U$. We then proceed as follows for $i = 2, \ldots, t$. Suppose that we have already computed $A_1, \ldots, A_{i-1}$. Computing A_i amounts to solving the problem

$$\max \{g(x, A) : A \in \mathcal{A} \setminus \{A_1, \ldots, A_{i-1}\}\} \ . \tag{7.4}$$

We claim that (7.4) admits an optimal solution that is a maximal proper subset of $A_{i'}$ for some $1 \leq i' < i$. Indeed, let A^* be an optimal solution of (7.4) with maximum cardinality, and suppose for a contradiction that it is not a maximal proper subset of $A_{i'}$ for any $1 \leq i' < i$. Note that since $A_1 = U$, we have $A^* \neq U$, so there is an element $e \in U \setminus A^*$. Now, consider the scenario $\overline{A} := A^* \cup \{e\}$. Since by assumption A^* is not a maximal subset of $A_{i'}$ for any $1 \leq i' < i - 1$, it follows that $\overline{A}$ is feasible for (7.4). By the monotonicity assumption, since $A^* \subseteq \overline{A}$, we have $g\big(x, \overline{A}\big) \geq g(x, A^*)$, and so $\overline{A}$ is also an optimal solution for (7.4). Since $\big|\overline{A}\big| > |A^*|$, this contradicts the definition of A^*.

We now utilize the observation above to show that given x and $A_1, \ldots, A_{i-1}$, we can solve (7.4) in $\mathsf{poly}(\mathcal{I}, i)$ time. This can be done by enumerating all maximal proper subsets of $A_1, \ldots, A_{i-1}$. Since each set $A_{i'}$ has $|A_{i'}|$ maximal proper subsets, we enumerate $\sum_{i'=1}^{i-1} |A_{i'}| \leq (i-1)\,|U| = \mathsf{poly}(\mathcal{I}, i)$ scenarios. Since evaluating $g(x, A)$ for a given scenario A takes $\mathsf{poly}(\mathcal{I})$ time, the claim follows. We conclude that we can solve problem (Υ) by solving (7.4) for $i = 2, \ldots, t$, which takes $\sum_{i=2}^{t} \mathsf{poly}(\mathcal{I}, i) = \mathsf{poly}(\mathcal{I}, t)$ time. □

Theorem 7.14. *Consider the two-stage DRS optimization problem* (DRSO_∞) *in the unrestricted setting (i.e., $\mathcal{A} = 2^U$). Suppose that for every fractional first-stage decision $x \in \mathcal{P}$ and every pair of scenarios $A, A' \in \mathcal{A}$ with $A \subseteq A'$ we have $g(x, A) \leq g(x, A')$. Moreover, suppose that we have a local ρ-approximation algorithm. Then, given $\varepsilon > 0$ and $\delta \in (0, 1)$, we can compute a $(2 + \varepsilon)\rho$-approximate solution for* (DRSO_∞) *with probability at least $1 - \delta$ in $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{r}, \frac{1}{\varepsilon}, \log \frac{1}{\delta}\big)$ time. In particular, we have the following approximation factors*

for specific applications: (a) $O(\log |U| + \varepsilon)$ *for set cover; (b)* $8 + O(\varepsilon)$ *for vertex cover; (c)* $6 + O(\varepsilon)$ *for edge cover; and (d)* $10.98 + O(\varepsilon)$ *for facility location.*

Proof. The general result follows immediately by combining Theorem 3.9 and Lemma 7.13.

Now, let us consider the four applications mentioned. They all satisfy assumptions (A1)–(A4), (A5'), and (A6). Recall that vertex cover and edge cover are special cases of set cover. For assumptions (A1)–(A4) and (A6), see Chapter 6, specifically Lemma 6.7 for set cover and Lemma 6.18 for facility location. Shmoys and Swamy [114] show that assumption (A5') holds for a broad class of two-stage problems that includes set cover and facility location. These are all covering problems, so the monotonicity assumption on $g(\cdot, \cdot)$ is satisfied. We have local ρ-approximation algorithms with $\rho = O(\log |U|)$ for set cover (see Shmoys and Swamy [114]); $\rho = 4$ for vertex cover (see Lemma 6.14-(iii)); $\rho = 3$ for edge cover (see Lemma 6.16-(iii)); and $\rho = 5.488$ for facility location (see Lemma 6.19-(ii)). □

Chapter 8

Conclusions and open directions

In this thesis, we developed a framework to solve distributionally robust stochastic (DRS) combinatorial-optimization problems when the ambiguity set of distributions arises as a ball in the Wasserstein metric, or the L_∞ metric, around a central distribution specified only by a sampling oracle. We showed that our framework is versatile and utilized it to obtain the first approximation guarantees for DRS versions of various combinatorial-optimization problems such as set cover, vertex cover, edge cover, facility location, and Steiner tree.

Our work opens up various directions for further research, and we conclude by listing some of these directions. We list the open questions below, roughly speaking, in order of more concrete questions that directly stem from our work and pertain to improving the guarantees that we obtain and/or expanding the scope of our work, followed by more open-ended questions related to distributionally robust stochastic optimization.

- Some of the approximation factors we obtained for specific applications (see Tables 3.1 and 3.2) can likely be improved by constant factors by utilizing stronger LP-relaxations and/or improved rounding algorithms.
- In the Wasserstein setting, can one compute a $(1+\varepsilon)$-approximate solution using only $\mathsf{poly}(\text{input size}, \lambda)$ independent samples from the central distribution? Whereas this is known to be achievable in the classical two-stage stochastic model under comparable assumptions (see Charikar, Chekuri, and Pál [24] and Shmoys and Swamy [114]), our framework incurs a factor-4 loss due to the use of the proxy function $\overline{h}(\mathring{p}\,;\cdot)$ that is (up to a constant term) within a factor 2 of the true objective function of the DRS problem (see Theorem 4.1 and Lemma 4.10). (As noted in Remark 4.15, this can be improved to

a factor-2 loss with some additional work.) In the L_∞ setting, can the $\frac{1}{r}$-dependence of the number of samples in Theorem 3.9 be avoided?

- In the Wasserstein setting, we obtained approximation algorithms for various applications with two choices of scenario metrics: the discrete metric ℓ^{disc} and the asymmetric metric $\ell_\infty^{\mathsf{asym}}$. Various other scenario metrics can be considered in future work, such as the asymmetric metric ℓ_1^{asym} and the symmetric metrics $\ell_\infty^{\mathrm{sym}}$ and ℓ_1^{sym} mentioned in the context of DRS facility location in Section 3.1.

- Is it possible to obtain constant-factor approximation algorithms for DRS Steiner tree under a Wasserstein ball in the k-bounded setting? One way to obtain such a result would be to provide a local O(1)-approximation algorithm for Steiner tree; it is also conceivable that this can be achieved using only a restricted local approximation algorithm. The lack of a local approximation algorithm for Steiner tree also prevented us from directly applying our framework to obtain an approximation algorithm for DRS Steiner tree under an L_∞ ball in the unrestricted setting. It is unclear how to obtain such an algorithm even in the polynomial-size central distribution setting.

- Another open question is whether we can apply or extend our framework for DRS optimization under an L_∞ ball to obtain approximation algorithms for specific applications in the k-bounded setting. The obstacle encountered in directly applying our framework is that, for the applications we considered, we do not have a suitable algorithm for problem (Υ) (which asks for the t most expensive scenarios under a given first-stage decision x).

- As mentioned in Section 2.3, in addition to Wasserstein balls and L_∞ balls, various other types of ambiguity sets have been considered in the literature, and it would be interesting to obtain approximation algorithms in those settings as well. We have obtained preliminary results for ambiguity sets consisting of a finite number of distributions given by sampling oracles.

- Another direction is obtaining approximation algorithms for the DRS versions of the applications that we considered via combinatorial techniques.

References

[1] Ajit Agrawal, Philip Klein, and R. Ravi. When trees collide: An approximation algorithm for the generalized Steiner problem on networks. *STOC* (1991), pp. 134–144.

[2] Shipra Agrawal, Yichuan Ding, Amin Saberi, and Yinyu Ye. Correlation Robust Stochastic Optimization. *arXiv.org* (2009). arXiv: 0902.1792v3 [cs.DS].

[3] Güzin Bayraksan and David K. Love. Data-Driven Stochastic Programming Using Phi-Divergences. *The Operations Research Revolution*. INFORMS, 2015, pp. 1–19.

[4] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

[5] Aharon Ben-Tal, A. P. Goryashko, E. Guslitzer, and Arkadi Nemirovski. Adjustable robust solutions of uncertain linear programs. *Math. Program.* 99.2 (2004), pp. 351–376.

[6] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Management Science* 59.2 (2013), pp. 341–357.

[7] Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. SIAM, 2001.

[8] Aharon Ben-Tal and Arkadi Nemirovski. Robust Convex Optimization. *Math. Oper. Res.* 23.4 (1998), pp. 769–805.

[9] Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of Linear Programming problems contaminated with uncertain data. *Math. Program.* 88.3 (2000), pp. 411–424.

[10] Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of uncertain linear programs. *Oper. Res. Lett.* 25.1 (1999), pp. 1–13.

[11] Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and Applications of Robust Optimization. *SIAM Review* 53.3 (2011), pp. 464–501.

[12] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Math. Program.* 167.2 (2018), pp. 235–292.

[13] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Robust sample average approximation. *Math. Program.* 171.1-2 (2018), pp. 217–282.

[14] Dimitris Bertsimas and Melvyn Sim. Robust discrete optimization and network flows. *Math. Program.* 98.1-3 (2003), pp. 49–71.

[15] Dimitris Bertsimas and Melvyn Sim. The Price of Robustness. *Operations Research* 52.1 (2004), pp. 35–53.

[16] Dimitris Bertsimas, Melvyn Sim, and Meilin Zhang. A practicable framework for distributionally robust linear optimization. *optimization-online.org* (2013).

[17] Dimitris Bertsimas, Melvyn Sim, and Meilin Zhang. Adaptive Distributionally Robust Optimization. *Management Science* 65.2 (2018), pp. 604–618.

[18] John R. Birge and François Louveaux. *Introduction to Stochastic Programming.* Springer Science & Business Media, 2011.

[19] Keith C. Brown. A note on the apparent bias of net revenue estimates for capital investment projects. *The Journal of Finance* 29.4 (1974), pp. 1215–1216.

[20] Jaroslaw Byrka and Aravind Srinivasan. Approximation Algorithms for Stochastic and Risk-Averse Optimization. *SIAM J. Discrete Math.* 32.1 (2018), pp. 44–63.

[21] John Gunnar Carlsson, Mehdi Behroozi, and Kresimir Mihic. Wasserstein Distance and the Distributionally Robust TSP. *Operations Research* 66.6 (2018), pp. 1603–1624.

[22] John Gunnar Carlsson and Erick Delage. Robust Partitioning for Stochastic Multivehicle Routing. *Operations Research* 61.3 (2013), pp. 727–744.

[23] Robert Carr and Santosh Vempala. Randomized Metarounding. *STOC* (2000), pp. 343–352.

[24] Moses Charikar, Chandra Chekuri, and Martin Pál. Sampling Bounds for Stochastic Optimization. *APPROX-RANDOM* (2005), pp. 257–269.

[25] Lin Chen, Nicole Megow, Roman Rischke, and Leen Stougie. Stochastic and Robust Scheduling in the Cloud. *APPROX-RANDOM* (2015), pp. 175–186.

[26] Xi Chen, Qihang Lin, and Guanglin Xu. Distributionally Robust Optimization with Confidence Bands for Probability Density Functions. *arXiv.org* (2019). arXiv: 1901.02169v1 [math.OC].

[27] Vasek Chvátal. A Greedy Heuristic for the Set-Covering Problem. *Mathematics of Operations Research* 3 (1979), pp. 233–235.

[28] George B. Dantzig. Discrete-Variable Extremum Problems. *Operations Research* 5.2 (1957), pp. 266–288.

[29] George B. Dantzig. Linear Programming Under Uncertainty. *Management Science* 1 (1951), pp. 197–206.

[30] Erick Delage. Distributionally robust optimization in context of data-driven problems. PhD Thesis. 2009.

[31] Erick Delage and Dan A. Iancu. Robust Multistage Decision Making. *The Operations Research Revolution*. INFORMS, 2015, pp. 20–46.

[32] Erick Delage and Yinyu Ye. Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research* 58.3 (2010), pp. 595–612.

[33] Kedar Dhamdhere, Vineet Goyal, R. Ravi, and Mohit Singh. How to Pay, Come What May: Approximation Algorithms for Demand-Robust Covering Problems. *STOC* (2005), pp. 367–378.

[34] Kedar Dhamdhere, R. Ravi, and Mohit Singh. On Two-Stage Stochastic Minimum Spanning Trees. *IPCO* (2005), pp. 321–334.

[35] Jitka Dupačová. On minimax solutions of stochastic linear programming problems. *Časopis pro pěstováni matematiky* 91.4 (1966), pp. 423–430.

[36] Shane Dye, Leen Stougie, and Asgeir Tomasgard. The Stochastic Single Resource Service-Provision Problem. *Naval Research Logistics* 50.8 (2003), pp. 869–887.

[37] Martin Dyer and Leen Stougie. Computational complexity of stochastic programming problems. *Math. Program.* 106.3 (2006), pp. 423–432.

[38] Martin Dyer and Leen Stougie. Erratum to: Computational complexity of stochastic programming problems. *Math. Program.* 153.2 (2015), pp. 723–725.

[39] Laurent El Ghaoui and Hervé Lebret. Robust Solutions to Least-Squares Problems with Uncertain Data. *SIAM Journal on Matrix Analysis and Applications* 18.4 (1997), pp. 1035–1064.

[40] Laurent El Ghaoui, Maksim Oks, and François Oustry. Worst-Case Value-At-Risk and Robust Portfolio Optimization: A Conic Programming Approach. *Operations Research* 51.4 (2003), pp. 543–556.

[41] Laurent El Ghaoui, François Oustry, and Hervé Lebret. Robust Solutions to Uncertain Semidefinite Programs. *SIAM Journal on Optimization* 9.1 (1998), pp. 33–52.

[42] Emre Erdoğan and Garud Iyengar. Ambiguous chance constrained problems and robust optimization. *Math. Program.* 107.1-2 (2005), pp. 37–61.

[43] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations. *arXiv.org* (2015). arXiv: 1505.05116v3 [math.OC].

[44] Hossein Esfandiari, Nitish Korula, and Vahab Mirrokni. Online Allocation with Traffic Spikes: Mixing Adversarial and Stochastic Models. *EC* (2015), pp. 169–186.

[45] James E. Falk. Technical Note - Exact Solutions of Inexact Linear Programs. *Operations Research* 24.4 (1976), pp. 783–787.

[46] Uriel Feige, Kamal Jain, Mohammad Mahdian, and Vahab Mirrokni. Robust Combinatorial Optimization with Exponential Scenarios. *IPCO* (2007), pp. 439–453.

[47] Moran Feldman, Guy Kortsarz, and Zeev Nutov. Improved Approximation Algorithms for Directed Steiner Forest. *Electronic Colloquium on Computational Complexity* (2007), pp. 279–292.

[48] Lisa Fleischer, Jochen Könemann, Stefano Leonardi, and Guido Schäfer. Simple Cost Sharing Schemes for Multicommodity Rent-or-Buy and Stochastic Steiner Tree. *STOC* (2006), pp. 663–670.

[49] Zachary Friggstad and Chaitanya Swamy. Approximation algorithms for regret-bounded vehicle routing and applications to distance-constrained vehicle routing. *STOC* (2014), pp. 744–753.

[50] Rui Gao, Xi Chen, and Anton J. Kleywegt. Wasserstein Distributional Robustness and Regularization in Statistical Learning. *arXiv.org* (2017). arXiv: 1712.06050v2 [cs.LG].

[51] Rui Gao and Anton J. Kleywegt. Distributionally Robust Stochastic Optimization with Dependence Structure. *arXiv.org* (2017). arXiv: 1701.04200v1 [math.OC].

[52] Rui Gao and Anton J. Kleywegt. Distributionally Robust Stochastic Optimization with Wasserstein Distance. *arXiv.org* (2016). arXiv: 1604.02199v2 [math.OC].

[53] Joel Goh and Melvyn Sim. Distributionally Robust Optimization and Its Tractable Approximations. *Operations Research* 58.4-part-1 (2010), pp. 902–917.

[54] Donald Goldfarb and Garud Iyengar. Robust Portfolio Selection Problems. *Math. Oper. Res.* 28.1 (2003), pp. 1–38.

[55] Daniel Golovin, Vineet Goyal, and R. Ravi. Pay Today for a Rainy Day: Improved Approximation Algorithms for Demand-Robust Min-Cut and Shortest Path Problems. *STACS* (2006), pp. 206–217.

[56] Vineet Goyal. Optimization Under Uncertainty. PhD Thesis. 2008.

[57] Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, 1988.

[58] Anupam Gupta and Amit Kumar. A Constant-Factor Approximation for Stochastic Steiner Forest. *STOC* (2009), pp. 659–668.

[59] Anupam Gupta, Viswanath Nagarajan, and R. Ravi. Robust and MaxMin Optimization under Matroid and Knapsack Uncertainty Sets. *arXiv.org* (2010). arXiv: 1012.4962v2 [cs.DS].

[60] Anupam Gupta, Viswanath Nagarajan, and R. Ravi. Thresholded Covering Algorithms for Robust and Max-Min Optimization. *ICALP* (2010), pp. 262–274.

[61] Anupam Gupta and Martin Pál. Stochastic Steiner Trees Without a Root. *ICALP* (2005), pp. 1051–1063.

[62] Anupam Gupta, Martin Pál, R. Ravi, and Amitabh Sinha. Boosted Sampling: Approximation Algorithms for Stochastic Optimization. *STOC* (2004), pp. 417–426.

[63] Anupam Gupta, Martin Pál, R. Ravi, and Amitabh Sinha. Sampling and Cost-Sharing: Approximation Algorithms for Stochastic Optimization Problems. *SIAM J. Comput.* 40.5 (2011), pp. 1361–1401.

[64] Anupam Gupta, R. Ravi, and Amitabh Sinha. An Edge in Time Saves Nine: LP Rounding Approximation Algorithms for Stochastic Network Design. *STOC* (2004), pp. 218–227.

[65] Shubham Gupta. Building Networks in the Face of Uncertainty. MSc Thesis. 2011.

[66] Eran Halperin and Robert Krauthgamer. Polylogarithmic inapproximability. *STOC* (2003), pp. 585–594.

[67] Grani A. Hanasusanto and Daniel Kuhn. Conic Programming Reformulations of Two-Stage Distributionally Robust Linear Programs over Wasserstein Balls. *Operations Research* 66.3 (2018), pp. 849–869.

[68] Grani A. Hanasusanto, Daniel Kuhn, and Wolfram Wiesemann. A comment on "computational complexity of stochastic programming problems". *Math. Program.* 159.1-2 (2016), pp. 557–569.

[69] J. Richard Harrison and James G. March. Decision Making and Postdecision Surprises. *Administrative Science Quarterly* 29.1 (1984), pp. 26–42.

[70] Lisa Hellerstein, Thomas Lidbetter, and Daniel Pirutinsky. Solving Zero-sum Games using Best Response Oracles with Applications to Search Games. *arXiv.org* (2017). arXiv: 1704.02657v4 [math.OC].

[71] Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* 58.301 (1963), pp. 13–30.

[72] Zhaolin Hu and L. Jeff Hong. Kullback-Leibler Divergence Constrained Distributionally Robust Optimization. *optimization-online.org* (2013).

[73] Zhiyi Huang, Yishay Mansour, and Tim Roughgarden. Making the Most of Your Samples. *EC* (2015), pp. 45–60.

[74] Nicole Immorlica, David Karger, Maria Minkoff, and Vahab Mirrokni. On the Costs and Benefits of Procrastination: Approximation Algorithms for Stochastic Combinatorial Optimization Problems. *SODA* (2004), pp. 691–700.

[75] Kamal Jain, Mohammad Mahdian, and Mohammad R. Salavatipour. Packing Steiner trees. *SODA* (2003), pp. 266–274.

[76] Klaus Jansen. Approximate Strong Separation with Application in Fractional Graph Coloring and Preemptive Scheduling. *STACS* (2002), pp. 239–256.

[77] J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica* 30 (1906), pp. 175–193.

[78] Narendra Karmarkar and Richard M Karp. An Efficient Approximation Scheme for the One-Dimensional Bin-Packing Problem. *STOC* (1982), pp. 312–320.

[79] Irit Katriel, Claire Kenyon-Mathieu, and Eli Upfal. Commitment under uncertainty: Two-stage stochastic matching problems. *Theor. Comput. Sci.* 408.2-3 (2008), pp. 213–223.

[80] Rohit Khandekar, Guy Kortsarz, Vahab Mirrokni, and Mohammad R. Salavatipour. Two-stage Robust Network Design with Exponential Scenarios. *Algorithmica* 65.2 (2013), pp. 391–408.

[81] Anton J. Kleywegt, Alexander Shapiro, and Tito Homem-de-Mello. The Sample Average Approximation Method for Stochastic Discrete Optimization. *SIAM Journal on Optimization* 12.2 (2002), pp. 479–502.

[82] Nan Kong and Andrew J. Schaefer. A factor 1/2 approximation algorithm for two-stage stochastic matching problems. *European Journal of Operational Research* 172.3 (2006), pp. 740–746.

[83] Shi Li. A 1.488 Approximation Algorithm for the Uncapacitated Facility Location Problem. *ICALP* (2011), pp. 77–88.

[84] Jeff Linderoth, Alexander Shapiro, and Stephen Wright. The empirical behavior of sampling methods for stochastic programming. *Annals OR* 142.1 (2006), pp. 215–241.

[85] André Linhares and Chaitanya Swamy. Approximation Algorithms for Distributionally Robust Stochastic Optimization with Black-Box Distributions. *To appear in STOC 2019. Detailed version on the CS arXiv* (2019). arXiv: 1904.07381v1 [cs.DS].

[86] Fengqiao Luo and Sanjay Mehrotra. Decomposition Algorithm for Distributionally Robust Optimization using Wasserstein Metric. *arXiv.org* (2017). arXiv: 1704.03920v1 [math.OC].

[87] Sanjay Mehrotra and Dávid Papp. A Cutting Surface Algorithm for Semi-Infinite Convex Programming with an Application to Moment Robust Optimization. *SIAM Journal on Optimization* 24.4 (2014), pp. 1670–1697.

[88] Sanjay Mehrotra and He Zhang. Models and algorithms for distributionally robust least squares problems. *Math. Program.* 146.1-2 (2014), pp. 123–141.

[89] Fanwen Meng, Jin Qi, Meilin Zhang, James Ang, Singfat Chu, and Melvyn Sim. A Robust Optimization Model for Managing Elective Admission in a Public Hospital. *Operations Research* 63.6 (2015), pp. 1452–1467.

[90] Richard O. Michaud. The Markowitz Optimization Enigma: Is 'Optimized' Optimal? *Financial Analysts Journal* 45.1 (1989), pp. 31–42.

[91] Vahab Mirrokni, Shayan Oveis Gharan, and Morteza Zadimoghaddam. Simultaneous approximations for adversarial and stochastic online budgeted allocation. *SODA* (2012), pp. 1690–1701.

[92] Arkadi Nemirovski and Alexander Shapiro. On complexity of Shmoys-Swamy class of two-stage linear stochastic programming problems. *optimization-online.org* (2006).

[93] Arkadi Nemirovski and David Yudin. Informational complexity and effective methods of solution for convex extremal problems. *Ekonomika i Matematicheskie Metody* 12.1 (1976), pp. 357–379.

[94] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2013.

[95] Martin Pál and Éva Tardos. Group Strategyproof Mechanisms via Primal-Dual Algorithms. *STOC* (2003), pp. 584–593.

[96] Ioana Popescu. Robust Mean-Covariance Solutions for Stochastic Optimization. *Operations Research* 55.1 (2007), pp. 98–112.

[97] András Prékopa. *Stochastic Programming*. Kluwer Academic Publishers, 1995.

[98] R. Ravi and Amitabh Sinha. Hedging Uncertainty: Approximation Algorithms for Stochastic Optimization Problems. *IPCO* (2004), pp. 101–115.

[99] R. T. Rockafellar. *Convex analysis*. Princeton University Press. 1970.

[100] Ward Romeijnders, Leen Stougie, and Martin H. van der Vlerk. Approximation in two-stage stochastic integer programming. *Surveys in Operations Research and Management Science* 19 (2014), pp. 17–33.

[101] Andrzej Ruszczyński and Alexander Shapiro. Stochastic Programming. *Handbook in Operations Research and Management Science* 10 (2003).

[102] Tjendera Santoso, Shabbir Ahmed, Marc Goetschalckx, and Alexander Shapiro. A stochastic programming approach for supply chain network design under uncertainty. *European Journal of Operational Research* 167.1 (2005), pp. 96–115.

[103] Herbert E. Scarf. A min-max solution of an inventory problem. *Studies in The Mathematical Theory of Inventory and Production* (1958), pp. 201–209.

[104] Frans Schalekamp and David B. Shmoys. Algorithms for the universal and a priori TSP. *Oper. Res. Lett.* 36.1 (2008), pp. 1–3.

[105] Peter Schütz, Asgeir Tomasgard, and Shabbir Ahmed. Supply chain design under uncertainty using sample average approximation and dual decomposition. *European Journal of Operational Research* 199.2 (2009), pp. 409–419.

[106] Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally Robust Logistic Regression. *NIPS* (2015), pp. 1576–1584.

[107] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via Mass Transportation. *arXiv.org* (2017). arXiv: 1710.10016v1 [math.OC].

[108] Alexander Shapiro. Monte Carlo Sampling Methods. *Handbooks in Operations Research and Management Science* 10 (2003), pp. 353–425.

[109] Alexander Shapiro and Shabbir Ahmed. On a Class of Minimax Stochastic Programs. *SIAM Journal on Optimization* 14.4 (2004), pp. 1237–1249.

[110] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. Lectures on Stochastic Programming: Modeling and Theory. *MOS-SIAM Series on Optimization* (2014).

[111] Alexander Shapiro and Arkadi Nemirovski. On Complexity of Stochastic Programming Problems. *Continuous Optimization*. Springer-Verlag, 2005, pp. 111–146.

[112] Cong Shi. Approximation Algorithms for Stochastic Optimization Problems in Operations Management. *Encyclopedia of Operations Research and Management Science*. Wiley, 2014.

[113] David B. Shmoys and Mauro Sozio. Approximation Algorithms for 2-Stage Stochastic Scheduling Problems. *IPCO* (2007), pp. 145–157.

[114] David B. Shmoys and Chaitanya Swamy. An Approximation Scheme for Stochastic Linear Programming and Its Application to Stochastic Integer Programs. *J. ACM* 53.6 (2006), pp. 978–1012.

[115] David B. Shmoys, Éva Tardos, and Karen Aardal. Approximation Algorithms for Facility Location Problems. *STOC* (1997), pp. 265–274.

[116] David Shmoys and Kunal Talwar. A Constant Approximation Algorithm for the a priori Traveling Salesman Problem. *IPCO* (2008), pp. 331–343.

[117] Naum Z. Shor. Utilization of the operation of space dilatation in the minimization of convex functions. *Kibernetika* 6.1 (1970), pp. 6–12.

[118] James E. Smith and Robert L. Winkler. The Optimizer's Curse: Skepticism and Postdecision Surprise in Decision Analysis. *Management Science* 52.3 (2006), pp. 311–322.

[119] Anthony Man-Cho So, Jiawei Zhang, and Yinyu Ye. Stochastic Combinatorial Optimization with Controllable Risk Aversion Level. *Math. Oper. Res.* 34.3 (2009), pp. 522–537.

[120] A. L. Soyster. Technical Note - Convex Programming with Set-Inclusive Constraints and Applications to Inexact Linear Programming. *Operations Research* 21.5 (1973), pp. 1154–1157.

[121] Chaitanya Swamy. Risk-Averse Stochastic Optimization: Probabilistically-Constrained Models and Algorithms for Black-Box Distributions. *SODA* (2011), pp. 1627–1646.

[122] Chaitanya Swamy and David B. Shmoys. Approximation algorithms for 2-stage stochastic optimization problems. *ACM SIGACT News* 37.1 (2006), pp. 33–46.

[123] Chaitanya Swamy and David B. Shmoys. Sampling-Based Approximation Algorithms for Multistage Stochastic Optimization. *SIAM J. Comput.* 41.4 (2012), pp. 975–1004.

[124] Chaitanya Swamy and David B. Shmoys. *The Sample Average Approximation Method for 2-stage Stochastic Optimization (unpublished manuscript)*. 2004.

[125] David J. Thuente. Technical Note - Duality Theory for Generalized Linear Programs with Computational Methods. *Operations Research* 28.4 (1980), pp. 1005–1011.

[126] Bart P. G. Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. From Data to Decisions: Distributionally Robust Optimization is Optimal. *arXiv.org* (2017). arXiv: 1704.04118v1 [math.OC].

[127] Bram Verweij, Shabbir Ahmed, Anton J. Kleywegt, George Nemhauser, and Alexander Shapiro. The Sample Average Approximation Method Applied to Stochastic Routing Problems: A Computational Study. *Computational Optimization and Applications* 24.2-3 (2003), pp. 289–333.

[128] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally Robust Convex Optimization. *Operations Research* 62.6 (2014), pp. 1358–1376.

[129] David P. Williamson and David B. Shmoys. *The Design of Approximation Algorithms*. Cambridge University Press, 2011.

[130] Chenchen Wu, Donglei Du, and Dachuan Xu. An Approximation Algorithm for the Two-Stage Distributionally Robust Facility Location Problem. *Advances in Global Optimization*. Springer International Publishing, 2014, pp. 99–107.

[131] Jinfeng Yue, Bintong Chen, and Min-Chiang Wang. Expected Value of Distribution Information for the Newsvendor Problem. *Operations Research* 54.6 (2006), pp. 1128–1136.

[132] Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Oper. Res. Lett.* 46.2 (2018), pp. 262–267.

[133] Zhisu Zhu, Jiawei Zhang, and Yinyu Ye. Newsvendor optimization with limited distribution information. *Optimization Methods and Software* 28.3 (2013), pp. 640–667.