Problem-driven Scenario Generation for Stochastic Programs

Submitted for the degree of Doctor of Philosophy
at Lancaster University

Jamie Fairbrother, MMath, MRes

November 2015

Abstract

Stochastic programming concerns mathematical programming in the presence of uncertainty. In a stochastic program uncertain parameters are modeled as random vectors and one aims to minimize the expectation, or some risk measure, of a loss function. However, stochastic programs are computationally intractable when the underlying uncertain parameters are modeled by continuous random vectors.

Scenario generation is the construction of a finite *discrete* random vector to use within a stochastic program. Scenario generation can consist of the discretization of a parametric probabilistic model, or the direct construction of a discrete distribution. There is typically a trade-off here in the number of scenarios that are used: one must use enough to represent the uncertainty faithfully but not so many that the resultant problem is computationally intractable. Standard scenario generation methods are *distribution-based*, that is they do not take into account the underlying problem when constructing the discrete distribution.

In this thesis we promote the idea of *problem-based* scenario generation. By taking into account the structure of the underlying problem one may be able to represent uncertainty in a more parsimonious way. The first two papers of this thesis focus on scenario generation for problems which use a tail-risk measure, such as the conditional value-at-risk, focusing in particular on portfolio selection problems. In the final paper we present a constraint

driven approach to scenario generation for simple recourse problems, a class of stochastic programs for minimizing the expected shortfall and surplus of some resources with respect to uncertain demands.

Contents

Abstract iii

Declaration ix

Preface xi

I Introduction 1

Background 3

1 Introduction . . . 3
2 Stochastic Programming . . . 3
2.1 General Stochastic Programs . . . 3
2.2 Two-stage stochastic linear programs . . . 6
3 The Basic Newsvendor Problem . . . 9
4 Risk Measures and Conditional Value-at-Risk . . . 12
4.1 General Risk Measures . . . 12
4.2 Conditional Value-at-Risk . . . 14
5 Scenario Generation . . . 20
5.1 Introduction . . . 20
5.2 Sampling Approaches . . . 21
5.3 Optimal Discretization . . . 30
5.4 Constructive Approaches . . . 34

5.5 Problem-Driven Scenario Generation 41

Thesis Summary **47**

References 50

II Papers 55

A Scenario Generation for Stochastic Programs with Tail Risk Measures **57**

1 Introduction 59

2 Tail risk measures and risk regions 63

2.1 Tail risk of random variables 63

2.2 Risk regions 66

3 Scenario generation 73

3.1 Aggregation sampling and reduction 74

3.2 Alternative approaches 76

4 Consistency of aggregation sampling 78

4.1 Uniform convergence of empirical β-quantiles 78

4.2 Equivalence of aggregation sampling with sampling from aggregated random vector 81

5 Risk regions for the portfolio selection problem 86

5.1 Problem statement and brute force aggregation 86

5.2 Non-risk region for elliptically distributed returns . . . 90

5.3 Non-risk region with convex constraints 93

6 Numerical tests 94

6.1 Experimental Set-up 95

6.2 Results 96

7 Conclusions 96

A Continuity of Distribution and Quantile Functions 99

B Convex cone results 103

References 108

B Scenario Generation for Portfolio Selection with Tail Risk Measures 113

1 Introduction . . . 115

2 Portfolio selection and risk regions . . . 119

2.1 Tail risk measures and risk regions . . . 119

2.2 Risk regions for elliptical distributions . . . 124

3 Projections and the conic hull . . . 126

3.1 Conic hull of feasible region . . . 126

3.2 Projection onto a finitely generated cone . . . 129

4 Scenario generation . . . 130

4.1 Aggregation sampling and reduction . . . 130

4.2 Approximation of risk regions . . . 133

4.3 Ghost constraints . . . 135

5 Probability of the non-risk region . . . 136

6 Numerical tests . . . 138

6.1 Experimental set-up . . . 139

6.2 Probability of non-risk region with quota constraints . . . 141

6.3 Aggregation sampling . . . 142

6.4 Aggregation reduction . . . 145

7 Case study . . . 146

8 Conclusions . . . 152

A Reduction proportion tables . . . 154

B Aggregation sampling tables . . . 157

C Reduction error tables . . . 162

References . . . 165

C Scenario Generation for Newsvendor Problems 169

1 Introduction . . . 171

2 Preliminaries . . . 173

2.1 Wasserstein distance . . . 174

2.2 Estimation of the optimality gap . . . 175

3 The univariate newsvendor problem 177
4 General case 179
4.1 Inactive Components 180
4.2 Scenario generation 182
5 Simple recourse problems 185
6 Numerical Test 188
7 Discussion and Future Work 190
A Numerical Test Problem Data 191
References 192

Thesis Conclusions **195**

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Jamie Fairbrother

Preface

Uncertainty is a key feature of many real-world decision making problems. In portfolio selection problems one has to choose how to invest capital in financial instruments with uncertain returns; in inventory problems one must choose quantities of stock without knowing the future demand. Scenario generation is concerned with the representation of uncertainty in a form appropriate for mathematical optimization. In particular, uncertain quantities must be represented by a finite number of possible future realizations, or *scenarios*, and one must specify a probability for each of these.

Typically the greater the number of scenarios one uses, the more reliable the solution that the optimization problem yields, but the more difficult the problem is to solve. It is therefore desirable to represent the uncertainty as concisely as possible. Standard scenario generation methods are *distribution-based*. That is, they construct scenario sets which faithfully reflect the set of future possibilities. The aim of this thesis is the design of *problem-based* scenario generation methods. These are methods which take advantage of the underlying structure of an optimization problem to provide a more parsimonious description of uncertainty. This may mean generating scenario sets which, in a probabilistic sense, do not accurately represent the distribution of future possibilities, but which yield near-optimal decisions to our problem.

The motivation for this thesis came from my supervisor Stein Wallace and research with his former PhD students Kjetil Høyland and Michal Kaut on

property-matching scenario generation. These methods consist of constructing scenario sets which have prescribed statistical properties. Crucially, these methods work on the premise that a given decision-problem will only react to certain statistical properties, and in this sense, can be considered to be problem-based. However, with these methods it is not usually clear *a priori* which properties are important to a particular decision problem and one has to resort to an empirical investigation to determine this. The aim of this project was therefore to develop methods which could be proven mathematically to be adapted to a particular problem. For this purpose my other supervisor, Amanda Turner, was enlisted to the project for her expertise in probability theory.

The first two papers of this thesis concern decision problems which involve tail risk measures. These are problems in which one attempts to mitigate or reduce the chance of extreme losses. The first paper is more general and theoretical in content, and was primarily written in collaboration with Amanda. The second paper, written primarily with Stein, is focused on portfolio selection, and how the methodology proposed in the first paper could be applied to realistic problems. The third and final paper of this thesis relates to a class of inventory problems, and although it was more of an independent piece of work, has benefited much from discussions with both of my supervisors.

And so a big thank to both of my supervisors. Stein, for his enthusiasm and insight, and whose flair for analogies would often be employed to make me see the bigger picture. Amanda, for her optimism and mathematical expertise, whose keen eye would often catch the flaws, subtle and unsubtle, in my own mathematical logic.

It has been a privilege to have undertaken this research at the STOR-i centre for doctoral training at Lancaster University. STOR-i has an engaged and collegial community of students who have enthusiastically developed and contributed to the activities, academic and social, of the centre. The

regular forums, training events, and masterclasses organized by them and staff have broadened my knowledge and skills well beyond the contents of this thesis. There are too many people to thank here individually for their work, collaboration and companionship, but I would like to mention my colleagues Chris Nemeth, Tim Park and Shreena Patel, with whom I joined STOR-i, for their friendship over these short few years.

Jamie Fairbrother

Lancaster University, April 21, 2016

Part I

Introduction

Background

1 Introduction

In this chapter we cover the preliminary material required for the reading of this thesis. This introduction is by no means exhaustive; its aim is to simply describe the general context of the research and provide some details on the results we will implictly rely upon. In Section 2 we give a brief overview of stochastic programming, in Sections 3 and 4 we present specific problems in stochastic programming: the newsvendor problem and conditional value-at-risk, two problems which feature prominently in our research papers. Finally, we end this chapter with a broad review of scenario generation methods in Section 5.

2 Stochastic Programming

2.1 General Stochastic Programs

Stochastic programming concerns optimization in the presence of uncertainty. In the most general form a stochastic program consists of a real-valued random vector $\tilde{\xi}(\omega) \in \Xi \subset \mathbb{R}^d$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, a deterministic set of feasible decisions $\mathcal{X} \subset \mathbb{R}^k$, and a *loss* function $\mathrm{f}_0 : \mathbb{R}^k \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and possibly a set of vector-valued functions $\mathrm{f}_i : \mathbb{R}^k \times \mathbb{R}^d \to \mathbb{R}^{m_i}$ for $i = 0, \ldots, m$ used to further constrain the problem. The aim of a stochastic pro-

gram is to minimize the expectation of the loss function subject to deterministic constraints and constraints in expectation:

$$\underset{x}{\text{minimize}} \;\; \mathbb{E}_{\mathbb{P}}\left[\mathrm{f}_0\left(x, \tilde{\xi}(\omega)\right)\right] \tag{1}$$

$$\text{subject to } \mathbb{E}_{\mathbb{P}}\left[\mathrm{f}_i\left(x, \tilde{\xi}(\omega)\right)\right] \leq 0, \qquad i = 1, \ldots, m$$

$$x \in \mathcal{X}. \tag{2}$$

Through the use of indicator functions, the constraints in expectation become probability constraints. These are useful in mitigating against extreme events which cannot reasonably be completely precluded (see [1] for instance).

This relatively simple form belies the modeling flexibility of stochastic programs and the difficulty of their solution. For instance, the two-stage stochastic linear program (SLP) has the following form:

$$\underset{x}{\text{minimize}} \; c^T x + \mathbb{E}_{\mathbb{P}}\left[Q(x, \tilde{\xi}(\omega))\right] \tag{3}$$

$$\text{subject to } Ax \leq b$$

$$x \geq 0$$

where

$$Q(x, \xi) = \min\left\{q^T y \;:\; Wy = h - Tx,\; y \geq 0\right\}, \tag{4}$$

and $y, q \in \mathbb{R}^t,\; h \in \mathbb{R}^s,\; W \in \mathbb{R}^{s \times t}, T \in \mathbb{R}^{s \times k}$ and finally $\xi = (q, W, h, T)$.

This type of problem models the situation where one has to make a strategic decision in the presence of uncertainty, followed by a corrective or *recourse* action once the values of uncertain parameters are fixed, and which incurs its own costs. The minimization in (4) is referred to as the recourse problem. The interpretation of the elements of the recourse problem is difficult as these themselves are constructed from underlying blocks of variables or parameters within the recourse problem. In addition, some of the components of the random vector $\tilde{\xi}(\omega) = (q(\omega), W(\omega), h(\omega), T(\omega))$ may be fixed.

For clarification we present a concrete example of a simple two-stage stochastic linear program. In this problem we have a group of facilities I

which must produce and delivery some commodity to a group of customers J. The aim of this problem is to decide on commodity capacities for each facility in such a way which minimizes the combined costs setting up the facilities, and the future costs of transporting the commodity and of rejected demand. The mathematical formulation follows:

Parameters:

c_i = unit cost of capacity for facility i

f_{ij} = unit cost of transporting commodity from facility i to customer j

r_j = unit rejection penalty for unsatisfied demand for customer j

$d_j(\omega)$ = stochastic demand of customer j

Decisions:

x_i = capacity of facility i

$y_{ij}(\omega)$ = amount of commodity to transport from facility i to customer j

$z_j(\omega)$ = rejected demand of customer j

$$\underset{x \geq 0}{\text{minimize}} \sum_{i \in I} c_i x_i + \mathbb{E}_{\mathbb{P}} \left[Q(x, d(\omega)) \right]$$

where $Q(x, d)$ is the optimal value to the following linear program:

$$\begin{aligned} \underset{y,z \geq 0}{\text{minimize}} \quad & \sum_{i \in I,\ j \in J} f_{ij} y_{ij} + \sum_{j \in J} r_j z_j \\ \text{subject to} \quad & \sum_{i \in I} y_{ij} + r_j = d_j \text{ for all } j \in J, && \text{(demand satisfied)} \\ & \sum_{j \in J} y_{ij} \leq x_i \text{ for all } i \in I. && \text{(capacity not exceeded)} \end{aligned}$$

The recourse problem of this stochastic program is the problem of minimizing the flow of the commodity from the facilities to the customers. Comparing this formulation to the general one in (4), we note that the only stochastic element of $\tilde{\xi}$ in this case is h.

For completeness, we mention also that the problem (1) also encompasses *multistage* stochastic programs in which the uncertainty takes the form of a

stochastic process $(\tilde{\xi}_1(\omega), \ldots, \tilde{\xi}_T(\omega))$, and one must make recourse decisions as each set of values $\tilde{\xi}_t(\omega)$ in the process is revealed. An example of this problem type is the multistage stochastic unit commitment problem [2]. The most general form of multistage stochastic program is the following:

$$\underset{x_1 \in \mathcal{X}_1}{\text{minimize}}\ f_{10}(x_1) + \mathbb{E}\left[\phi_1(x_1, \tilde{\xi}_1)\right]$$

where for $t = 2, \ldots, T$ the function $\phi_{t-1}\left(x_1, \ldots, x_{t-1}, \tilde{\xi}_1, \ldots, \tilde{\xi}_t\right)$ is defined implicitly as the optimal value to the following stochastic program:

$$\begin{aligned} \underset{x_t}{\text{minimize}}\ & \mathrm{f}_{t0}\left(x_t\right) + \mathbb{E}_{\tilde{\xi}_t|\tilde{\xi}_{t-1}}\left[\phi_t\left(x_1, \ldots, x_t, \tilde{\xi}_1, \ldots, \tilde{\xi}_t\right)\right] \\ \text{subject to}\ & \mathrm{f}_{ti}\left(x_1, \ldots, x_{t-1}, \tilde{\xi}_1, \ldots, \tilde{\xi}_{t-1}\right) \leq 0, \qquad i = 1, \ldots, m_t \\ & x_t \in \mathcal{X}_t. \end{aligned}$$

where $\tilde{\xi} = (\tilde{\xi}_1, \ldots, \tilde{\xi}_{t-1})$ and $\mathcal{X}_t$ are deterministic sets of feasible decisions.

2.2 Two-stage stochastic linear programs

The research in this thesis mostly relates to two-stage SLPs. Here we present some terminology related to and properties of this type of problem. For a detailed overview of this type of problem see [3, Chapter 3].

The function $Q(x, \xi)$ as defined in (4) is referred to as the *recourse function*, while $\mathcal{Q}(x) := \mathbb{E}\left[Q(x, \xi)\right]$ is the expected recourse function. By convention, when the mathematical program which defines the recourse function in (4) is infeasible, we set its value to $+\infty$. We denote the set of solutions x for which the $Q(x, \xi)$ is feasible for all $\xi \in \Xi$ by K, that is $K = \{x \in \mathbb{R}^k : Q(x, \xi) < +\infty \text{ for all } \xi \in \Xi\}$.

Similarly, if problem (4) is unbounded below then we set the value to be $-\infty$. Note that if $Q(x', \xi) = -\infty$ for some $x' \in \mathbb{R}^k$ then $Q(x, \xi) = -\infty$ for all $x \in \mathbb{R}^k$. To see this, note that dual to the linear program defined in (4) is as

follows:

$$\begin{aligned} &\underset{\pi \in \mathbb{R}^s}{\text{maximize}} \ (h - Tx)^T \pi \\ &\text{subject to } W^T \pi \geq q. \end{aligned}$$

If we have $Q(x', \xi) = -\infty$ for some $x' \in \mathbb{R}^k$ then this dual program is infeasible for x', but given that the constraints of this problem do not involve x, it must then be infeasible for all $x \in \mathbb{R}^k$, in which case $Q(x, \xi) = -\infty$ for all $x \in \mathbb{R}^k$.

A decision x is considered to be feasible for the problem (3) if $Q(x, \tilde{\xi}) < +\infty$ with probability 1, or equivalently, if $\mathcal{Q}(x) < +\infty$. Note that this condition is slightly weaker than the constraint $x \in K$.

The following result concerns the convexity of the recourse function:

Theorem 2.1. *Assuming the recourse function $Q(x, \xi)$ defined in (4) is not identically $-\infty$, it is:*

1. *a piecewise linear convex function in (h, T);*
2. *a piecewise linear concave function in q;*
3. *a piecewise linear convex function in x for all $x \in K$.*

Proof. We just prove that Q is convex in x. The proofs that Q is convex in (h, T) and concave in q are similar. For the proofs of piecewise linearity, see [3].

Fix $\xi \in \Xi$. If $Q(x, \xi) = -\infty$ for some x then the result is immediate as the function is identically $-\infty$, so we assume that this is not the case. Now, let $x_1, x_2 \in K$ and y_1, y_2 be corresponding solutions to the problem (4). We first show that the problem (4) is feasible for $x = \lambda x_1 + (1 - \lambda) x_2$:

$$\begin{aligned} W(\lambda y_1 + (1 - \lambda) y_2) &= \lambda W y_1 + (1 - \lambda) W y_2 \\ &= \lambda (h - T x_1) + (1 - \lambda)(h - T x_2) \\ &= h - T(\lambda x_1 + (1 - \lambda) x_2). \end{aligned}$$

Finally,

$$\begin{aligned} Q(\lambda x_1 + (1-\lambda)x_2, \xi) &\leq q^T(\lambda y_1 + (1-\lambda)y_2) \\ &= \lambda q^T y_1 + (1-\lambda) q^T y_2 \\ &= \lambda Q(x_1, \xi) + (1-\lambda) Q(x_2, \xi), \end{aligned}$$

where the first inequality follows from the fact that $\lambda y_1 + (1-\lambda)y_2$ is a feasible solution to the recourse problem. □

The recourse function is not convex or concave as a function of the matrix W. Thus, if the matrix W is non-stochastic, that is $W(\omega) \equiv W$, the stochastic program is more tractable. The problem in this case is said to have *fixed* recourse. In particular, if a problem has fixed recourse, it follows from Theorem 2.1, that the expected recourse function is convex.

The evaluation of the expected recourse function, and thus solving the problem (3), is typically analytically and numerically intractable when the random vector $\tilde{\xi}$ has a continuous distribution. However, when the distribution is discrete with mass points $\xi_s = (q_s, h_s, T^s)$ for $s = 1, \ldots, n$ and corresponding probabilities $(p_s)_{s=1}^n$, this evaluation reduces to a summation, and the optimization problem to a linear program:

$$\begin{aligned} \underset{x}{\text{minimize}}\ & c^T x + \sum_{i=1}^{n} p_s q_s^T y_s \\ \text{subject to}\ & Ax \leq b \\ & Wy_s = h_s - T^s \text{ for } s = 1, \ldots, n \\ & x, y \geq 0 \end{aligned}$$

Although this can be solved using standard linear programming, specialized algorithms exist which exploit the structure of this program, for example the L-shaped decomposition [4].

In paper C of this thesis we study a particular type of fixed recourse called *simple recourse*, which has the following form:

$$Q(x, \xi) = \min\{q_+^T y_+ + q_-^T y_- : Tx - \xi = Iy_+ - Iy_-,\ y_+,\ y_- \geq 0\}.$$

This function can be trivially rewritten as follows:

$$Q(x,\xi) = q_+^T (Tx - \xi)^+ + q_-^T (\xi - Tx)^+$$

where $x^+ = \max(x, 0)$ and the operator is applied element-wise. A simple recourse problem can thus be interpreted as follows: the vector Tx can be thought of as the availability of a set of resources, ξ the corresponding random demands for those resources, and y_+, y_- the surplus and shortfalls, respectively of the resources with respect to this demand. The vectors q_+ and q_- are then considered to be unit holding costs, and rejection costs, respectively.

An important property of simple recourse is their *separability*. That is, the recourse function can be decomposed as follows:

$$Q(x,\xi) = \sum_{i=1}^{d} Q_i(x,\xi)$$

where

$$Q_i(x,\xi) = q_{+i} (T_i x - \xi_i)^+ + q_{-i} (\xi_i - T_i x)^+ ,$$

and T_i denotes the i-th row of the matrix T. This feature is exploited in more specialized solution algorithms, see [5] for instance. We also make use of this property in Paper C.

3 The Basic Newsvendor Problem

The newsvendor problem is a univariate decision problem which concerns the inventory level of some product subject to an uncertain demand. The name newsvendor problem has been given to this as it aptly models the situation of a newsvendor who must decide upon a stock of newspapers to order to satisfy a daily uncertain demand. This problem is an example of a two-stage stochastic linear program with simple recourse, and it is used to illustrate our scenario generation methodology in Paper C of this thesis.

However, as will be seen in Section 4 the newsvendor problem is also intimately related to conditional-value-at-risk. In this section we define this problem, state its optimal solution, and give a detailed proof of this. A more in-depth study of this model, including its applications and extensions can be found in the classic textbook [6].

In the newsvendor problem, shortfall of stock relative to the demand incurs a unit rejection cost of $R > 0$. Similarly, a surplus of stock incurs a holding cost $h > 0$. The aim of the problem is to choose an inventory level which will minimize the total expected cost. If $\tilde{\xi}$ is a random variable representing demand, and $x \in \mathbb{R}$ is the inventory of the product, then this problem can be written as follows[1]:

$$\begin{aligned} &\underset{x\in\mathbb{R}}{\text{minimize}}\ \mathbb{E}\left[Q(x,\tilde{\xi})\right] \\ &\text{where } Q(x,\tilde{\xi}) = \min\{hz_+ + Rz_- :\ x - \tilde{\xi} = z_+ - z_-,\ z_+, z_- \geq 0\}. \end{aligned}$$

For convenience, we rewrite the recourse function $Q(x,\tilde{\xi})$ in the following form:

$$\underset{x\in\mathbb{R}}{\text{minimize}}\ h\ \mathbb{E}\left[\left(x-\tilde{\xi}\right)^+\right] + R\ \mathbb{E}\left[\left(\tilde{\xi}-x\right)^+\right] \tag{5}$$

Note that the objective function $\mathbb{E}\left[Q(x,\tilde{\xi})\right]$ is convex by the results in Section 2.2. The set of minimizers of $\mathbb{E}\left[Q(x,\tilde{\xi})\right]$ can be written in terms lower and upper quantiles of the random variable $\tilde{\xi}$. The lower quantile, or simply the quantile [2]of a random variable $\tilde{\xi}$ for $0 < \beta < 1$ is defined to be:

$$\xi_\beta = \inf\{x \in \mathbb{R} : \mathbb{P}\left(\tilde{\xi} \leq x\right) \geq \beta\},$$

similarly upper quantile is defined as follows:

$$\bar{\xi}_\beta = \inf\{x \in \mathbb{R} : \mathbb{P}\left(\tilde{\xi} \leq x\right) > \beta\}.$$

[1]The above interpretation of this problem requires that the solution satisfies $x \geq 0$. However, as will be seen, if the random variable $\tilde{\xi}$ is almost surely non-negative then the solution is guaranteed to be non-negative so we do not need to explicitly enforce this constraint.

[2]The lower quantile of a random variable when considered as a function of β is also referred to as the generalized inverse distribution function.

Proposition 3.1. *The set of minimizers of* $\mathbb{E}\left[Q(x,\tilde{\xi})\right]$ *is the following compact interval:*

$$I = [\xi_{\frac{R}{R+h}},\ \bar{\xi}_{\frac{R}{R+h}}].$$

Proof. Note first that for $x' < x$ we have

$$\begin{aligned}
\mathbb{E}\left[\left(x-\tilde{\xi}\right)^+\right] &= \int_{(-\infty,x]}(x-\xi)\ \mathbb{P}\left(d\xi\right) = \int_{(-\infty,x]}\left((x-x')-(\xi-x')\right)\mathbb{P}\left(d\xi\right)\\
&= (x-x')\ \mathbb{P}\left(\tilde{\xi}\le x\right) + \int_{(-\infty,x']}(x'-\xi)\ \mathbb{P}\left(d\xi\right) - \int_{(x',x]}(\xi-x')\ \mathbb{P}\left(d\xi\right)\\
&= (x-x')\ \mathbb{P}\left(\tilde{\xi}\le x\right) + \mathbb{E}\left[\left(x'-\tilde{\xi}\right)^+\right] - \int_{(x',x]}(\xi-x')\ \mathbb{P}\left(d\xi\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}\left[\left(\tilde{\xi}-x\right)^+\right] &= \int_{(x,+\infty)}(\xi-x)\ \mathbb{P}\left(d\xi\right) = \int_{(x,+\infty)}\left((\xi-x')-(x-x')\right)\ \mathbb{P}\left(d\xi\right)\\
&= \int_{(x',+\infty)}\left(\xi-x'\right)\ \mathbb{P}\left(d\xi\right) - \int_{(x',x]}(\xi-x')\ \mathbb{P}\left(d\xi\right) - (x-x')\mathbb{P}\left(\tilde{\xi}>x\right)\\
&= \mathbb{E}\left[(\tilde{\xi}-x')^+\right] - \int_{(x',x]}(\xi-x')\ \mathbb{P}\left(d\xi\right) - (x-x')\mathbb{P}\left(\tilde{\xi}>x\right).
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}\left[Q(x,\tilde{\xi})\right] &= \mathbb{E}\left[Q(x',\tilde{\xi})\right] + (x-x')\left(h\mathbb{P}\left(\tilde{\xi}\le x\right) - R\mathbb{P}\left(\tilde{\xi}>x\right)\right)\\
&\qquad - (h+R)\int_{(x',x]}(\xi-x')\ \mathbb{P}\left(d\xi\right)\\
&= \mathbb{E}\left[Q(x',\tilde{\xi})\right] + (h+R)(x-x')\left(\mathbb{P}\left(\tilde{\xi}\le x\right) - \frac{R}{h+R}\right)\\
&\qquad - (h+R)\int_{(x',x]}(\xi-x')\ \mathbb{P}\left(d\xi\right). \qquad (6)
\end{aligned}$$

To show that the values in the above interval minimize $\mathbb{E}\left[Q(x,\tilde{\xi})\right]$ we compare the value of this function for different values of x and x' using (6). First, let $x' \in I$ and $x > \bar{\xi}_{\frac{R}{R+h}}$. By the definition of the upper quantile function, the second term in (6) is strictly positive, also

$$\begin{aligned}
\int_{(x',x]}(y-x')\ \mathbb{P}\left(dy\right) &< (x-x')\ \mathbb{P}\left(x'<\tilde{\xi}\le x\right)\\
&= (x-x')\left(\mathbb{P}\left(\tilde{\xi}\le x\right) - \frac{R}{R+h}\right),
\end{aligned}$$

hence $\mathbb{E}\left[Q(x',\tilde{\xi})\right] < \mathbb{E}\left[Q(x,\tilde{\xi})\right]$. Similarly, if $x' < \xi_{\frac{R}{R+h}}$ and $x \in I$, it can be shown that $\mathbb{E}\left[Q(x,\tilde{\xi})\right] \leq \mathbb{E}\left[Q(x',\tilde{\xi})\right]$. Since $\mathbb{E}\left[Q(x,\tilde{\xi})\right]$ is convex, it just remains to be shown that it is constant on I. Suppose I is not a single point, that $x' = \xi_{\frac{R}{R+h}}$ and $x \in I$ with $x > x'$. Note that if I is not a single point then we must have $\mathbb{P}\left(\tilde{\xi} \leq \xi_{\frac{R}{R+h}}\right) = \frac{R}{R+h}$. If $x < \bar{\xi}_{\frac{R}{R+h}}$ then $\mathbb{P}\left(x' < \tilde{\xi} \leq x\right) = 0$ and by (6) we see that we must have $\mathbb{E}\left[Q(x,\tilde{\xi})\right] = \mathbb{E}\left[Q(x',\tilde{\xi})\right]$. If $x = \xi_{\frac{R}{R+h}}$, then $\int_{(x',x]}(\xi - x')\ \mathbb{P}\left(d\xi\right) = (x - x')\ \left(\mathbb{P}\left(\tilde{\xi} \leq x\right) - \frac{R}{R+h}\right)$ and again using (6) we have that $\mathbb{E}\left[Q(x,\tilde{\xi})\right] = \mathbb{E}\left[Q(x',\tilde{\xi})\right]$ as required. □

4 Risk Measures and Conditional Value-at-Risk

4.1 General Risk Measures

Throughout this section will denote by Z a random variable in $\mathbb{R}$ which represents some loss. For our purposes, a risk measure is simply a functional on a space of random variables.

Definition 4.1 (Risk Measure)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and V be a non-empty set of $\mathcal{F}$-measurable real-valued random variables. Then, a risk measure is some function $\rho : V \to \mathbb{R} \cup \{\infty\}$.*

However, for a risk measure to be useful it should in some way quantify the danger of large losses[3]. The quintessential example of a risk measure is the variance of a random variable and was first used in [8] for portfolio selection problems. A small variance implies a small probability of extreme losses by Chebyshev's inequality:

$$\mathbb{P}\left(|Z - E|Z|| \geq \alpha\right) \leq \frac{\mathrm{Var}(Z)}{\alpha^2}.$$

[3]The recent paper [7] which proposes a more general framework for measures of risk and deviation, gives the following more specific characterization: a risk measure ρ should "model X is "adequately" $\leq C$ by the inequality $\rho(Z) \leq C$", where C is some loss one wishes not to exceed.

The use of variance as a measure of risk is problematic for a few reasons. The foremost of these is perhaps that variance penalizes large profits as well as large losses. As a consequence, in the case where the returns of financial assets are not symmetrically distributed, using the variance can lead to patently bad decisions; for instance, a portfolio can be chosen in favor of one which always has higher returns (see [9] for an example of this). This particular issue can be overcome by using a "downside" risk measure, that is one which only depends on losses greater than the mean, or some other specified threshold. For example the semi-variance [10, Chapter 9], or mean regret [11]:

$$\begin{aligned} \text{SemiVar}(Z) &= \mathbb{E}\left[|Z - \mathbb{E}[Z]|_+^2\right] \\ \text{MeanRegret}_\tau(Z) &= \mathbb{E}\left[|Z - \tau|_+\right] \end{aligned}$$

The semi-variance measures the deviation of losses greater than the mean, whereas the mean-regret calculates the average loss exceeding some level τ.

The paper [12] introduced the idea of a *coherent* risk measure which is a risk measure which satisfies the following properties:

- (Positive homogeneity) $\rho(\lambda Z) = \lambda\rho(Z)$ for $\lambda \geq 0$
- (Translation invariant) $\rho(Z + a) = \rho(Z) + a$ for any $a \in \mathbb{R}$
- (Subadditivity) $\rho(Z_1 + Z_2) \leq \rho(Z_1) + \rho(Z_2)$
- (Monotonicity) If $Z \geq 0$ then $\rho(Z) \geq 0$[4]

Each of these has interpretations in finance, for instance if Z represents the loss associated with the return of a portfolio, the subadditivity property ensures that a risk measure favors diversification of portfolios. See [12] for more details. These properties also ensure that a risk measure has desirable mathematical properties. In particular subadditivity and positive homogeneity directly imply that a risk measure is convex.

[4]This differs from the corresponding axiom in [12] where Z is interpreted as utility rather than loss

4.2 Conditional Value-at-Risk

We now concentrate on a risk measure known as the *conditional value-at-risk*, as this is the risk measure we use for our numerical experiments.

Denote by $0 < \beta < 1$ a risk level. By G_Z we denote the distribution function of Z, that is

$$G_Z(z) = \mathbb{P}(Z \leq z).$$

By G_Z^{-1} we denote the *generalized inverse distribution function*, or *quantile function*, of Z, that is

$$G_Z^{-1}(\beta) = \inf\{z \in \mathbb{R} : G_Z(z) \leq \beta\}.$$

We will assume that the random variable Z has finite mean.

The β Value-at-Risk , or β-VaR, is a risk measure simply defined to be the β-quantile of a random variable, that is $\beta\text{-VaR}(Z) = G_Z^{-1}(\beta)$. The β-VaR has been widely used in finance [13], and it has the convenient interpretation of representing the amount of capital required to cover up to $\beta \times 100\%$ of potential losses. However, the β-VaR has some undesirable properties: it is not coherent and is generally intractable in an optimization context.

The β Conditional Value-at-Risk, or β-CVaR, is a risk measure which dominates the β-VaR and overcomes its major deficiencies. It can be thought of as the conditional expectation of a random variable above the β-VaR, which is indeed the case for continuous random variables, but the general definition is more technical. The β-VaR and β-CVaR for a continuous random variable are illustrated in Figure 1.

The β-CVaR was first proposed in [14], and can be defined in several ways. We use the following definition which is the most relevant in the context of optimization.

Definition 4.2 (β-CVaR)**.**

$$\beta\text{-CVaR}(Z) = \min_{\alpha \in \mathbb{R}} \{\alpha + \frac{1}{1-\beta}\mathbb{E}\left[(Z-\alpha)^+\right]\} \tag{7}$$

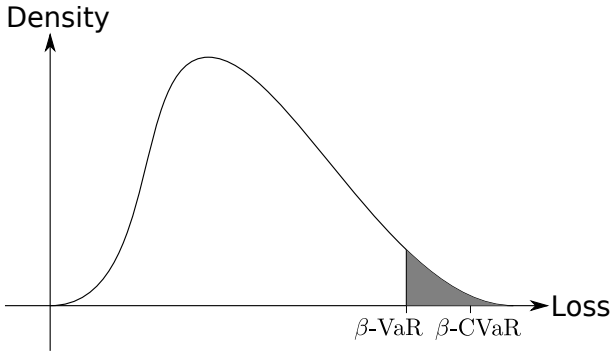


Fig. 1: The β-VaR and β-CVaR of a continuous random variable

The definition of β-CVaR given in (7) is intimately related to the newsvendor problem presented in Section 3. Setting $h = (1-\beta)$ and $R = \beta$ in (5) and using α in place of x, the objective function of the newsboy problem can be rewritten as follows:

$$\begin{aligned}
&(1-\beta)\ \mathbb{E}\left[(\alpha - Z)^+\right] + \beta\ \mathbb{E}\left[(Z-\alpha)^+\right] \\
&\qquad = (1-\beta)\ \mathbb{E}\left[(Z-\alpha)^+ - (Z-\alpha)\right] + \beta\ \mathbb{E}\left[(Z-\alpha)^+\right] \\
&\qquad = -(1-\beta)\ \mathbb{E}\left[Z\right] + (1-\beta)\ \alpha + \mathbb{E}\left[(Z-\alpha)^+\right] \\
&\qquad = -(1-\beta)\ \mathbb{E}\left[Z\right] + (1-\beta)\left(\alpha + \frac{1}{1-\beta}\ \mathbb{E}\left[(Z-\alpha)^+\right]\right).
\end{aligned}$$

Thus, calculating the β-CVaR is equivalent to solving a newsvendor problem.

Sometimes the conditional value-at-risk is referred to as the *expected shortfall*. As the name suggests, this quantity is usually defined with respect to lower tail of a random variable representing profit, rather than the upper tail of a random variable representing loss as we have done. The following alternative characterizations of β-CVaR were originally given in [15] in relatation to the expected shortfall. We restate and prove these results with respect to the upper tail of the distributions rather than lower tails.

Proposition 4.3. *The following are alternative characterizations of the conditional value-at-risk:*

i $\frac{1}{1-\beta}\left(\mathbb{E}\left[Z\mathbb{1}_{\{Z\geq G_Z^{-1}(\beta)\}}\right]-G_Z^{-1}(\beta)\left(\beta-\mathbb{P}\left(Z<G_Z^{-1}(\beta)\right)\right)\right)$

ii $\frac{1}{1-\beta}\int_\beta^1 G_Z^{-1}(u)\,du$

From the first of these characterizations it is clear that when Z is continuous, we have β-CVaR$(Z) = \mathbb{E}[Z|Z \geq \beta\text{-VaR}(Z)]$. The second characterization is written purely in terms of the quantile function of the distribution and allows us to easily place β-CVaR in a wider collection of risk measures we call β-tail risk measures as will be seen in Paper A of this thesis.

Proof. We first show that the first characterization is equivalent to (7). Noting the equivalence of calculating the β-CVaR with the newsvendor problem, and using Proposition 3.1, we see that the minimization in (7) is achieved for $\alpha = G_Z^{-1}(\beta)$. Hence,

$$\begin{aligned}
\beta\text{-CVaR} &= G_Z^{-1}(\beta)+\frac{1}{1-\beta}\mathbb{E}\left[\left(Z-G_Z^{-1}(\beta)\right)^+\right]\\
&= G_Z^{-1}(\beta)+\frac{1}{1-\beta}\int_{[G_Z^{-1}(\beta),+\infty)}\left(z-G_Z^{-1}(\beta)\right)\,\mathbb{P}(dz)\\
&= G_Z^{-1}(\beta)+\frac{1}{1-\beta}\left(\mathbb{E}\left[Z\mathbb{1}_{\{Z\geq G_Z^{-1}(\beta)\}}\right]-G_Z^{-1}(\beta)\,\mathbb{P}\left(Z\geq G_Z^{-1}(\beta)\right)\right)\\
&= \frac{1}{1-\beta}\left(\mathbb{E}\left[Z\mathbb{1}_{\{Z\geq G_Z^{-1}(\beta)\}}\right]+G_Z^{-1}(\beta)\left(1-\beta-\mathbb{P}\left(Z\geq G_Z^{-1}(\beta)\right)\right)\right)\\
&= \frac{1}{1-\beta}\left(\mathbb{E}\left[Z\mathbb{1}_{\{Z\geq G_Z^{-1}(\beta)\}}\right]-G_Z^{-1}(\beta)\left(\beta-\mathbb{P}\left(Z<G_Z^{-1}(\beta)\right)\right)\right).
\end{aligned}$$

Thus, the first alternative characterization is proved.

To verify the second alternative formulation, we show that it is equivalent to the first. Let $U \sim \text{Uniform}(0,1)$ and define $Z' = G_Z^{-1}(U) \sim Z$. Note that,

$$\{U\geq\beta\} = \{Z'\geq G_Z^{-1}(\beta)\}\setminus\left(\{U<\beta\}\cap\{Z'=G_Z^{-1}(\beta)\}\right) \tag{8}$$

and so

$$\mathbb{1}_{\{U\geq\beta\}} = \mathbb{1}_{\{Z'\geq G_Z^{-1}(\beta)\}}-\mathbb{1}_{\{U<\beta\}\cap\{Z'=G_Z^{-1}(\beta)\}}. \tag{9}$$

Now,

$$\begin{aligned}\int_\beta^1 G_Z^{-1}(u)\,du &= \mathbb{E}\left[Z'\mathbb{1}_{\{U\geq\beta\}}\right]\\ &= \mathbb{E}\left[Z'\mathbb{1}_{\{Z'\geq G_Z^{-1}(\beta)\}}\right] - \mathbb{E}\left[Z'\mathbb{1}_{\{Z'=G_Z^{-1}(\beta)\}\cap\{U<\beta\}}\right]\\ &= \mathbb{E}\left[Z'\mathbb{1}_{\{Z'\geq G_Z^{-1}(\beta)\}}\right] - G_Z^{-1}(\beta)\left(\beta - \mathbb{P}\left(Z' < G_Z^{-1}(\beta)\right)\right)\\ &= \mathbb{E}\left[Z\mathbb{1}_{\{Z\geq G_Z^{-1}(\beta)\}}\right] - G_Z^{-1}(\beta)\left(\beta - \mathbb{P}\left(Z < G_Z^{-1}(\beta)\right)\right)\end{aligned}$$

as required. □

Another definition of β-CVaR is given in [16] where it is defined to be the expectation with respect to an appropriately modified tail distribution function. The β-CVaR was shown to be a coherent risk measure in [16], and [17].

The main reason for the popularity of β-CVaR is that it is tractable in an optimization setting. Like in Section 2 denote by $\mathcal{X} \subset \mathbb{R}^k$ a set of feasible decisions, by $\tilde{\xi}$ a random vector with support $\Xi \subset \mathbb{R}^d$, and our loss function by $\mathrm{f} : \mathbb{R}^k \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. We make the technical assumptions that for all $x \in \mathcal{X}$ we have that $\xi \mapsto \mathrm{f}(x,\xi)$ is measurable and $\mathbb{E}\left[\mathrm{f}\left(x,\tilde{\xi}\right)^+\right] < +\infty$.

Define the following auxiliary function,

$$F_\beta(x,\alpha) = \alpha + \frac{1}{1-\beta}\,\mathbb{E}\left[\left(\mathrm{f}\left(x,\tilde{\xi}\right)-\alpha\right)^+\right],$$

so that β-CVaR$(\mathrm{f}\left(x,\tilde{\xi}\right)) = \min_{\alpha\in\mathbb{R}}\{F_\beta(x,\alpha)\}$. Now, the basic theory of optimization ensures that minimizing $F_\beta(x,\alpha)$ with respect to $\alpha \in \mathbb{R}$ and then minimizing the residual function with respect to $x \in \mathcal{X}$ is equivalent to minimizing $F_\beta(x,\alpha)$ with respect to $(x,\alpha) \in \mathcal{X}\times\mathbb{R}$. That is, minimizing the β-CVaR $\left(\mathrm{f}\left(x,\tilde{\xi}\right)\right)$ is equivalent to minimizing the much more tractable function $F_\beta(x,\alpha)$. Moreover, since $F_\beta(x,\cdot)$ achieves its minimum for each $x \in \mathcal{X}$ the solution sets coincide. This is summarized in the following theorem.

Theorem 4.4. *The minimization of β-CVaR$(\mathrm{f}\left(x,\tilde{\xi}\right))$ with respect to $x \in \mathcal{X}$ is equivalent to minimizing $F_\beta(x,\alpha)$ over $\mathcal{X}\times\mathbb{R}$:*

$$\min_{x\in\mathcal{X}}\ \beta\text{-CVaR}(\mathrm{f}\left(x,\tilde{\xi}\right)) = \min_{(x,\alpha)\in\mathcal{X}\times\mathbb{R}} F_\beta(x,\alpha). \tag{10}$$

and moreover the sets of solutions coincide:

$$(x^*, \alpha^*) \in \underset{(x,\alpha)\in\mathcal{X}\times\mathbb{R}}{\operatorname{argmin}} \; F_\beta(x,\alpha) \iff$$

$$x^* \in \underset{x\in\mathcal{X}}{\operatorname{argmin}} \; \beta\text{-CVaR}\left(\mathrm{f}\left(x,\tilde{\xi}\right)\right), \; \alpha^* \in \underset{\alpha\in\mathbb{R}}{\operatorname{argmin}} \; F_\beta(x^*,\alpha).$$

The minimization of the auxiliary function $F_\beta(x,\alpha)$ with respect to $(x,\alpha) \in \mathcal{X}\times\mathbb{R}$ is particularly tractable when the underlying cost function is convex.

Corollary 4.5. *Suppose that the loss function $x \mapsto \mathrm{f}(x,\xi)$ is convex for all $\xi \in \Xi$. Then, the function $F_\beta(x,\alpha)$ is jointly convex in $(x,\alpha) \in \mathcal{X}\times\mathbb{R}$, and moreover, β-CVaR$\left(\mathrm{f}\left(x,\tilde{\xi}\right)\right)$ is convex as a function of $x \in \mathcal{X}$.*

Proof. If $\mathrm{f}(x,\xi)$ is a convex function, then the function $(\mathrm{f}(x,\xi)-\alpha)^+$ is also convex as a function of (x,α), and since the expectation of a convex function is convex, the function $F_\beta(x,\alpha)$ is a convex function of (x,α).

The function β-CVaR$\left(\mathrm{f}\left(x,\tilde{\xi}\right)\right)$ is the residual of $F_\beta(x,\alpha)$ when we have minimized over α. A standard result from convex analysis [18, Proposition 2.22] tells us that when convex function is minimized with respect to some of its variables, the residual function is convex. Thus, β-CVaR$(\mathrm{f}\left(x,\tilde{\xi}\right))$ is also convex function of $x \in \mathcal{X}$. □

When the loss function is convex in $x \in \mathcal{X}$ we can thus use standard algorithms from convex optimization to minimize the β-CVaR. In the case where the random vector $\tilde{\xi}$ is discrete, and $\mathrm{f}\left(x,\tilde{\xi}\right)$ is the recourse function of the stochastic linear program in (4), we can write the problem in (10) as a linear program. Suppose the random vector $\tilde{\xi}$ has mass points $\xi_s = (q_s, h_s, T^s)$ with associated probabilities p_s, for $s = 1,\ldots,n$. We introduce non-negative auxiliary decision variables $z_s \geq 0$, along with the constraints $z_s \geq q_s^T y_s - \alpha$ for $s = 1,\ldots,n$, so that z_s models the exceedance of the loss over the variable α in scenario s. The problem of minimizing the β-CVaR of loss function of

two-stage SLP can now be written as follows:

$$\begin{aligned}
\underset{x,\ \alpha}{\text{minimize}}\ & c^T x + \alpha + \frac{1}{1-\beta}\sum_{s=1}^{n} p_s z_s \\
\text{subject to}\ & z_s \geq q_s^T y_s - \alpha \qquad \forall\ s = 1, \ldots, n \\
& Wy = h_s - T_s x \qquad \forall\ s = 1, \ldots, n \\
& Ax \leq b \\
& x, y, z \geq 0.
\end{aligned}$$

5 Scenario Generation

5.1 Introduction

In Section 2 we stated that stochastic programming problems were generally intractable when the underlying random vector was continuous. Scenario generation is the construction of a discrete random vector to use within a stochastic program. This discrete random vector is usually referred to as a *scenario set* and the individual atoms of the distribution as the *scenarios*. Generally, the more scenarios in a scenario set, the better the representation of the uncertainty, and so the more reliable the solutions they yield. However, the more scenarios one uses, the more computationally expensive the problem is to solve. Scenario generation is therefore a trade-off between accuracy and tractability.

Scenario generation methods can be categorized as distribution-driven or problem-driven. The first three subsection present the main three families of standard distribution-driven methods. In Section 5.2, we present sampling approaches where one simply uses a sample from an underlying probabilistic model of the uncertainty as a scenario set. In Section 5.3, we present the optimal discretization approach to scenario generation where one attempts to explicitly minimize the distance between a probabilistic model and the constructed scenario set. In Section 5.4, we cover constructive approaches to scenario generation where one directly models uncertain parameters with a discrete distribution.

The focus of this thesis is the developement of problem-driven scenario generation methods which have not received much study. In Section 5.5 we we present two heuristic examples problem-driven approaches to scenario generation from the literature.

5.2 Sampling Approaches

The simplest way to construct a scenario set is to sample from a probabilistic model for the uncertain quantities in the stochastic program.

In this section we suppose our stochastic optimization problem is of the following form:

$$\underset{x\in\mathcal{X}}{\text{minimize}} \;\; \{\mathrm{F}_{\tilde{\xi}}(x) := \mathbb{E}_{\tilde{\xi}}\left[\mathrm{f}(x,\tilde{\xi})\right]\} \tag{11}$$

where x is a vector of decision variables with deterministic feasible region $\mathcal{X} \subset \mathbb{R}^k$, $\tilde{\xi}$ is a random vector with support $\Xi \subset \mathbb{R}^m$, and $\mathrm{f} : \mathbb{R}^k \times \mathbb{R}^d \to \mathbb{R}$ is the loss function. Unlike the general model given in (1) we assume here that there are no expectation constraints, that is the feasible region does not depend on the distribution of $\tilde{\xi}$. Not only does this simplify the theory of sampling in stochastic problems, as will be seen in Section 5.2, it also allows one to easily assess the quality of solutions. For a more detailed treatment of this subject, including the general case, see [19, Chapter 5].

We denote the set of optimal solutions and the optimal solution value to problem (11) respectively as follows:

$$S := \underset{x\in\mathcal{X}}{\operatorname{argmin}} \mathrm{F}_{\tilde{\xi}}(x), \qquad z^* = \min_{x\in\mathcal{X}} \mathrm{F}_{\tilde{\xi}}(x).$$

we also denote an optimal solution as follows:

$$x^* \in \underset{x\in\mathcal{X}}{\operatorname{argmin}} \mathrm{F}_{\tilde{\xi}}(x).$$

Suppose now that $\tilde{\xi}_1, \tilde{\xi}_2, \ldots$ are a sequence of independently, identically distributed copies of $\tilde{\xi}$, on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The *sample average approximation* (SAA) of the problem is defined as follows:

$$\underset{x\in\mathcal{X}}{\text{minimize}} \;\; \{\mathrm{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathrm{f}(x,\tilde{\xi}_i)\} \tag{12}$$

Similarly to the above, we denote the set of optimal solutions and the optimal solution value for the SAA as follows:

$$S_n := \underset{x\in\mathcal{X}}{\operatorname{argmin}} \; \mathrm{F}_n(x), \qquad z_n^* := \min_{x\in\mathcal{X}} \; \mathrm{F}_n(x)$$

and an optimal solution is denoted as follows:

$$x_n^* \in \underset{x \in \mathcal{X}}{\operatorname{argmin}} \; \mathrm{F}_n(x).$$

The quality of a solution x_n^* with respect to the original problem (11) is not guaranteed. Indeed, z_n^* and x_n^* are random[5]since they depend on the realizations of the random vectors $\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_n$. All we can hope to do is make probabilistic statements about the distributions of z_n^* and x_n^*. In this section we present some theorems concerning the asymptotic behavior of solutions from SAAs, and also how sampling can be used to assess the quality of a feasible solution.

Before moving on to the asymptotic theory, we present now two basic results, taken from [20], which provide some intuition about the behavior of solutions obtained from the SAA.

In a stochastic program, the objective is to find a decision which minimizes some *expected* future loss. In effect, this means that we must find a decision which leads to relatively low losses for all likely future scenarios. In a SAA, we are minimizing our costs with respect to only a subset of possible future scenarios. Hedging over a smaller set of scenarios, we are liable to 'over-optimize', and so we may expect the optimal costs with respect to the approximated problem to be lower. This observation is formalized in the following proposition.

Proposition 5.1. *Let $\tilde{\xi}_1, \ldots, \tilde{\xi}_n$ be independently, identically distributed, with the distribution of $\tilde{\xi}$; then,*

$$\mathbb{E}\left[z_n^*\right] \leq z^* \tag{13}$$

[5]Given that $f(x, \xi)$ is continuous in x and measurable in ξ, it can be shown that $z_n^*(\omega)$ and the set of optimal solutions $S_n(\omega) = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \mathrm{F}_n(\omega, x)$ are measurable functions. Viewing x_n^* as a measurable selection of S_n, it can be considered alongside z_n^* to be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. See [19] for more details.

Proof.

$$\begin{aligned} z^* &= \min_{x\in\mathcal{X}} \mathbb{E}_{\tilde{\xi}} \left[f(x, \tilde{\xi}) \right] \\ &= \min_{x\in\mathcal{X}} \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^{n} f(x, \tilde{\xi}_i) \right] \\ &\geq \mathbb{E} \left[\min_{x\in\mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} f(x, \tilde{\xi}_i) \right] \\ &= \mathbb{E} \left[z_n^* \right] \end{aligned}$$

□

The greater the sample size, the more scenarios against which we have to hedge against in our SAA. Thus, we may expect the optimal costs of the SAA increase as we increase our sample size.

Proposition 5.2. *Let $\tilde{\xi}_1, \ldots, \tilde{\xi}_{n+1}$ independently, identically distributed with the distribution of $\tilde{\xi}$; then,*

$$\mathbb{E} \left[z_n^* \right] \leq \mathbb{E} \left[z_{n+1}^* \right]$$

Proof.

$$\begin{aligned} \mathbb{E} \left[z_{n+1}^* \right] &= \mathbb{E} \left[\min_{x\in\mathcal{X}} \frac{1}{n+1} \sum_{i=1}^{n} f(x, \tilde{\xi}_i) \right] \\ &= \mathbb{E} \left[\min_{x\in\mathcal{X}} \frac{1}{n+1} \sum_{i=1}^{n+1} \frac{1}{n} \sum_{j=1, j\neq i}^{n+1} f(x, \tilde{\xi}_i) \right] \\ &\geq \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E} \left[\min_{x\in\mathcal{X}} \frac{1}{n} \sum_{j=1, j\neq i}^{n+1} f(x, \tilde{\xi}_i) \right] \\ &= \mathbb{E} \left[z_n^* \right] \end{aligned}$$

□

The preceding propositions are instructive: they tell us that our expected solution value is optimistic and improves as we increase our sample size. In addition, they hold in full generality, unlike the main theorems in this section.

Consistency

A sequence of estimators (random variables) $\tilde{\zeta}_1, \tilde{\zeta}_2, \ldots$ is said to be *consistent* with the parameter (value) ζ if $\tilde{\zeta}_n$ converges to ζ with probability 1, that is, if $\mathbb{P}\left(\lim_{n\to\infty} \tilde{\zeta}_n = \zeta\right) = 1$. For the SAA to be a useful approximation to (11) the estimators z_n^* and x_n^* must be consistent with z^* and x^* respectively.

For the sake of generality, in the following results, taken from [19], we take $F : \mathcal{X} \to \mathbb{R}$ to be an arbitrary function, and $F_n : \mathcal{X} \to \mathbb{R}$ a sequence of random functions defined on the common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We will assume that F_n converges uniformly to F with probability 1 as $n \to \infty$. Although this is not strictly required for consistency, this assumption allows for more elementary proofs of the following two theorems.

Theorem 5.3. *Suppose that* F_n *converges to* F *with probability 1 as* $n \to \infty$ *uniformly on* $\mathcal{X}$*. Then* z_n^* *converges to* z^* *with probability 1 as* $n \to \infty$*.*

Proof. For $\omega \in \Omega$, we modify our notation for the sample average function to be $\mathrm{F}_n(\omega, x) := \frac{1}{n}\sum_{i=1}^{n} \mathrm{f}(x, \xi_i(\omega))$ to make explicit the dependence of its value on the underlying probability space. The uniform convergence with probability 1 means that for all $\epsilon > 0$, and almost every $\omega \in \Omega$ there exists $N(\epsilon, \omega) \in \mathbb{N}$ such that for all $n > N(\epsilon, \omega)$ we have

$$\sup_{x\in\mathcal{X}} |\mathrm{F}_n(\omega, x) - \mathrm{F}(x)| < \epsilon. \tag{14}$$

Fix $\omega \in \Omega$ such that (14) holds and $n > N(\epsilon, \omega)$. Also, let $x_n^* \in \operatorname*{argmin}_{x\in\mathcal{X}} \mathrm{F}_n(\omega, x)$ and $x^* \in \operatorname*{argmin}_{x\in\mathcal{X}} \mathrm{F}(x)$, and without loss of generality suppose that $z_n^* \leq z^*$. Now,

$$\begin{aligned} z^* - z_n^*(\omega) &= \mathrm{F}(x^*) - \mathrm{F}_n(\omega, x_n^*) \\ &\leq \mathrm{F}(x_n^*) - \mathrm{F}_n(\omega, x_n^*) \\ &< \epsilon \qquad \text{by (14).} \end{aligned}$$

Hence

$$|z_n^*(\omega) - z^*| < \epsilon$$

for almost all $\omega \in \Omega$ as required. □

The above Theorem guarantees the convergence of solution values. For the convergence of solutions, we need some notion of convergence of solution sets. For this we use a measure of distance between sets, the *deviation*. For $A, B \subset \mathbb{R}^k$ this is defined as follows:

$$\mathbb{D}(A, B) = \sup_{x \in A} \operatorname{dist}(x, B)$$
$$\text{where } \operatorname{dist}(x, B) = \inf_{x' \in B} \left\| x - x' \right\|$$

The following theorem, which has slightly stronger conditions than Theorem 5.3, guarantees the convergence of the set of optimal solutions.

Theorem 5.4. *Suppose that there exists a compact set $C \subset \mathbb{R}^k$ such that:*

i the set S of optimal solutions to the true problem is contained in C

ii the function $\mathrm{F}(x)$ is finite-valued and continuous on C

iii $\mathrm{F}_n(x)$ converges to $\mathrm{F}(x)$ with probability 1 uniformly on C

iv With probability 1 for n large enough the set S_n is non-empty and $S_n \subset C$.

Then $z_n^* \to z^*$ and $\mathbb{D}(S_n, S) \to 0$ with probability 1 as $n \to \infty$.

Proof. Given that $S \subset C$ we can assume without loss of generality that $\mathcal{X}$ is compact. From assumptions (i) and (iii), we have by Theorem 5.3 that $z_n^* \to z^*$ with probability 1. To show that $\mathbb{D}(S_n, S) \to 0$ with probability 1 as $n \to \infty$ it thus suffices to show that $\mathbb{D}(S_n(\omega), S) \to 0$ for all $\omega \in \Omega$ such that $z_n^*(\omega) \to z^*$. We prove this by contradiction.

Suppose $z_n^*(\omega) \to z^*$ but $\mathbb{D}(S_n(\omega), S) \nrightarrow 0$. Then, there exists $\epsilon > 0$ such that for each $n \in \mathbb{N}$ there is $x_n^*(\omega) \in S_n(\omega)$ such that $\|z_n^*(\omega) - z^*\| \geq \epsilon$. By the compactness of $\mathcal{X}$ we may assume (taking a subsequence if necessary) that $x_n^* \to x^*$ for some $x^* \in \mathcal{X}$. Note that $x^* \notin S$ hence $\mathrm{F}(x^*) > z^*$. Now,

$$\mathrm{F}_n(x_n^*) - \mathrm{F}(x^*) = [\mathrm{F}_n(x_n^*) - \mathrm{F}(x_n^*)] + [\mathrm{F}(x_n^*) - \mathrm{F}(x^*)].$$

The first term on the RHS of this expression tends to zero by assumption (iii). The second term on the RHS of this expression tends to zero by assumption (ii). Hence, $z_n^* = \mathrm{F}_n(x^*) \to \mathrm{F}(x^*) > z^*$ which is a contradiction. □

Despite the strength of the assumption, uniform convergence of the SAA holds for an important class of stochastic programs. Given a two-stage stochastic linear program with fixed recourse, we have uniform convergence of the sample average function if the set of feasible decisions $\mathcal{X}$ is compact [19, Theorem 7.48]. If the the loss function is convex, then there exist similar consistency results which only require point-wise convergence, for example see [21].

Asymptotic Distributions

The previous results did not tell us anything about the rate of convergence of the optimal solution values of the SAA. The following result due to Shapiro in [22] gives a central limit theorem for the optimal solution values when the stochastic program has a unique minimizer.

Theorem 5.5. *Suppose that $\mathcal{X}$ is compact and the following conditions hold:*

i *For all $x \in \mathcal{X}$, $\tilde{\xi} \mapsto \mathrm{f}(x, \tilde{\xi})$ is measurable.*

ii *There exists a point $\tilde{x} \in \mathcal{X}$ such that $\mathbb{E}\left[\mathrm{f}(\tilde{x}, \tilde{\xi})^2\right] < \infty$.*

iii *There exists $b : \Xi \to \mathbb{R}$ such that $\mathbb{E}\left[b(\tilde{\xi})^2\right] < \infty$ and $|\mathrm{f}(x, \tilde{\xi}) - \mathrm{f}(y, \tilde{\xi})| \leq b(\tilde{\xi}) \|x - y\|$.*

If the stochastic program (11) has a unique minimizer $S = \{x^*\}$, then

$$n^{\frac{1}{2}}(z_n^* - z^*) \xrightarrow{d} N(0, \sigma^2) \text{ as } n \to \infty$$

where $\sigma^2 = \mathrm{Var}\left(\mathrm{f}(x^*, \tilde{\xi})\right)$.

Notice that most of these assumptions will generally hold for a two-stage linear stochastic program: the deterministic feasible region is defined by linear equalities (or inequalities) and so is closed, and can also be made bounded

and thus compact by introducing artificial bounds on the decision variables, something which is not unrealistic in most real-world problems; $\xi \mapsto \mathrm{f}(x, \xi)$ is piecewise linear (by Theorem 2.1) and thus measurable. Assumptions (ii) will hold for instance if the random vector $\tilde{\xi}$ has bounded support. Under stronger assumptions, it was shown that a similar central limit theorem also holds for the optimal solutions x_n^*. See [22] for details.

Shapiro has also derived bounds on the probabilities of a solution to an SAA having a value close to the optimal solution. Under stronger conditions, it has been shown that these probabilities converge at an exponential rate to one (see [23] for instance).

Assessing Solution Quality

Given a candidate solution $x_0 \in \mathcal{X}$ to the stochastic program (11) we show how one can construct approximate confidence intervals for the true objective function value $\mathbb{E}\left[\mathrm{f}(x_0, \tilde{\xi})\right]$ (also known as the out-of-sample value) and the optimality gap.

We will assume that for all $x \in \mathcal{X}$ that $\mathbb{E}\left[\mathrm{f}(x, \tilde{\xi})^2\right] < \infty$. This allows us to appeal to the central limit theorem (CLT).

Suppose we have a feasible solution x_0 to the problem (11). Let $X = \mathrm{f}(x_0, \tilde{\xi})$ and $X_i = \mathrm{f}(x_0, \tilde{\xi}_i)$ for $i = 1, \ldots, n$. Now, the random variables X_i are i.i.d. with the distribution of X, and by our assumptions the mean and variance of X exist and are finite. For large n we can therefore apply the CLT.

Fix a confidence level $0 < \beta < 1$, and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean and $\sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ the sample variance. Using standard results from statistics, $\left(\bar{X}_n - \epsilon_\beta, \bar{X}_n + \epsilon_\beta\right)$ is a $(1-\beta)$ approximate confidence interval for $\mathbb{E}[X]$ where $\epsilon_\beta = \frac{\sigma_n}{\sqrt{n}}\,\Phi^{-1}\left(1 - \frac{\beta}{2}\right)$ and Φ is the distribution function of the standard Normal distribution. That is, for large n,

$$\left(\frac{1}{n}\sum_{i=1}^n \mathrm{f}(x_0, \tilde{\xi}_i) - \epsilon_\beta, \frac{1}{n}\sum_{i=1}^n \mathrm{f}(x_0, \tilde{\xi}_i) + \epsilon_\beta\right)$$

is a $(1-\beta)$ approximate confidence interval for $\mathbb{E}\left[\mathrm{f}(x_0, \tilde{\xi})\right]$.

A confidence interval can be similarly constructed for the optimality gap of a given feasible solution. The method presented originates from [20]. The optimality gap of a feasible solution $x_0 \in \mathcal{X}$ is defined as follows:

$$G = \mathbb{E}\left[\mathrm{f}(x_0, \tilde{\xi})\right] - z^*.$$

Now, define

$$G_n = \frac{1}{n} \sum_{i=1}^{n} \mathrm{f}(x_0, \tilde{\xi}_i) - z_n^*.$$

Note that,

$$\begin{aligned}
\mathbb{E}\left[G_n\right] = & \ \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \mathrm{f}(x_0, \tilde{\xi}_i) - z_n^*\right] \\
\geq & \ \mathbb{E}\left[\mathrm{f}(x_0, \tilde{\xi})\right] - z^* \qquad \text{by Proposition 5.1} \\
= & \ G.
\end{aligned}$$

Since $0 \leq G \leq \mathbb{E}\left[G_n\right]$, a conservative confidence interval on G can be made by constructing a confidence interval for $\mathbb{E}\left[G_n\right]$. We make the assumption that the central limit theorem holds for the random variable G_n and construct an approximate confidence interval for $\mathbb{E}\left[G_n\right]$ in a similar fashion to that above.

Let $\tilde{\xi}ij$ for $1 \leq i \leq n_g$, and $1 \leq j \leq n$ be i.i.d. random variables with the distribution of $\tilde{\xi}$, and define

$$G_n^i = z_{n,i}^* - \frac{1}{n} \sum_{j=1}^{n} \mathrm{f}(x_0, \tilde{\xi}_{ij})$$

where $z_{n,i}^* = \min_{x \in \mathcal{X}} \frac{1}{n} \sum_{j=1}^{n} \mathrm{f}(x, \tilde{\xi}_{ij})$. The random variables G_n^i are independently identically distributed with the distribution of G_n. Let $\bar{G}_n = \frac{1}{n_g} \sum_{i=1}^{n_g} G_n^i$ and $\sigma_{G,n_g}^2 = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} \left(G_n^i - \bar{G}_n\right)$. Since each evaluation of G_n^i may be expensive, the number of batches n_g used may be relatively small and so the random variable

$$\frac{\sqrt{n_g}\left(\bar{G}_n - \mathbb{E}\left[G_n\right]\right)}{\sigma_{G,n_g}}$$

is best approximated by a t-distribution with $n_g - 1$ degrees of freedom. For $0 < \alpha < 1$, let $t_{n_g - 1, \alpha}$ be the $(1 - \alpha)$-quantile of the t-distribution with $n_g - 1$

degrees of freedom. That is,

$$\mathbb{P}\left(T_n \leq t_{n_g-1,\alpha}\right) = 1-\alpha.$$

Now, setting $\epsilon_{n_g,\alpha} = t_{n_g-1,\alpha}\frac{\sigma_{G,n_g}}{\sqrt{n_g}}$, we have:

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{E}\left[G_n\right] \leq \bar{G}_n + \epsilon_{n_g,\alpha}\right) \\
&= \mathbb{P}\left(\frac{\bar{G}_n - \mathbb{E}\left[G_n\right]}{\frac{\sigma_{G,n_g}}{\sqrt{n_g}}} \geq -t_{n_g-1,\alpha}\right) \\
&\approx \mathbb{P}\left(\frac{\bar{G}_n - \mathbb{E}\left[G_n\right]}{\frac{\sigma_{G,n_g}}{\sqrt{n_g}}} \leq t_{n_g-1,\alpha}\right) \qquad \text{(by symmetry of t-distribution)} \\
&\approx 1-\alpha.
\end{aligned}$$

Hence, $(0, \bar{G}_n + \epsilon_{n_g,\alpha})$ is a $(1-\alpha)$-approximate confidence interval for $\mathbb{E}\left[G_n\right]$ and thus a $(1-\alpha)$-approximate confidence interval for G. Note that the size of this interval decreases if we increase either the number of batches n_g since $\epsilon_{n_g,\alpha}$ decreases as n_g increases, or the batch size n, since $\mathbb{E}\left[G_n\right]$, an upper bound on G, will decrease as we increase n by Proposition 5.2.

The main drawback of the above method for estimating a confidence interval for the optimality gap is that it involves solving multiple problems. Other procedures have been proposed which require require only one or two replications [24], [25].

The estimation technique presented here does not require i.i.d. samples. Any sampling technique which produces unbiased estimates of the expected loss function is also valid, we only require independence between the batches of samples. This opens up the possibility of using variance reduction techniques, such as Latin hypercube sampling, or antithetic sampling, to reduce the size of the error in our estimates of the optimality gap. See [26] and [27] for instance.

5.3 Optimal Discretization

One might expect for a stochastic program, whose loss function satisfies certain continuity properties, that if the underlying random vector is slightly perturbed then the expected loss function would only experience a small change. The distance between two random vectors can be measured using a probability metric, and it has been shown that under certain conditions the effect of using a different random vector in a stochastic program can be bounded by the distance between the original and new random vector.

By *optimal discretization* we mean the discretization of a random vector so as to explicitly minimize the distance between the original and discretized random vectors, with respect to some probability metric. The probability metric which should be used depends upon the type of problem. For instance, it has been shown that discrepancy distances are a natural metric to use for probabilistically constrained problems and mixed integer recourse problems [28]; Fortet-Mourier metrics are a natural choose for two-stage recourse problems [29].

In this section we introduce a probability metric called the Wasserstein distance and show that this is a natural metric to use for discretization with linear fixed recourse problems. This metric is used in Paper C to analyze the behavior of the proposed scenario generation methodology.

Approximation Error and Wasserstein Distance

The discretization of a continuous random vector to solve a stochastic program leads to another stochastic program which is an approximation of the original. The error is most meaningfully quantified by the optimality gap of the solution that the approximate problem yields.

Definition 5.6 (Approximation error)**.** *The* approximation error *induced by using the random vector* $\check{\xi}$ *in the place of* $\tilde{\xi}$ *with respect to the problem* (11) *is as*

follows:

$$e(\tilde{\xi}, \breve{\xi}) = \sup_{x_0 \in \underset{x \in \mathcal{X}}{\operatorname{argmin}} F_{\breve{\xi}}(x)} \{\min_{x \in \mathcal{X}} F_{\tilde{\xi}}(x) - F_{\tilde{\xi}}(x_0)\}$$

A convenient way to bound the approximation error is to use the sup-distance between the true and approximate expected cost functions. The following elementary lemma is taken from [29].

Lemma 5.7.

$$e(\tilde{\xi}, \breve{\xi}) \leq 2 \, ||F_{\tilde{\xi}} - F_{\breve{\xi}}||_\infty$$

Proof. Set $\epsilon = ||F_{\tilde{\xi}} - F_{\breve{\xi}}||_\infty$, let $x^* \in \operatorname{argmin} F_{\tilde{\xi}}$ and $\tilde{x}^* \in \operatorname{argmin} F_{\breve{\xi}}$. We assume that $F_{\tilde{\xi}}(x^*) \leq F_{\breve{\xi}}(\tilde{x}^*)$ and derive a contradiction by supposing that $F_{\tilde{\xi}}(x^*) + 2\epsilon < F_{\tilde{\xi}}(\tilde{x}^*)$. A similar argument holds for the reverse case.

$$\begin{aligned} F_{\tilde{\xi}}(x^*) + 2\epsilon &< F_{\tilde{\xi}}(\tilde{x^*}) \\ &\leq F_{\breve{\xi}}(\tilde{x}^*) + \epsilon \qquad \text{by definition of } \epsilon \\ &\leq F_{\breve{\xi}}(x^*) + \epsilon \\ &\leq F_{\tilde{\xi}}(x^*) + 2\epsilon \qquad \text{by definition of } \epsilon \end{aligned}$$

A contradiction is established and so the result holds. □

Minimizing the sup-distance is thus a good proxy to minimize the approximation error. For a stochastic linear program with fixed recourse, this sup-distance can be bounded in turn by the Wasserstein distance between $\tilde{\xi}$ and $\breve{\xi}$ which we now define.

Definition 5.8. *Suppose $\tilde{\xi}$ and $\breve{\xi}$ are random vectors in $\mathbb{R}^d$. Then, the Wasserstein distance between $\tilde{\xi}$ and $\breve{\xi}$ (with respect to the 1-norm) is as follows:*

$$d_W(\tilde{\xi}, \breve{\xi}) = \inf_{Y_1, Y_2} \{\mathbb{E}\left[\|Y_1 - Y_2\|\right]\} \tag{15}$$

where the infimum is taken over all pairs of random vectors Y_1, Y_2 defined on the same probability space such that $Y_1 \sim \tilde{\xi}$ and $Y_2 \sim \breve{\xi}$.

The Wasserstein distance is strongly related to the optimal transportation problem. To see this, we restate the definition in terms of probability measures:

$$d_W(\tilde{\xi}, \check{\xi}) = \inf_{\pi} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|y_1 - y_2\| \, d\pi(y_1, y_2)$$

where the infimum is taken over all probability measures π on the product space $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are such that for all measurable $A, B \subset \mathbb{R}^d$:

$$\pi(A \times \mathbb{R}^n) = \mu_1(A)$$
$$\pi(\mathbb{R}^n \times B) = \mu_2(B)$$

where μ_1 and μ_2 are the probability measures for the random vectors $\tilde{\xi}$ and $\check{\xi}$ respectively. Now, for a fixed measure π the quantity $\pi(A \times B)$ can be viewed as the amount of mass one is transporting from A to B, and $\int_{A \times B} \|y_1 - y_2\| \; d\pi(y_1, y_2)$ the cost of this transportation. The calculation of the Wasserstein distance thus amounts to finding a transportation plan of minimal cost. See [30] for more details.

The key property of fixed recourse problems that allows us to use the Wasserstein distance to bound the sup-distance between the true expected loss function and an approximation is that the loss function in such a problem has the Lipschitz property[6], whose definition we now recall.

Definition 5.9 (Lipschitz). *For a function* $\mathrm{g} : \Xi \subset \mathbb{R}^m \to \mathbb{R}$*, its Lipschitz constant is defined as follows:*

$$L(\mathrm{g}) = \inf\{L : |\mathrm{g}(u) - \mathrm{g}(v)| \leq L \|u - v\| \text{ for all } u, v \in \mathbb{R}^m\} \tag{16}$$

The function g is said to be Lipschitz if $L(g) < \infty$*.*

The Wasserstein distance is related to Lipschitz functions via the Kantorovich-Rubinstein Theorem.

[6]This follows from Theorem 2.1 which says that the loss function for a stochastic program with fixed recourse is piecewise-linear

Theorem 5.10 (Kantorovich-Rubinstein)**.**

$$d_W(\tilde{\xi}, \breve{\xi}) = \sup\{\mathbb{E}_{\tilde{\xi}}\left[g\left(\tilde{\xi}\right)\right] - \mathbb{E}_{\breve{\xi}}\left[g\left(\breve{\xi}\right)\right] : g : \mathbb{R}^n \to \mathbb{R} \textit{ is Lipschitz } \}$$

For a proof of this see [30, Chapter 1]. Suppose now that $\bar{L} > 0$ is a Lipschitz constant for our loss function, uniform across all decisions $x \in \mathcal{X}$, that is

$$|f(x, \xi_1) - f(x, \xi_2)| \leq \bar{L} \|\xi_1 - \xi_2\| \qquad \text{for all } x \in X, \text{ and } \xi_1, \xi_2 \in \Xi.$$

Hence, $\xi \mapsto \frac{1}{\bar{L}} f(x, \xi)$ is Lipschitz with constant 1 for all $x \in \mathcal{X}$ and so applying the Kantorovich-Rubinstein Theorem we have

$$\begin{aligned} ||F_{\tilde{\xi}} - F_{\breve{\xi}}||_\infty &= \sup_{x \in \mathcal{X}} \{\mathbb{E}_{\tilde{\xi}}\left[f(x, \tilde{\xi})\right] - \mathbb{E}\left[f(x, \breve{\xi})\right]\} \\ &\leq \bar{L} d_W(\tilde{\xi}, \breve{\xi}). \end{aligned}$$

In particular, we have

$$e(\tilde{\xi}, \breve{\xi}) \leq 2\bar{L} d_W(\tilde{\xi}, \breve{\xi}).$$

Scenario Reduction and Generation

In the section above we showed that the error of approximating a random vector in a stochastic program can be bounded by the Wasserstein distance between the true and approximate random vectors. Hence, when approximating a random vector with a discrete one, we should try to minimize this distance.

Suppose we are trying to approximate the random vector $\tilde{\xi}$ with the discrete random vector $\breve{\xi}$ which has mass points $\{\xi_1, \ldots, \xi_N\}$ and probabilities $\{p_1, \ldots, p_N\}$. These mass points induce a (Voronoi) partition[7] on the space $\mathbb{R}^d$:

$$A_i = \{\xi \in \mathbb{R}^d : \|\xi - \xi_i\| = \min_{1 \leq i \leq N} \|\xi - \xi_i\|\}$$

[7] These partitions can be made disjoint using the following convention: if ξ belongs to more than one set assign it to the one with minimal i

Now, the probabilities which minimizes the Wasserstein distance between $\tilde{\xi}$ and this discrete random vector are $p_i = \mathbb{P}\left(\tilde{\xi} \in A_i\right)$ for $i = 1, \ldots, N$. In this case the Wasserstein distance is as follows:

$$d_W(\tilde{\xi}, \breve{\xi}) = \sum_{i=1}^{N} \int_{A_i} \left\| \tilde{\xi} - \xi_i \right\| \mathbb{P}\left(d\tilde{\xi}\right)$$

This fact can be seen by viewing the definition of the Wasserstein distance as a mass transportation problem. The most efficient way of transporting the mass in the partition set A_i is to transport it to the closest mass point ξ_i. See [29] for more details.

To minimize the Wasserstein distance of a discrete approximation, scenario generation methods thus seek to solve the following problem:

$$\underset{\xi_1, \ldots, \xi_N}{\text{minimize}} \quad \sum_{i=1}^{N} \int_{A_i} \left\| \tilde{\xi} - \xi_i \right\| \mathbb{P}\left(d\tilde{\xi}\right) \tag{17}$$

This problem is highly non-convex and one typically must resort to heuristics. The paper [29] suggests a variant of the *k*-means clustering algorithm [31] to converge to a local optimum. The paper [32] suggests two heuristics, forwards and backwards reduction, for the case of scenario reduction where one is attempting delete a given proportion of scenarios from a large scenario set in a way that minimizes the distance between the Wasserstein distance between the original and reduced sets.

More recently, a *nested distance* has been proposed in [33], which is specially adapted for multistage stochastic programs where one must discretize a stochastic process.

5.4 Constructive Approaches

When formulating a stochastic program, uncertain parameters must be described by a full multivariate probability distribution. Expertise and analysis of historical data may lead us to compile a list of properties we would like our distribution to have. For instance, if we wanted to model the distribution of stock returns, we may want to specify the first four moments along

with the correlation structure to adequately describe the body and tails. If our uncertain parameters are described by a stochastic process we may wish to prescribe auto-correlations. However, finding a parametric distribution which has all our given properties may be difficult or impossible. For example, with the Normal distribution one has direct control over the mean and covariance structure, and no control over skewness or kurtosis.

In constructive methods, one aims to directly construct a discrete distribution which has certain statistical properties equal or approximately equal to some given target values. The approach was first advocated in [34] where it is postulated that a given stochastic program will only be sensitive to particular statistical properties of the distribution. A concrete example of this idea is the Markowitz model [8]. This is an optimization problem used in portfolio selection where one must choose a portfolio that balances its expected return against its variance. One formulation of this problem is the following:

$$\begin{aligned} \underset{x\in\mathbb{R}^d}{\text{minimize}} \quad & \alpha \operatorname{Var}\left(x^T\tilde{\xi}\right) - (1-\alpha)\mathbb{E}\left[x^T\tilde{\xi}\right] \\ \text{subject to} \quad & \sum_{i=1}^{d} x_i = 1, \\ & x \geq 0, \end{aligned}$$

where the decision vector x represents the portfolio allocation, $\tilde{\xi}$ is a random vector which represents the returns of the assets, and $0 < \alpha < 1$ is a parameter controlling risk aversion. This model can be rewritten as follows:

$$\begin{aligned} \underset{x\in\mathbb{R}^d}{\text{minimize}} \quad & \alpha\ x^T \operatorname{Cov}(\tilde{\xi})x - (1-\alpha)\ x^T\mathbb{E}\left[\tilde{\xi}\right] \\ \text{subject to} \quad & \sum_{i=1}^{d} x_i = 1, \\ & x \geq 0. \end{aligned}$$

From this restatement, it is clear that any random vectors with the same mean and covariance matrix will yield an equivalent problem. However, in general, it is not clear which are the important properties to match for

a given stochastic program, and this must be investigated through stability analysis [35].

We now review some existing constructive scenario generation methods. Special emphasis is given to the moment matching method which is used for numerical tests in Paper B.

Property Optimization

In this approach we take our scenarios and their associated probabilities as decision variables and try to minimize the squared error of specified statistical properties from target values. This approach was first proposed in [34]. For clarity, we just describe it for two-stage problems. The construction of a scenario set is done by solving the following optimization problem:

$$\begin{aligned} \underset{\xi,p}{\text{minimize}} \quad & \sum_{i=1}^{l} w_i \left(g_i(\xi, p) - t_i\right)^2 \\ \text{subject to} \quad & \sum_{s=1}^{n} p_s = 1 \\ & p \geq 0, \end{aligned}$$

where w_i is the weighting of property i, for $i = 1, \ldots, l$; t_i is the target value of property i; p_s is the probability of scenario s; ξ_i^s is the realization of the $i-$th random variable in scenario s; and $\mathrm{g}_i(\xi, p)$ is a function which gives the value of the $i-$th specified statistical property. For example, if the $i-$th statistical property is the mean of the j-th random variable then $\mathrm{g}_i(\xi, p) = \sum_{s=1}^{n} \xi_j^s p_s$. Depending on the properties specified, the above problem may be non-convex, in which case it must be solved using heuristic methods. The problem may be simplified by fixing p^s as parameters rather than decision variables.

The paper identifies three main issues with this method:

- **Inconsistent specifications** Many statistical properties are related, and target values of some statistical properties may be inconsistent with

others. In this case, the property optimization will lead to a scenario set which satisfies neither of the inconsistent properties exactly. The degree of the match will depend on the relative weight assigned to that property.

- **Over-specifications** If the number of scenarios chosen for the optimization is too small, then there will be no scenario set which satisfies the specified properties. In this case, one must increase the number of scenarios used to yield a good match.

- **Under-specifications** If the number of scenarios is large with respect to the number of specified properties then the scenario set may be a good match but there could also be undesired side effects. In [34] it was noted that under-specification leads to many probabilities being set to zero.

As a rough guide one should choose the number of scenarios to be approximately the number of specifications. See [34] for a deeper discussion.

Moment-matching

The first four moments of a probability distribution (mean, variance, skewness and kurtosis) give one vital information about a distribution. Visually, they tell one about the location, spread, symmetry, and the thickness of the tails of the tails of a distribution. For a multivariate distribution, in addition to the moments, the correlation matrix gives a visual description of the shape and orientation of the distribution. The descriptive power of these statistics relies on the distribution being uni-modal and near-elliptical, a realistic assumption when modeling many real-world phenomena. In [36] the authors present an heuristic to construct a discrete distribution whose margins have specified values for their first four moments with specified values, and whose correlation structure is also specified. This has grown in popularity because

of its simplicity of application and has been used in different domains including finance [37] and inventory management [38]. The algorithm is based around two transformations: a cubic transformation which corrects the first four moments, and a linear transformation to correct the correlations.

Cubic Transformation Suppose we have a random variable, X, which we would like to transform to have first four non-central moments $\mu_1, \mu_2, \mu_3, \mu_4$. Let $Y = a + bX + cX^2 + dX^3$. Now Y has the specified moments if the coefficients in this transformation a, b, c, d satisfy the following system of non-linear equations:

$$\mu_i = \left(a + b\ \mathbb{E}\left[X\right] + c\ \mathbb{E}\left[X^2\right] + d\ \mathbb{E}\left[X^3\right]\right)^i \qquad \text{for i = 1, \ldots, 4.} \tag{18}$$

In [36] this system is solved by reformulating this problem as an unconstrained optimization problem where the coefficients a, b, c, d are decision variables, and the objective is to minimize the total distance of the moments of Y from their target values $\mu_1, \ldots, \mu_4$. This approach ensures that if the system of equations (18) does not have a solution, the solution algorithm to the optimization problem will return the best available one rather than just fail.

Linear Transformation Let R be a correlation matrix. Now, R has a Cholesky decomposition $R = LL^T$ where L is an lower-triangular matrix. A basic result from statistics states that if $Z \sim N(0, I)$, then $LZ \sim N(0, R)$. More generally, we have the following theorem, adapted from [36]:

Theorem 5.11. *Let R be a correlation matrix, $R = LL^T$ the Cholesky decomposition. Suppose $X = (X_1, \ldots, X_n)$ is a random vector with the following properties:*

1. $\mathbb{E}\left[X_i\right] = 0$ *and* $\mathbb{E}\left[X_i^2\right] = 1 \qquad$ *for* $i = 1 \ldots n$.
2. *The marginals of X are pairwise independent.*

Then, for $Y = LX$, we have

1. $\mathbb{E}[Y] = 0$ *and* $\mathbb{E}[Y^2] = 1$.

2. *Y has correlation matrix* $R = LL^T$.

Suppose we have a discrete random vector $X = (X_1, \ldots, X_n)$. We can assume without loss of generality that $\mathbb{E}[X] = 0$ and $\mathbb{E}[X]^2 = 1$ as this can be corrected by a simple translation and scaling $X_i \mapsto \alpha_i X_i + \beta_i$. However, for X to have independent margins it would need an exponential number of scenarios. Such a large number of scenarios would be computationally intractable. The scenario generation method presented in [36] is initialized by sampling the marginals from a standard normal distribution and combining these appropriately to construct a discrete distribution. The marginals are unlikely to be independent and so the two transformations presented above cannot be used alone to construct a discrete distribution with exact target moments and correlations. The heuristic is instead an iterative procedure in which the above two transformations are repeatedly applied until the moments and correlations of our constructed distribution are within a certain distance of their target values. The basic structure of the heuristic is given below.

Although it lacks the flexibility of the property optimization method discussed above, this method is faster and doesn't suffer from the same underspecification problems discussed above. A simpler but less flexible moment matching method was more recently proposed in [40].

Other methods

Matching moments in a discrete distribution is just one way of controlling the marginals. In [39] a heuristic is presented which constructs a distribution which has marginal distributions which are close, in a probabilistic sense, to some target distributions as well as having specified correlations. This moves away from the idea of choosing properties which are important to the underlying stochastic program towards a convenient way of modeling the

input : Target moments and correlations

output: Scenario set with target moments and correlations

Generate initial sample;

while *errorOfCorrelations* $>$ *MaxErrCorr* **or** *errorOfMoments* $>$ *MaxErrorMoms* **do**

  if *errorOfCorrelations* $>$ *MaxErrorCorr* **then**

    Correct correlations with linear transformation;

  end

  for $i = 1, \cdots, d$ **do**

    if $errorOfMoments_i > MaxErrorMom$ **then**

      Correct moments of margin i with cubic transformation;

    end

  end

end

Algorithm 1: The moment matching algorithm, taken from [39].

distribution of uncertain parameters.

Like the moment matching heuristic, this marginal matching method depends on two transformations, one to fix the marginals and one to fix the correlations of the distribution. The marginals are transformed using cumulative distribution functions, and the correlations are fixed using the same linear transformation presented in Section 5.4.

In [41], the authors provide an alternative approach to generating scenario sets with specified marginals, which attempts to minimize directly the distance between specified marginal distributions and the constructed scenario set. The paper [42] advocates the use of copulas to allow the user to specify a more arbitrary dependency structure between the random variables.

5.5 Problem-Driven Scenario Generation

The approaches to scenario generation in the previous sections were *distribution-driven*, that is, they were primarily concerned with the accurate representation of future uncertainty without taking into account the underlying stochastic program. The aim of this thesis is to promote the idea of *problem-driven* scenario generation. By taking into account the structure of the problem it may be possible to construct a more parsimonious representation of the uncertainty. Crucially, a scenario set constructed in such a way may not be close to the true distribution of uncertainty as measured by a probability metric such as the Wasserstein distance (see Section 5.3), but will yield a solution which is near optimal with respect to the "true" distribution.

There are only a few cases of problem-driven scenario generation in the literature and these are somewhat heuristic in nature. The property-matching scenario generation methods of Section 5.4 can be considered problem-driven in the sense that a given stochastic program may only react to certain statistical properties. However, as we explained, an empirical investigation must be carried out to to identify which properties are important, and in reality these methods are often used as a convenient way of modeling the uncertain

quantities.

In this section we present two very different examples of problem-driven scenario generation from the literature: the first based on sampling, the second derived from the optimal discretization approaches discussed in Section 5.3.

Importance Sampling At the end of Section 5.2 we briefly discussed the use of variance reduction techniques to improve performance of sampling as a scenario generation method. One such technique is *importance sampling*. In importance sampling, one draws samples from a proxy distribution, and appropriately adjusts the weights of each sample to approximate the required expectation. This technique was first used in stochastic programming as an internal sampling method within a Bender's decomposition algorithm in [43], and was further developed in [44]. It can be considered a problem-driven scenario generation approach as the construction of the proxy distribution depends on the underlying loss function of the stochastic program.

We suppose the random vector $\tilde{\xi}$ is discrete and has probability mass function p, and for fixed $x \in \mathcal{X}$ we would like to estimate the expected loss function $\mathbb{E}\left[\mathrm{f}(x, \tilde{\xi})\right]$. Suppose $q : \Omega \mapsto \mathbb{R}$ is another probability mass function with the same support, then:

$$\begin{aligned}
\mathbb{E}\left[\mathrm{f}(x, \tilde{\xi})\right] &= \sum_{\omega \in \Omega} \mathrm{f}(x, \tilde{\xi}(\omega))\ p(\tilde{\xi}(\omega)) \\
&= \sum_{\omega \in \Omega} \frac{\mathrm{f}(x, \tilde{\xi}(\omega)) p(\tilde{\xi}(\omega))}{q(\tilde{\xi}(\omega))} q(\tilde{\xi}(\omega)) \\
&= \mathbb{E}\left[\frac{\mathrm{f}(x, \breve{\xi}) p(\breve{\xi})}{q(\breve{\xi})}\right]
\end{aligned}$$

where the final expectation is taken with respect to the random vector $\breve{\xi}$ which has probability mass function q. Therefore the following estimator can be used for estimating the expected loss function:

$$\bar{z}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{f}(x, \xi_i) p(\xi_i)}{q(\xi_i)}$$

where $\xi_1, \ldots, \xi_n$ are independent samples from the distribution with mass function q. The mean and variance of this estimator are as follows:

$$\mathbb{E}\left[\bar{z}_n\right] = \mathbb{E}\left[\mathrm{f}(x, \tilde{\xi})\right]$$
$$\mathrm{Var}(\bar{z}_n) = \frac{1}{n} \sum_{\omega \in \Omega} \left(\frac{\mathrm{f}(x, \tilde{\xi}(\omega)) p(\omega)}{q(\omega)} - \mathbb{E}\left[\mathrm{f}(x, \tilde{\xi})\right] \right)^2 q(\omega)).$$

Note that unlike the other variance reduction techniques mentioned at the end of Section 5.2, the variance is not guaranteed to be reduced, and so one much choose the proxy distribution q wit care. Assuming that the loss function at x is non-negative, then the ideal choice for q would be

$$q^*(\xi) = \frac{\mathrm{f}(x, \xi) p(\xi)}{\mathbb{E}\left[\mathrm{f}(x, \tilde{\xi})\right]} \tag{19}$$

for which we would have $\mathrm{Var}(\bar{z}_n) = 0$ for any n. However, this density requires knowledge of $\mathbb{E}\left[\mathrm{f}(x, \tilde{\xi})\right]$ which is what we are trying to estimate in the first place. The important observation here is that to reduce the variance we should construct q to be close to to q^*, which we can do by approximating $\mathrm{f}(x, \xi)$ in the expression (19).

The following "additive" approximation for $\mathrm{f}(x, \xi)$ was employed in [43]:

$$\mathrm{f}(x, \xi) \approx \mathrm{f}(x, \tau) + \sum_{k=1}^{d} \Delta \mathrm{f}_k(x, \xi_i) \tag{20}$$

where

$$\Delta \mathrm{f}_i(x, \xi_i) = \mathrm{f}(x, \tau_i, \ldots, \xi_i, \ldots, \tau_d) - \mathrm{f}(x, \tau), \tag{21}$$

and τ is some fixed point $\tau \in \Xi$. If τ is chosen such that $\mathrm{f}(x, \tau) \leq \mathrm{f}(x, \tilde{\xi}(\omega))$ for all $\omega \in \Omega$ then each $\Delta \mathrm{f}_i(x, \xi_i)$ is non-negative. If $\mathrm{f}(x, \tau) \geq 0$ then the approximation in (20) is also non-negative and so can be used to construct a probability mass function. Now, the calculation of the coefficient of proportionality using the above approximation only involves the marginal expectations $\mathbb{E}\left[\Delta \mathrm{f}_k(x, \tilde{\xi})\right]$.

As compared to other variance reduction techniques, this method of importance sampling is computationally expensive, as the definition of the

proxy distribution requires the evaluation of $\Delta f_i(x, \xi(\omega)_i)$ for all $\omega \in \Omega$ and $i = 1, \ldots, d$. In addition if $f(x, \tau) + \sum_{k=1}^{d} \Delta f_k(x, \xi(\omega)_i) = 0$ for some $\omega \in \Omega$ then the proxy distribution will have zero mass at this point, and so one will be unable to use it for importance sampling.

Forward Selection in Recourse Clusters The paper [45] proposes a modification to the fast forward selection algorithm (FFS) of [32] which is a method of scenario reduction. This method, called *forward selection in recourse clusters* (FSRC), attempts to avoid redundancy in scenarios by first clustering the scenarios according to their behavior with respect to the problem, and then using a standard reduction technique, called forward selection, to select one scenario per cluster in the reduced set.

Suppose $Q(x, \xi)$ is the recourse function from (4) and we would like to reduce the set of (equiprobable) scenarios $\{\xi_s\}_{s \in S}$. For each $s \in S$, let y_s^* denote the corresponding solution to the recourse problem. Now, for our problem we define *sensitivity indices* $\mathcal{F}_i(x, y)$ for $i = 1, \ldots, v$. Note that these depend on a feasible first stage decision $x \in \mathcal{X}$ and a second-stage decision y. These sensitivity indices are problem-dependent and are used to characterize scenarios. For example, in the paper [45], this method is applied to a stochastic unit commitment problem and the three sensitivity indices are used: the total cumulative generation cost, the costs associated to a shortfall and excess of power supply.

The following outline of the FSRC algorithm was taken directly from [45] with minor adaptations.

Algorithm: Forward Selection in Recourse Clusters

1. *Evaluate:* For each $s \in S$, identify an optimal y_s^*, given a feasible $\hat{x}$, by solving the recourse problem:

$$Q(\hat{x}, \xi_s) = \min_{y_s} \{q^T y_s | W y_s = h_s - T_s \hat{x}\}$$

2. *Summarize:* Compute solution sensitivity indices:

$$\mathcal{N}_s := [\mathcal{F}_1(\hat{x}, y_s^*), \ldots, \mathcal{F}_v(\hat{x}, y_s^*)]$$

3. *Cluster:* Scale $\mathcal{F}_i$, $i = 1, \ldots, v$, into similar magnitudes, denoted as $\hat{\mathcal{F}}_i$, $i = 1, \ldots, v$. Assign weight w_i to each $\hat{\mathcal{F}}_i$, and then set

$$V^s = [w_1 \hat{\mathcal{F}}_1(\hat{x}, y_s^*), \ldots, w_v \hat{\mathcal{F}}_v(\hat{x}, y_s^*)].$$

   Form n clusters on $\{V_s\}_{s \in S}$ by the k-means method using an appropriate norm, and create the corresponding n clusters in S.

4. *Select:* Use FFS to select one scenario from each cluster of the original scenarios.

Like the importance sampling method, this method is somewhat expensive as one has to evaluate the recourse function for every scenario. However, this step can be easily parallelized. The biggest obstacle in applying this method is the selection of sensitivity indices, and the relative weights of each of these; these have to be customized to the problem one is solving.

The paper [46] presents a similar method of scenario reduction where again the measure of similarity between scenarios has again been modified to take into account the behavior of the loss function.

Thesis Summary

This thesis concerns the development of new methods of problem-driven scenario generation. Unlike the problem-driven methods discussed in Section 5.5 which are somewhat heuristic, our methods are mathematically adapted to specific classes of problem. They are also largely constraint-driven, that is, the more the problem is constrained, the more effective our methods. The performance of the methods can therefore be improved by the addition of "ghost" constraints to a problem, that is, artificial constraints which reduce the set of feasible solutions but which do not affect the set of optimal solutions.

The first two papers of this thesis concern stochastic programs with tail-risk measures: the first paper provides the general mathematical foundations for our methodology, and the second paper concerns the practical application of the theory to portfolio selection problems. The third and final paper of this thesis describes an approach to scenario generation which exploits a special type of decomposition of the loss function, and is in particular demonstrated on simple recourse problems. The contents of each paper are summarized below.

Paper A: Scenario generation for stochastic programs with tail risk measures Tail risk measures such as Value-at-Risk and Conditional Value-at-Risk are used in stochastic programming to mitigate or reduce the probability of large losses. However, these are problematic in stochastic programs.

Because the value of a tail risk measure only depends on a small subset of the support of the distribution of asset returns, traditional scenario generation methods, which spread scenarios evenly across the whole support of the distribution, yield very unstable solutions unless we use a very large number scenarios.

In this paper we propose a scenario generation methodology for stochastic programs which uses tail risk measure. In this methodology we identify a region in the support of the distribution, which we call the *risk region*, in which all outcomes lead to a loss in tail for *some* feasible decision. We demonstrate that under mild conditions, the distribution outside the risk region can be represented with a single point while preserving the value of any tail-risk measure. This approach can thus reduce considerably the size of the resulting scenario-based problem. We propose a simple sampling algorithm which takes advantage of this idea and prove that it is asymptotically consistent with sampling.

The characterization of the risk region is difficult in general as it depends on the loss function, problem constraints, and probability distribution of the stochastic parameters. In this paper, we demonstrate this approach for portfolio selection problems where the returns of the assets have elliptical distributions where we are able to give convenient characterization of the risk region.

Paper B: Scenario generation for portfolio selection with tail risk measure
In this paper we develop further the application of our risk region methodology to the portfolio selection problem. Several issues are addressed: we provide additional details on the computations required to test whether or not a point lies is the the risk region; we investigate empirically under what circumstances the methodology performs well; we investigate the possibility of approximating the risk regions of non-elliptical distributions; and finally we investigate the use of "ghost constraints". Ghost constraints are constraints

added to the problem which do not affect the set of optimal solutions to a problem, but which improve the performance of our methodology by reducing the size of the risk region. The conclusions of this paper are founded on a variety of numerical tests, the distributions of which are constructed from real-financial data.

Paper C: Scenario generation for simple recourse problems In this paper, we present a general constraint-driven approach to scenario generation which exploits a special type of decomposition of the loss function to partition the support of the distribution into active and inactive components. The inactive components can typically be represented by a single scenario, which reduces the computational burden of solving the problem. Like with risk regions for stochastic programs with tail risk measure, the partition of the support into active and non-active depends on the form of the loss function and problem constraints. However, unlike the former approach it does not depend of the underlying probability distribution which simplifies the application of this approach. We demonstrate this method for simple recourse problems, a class of stochastic programs which aims to minimize the deviation between the availability of a set of resources and their corresponding stochastic demands.

References

[1] A. Prékopa, "Probabilistic programming," in *Stochastic Programming*, ser. Handbooks in Operations Research and Management Science, A. Ruszczyński and A. Shapiro, Eds. Amsterdam: Elsevier Science B.V., 2003, vol. 10, ch. 5, pp. 267–351.

[2] R. Gollmer, M. N. andW. Römisch, and R. Schultz, "Unit commitment in power generation—a basic model and some extensions," *Annals of Operations Research*, vol. 96, pp. 167–189, 2000.

[3] J. R. Birge and F. Louveaux, *Introduction to stochastic programming*. New York: Springer-Verlag, 1997.

[4] R. Van Slyke and R. J.-B. Wets, "L-shaped linear programs with applications to optimal control and stochastic programming," *SIAM Journal of Applied Mathematics*, vol. 17, pp. 638–663, 1969.

[5] R. J.-B. Wets, "Solving stochastic programs with simple recourse," *Journal of Stochastics*, vol. 10, pp. 219–242, 1983.

[6] G. Hadley and T. Whitin, *Analysis of inventory systems*, ser. Prentice-Hall international series in management. Prentice-Hall, 1963. [Online]. Available: https://books.google.co.uk/books?id=TrQ-AAAAIAAJ

[7] R. T. Rockafellar and S. Uryasev, "The fundamental risk quadrangle in risk management, optimization and statistical estimation," *Surveys in Operations Research and Management Science*, vol. 18, no. 1–2, pp. 33–53, 2013.

[8] H. Markowitz, "Portfolio selection," *Journal of Finance*, vol. 7, pp. 77–91, 1952.

[9] M. R. Young, "A minimax portfolio selection rule with linear programming solution," *Management Science*, vol. 44, no. 5, pp. 673–683,

1998. [Online]. Available: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.44.5.673

[10] H. Markowitz, *Portfolio selection: Efficient diversification of investment*. New Haven: Yale University Press, 1959.

[11] R. Dembo and D. Rosen, "The practice of portfolio replication. A practical overview of forward and inverse problems," *Annals of Operations Research*, vol. 85, pp. 267–284, 1999.

[12] P. Artzner, F. Delbaen, J. Eber, and D. Heath, "Coherent measures of risk," *Mathematical Finance*, vol. 9, no. 3, pp. 203–228, 1999.

[13] P. Jorion, *Value at Risk: The New Benchmark for Controlling Market Risk*. Irwin Professional, 1996.

[14] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *The Journal of Risk*, vol. 2, no. 3, pp. 21–41, 2000.

[15] D. Tasche, "Expected shortfall and beyond," *Journal of Banking & Finance*, vol. 26, no. 7, pp. 1519–1533, 2002.

[16] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *Journal of Banking & Finance*, vol. 26, no. 7, pp. 1443–1471, 2002.

[17] C. Acerbi and D. Tasche, "On the coherence of expected shortfall," *Journal of Banking & Finance*, vol. 26, no. 7, pp. 1487–1503, 2002.

[18] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.

[19] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, ser. MPS-SIAM Series on Optimization. Philadelphia: SIAM, 2009, vol. 9.

[20] W. Mak, D. Morton, and R. Wood, "Monte Carlo bounding techniques for determining solution quality in stochastic programs," *Operations Research Letters*, vol. 24, pp. 47–56, 1999.

[21] A. J. King and R. J.-B. Wets, "Epi-consistency of convex stochastic programs," *Stochastics and Stochastic Reports*, vol. 34, no. 1-2, pp. 83–92, 1991.

[22] A. Shapiro, "Asymptotic analysis of stochastic programs," *Annals of Operations Research*, vol. 30, no. 1, pp. 169–186, 1991.

[23] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 479–502, 2001.

[24] G. Bayraksan and D. P. Morton, "Assessing solution quality in stochastic programs," *Mathematical Programming*, vol. 108, no. 2–3, pp. 495–514, sep 2006.

[25] R. Stockbridge and G. Bayraksan, "A probability metrics approach for reducing the bias of optimality gap estimators in two-stage stochastic linear programming," *Mathematical Programming*, vol. 142, no. 1–2, pp. 107–131, 2013.

[26] J. L. Higle, "Variance reduction and objective function evaluation in stochastic linear programs," *INFORMS Journal on Computing*, vol. 10, no. 2, pp. 236–247, 1998.

[27] J. Linderoth, A. Shapiro, and S. Wright, "The empirical behavior of sampling methods for stochastic programming," *Annals of Operations Research*, vol. 142, no. 1, pp. 215–241, 2006.

[28] R. Henrion, C. Küchler, and W. Römisch, "Scenario reduction in stochastic programming with respect to discrepancy distances," *Computational Optimization and Applications*, vol. 43, no. 1, pp. 67–93, 2009.

[29] G. C. Pflug, "Scenario tree generation for multiperiod financial optimization by optimal discretization," *Mathematical Programming*, vol. 89, no. 2, pp. 251–271, 2001.

[30] C. Villani, *Topics in Optimal Transportation*, ser. Graduate studies in mathematics. American Mathematical Society, 2003. [Online]. Available: https://books.google.be/books?id=GqRXYFxe0l0C

[31] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1. Oakland, CA, USA., 1967, pp. 281–297.

[32] J. Dupačová, N. Gröwe-Kuska, and W. Römisch, "Scenario reduction in stochastic programming: An approach using probability metrics," *Mathematical Programming*, vol. 95, no. 3, pp. 493–511, 2003.

[33] G. C. Pflug and A. Pichler, "A distance for multistage stochastic optimization models," *SIAM Journal on Optimization*, vol. 22, no. 1, pp. 1–23, 2012.

[34] K. Høyland and S. W. Wallace, "Generating scenario trees for multistage decision problems," *Management Science*, vol. 47, no. 2, pp. 295–307, 2001.

[35] M. Kaut and S. W. Wallace, "Evaluation of scenario-generation methods for stochastic programming," *Pacific Journal of Optimization*, vol. 3, no. 2, pp. 257–271, 2007.

[36] K. Høyland, M. Kaut, and S. W. Wallace, "A heuristic for moment-matching scenario generation," *Computational Optimization and Applications*, vol. 24, no. 2–3, pp. 169–185, 2003.

[37] M. Kaut and S. W. Wallace, "Multi-period scenario tree generation using moment-matching: Example from option pricing," Apr 2003, available from http://michalkaut.net.

[38] H. Vaagen and S. W. Wallace, "Product variety arising from hedging in the fashion supply chains," *International Journal of Production Economics*, vol. 114, no. 2, pp. 431–455, 2008.

[39] M. Kaut and A.-G. Lium, "Scenario generation with distribution functions and correlations," *Kybernetika*, vol. 50, no. 6, pp. 1049–1064, 2014.

[40] K. Ponomareva, D. Roman, and P. Date, "An algorithm for moment-matching scenario generation with application to financial portfolio optimization," *European Journal of Operational Research*, vol. In press, 2014.

[41] B. Calfa, A. Agarwal, I. Grossmann, and J. Wassick, "Data-driven multi-stage scenario tree generation via statistical property and distribution matching," *Computers & Chemical Engineering*, vol. 68, pp. 7–23, 2014.

[42] M. Kaut and S. W. Wallace, "Shape-based scenario generation using copulas," *Computational Management Science*, vol. 8, no. 1–2, pp. 181–199, 2011.

[43] G. B. Dantzig and P. W. Glynn, "Parallel processors for planning under uncertainty," *Annals of Operations Research*, vol. 22, no. 1, pp. 1–21, 1990.

[44] G. Infanger, "Monte carlo (importance) sampling within a benders decomposition algorithm for stochastic linear programs," *Annals of Operations Research*, vol. 39, no. 1, pp. 69–95, 1992.

[45] Y. Feng and S. M. Ryan, "Solution sensitivity-based scenario reduction for stochastic unit commitment," *Computational Management Science*, pp. 1–34, 2014. [Online]. Available: http://dx.doi.org/10.1007/s10287-014-0220-z

[46] Z. Li and C. A. Floudas, "Optimal scenario reduction framework based on distance of uncertainty distribution and output performance: I. Single reduction via mixed integer linear optimization," *Computers & Chemical Engineering*, vol. in press, 2014.

Part II

Papers

Paper A

Scenario Generation for Stochastic Programs with Tail Risk Measures

Jamie Fairbrother, Amanda Turner, Stein W. Wallace

1 Introduction

Stochastic programming is a tool for making decisions under uncertainty. Stochastic programs are used to model situations where an initial decision must be taken with some information unknown until after the decision has been made. For example, one may want to know how much to invest in a new production technology without knowing exactly the future demand for the product. In stochastic programming, uncertain parameters are modeled as random variables, and one attempts to minimize the expectation or risk measure of some loss function which depends on the initial decision. However, what distinguishes stochastic programming from other stochastic modeling approaches is the ability to explicitly model future decisions based on outcomes of stochastic parameters and initial decisions, and the associated costs of these future decisions. In our example, given an investment decision and a demand, we could model how to distribute this product and the costs of this distribution. The power and flexibility of the stochastic programming approach comes at a price: stochastic programs are usually analytically intractable, and not susceptible to deterministic optimization techniques. See [1] for a guide to how stochastic programs are used to model real problems, and [2], [3] for more general overviews of the subject.

Typically, a stochastic program can only be solved when it is *scenario-based*, that is when the random variables of the problem have finite discrete distributions. For example, stochastic linear programs just become linear programs when the underlying random variables are discrete. In the stochastic programming literature, the mass points of these random variables are referred to as *scenarios*, the discrete distribution as the *scenario set* and the construction of this as *scenario generation*. Scenario generation can consist of discretizing a continuous probability distribution, or directly modeling the uncertain quantities as discrete random variables. The more scenarios in a set, the more computational power that is required to solve the problem. The key issue

of scenario generation is how to represent the uncertainty to ensure that the solution to the problem is reliable, while keeping the number of scenarios low so that the problem is computational tractable. See [4] for methods of evaluating scenario generation methods and a discussion of what constitutes a reliable solution.

Minimizing the expectation of a loss function can be thought of as minimizing the long-term costs of a system. This is appropriate when the initial decision is going to be used again and again, and large losses do not matter in the short term. For example, a news vendor may have to decide on a daily order of items to which they are committed for some period of time. In other cases, the decision may be only used a few times, and the occurrence of large losses may be lead to bankruptcy. In this latter case, minimizing the expectation alone is not appropriate as this does not necessarily mitigate against large losses. The usual action of recourse in this case is to use some sort of *risk measure* which quantifies in some way the likelihood and severity of potential large losses. In these problems we try to find a decision which appropriately balances in the expectation against risk.

In this paper we are interested in problems which use *tail risk measures*. A precise definition of a tail-risk measure will be given in Section 2 but for now, one can think of a tail risk measure as a function of a random variable which only depends on the upper tail of its distribution function. Examples of tail risk measure include the Value-at-Risk [5] and the Conditional Value-at-Risk [6], both of which are commonly used in financial contexts. The problem of scenario generation is particularly acute when the scenarios are being used to calculate the value of a tail risk measure. This is because standard scenario generation methods will not produce many scenarios in the tail of the loss function and so it is inadequately represented.

The most basic approach to discretization is to simply use a random sample from the true distribution. This has desirable asymptotic properties [7], [8], but may require large sample sizes to ensure the reliability of

the solutions it yields. This can be mitigated somewhat by using variance reduction techniques such as stratified sampling and importance sampling [9]. Sampling also has the advantage that it can be used to construct confidence intervals on the true solution value [10]. Another approach to discretization is to construct a distribution whose distance from the true distribution, with respect to some probability metric, is small [11], [12]. These approaches tend to yield better and much more stable solutions to stochastic programs than does sampling.

A characteristic of both of these approaches to scenario generation is that they are *distribution-based*; that is, they only aim to approximate a distribution and are divorced from the stochastic program for which they are producing scenarios. By exploiting the structure of a problem, it may be possible to find a more parsimonious representation of the uncertainty. Note that such a *problem-based* approach may not yield a discrete distribution which is close to the true distribution in a probabilistic sense; the aim is only to find a discrete distribution which yields a high quality solution to our problem.

A set of approaches which move away from the purely distribution-based paradigm of scenario generation are *constructive methods*. In these approaches, the modeler does not use a full probability distribution for the uncertain problem parameters but specifies a set of target statistical properties they believe the distribution satisfies, and generates a scenario set with these target properties. This approach was first proposed in [13], where it is postulated that the solution to a stochastic program will depend largely on a small set of statistical properties of the random variables, specific to that problem. That is, if we can generate a scenario set with the required properties, this should yield good solutions in our stochastic program even if the true distribution is significantly different. For example, it is known that for the classical Markowitz problem [14] the first two moments of the return distributions determine exactly the solution. Constructive approaches have gained much popularity because they simplify the stochastic modeling of the uncertain

parameters. In particular they eliminate the need to fit parametric stochastic models. Other constructive approaches can be found in [15], and [16]. However, the major draw-back with constructive approaches is that it is not always clear which properties are important for a given problem. Finding out which properties are important is therefore an important part of the analysis.

In this paper, we present a general problem-based approach to scenario generation for stochastic programs which use tail risk measures. We observe that the value of any tail risk measure depends only on scenarios confined to an area that we call the *risk region*. This means that all scenarios not in the risk region can be aggregated into a single point. By concentrating almost all scenarios in the risk region, we can calculate the value a tail risk measure more accurately. One feature of the risk region is that the more constrained our problem, the smaller it becomes, and so the more useful our methodology. However, finding the risk region is difficult as it is determined by both the problem and the distribution of the uncertain parameters.

We demonstrate our methodology for portfolio selection problems where the assets are assumed to have returns which are elliptically distributed. For this type of problem we are able to characterize the risk region in a convenient way. We will show that the risk region depends only on the conic hull of our feasible region. Another useful property of the portfolio selection problem is the linearity (affinity) of the loss function. This means that all scenarios not in the risk region can be aggregated while preserving the overall expected return.

Some ideas in this paper are similar to those in [17]. In that paper, the authors, like us, observe that only scenarios which have a loss in the tail of the distribution are used in the calculation of the tail risk measure. However, while we use this observation to construct a scenario set, they exploit this property to solve a problem which uses the β-CVaR risk measure for a given scenario set. Their approach is to iteratively solve the problem with a subset of scenarios, identify the scenarios which have loss in the tail, update their

scenario set appropriately and resolve, until the true solution has been found.

This paper is organized as follows: in Section 2 we define tail risk measures and their associated risk regions; in Section 3 we discuss how these risk regions can be exploited for the purposes of scenario generation and scenario reduction; in Section 4 we prove that our scenario generation method is consistent with sampling, in Section 5 and Section 6 we provide a proof of concept for our methodology: we give convenient characterizations for risk regions for a class of portfolio selection problems and present numerical tests which compare our methodology against basic sampling; finally in Section 7 we summarize our results make some concluding remarks.

2 Tail risk measures and risk regions

In this section we define the core concepts related to our scenario generation methodology, and prove some results relating to these. Specifically, in Section 2.1 we formally define tail-risk measures of random variables and in Section 2.2 we define risk regions and present some key results related to these.

2.1 Tail risk of random variables

Suppose that we have an uncertain quantity representing some loss, and we would like to somehow quantify the riskiness of this quantity. We model the uncertain quantity as a random variable and take a risk measure to be any function of a random variable. The following definition is taken from [18].

Definition 2.1 (Risk Measure)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space, and* V *be a non-empty set of* $\mathcal{F}$*-measurable real-valued random variables*[1]*. Then, a risk measure is some function* $\rho : V \to \mathbb{R} \cup \{\infty\}$.

[1]We implicitly assume throughout that V is large enough to contain all constructed random variables

However, for a risk measure to be useful, it should in some way penalize potential large losses. For example, in the classical Markowitz problem [14], the uncertain quantity is the return of a portfolio of financial assets, and the measure of risk is the variance of that return. By choosing a portfolio with a low variance, we reduce the probability of larges losses as a direct consequence of Chebyshev's inequality (see for instance [19]). Various criteria for risk measures have been proposed; in [20] a *coherent risk measure* is defined to be a risk measure which satisfies axioms such as positive homogeneity and subadditivity; another perhaps desirable criterion for risk measures is that the risk measure is consistent with respect to first and second order stochastic dominance, see [21] for instance.

Besides not satisfying some of the above criteria, a major drawback with using variance as a measure is that it penalizes all large deviations from the mean, that is, it penalizes large profits as well as large losses. This problem can be overcome by using *downside* risk measures such as the semi-variance, which only penalize losses above the mean. However, if we are truly interested in rare or extreme losses, using a risk measure which still depends on the main body of the distribution such as semi-variance may give us distorted or over-optimistic results.

These considerations motivate the idea of using risk measures which depend only on the upper tail of the distribution. To be more precise, the upper tail of a distribution consists of outcomes with a loss greater than or equal to some quantile of the underlying distribution function.

Definition 2.2 (Quantile Function)**.** *Suppose Z is a random variable with distribution function F_Z. Then the generalized inverse distribution function, or* quantile function *is defined as follows:*

$$\begin{aligned} F_Z^{-1} : (0,1] &\to \mathbb{R} \cup \{\infty\} \\ \beta &\mapsto \inf\{x \in \mathbb{R} : F_Z(x) \geq \beta\}. \end{aligned}$$

Definition 2.3 (Tail Risk Measure)**.** *Let $\rho_\beta : V \to \mathbb{R} \cup \{\infty\}$ be a risk measure*

as above, then ρ_β is a β-tail risk measure if $\rho_\beta(Z)$ depends only on the restriction of quantile function of Z above β, that is $F_Z^{-1}|_{[\beta,1]}$.

To show that ρ_β is a β-tail risk measure, we must show that $\rho_\beta(Z)$ can be written as a function of the quantile function above or equal to β. Two very popular tail risk measures are the value-at-risk [5] and the conditional value-at-risk [22]:

Example 2.4 (Value at risk)**.** *Let Z be a random variable, and $0 < \beta < 1$. Then, the $\beta-$VaR for Z is defined to be the β-quantile of Z:*

$$\beta\text{-VaR}(Z) := F_Z^{-1}(\beta).$$

Example 2.5 (Conditional value at risk)**.** *Let Z be a random variable, and $0 < \beta < 1$. Then, the β-CVaR can be thought roughly as the conditional expectation of a random variable above its β-quantile. The following alternative characterization of β-CVaR [23] shows directly that it is a β-tail risk measure.*

$$\beta\text{-CVaR}(Z) = \int_\beta^1 F_Z^{-1}(u)\, du.$$

The observation that we exploit for this work is that very different random variables will have the same β-tail risk measure as long as their β-tails are the same. Such a situation is illustrated in Figure B.1 for two discrete random variables.

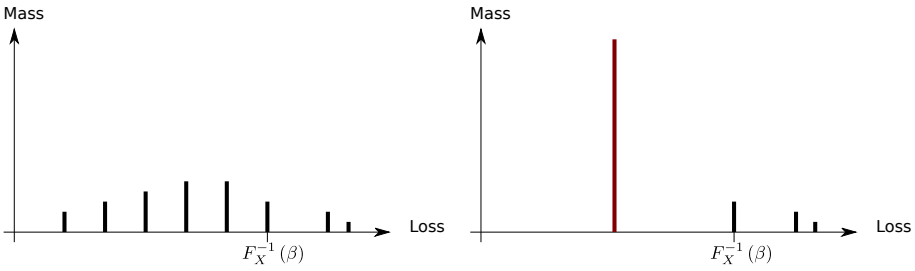


Fig. A.1: Two very different random variables with identical β-tails

When showing that two distributions have the same β-tails, it is convenient to use distribution functions rather than quantile functions. An equiv-

alent condition for showing that two random variables Z_1 and Z_2 have the same β-tail, that is $F_{Z_1}^{-1}(u) = F_{Z_2}^{-1}(u)$ for all $\beta \leq u \leq 1$, is the following:

$$F_{Z_1}^{-1}(\beta) = F_{Z_2}^{-1}(\beta) \text{ and } F_{Z_1}(z) = F_{Z_2}(z) \text{ for all } z \geq F_{Z_1}^{-1}(\beta). \tag{A.1}$$

2.2 Risk regions

In the optimization context we suppose that the loss depends on some decision $x \in \mathcal{X} \subseteq \mathbb{R}^k$ and the outcome of some latent random vector Y with support $\mathcal{Y} \subseteq \mathbb{R}^d$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and which is independent of x. That is, we suppose our loss is determined by some function, $f : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}$, which we refer to as the *loss function*. For a given decision $x \in \mathcal{X}$, the random variable associated with the loss is thus $f(x, Y)$. We are typically interested in optimization problems where the aim is to find some decision $x \in \mathcal{X}$ which in some way minimizes both the expected loss $\mathbb{E}[f(x, Y)]$ and the value of some β-tail risk measure $\rho_\beta(f(x, Y))$. See Section 5.1 for some explicit formulations of such an optimization problem.

To avoid repeated use of cumbersome notation we introduce the following short-hand for distribution and quantile functions:

$$\begin{aligned} F_x(z) &:= F_{f(x,Y)}(z) = \mathbb{P}(f(x, Y) \leq z), \\ F_x^{-1}(\beta) &:= F_{f(x,Y)}^{-1}(\beta) = \inf\{z \in \mathbb{R} : \ F_x(z) \geq \beta\}. \end{aligned}$$

Since tail risk measures depend only on those outcomes which are in the β-tail, we aim to identify the region of the support which lead to a loss in the β-tails for some decision.

Definition 2.6 (Risk region)**.** *For $0 < \beta < 1$ the β-risk region with respect to the decision $x \in \mathcal{X}$ is defined as follows:*

$$\mathcal{R}_x(\beta) = \{y \in \mathbb{R}^d : F_x(f(x, y)) \geq \beta\},$$

or equivalently

$$\mathcal{R}_x(\beta) = \{y \in \mathbb{R}^d : f(x, y) \geq F_x^{-1}(\beta)\}. \tag{A.2}$$

The risk region with respect to the feasible region $\mathcal{X} \subset \mathbb{R}^k$ *is defined to be:*

$$\mathcal{R}_{\mathcal{X}}(\beta) = \bigcup_{x \in \mathcal{X}} \mathcal{R}_x(\beta).$$

The complement of this region is called the non-risk region*. This can also be written*

$$\mathcal{R}_{\mathcal{X}}(\beta)^c = \bigcap_{x \in \mathcal{X}} \mathcal{R}_x(\beta)^c. \tag{A.3}$$

The definition above says that the risk region $\mathcal{R}_x$ consists of all points $y \in \mathbb{R}^d$ which lead to a loss in the β-tail for the decision $x \in \mathcal{X}$. The risk region $\mathcal{R}_{\mathcal{X}}$ is the set of all points $y \in \mathbb{R}^d$ which can lead to a loss in the β-tail for *any* decision $x \in \mathcal{X}$.

The following basic properties of the risk region follow directly from the definition.

(i) $0 < \beta' < \beta < 1 \;\Rightarrow\; \mathcal{R}_{\mathcal{X}}(\beta) \subseteq \mathcal{R}_{\mathcal{X}}(\beta')$; (A.4)

(ii) $\mathcal{X}' \subset \mathcal{X} \;\Rightarrow \mathcal{R}_{\mathcal{X}'}(\beta) \subseteq \mathcal{R}_{\mathcal{X}}(\beta)$; (A.5)

(iii) If $y \mapsto f(x, y)$ is continuous then $\mathcal{R}_x(\beta)$ is closed and $\mathcal{R}_x(\beta)^c$ is open. (A.6)

We now state a technical property and prove that this ensures the distribution of the random vector in a given region completely determines the value of a tail risk measure. In essence, this condition ensures that there is enough mass in the set to ensure that the β-quantile does not depend on the probability distribution outside of it.

Definition 2.7 (Aggregation condition)**.** *Suppose that* $\mathcal{R}_{\mathcal{X}}(\beta) \subseteq \mathcal{R} \subset \mathbb{R}^d$ *and that for all* $x \in \mathcal{X}$*,* $\mathcal{R}$ *satisfies the following condition:*

$$\mathbb{P}\left(Y \in \{y : z' < f(x, y) \leq F_x^{-1}(\beta)\} \cap \mathcal{R}\right) > 0 \qquad \forall\, z' < F_x^{-1}(\beta). \tag{A.7}$$

Then $\mathcal{R}$ *is said to satisfy the* β-aggregation condition*.*

The motivation for the term *aggregation condition* comes from Theorem 2.8 which follows. This result ensures that if a set satisfies the aggregation condition then we can transform the probability distribution of Y so that all the mass in the complement of this set is aggregated into a single point without affecting the value of the tail risk measure. This property is particularly relevant to scenario generation as if we have such a set, then all scenarios which it does not contain can be aggregated, reducing the size of the stochastic program.

Theorem 2.8. *Suppose that $\mathcal{R}_{\mathcal{X}}(\beta) \subseteq \mathcal{R} \subset \mathbb{R}^d$ satisfies the β-aggregation condition and that $\tilde{Y}$ is a random vector for which*

$$\mathbb{P}\left(Y \in \mathcal{A}\right) = \mathbb{P}\left(\tilde{Y} \in \mathcal{A}\right) \qquad \text{for any measurable } \mathcal{A} \subseteq \mathcal{R}. \tag{A.8}$$

Then for any tail risk measure ρ_β we have $\rho_\beta\left(f(x,Y)\right) = \rho_\beta\left(f(x,\tilde{Y})\right)$ for all $x \in \mathcal{X}$.

Proof. Fix $x \in \mathcal{X}$. To show that $\rho_\beta\left(f(x,Y)\right) = \rho_\beta\left(f(x,\tilde{Y})\right)$ we must show that the β-quantile and the β-tail distributions of $f(x,Y)$ and $f(x,\tilde{Y})$ are the same. The following two conditions are necessary and sufficient for this to occur:

$$\begin{aligned} F_x(z) = F_{f(x,\tilde{Y})}(z) \qquad & \forall\, z \geq F_x^{-1}\left(\beta\right), \\ F_{f(x,\tilde{Y})}(z) < \beta \qquad & \forall z < F_x^{-1}\left(\beta\right). \end{aligned}$$

Suppose $z' \geq F_x^{-1}(\beta)$. First note as a direct consequence of (A.8) we have

$$\mathbb{P}\left(Y \in \mathcal{B}\right) = \mathbb{P}\left(\tilde{Y} \in \mathcal{B}\right) \qquad \text{for any } \mathcal{B} \supseteq \mathcal{R}^c. \tag{A.9}$$

Now,

$$\begin{aligned}
F_{f(x,\tilde{Y})}(z') &= \mathbb{P}\left(\tilde{Y} \in \{y:\ f(x,y) \le z'\}\right) \\
&= \mathbb{P}\left(\tilde{Y} \in \underbrace{\mathcal{R}^c \cap \{y:\ f(x,y) \le z'\}}_{=\mathcal{R}^c}\right) + \mathbb{P}\left(\tilde{Y} \in \underbrace{\mathcal{R} \cap \{y:\ f(x,Y) \le z'\}}_{\subset \mathcal{R}}\right) \\
&= \mathbb{P}\left(Y \in \mathcal{R}^c\right) + \mathbb{P}\left(Y \in \mathcal{R} \cap \{y : f(x,y) \le z'\}\right) \qquad \text{by (A.8) and (A.9)} \\
&= \mathbb{P}\left(Y \in \{y : f(x,y) \le z'\}\right) \\
&= F_x(z')
\end{aligned}$$

as required.

Now suppose $z' < F_x^{-1}(\beta)$. There are two cases; in the first instance suppose $\mathbb{P}\left(f(x,Y) = F_x^{-1}(\beta)\right) > 0$, then we have:

$$\begin{aligned}
F_{f(x,\tilde{Y})}(z') &\le \mathbb{P}\left(f(x,\tilde{Y}) < F_x^{-1}(\beta)\right) \\
&= \mathbb{P}\left(f(x,Y) < F_x^{-1}(\beta)\right) \\
&< \beta,
\end{aligned}$$

as required. In the case where $\mathbb{P}\left(f(x,Y) = F_x^{-1}(\beta)\right) = 0$ we have:

$$\begin{aligned}
F_{f(x,\tilde{Y})}(z') &= \mathbb{P}\left(\tilde{Y} \in \{y : f(x,y) \le z'\}\right) \\
&\le \mathbb{P}\left(\tilde{Y} \in \mathcal{R}^c \cup \{y : f(x,y) \le z'\}\right) \\
&= \mathbb{P}\left(\tilde{Y} \in \underbrace{\{y : f(x,y) \le F_x^{-1}(\beta)\}}_{\supseteq \mathcal{R}^c}\right) - \mathbb{P}\left(\tilde{Y} \in \underbrace{\mathcal{R} \cap \{y : z' < f(x,y) \le F_x^{-1}(\beta)\}}_{\subseteq \mathcal{R}}\right) \\
&= \mathbb{P}\left(Y \in \{y : f(x,y) \le F_x^{-1}(\beta)\}\right) - \mathbb{P}\left(Y \in \mathcal{R} \cap \{y : z' < f(x,y) \le F_x^{-1}(\beta)\}\right) \\
&< \mathbb{P}\left(Y \in \{y : f(x,y) \le F_x^{-1}(\beta)\}\right) \qquad \text{by (A.7)} \\
&= \beta \qquad \text{since } \mathbb{P}\left(f(x,Y) = F_x^{-1}(\beta)\right) > 0
\end{aligned}$$

as required. □

The β-aggregation condition is difficult to verify directly. The following shows that it immediately holds for $\mathcal{R}_{\mathcal{X}}(\beta')$ when $\beta' < \beta$.

Proposition 2.9. *Suppose $\beta' < \beta$. Then, $\mathcal{R}_{\mathcal{X}}(\beta')$ satisfies the β-aggregation condition. That is for all $x \in \mathcal{X}$*

$$\mathbb{P}\left(Y \in \{y : z' \leq f(x,y) \leq F_x^{-1}(\beta)\} \cap \mathcal{R}_{\mathcal{X}}\left(\beta'\right)\right) > 0 \qquad \forall\, z' < F_x^{-1}(\beta).$$

Proof. Fix $x \in \mathcal{X}$.

Case 1: $F_x^{-1}(\beta') = F_x^{-1}(\beta)$.

In this case, the distribution function F_x has a discontinuity at $z = F_x^{-1}(\beta)$, that is $\mathbb{P}(f(x,Y) = z) > 0$. Therefore, for $z' < z$ we have

$$\begin{aligned}\mathbb{P}\left(Y \in \{y : z' \leq f(x,y) \leq F_x^{-1}(\beta)\} \cap \mathcal{R}_{\mathcal{X}}\left(\beta'\right)\right) &\geq \mathbb{P}(f(x,Y) = z)\\ &> 0\end{aligned}$$

as required.

Case 2: $F_x^{-1}(\beta') < F_x^{-1}(\beta)$.

In this case for all $F_x^{-1}(\beta') < z' < F_x^{-1}(\beta)$, we have $\{y : z' < f(x,y) \leq F_x^{-1}(\beta)\} \subset \mathcal{R}_{\mathcal{X}}(\beta')$ and so

$$\begin{aligned}\mathbb{P}\left(Y \in \{y : z' \leq f(x,y) \leq F_x^{-1}(\beta)\} \cap \mathcal{R}_{\mathcal{X}}\left(\beta'\right)\right) &= \mathbb{P}\left(z' \leq f(x,Y) \leq F_x^{-1}(\beta)\right)\\ &> 0.\end{aligned}$$

□

For convenience, we now drop β from our notation and terminology. Thus, we refer to the β-risk region and β-aggregation condition as simply the risk region and aggregation condition respectively, and write $\mathcal{R}_{\mathcal{X}}(\beta)$ as $\mathcal{R}_{\mathcal{X}}$.

All sets satisfying the aggregation condition must contain the risk region, however, the aggregation condition does not necessarily hold for the risk region itself. It is guaranteed to hold if Y has a discrete distribution, since in this case for all $x \in \mathcal{X}$ and $z' < F_x^{-1}(\beta)$ we have:

$$\begin{aligned}\mathbb{P}\left(Y \in \{y : z' < f(x,y) \leq F_x^{-1}(\beta)\} \cap \mathcal{R}_{\mathcal{X}}\right) &\geq \mathbb{P}\left(f(x,Y) = F_x^{-1}(\beta)\right)\\ &> 0.\end{aligned}$$

In the non-discrete case we must impose extra conditions on the problem to avoid some degenerate cases. Recall that $\mathcal{Y}$ denotes the support of the random vector Y.

Proposition 2.10. *Suppose the following conditions hold:*

(i) $\operatorname{int}(\mathcal{Y}) \cap \operatorname{int}(\mathcal{R}_{\mathcal{X}})$ *is connected*

(ii) $y \mapsto f(x,y)$ *is continuous for all* $x \in \mathcal{X}$

(iii) For each $x \in \mathcal{X}$ *there exists* $x' \in \mathcal{X}$ *such that*

$$\operatorname{int}(\mathcal{Y}) \cap \operatorname{int}(\mathcal{R}_x \cap \mathcal{R}_{x'}) \neq \emptyset \text{ and } \operatorname{int}(\mathcal{Y}) \cap \operatorname{int}(\mathcal{R}_{x'} \setminus \mathcal{R}_x) \neq \emptyset \quad \text{(A.10)}$$

Then the risk region $\mathcal{R}_{\mathcal{X}}$ *satisfies the aggregation condition.*

Proof. Fix $x \in \mathcal{X}$ and $z' < F_x^{-1}(\beta)$. Pick $x' \in \mathcal{X}$ such that (A.10) holds. Also, let $y_0 \in \operatorname{int}(\mathcal{Y}) \cap \operatorname{int}(\mathcal{R}_{x'} \setminus \mathcal{R}_x)$ and $y_1 \in \operatorname{int}(\mathcal{Y}) \cap \operatorname{int}(\mathcal{R}_x \cap \mathcal{R}_{x'})$. Since $\operatorname{int}(\mathcal{Y}) \cap \operatorname{int}(\mathcal{R}_{\mathcal{X}})$ is connected there exists continuous path from y_0 to y_1. That is, there exists

$$\gamma : [0,1] \to \operatorname{int}(\mathcal{Y}) \cap \operatorname{int}(\mathcal{R}_{\mathcal{X}})$$

such that $\gamma(0) = y_0$ and $\gamma(1) = y_1$. Now, $f(x,y_0) < F_x^{-1}(\beta)$ and $f(x,y_1) \geq F_x^{-1}(\beta)$ and so given that $t \mapsto f(x,\gamma(t))$ is continuous there must exist $0 < t < 1$ such that $z' < f(x,\gamma(t)) < F_x^{-1}(\beta)$. That is,

$$\operatorname{int}(\mathcal{Y}) \cap \operatorname{int}(\mathcal{R}_{\mathcal{X}}) \cap \{y : z' < f(x,y) < F_x^{-1}(\beta)\}$$

is non-empty. This is a non-empty open set contained in the support of Y and so has positive probability, hence the aggregation condition holds. □

The following Proposition gives a condition under which the non-risk region is convex. This is useful as if we can find some points in the non-risk region, then the the convex hull of these points will be contained in the non-risk region, and the complement of this convex hull will thus contain the risk region.

Proposition 2.11. *Suppose that for each $x \in \mathcal{X}$ the function $y \mapsto f(x,y)$ is convex. Then the non-risk region $\mathcal{R}^c_{\mathcal{X}}$ is convex.*

Proof. For $x \in \mathcal{X}$, if $y \mapsto f(x,y)$ is convex then the set $\mathcal{R}^c_x = \{y \in \mathbb{R}^d : f(x,y) < F_x^{-1}(\beta)\}$ must be convex. The arbitrary intersection of convex sets is convex, hence $\mathcal{R}^c_{\mathcal{X}} = \bigcap_{x\in\mathcal{X}} \mathcal{R}^c_x$ is convex. □

This convexity condition is held by a large class of stochastic programs, for instance, all two-stage linear recourse problems with fixed recourse will have this property (see, for instance, [3]).

The random vector in the following definition plays a special role in our theory.

Definition 2.12 (Aggregated random vector)**.** *For some set $\mathcal{R} \subset \mathbb{R}^d$ satisfying the aggregation condition, the* aggregated random vector *is defined as follows:*

$$\psi_{\mathcal{R}}(Y) := \begin{cases} Y & \text{if } Y \in \mathcal{R}, \\ \mathbb{E}\left[\, Y | Y \in \mathcal{R}^c \,\right] & \text{otherwise.} \end{cases}$$

If we have $\mathbb{E}\left[Y | Y \in \mathcal{R}^c\right] \in \mathcal{R}^c$ then Theorem 2.8 guarantees that $\rho_\beta\left(f\left(x, \psi_{\mathcal{R}}(Y)\right)\right) = \rho_\beta\left(f\right.$ for all $x \in \mathcal{X}$. For example, the conditions of Proposition 2.11 will guarantee this. As well as preserving the value of the tail risk measure, the function $\psi_{\mathcal{R}}$ will preserve the expectation for affine cost functions.

Corollary 2.13. *Suppose for each $x \in \mathcal{X}$ the function $y \mapsto f(x,y)$ is affine and for a set $\mathcal{R} \subset \mathbb{R}^d$ satisfying the aggregation condition we have that*

$$\mathbb{E}\left[Y | Y \in \mathcal{R}^c\right] \in \mathcal{R}^c$$

Then,

$$\rho_\beta\left(f\left(x, \psi_{\mathcal{R}}(Y)\right)\right) = \rho_\beta\left(f\left(x, Y\right)\right), \tag{A.11}$$

$$\mathbb{E}\left[f\left(x, \psi_{\mathcal{R}^c}\left(Y\right)\right)\right] = \mathbb{E}\left[f(x, Y)\right], \tag{A.12}$$

for all $x \in \mathcal{X}$.

Proof. The equality (A.11) follows immediately from Theorem 2.8. For the expectation function we have

$$\begin{aligned}\mathbb{E}\left[\psi_{\mathcal{R}}(Y)\right] &= \mathbb{P}\left(Y \in \mathcal{R}\right)\mathbb{E}\left[\psi_{\mathcal{R}}(Y)|Y \in \mathcal{R}\right] + \mathbb{P}\left(Y \in \mathcal{R}^c\right)\mathbb{E}\left[\psi_{\mathcal{R}}(Y)|Y \in \mathcal{R}^c\right] \\ &= \mathbb{P}\left(Y \in \mathcal{R}\right)\mathbb{E}\left[Y|Y \in \mathcal{R}\right] + \mathbb{P}\left(Y \in \mathcal{R}^c\right)\mathbb{E}\left[Y|Y \in \mathcal{R}^c\right] \\ &= \mathbb{E}\left[Y\right].\end{aligned}$$

Since $y \mapsto f(x, y)$ is affine this means that

$$\begin{aligned}\mathbb{E}\left[f(x, \psi_{\mathcal{R}}(Y))\right] &= f(x, \mathbb{E}\left[\psi_{\mathcal{R}}(Y)\right] \\ &= f(x, \mathbb{E}\left[Y\right]) \\ &= \mathbb{E}\left[f(x, Y)\right].\end{aligned}$$

□

3 Scenario generation

In the previous section, we showed that under mild conditions the value of a tail risk measure only depends on the distribution of outcomes in the risk region. In this section we demonstrate how this feature may be exploited for the purposes of scenario generation and scenario reduction.

We assume throughout this section that our scenario sets are constructed from some underlying probabilistic model from which we can draw independent identically distributed samples. We also assume we have a set $\mathcal{R} \subset \mathbb{R}^d$ which satisfies the aggregation condition and for which we can easily test membership. In Section 5 we show such a convenient characterization is available for the risk region of the portfolio selection problem. However, in general finding such a set is difficult as the risk region depends both on the loss function and the distribution of the random vector Y.

Our general approach is as follows: for scenario generation we prioritize the construction of scenarios in the risk region to allow one to better approximate the value of the β-tail risk measure; for scenario reduction we reduce

the number of scenarios in the non-risk region which are in some sense redundant for computing the value of the β-tail risk measure.

In Section 3.1 we present and analyse two concrete approaches: aggregation sampling and aggregation reduction. In Section 3.2 we briefly discuss alternative ways of exploiting risk regions for scenario generation.

3.1 Aggregation sampling and reduction

In *aggregation sampling*, the user specifies a number of scenarios to be in the risk region. The algorithm then draws samples from the distribution, storing those samples which lie in the risk region and aggregating those in the non-risk region into a single point. In particular, the samples in the non-risk region are aggregated into their mean. The algorithm terminates when the specified number of risk scenarios has been reached. This is detailed in Algorithm 1. In *aggregation reduction* one draws a fixed number of samples from the distribution and then aggregates all those in the non-risk region.

Aggregation sampling and aggregation reduction can be thought of as equivalent to sampling from the aggregated random vector for large sample sizes. Therefore, aggregation sampling and aggregation reduction are consistent with sampling only if $\mathcal{R}$ satisfies the aggregation condition and $\mathbb{E}\left[Y|Y \in \mathcal{R}^c\right] \in \mathcal{R}^c$. For the precise conditions required for consistency and proofs of these see Theorem 4.4.

We now study the performance of our methodology. Let q the probability of the non-risk region, and n the desired number of risk scenarios. Let $N(n)$ denote the *effective sample size* for aggregation sampling, that is, the number of samples drawn until the algorithm terminates[2]. The aggregation sampling algorithm can be viewed as a sequence of Bernoulli trials where a trial is a success if the corresponding sample lies in the non-risk region, and which terminates once we have reached n failures, that is, once we have sampled n

[2]For simplicity of exposition we discount the event that the while loop of the algorithm terminates with $n_{\mathcal{R}^c} = 0$ which occurs with probability q^n

input : $\mathcal{R} \subset \mathbb{R}^d$ set satisfying aggregation condition, $N_{\mathcal{R}}$ number of required risk scenarios

output: $\{(y_s, p_s)\}_{s=1}^{N_{\mathcal{R}}+1}$ scenario set

$n_{\mathcal{R}^c} \leftarrow 0$, $n_{\mathcal{R}} \leftarrow 0$, $y_{\mathcal{R}^c} = \mathbf{0}$;

while $n_{\mathcal{R}} < N_{\mathcal{R}}$ **do**

  Sample new point y;

  if $y \in \mathcal{R}$ **then**

    $n_{\mathcal{R}} \leftarrow n_{\mathcal{R}} + 1$;

    $y_{n_{\mathcal{R}}} \leftarrow y$;

  end

  else

    $n_{\mathcal{R}^c} \leftarrow n_{\mathcal{R}^c} + 1$;

    $y_{\mathcal{R}^c} \leftarrow \frac{1}{n_{\mathcal{R}^c}+1} \left(n_{\mathcal{R}^c} y_{\mathcal{R}^c} + y\right)$

  end

end

foreach i *in* $1, \ldots, N_{\mathcal{R}}$ **do** $p_i \leftarrow \frac{1}{(n_{\mathcal{R}^c}+N_{\mathcal{R}})}$;

if $n_{\mathcal{R}^c} > 0$ **then**

  $p_{n_{\mathcal{R}^c}+1} \leftarrow \frac{n_{\mathcal{R}^c}}{n_{\mathcal{R}^c}+N_{\mathcal{R}}}$;

end

else

  Sample new point y;

  $n_{\mathcal{R}^c} \leftarrow 1$;

  $y_{N_{\mathcal{R}}+1} \leftarrow y$;

end

$p_{N_{\mathcal{R}}+1} \leftarrow \frac{n_{\mathcal{R}^c}}{n_{\mathcal{R}^c}+N_{\mathcal{R}}}$

Algorithm 1: Aggregation sampling

scenarios from the risk region. We can therefore write down the distribution of $N(n)$:

$$N(n) \sim n + \mathcal{NB}(n, q),$$

where $\mathcal{NB}(N, q)$ denotes a *negative binomial* random variable whose probability mass function is as follows:

$$\binom{k+n-1}{k}(1-q)^n q^k, \qquad k \geq 0.$$

The expected effective sample size of aggregation sampling is thus:

$$\mathbb{E}\left[N(n)\right] = n + n\frac{q}{1-q} \tag{A.13}$$

Let $R(n)$ denote the number of scenarios which are aggregated in the aggregation reduction method. Aggregation reduction can similarly be viewed as a sequence of n Bernoulli trials, where success and failure are defined in the same way as described above. The number of aggregated scenarios in aggregation reduction is therefore distributed as follows:

$$R(n) \sim \mathcal{B}(n, q)$$

where $\mathcal{B}(n, q)$ denotes a binomial random variable and so we have

$$\mathbb{E}\left[R(n)\right] = nq. \tag{A.14}$$

From (A.13) and (A.14) we can see that for both aggregation sampling and aggregation reduction the effectiveness of the method improves as the probability of the non-risk region q increases. In particular, given the properties of risk regions in (A.4) and (A.5), we can expect the performance of our methods to improve as β, the level of tail risk measure increases, and as $\mathcal{X}$, our feasible region of decisions becomes more constrained.

3.2 Alternative approaches

The above algorithms and analyses assume that the samples of Y were identically, indepenedently distribution. However, in principle the algorithms will

work for any unbiased sequence of samples. This opens up the possibility of enhancing the scenario aggregation and reduction algorithms by using them in conjuction with variance reduction techniques such as importance sampling, latin hypercube sampling or antithetic sampling [24][3]. The formulae (A.13) and (A.14) will still hold, but q will the probability of a *sample* occuring in the risk region rather than the actual probability of the risk region itself.

The above algorithms can also be generalized in how they represent the non-risk region. Because aggregation sampling and aggregating reduction only represent the non-risk region with a single scenario, they do not in general preserve the overall expectation of the cost function, or any other statistics of the loss function except for the value of a β-tail risk measure. These algorithms should therefore generally only be used for problems which only involve β-tail risk measures. However, if the cost function is affine (in the sense of Corollary 2.13), then collapsing all points in the non-risk region to the conditional expectation preserves the overall expectation.

If expectation or any other statistic of the cost function is used in the optimization problem then one could represent the non-risk region region with many scenarios. For example, instead of aggregating all scenarios in the non-risk region into a single point we could apply a clustering algorithm to them such as k-means. Such a clustered scenario set for the portfolio selection problem is illustrated for the arbitrarily chosen value $k = 10$ in Figure A.2; see Section 5 for details of this problem. The ideal allocation of points between the risk and non-risk regions will be problem dependent and is beyond the scope of this paper.

[3]Batch sampling methods such as stratified sampling will not work with aggregation sampling which requires samples to be drawn sequentially.

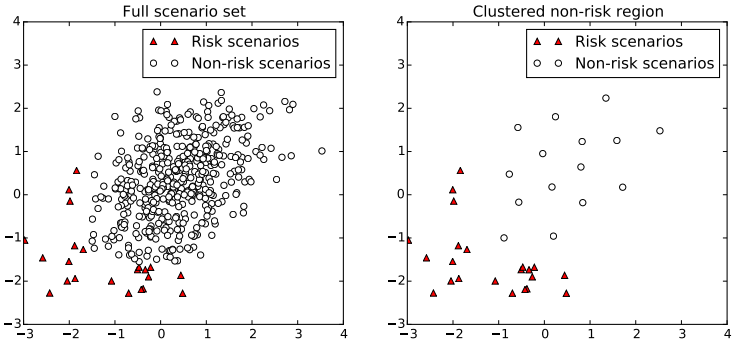


Fig. A.2: Scenario reduction via k-means clustering on a non-risk region for a portfolio selection problem.

4 Consistency of aggregation sampling

The reason that aggregation sampling and aggregation reduction work is that for large sample sizes, they are equivalent to sampling from the aggregated random vector, and if the aggregation condition holds then the aggregated random vector yields the same optimization problem as the original random vector. We only prove consistency for aggregation sampling and not aggregation reduction as the proofs are very similar. Essentially, the only difference is that aggregation sampling has the additional complication of terminating after a random number of samples.

We suppose in this section that we have a sequence of independently identically distributed (i.i.d.) random vectors $Y_1, Y_2, \ldots$ with the same distribution as Y, and which are defined on the product probability space Ω^∞.

4.1 Uniform convergence of empirical β-quantiles

The i.i.d. sequence of random vectors $Y_1, Y_2, \ldots$ can be used to estimate the distribution and quantile functions of Y. We introduce the additional short-

hand for the empirical distribution and quantile functions:

$$F_{n,x}(z) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{f(x,Y_i)\leq z\}},$$
$$F_{n,x}^{-1}(\beta) := \inf\{z \in \mathbb{R} : \ F_{n,x}(z) \geq \beta\}.$$

Note that these are random-valued functions on the probability space Ω^∞. It is immediate from the strong law of large numbers that for all $\bar{x} \in \mathbb{R}$ and $z \in \mathbb{R}$, we have $F_{n,x}(z) \overset{\text{w.p.1}}{\to} F_{\bar{x}}(z)$ as $n \to \infty$. In addition, if $F_{\bar{x}}$ is strictly increasing at $z = F_{\bar{x}}^{-1}$ then we also have $F_{n,\bar{x}}^{-1}(\beta) \overset{\text{w.p.1}}{\to} F_{\bar{x}}^{-1}(\beta)$ as $n \to \infty$; see for instance [25][Chapter 2]. The following result extends this pointwise convergence to a convergence result which is uniform with respect to $x \in \mathcal{X}$.

Theorem 4.1. *Suppose the following hold:*

(i) For each $x \in \mathcal{X}$, F_x is strictly increasing and continuous in some neighborhood of $F_x^{-1}(\beta)$

(ii) For all $\bar{x} \in \mathcal{X}$ the mapping $x \mapsto f(x, Y)$ is continuous at $\bar{x}$ with probability 1.

(iii) $\mathcal{X} \subset \mathbb{R}^k$ is compact

then $F_{n,x}^{-1}(\beta) \to F_x^{-1}(\beta)$ uniformly on $\mathcal{X}$ with probability 1.

The proof of this result relies on various continuity properties of the distribution and quantile functions which are provided in Appendix A. Some elements of the proof below have been adapted from [26, Theorem 7.48], a result which concerns the uniform convergence of expectation functions.

Proof. Fix $\epsilon_0 > 0$ and $\bar{x} \in \mathcal{X}$. Since $F_{\bar{x}}$ is continuous in a neighborhood of $F_{\bar{x}}^{-1}(\beta)$, there exists $0 < \epsilon < \epsilon_0$ such $F_{\bar{x}}$ is continuous at $F_{\bar{x}}^{-1}(\beta) \pm \epsilon$. Since $F_{\bar{x}}$ is strictly increasing at $F_{\bar{x}}^{-1}(\beta)$,

$$\delta := \min\{\beta - F_{\bar{x}}\left(F_{\bar{x}}^{-1}(\beta) - \epsilon\right),\ F_{\bar{x}}\left(F_{\bar{x}}^{-1}(\beta) + \epsilon\right) - \beta\} > 0.$$

By Corollary A.2 the mapping $x \mapsto F_x\left(F_{\bar{x}}^{-1}(\beta) - \epsilon\right)$ is continuous at $\bar{x}$ with probability 1. Applying Lemma A.4, there exists a neighborhood W of $\bar{x}$ such

that with probability 1, for n large enough

$$\sup_{x \in W \cap \mathcal{X}} \left| F_{n,x}(F_{\bar{x}}^{-1}(\beta) - \epsilon) - F_{n,\bar{x}}(F_{\bar{x}}^{-1}(\beta) - \epsilon) \right| < \frac{\delta}{2}.$$

In addition, by the strong law of large numbers, with probability 1, for n large enough

$$\left| F_{n,\bar{x}} \left(F_{\bar{x}}^{-1}(\beta) - \epsilon \right) - F_{\bar{x}} \left(F_{\bar{x}}^{-1}(\beta) - \epsilon \right) \right| < \frac{\delta}{2} \tag{A.15}$$

Thus, for all $x \in W \cap \mathcal{X}$ we have that

$$\left| F_{n,x} \left(F_{\bar{x}}^{-1}(\beta) - \epsilon \right) - F_{\bar{x}} \left(F_{\bar{x}}^{-1}(\beta) - \epsilon \right) \right| < \delta.$$

Similarly, we we can choose W so that we also have

$$\left| F_{n,x} \left(F_{\bar{x}}^{-1}(\beta) + \epsilon \right) - F_{\bar{x}} \left(F_{\bar{x}}^{-1}(\beta) + \epsilon \right) \right| < \delta.$$

and so

$$F_{n,x} \left(F_{\bar{x}}^{-1}(\beta) - \epsilon \right) < \beta < F_{n,x} \left(F_{\bar{x}}^{-1}(\beta) + \epsilon \right).$$

Hence, we have that with probability 1, for n large enough

$$\sup_{x \in W \cap \mathcal{X}} \left| F_{n,x}^{-1}(\beta) - F_{\bar{x}}^{-1}(\beta) \right| \leq \epsilon < \epsilon_0. \tag{A.16}$$

Also, by Proposition A.3 the function $x \mapsto F_x^{-1}(\beta)$ is continuous and so the neighborhood can also be chosen so that

$$\sup_{x \in W \cap \mathcal{X}} \left| F_{\bar{x}}^{-1}(\beta) - F_x^{-1}(\beta) \right| < \epsilon_0, \tag{A.17}$$

and so combining (A.16) and (A.17) we have

$$\sup_{x \in W \cap \mathcal{X}} \left| F_{n,x}^{-1}(\beta) - F_x^{-1}(\beta) \right| < 2\epsilon_0.$$

Finally, since $\mathcal{X}$ is compact, there exists a finite number of points $x_1, \ldots, x_m \in \mathcal{X}$ with corresponding neighborhoods $W_1, \ldots, W_m$ covering $\mathcal{X}$, such that with probability 1, for n large enough the following holds:

$$\sup_{x \in W_j \cap \mathcal{X}} \left| F_{n,x}^{-1}(\beta) - F_x^{-1}(\beta) \right| < 2\epsilon_0 \qquad \text{for } i = 1, \ldots, m$$

that is, with probability 1, for n large enough

$$\sup_{x \in \mathcal{X}} \left| F_{n,x}^{-1}(\beta) - F_x^{-1}(\beta) \right| < 2\epsilon_0.$$

□

In the next subsection this result will be used to show that any point in the interior of the non-risk region will, with probability 1, be in the non-risk region of the sampled scenario set for a large enough sample size.

4.2 Equivalence of aggregation sampling with sampling from aggregated random vector

The main obstacle in showing that aggregation sampling is equivalent to sampling from the aggregated random vector is to show that the aggregated scenario in the non-risk region converges almost surely to the conditional expectation of the non-risk region as the number of specified risk scenarios tends to infinity. Recall from Section 3 that $N(n)$ denotes the effective sample size in aggregation sampling when we require n risk scenarios and is distributed as $n + \mathcal{NB}(n, q)$ where q is the probability of the non-risk region. The purpose of the next Lemma is to show that as $n \to \infty$ the number of samples drawn from the non-risk region almost surely tends to infinity.

Lemma 4.2. *Suppose $M(n) \sim \mathcal{NB}(n, p)$ where $0 < p < 1$. Then with probability 1 we have that $\lim_{n\to\infty} M(n) = \infty$.*

Proof. First note that,

$$\begin{aligned}\{\lim_{n\to\infty} M(n) = \infty\}^c &= \bigcup_{k\in\mathbb{N}} \left(\bigcap_{n\in\mathbb{N}} \bigcup_{t>n} \{M(t) > k\}^c \right) \\ &= \bigcup_{k\in\mathbb{N}} \limsup_{n\to\infty} \{M(n) \leq k\}.\end{aligned}$$

Hence, to show that $\mathbb{P}\left(\{\lim_{n\to\infty} M(n) = \infty\}\right) = 1$ it is enough to show for each $k \in \mathbb{N}$ we have that

$$\mathbb{P}\left(\limsup_{n\to\infty} \{M(n) \leq k\} \right) = 0. \tag{A.18}$$

Now, fix $k \in \mathbb{N}$. Then for all $n \in \mathbb{N}$ we have that

$$\mathbb{P}\left(M(n) = k\right) = \binom{k+n-1}{k}(1-p)^n \; p^k,$$

and in particular,

$$\begin{aligned}\mathbb{P}\left(M(n+1) = k\right) &= \binom{k+n}{k}(1-p)^{n+1}p^k \\ &= \frac{k+n}{n}(1-p)\,\mathbb{P}\left(M(n) = k\right).\end{aligned}$$

For large enough n we have that $\frac{k+n}{n}(1-p) < 1$, hence $\sum_{n=1}^{\infty}\mathbb{P}\left(M(n) = k\right) < +\infty$ and so

$$\begin{aligned}\sum_{n=1}^{\infty}\mathbb{P}\left(M(n) \leq k\right) &= \sum_{n=1}^{\infty}\sum_{j=1}^{k}\mathbb{P}\left(M(n) = j\right) \\ &= \sum_{j=1}^{k}\sum_{n=1}^{\infty}\mathbb{P}\left(M(n) = j\right) \\ &< \infty.\end{aligned}$$

The result (A.18) now holds by the first Borel-Cantelli Lemma [19, Section 4].

□

The next Corollary shows that the strong law of large numbers still applies for the conditional expectation of the non-risk region in aggregation sampling despite the sample size being a random quantity.

Corollary 4.3. *Suppose* $\mathbb{E}\left[|Y|\right] < +\infty$ *and* $\mathbb{P}\left(Y \in \mathcal{R}^c\right) > 0$, *then*

$$\frac{1}{N(n)-n}\sum_{i\in 1\ldots,N(n):\; Y_i\in\mathcal{R}^c} Y_i \to \mathbb{E}\left[\; Y|Y \in \mathcal{R}^c\;\right] \text{ with probability 1 as } n \to \infty$$

This theorem could be proved by viewing the random variable $\sum_{i\in 1\ldots,N(n):\; Y_i\in\mathcal{R}^c} Y_i \to \mathbb{E}\left[\; Y|Y \in \mathcal{R}^c\;\right]$ as part of an appropriately defined renewal-reward process, and then using standard asymptotic results which apply to these; see [27, Chapter 10]. To keep this paper self-contained, we provide an elementary proof.

Proof. Define the following measurable subsets of Ω^∞:

$$\Omega_1 = \{\omega \in \Omega : \lim_{n\to\infty} N(n)(\omega) - n = \infty\},$$
$$\Omega_2 = \{\omega \in \Omega : \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i(\omega)\in\mathcal{R}^c\}} Y_i(\omega) = \mathbb{E}\left[\mathbb{1}_{\{Y\in\mathcal{R}^c\}} Y\right]\},$$
$$\Omega_3 = \{\omega \in \Omega : \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i(\omega)\in\mathcal{R}^c\}} = \mathbb{P}\left(Y \in \mathcal{R}^c\right)\}.$$

By the strong law of large numbers Ω_2 and Ω_3 have probability one. Since $N(n) - n \sim \mathcal{NB}(n, q)$, where $q = \mathbb{P}\left(Y \in \mathcal{R}^c\right)$, Ω_1 has probability 1 by Lemma 4.2. Therefore, $\Omega_1 \cap \Omega_2 \cap \Omega_3$ has probability 1 and so it is enough to show that for any $\omega \in \Omega_1 \cap \Omega_2 \cap \Omega_3$ we have that

$$\frac{1}{N(n)(\omega) - n} \sum_{i\in 1\ldots, N(n):\ Y_i(\omega)\in\mathcal{R}^c} Y_i(\omega) \to \mathbb{E}\left[\ Y | Y \in \mathcal{R}^c\ \right] \text{ as } n \to \infty.$$

Let $\omega \in \Omega_1 \cap \Omega_2 \cap \Omega_3$. Since $\omega \in \Omega_2 \cap \Omega_3$, we have that as $m \to \infty$:

$$\frac{1}{\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{Y_i(\omega)\in\mathcal{R}^c\}}} \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{Y_i(\omega)\in\mathcal{R}^c\}} Y_i \to \frac{1}{\mathbb{P}\left(Y \in \mathcal{R}^c\right)} \mathbb{E}\left[\mathbb{1}_{\{Y\in\mathcal{R}^c\}} Y\right]$$
$$= \mathbb{E}\left[Y | Y \in \mathcal{R}^c\right].$$

Now, fix $\epsilon > 0$. Then there exists $N_1(\omega) \in \mathbb{N}$ such

$$m > N_1(\omega) \implies \left| \frac{1}{\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{Y_i(\omega)\in\mathcal{R}^c\}}} \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{Y_i(\omega)\in\mathcal{R}^c\}} Y_i - \mathbb{E}\left[Y | Y \in \mathcal{R}^c\right] \right| < \epsilon.$$

Since $\omega \in \Omega_1$ there exists $N_2(\omega)$ such that

$$n > N_2(\omega) \implies N(n)(\omega) > N_1(\omega).$$

Noting that

$$\frac{1}{\frac{1}{N(n)(\omega)}\sum_{i=1}^{N(n)(\omega)} \mathbb{1}_{\{Y_i(\omega)\in\mathcal{R}^c\}}} \frac{1}{N(n)(\omega)} \sum_{i=1}^{N(n)(\omega)} \mathbb{1}_{\{Y_i(\omega)\in\mathcal{R}^c\}} Y_i(\omega)$$
$$= \frac{1}{\frac{N(n)(\omega)-n}{N(n)(\omega)}} \frac{1}{N(n)(\omega)} \sum_{i=1}^{N(n)(\omega)} \mathbb{1}_{\{Y_i(\omega)\in\mathcal{R}^c\}} Y_i(\omega)$$
$$= \frac{1}{N(n)(\omega) - n} \sum_{i:Y_i(\omega)\in\mathcal{R}^c} Y_i,$$

we have that

$$n > N_2 \implies \left| \frac{1}{N(n)(\omega) - n} \sum_{i:Y_i(\omega) \in \mathcal{R}^c} Y_i(\omega) - \mathbb{E}\left[Y | Y \in \mathcal{R}^c\right] \right| < \epsilon$$

and so $\frac{1}{N(n)(\omega)-n} \sum_{i:Y_i(\omega)\in\mathcal{R}^c} Y_i(\omega) \to \mathbb{E}\left[Y|Y \in \mathcal{R}^c\right]$ as $n \to \infty$. □

To show that aggregation sampling yields solutions consistent with the underlying random vector Y, we show that with probability 1, for n large enough, it is equivalent to sampling from the aggregated random vector $\psi_{\mathcal{R}}(Y)$, as defined in Definition 2.5. If the region $\mathcal{R}$ satisfies the aggregation condition, and $\mathbb{E}\left[Y|Y \in \mathcal{R}^c\right] \in \mathcal{R}^c$, Theorem 2.8 tells us that $\rho_\beta\left(f(x, \psi_{\mathcal{R}}(Y))\right) = \rho_\beta\left(f(x, Y)\right)$ for all $x \in \mathcal{X}$. Hence, if sampling is consistent for the risk measure ρ_β, then aggregation sampling also consistent.

Denote by $\tilde{F}_{n,x}, \tilde{F}^{-1}_{n,x}$, the empirical distribution, and quantile functions respectively and by $\tilde{\rho}_{n,\beta}(x)$ the value of the tail-risk measure for the decision $x \in \mathcal{X}$ for the sample from the aggregated random vector: $\psi_{\mathcal{R}}(Y_1), \ldots, \psi_{\mathcal{R}}(Y_n)$. Similarly, denote by $\hat{F}_{n,x}, \hat{F}^{-1}_{n,x}$, and $\hat{\rho}_{n,\beta}$ the analogous functions for the scenario set constructed by aggregation sampling with n risk scenarios. Note that these latter functions will depend on the sample $Y_1, \ldots, Y_{N(n)}$. Note also that like $F_{n,x}$ and $F^{-1}_{n,x}$, all these functions are random and defined on the same sample space Ω^∞.

Theorem 4.4. *Suppose the following conditions hold:*

(i) $(x, y) \mapsto f(x, y)$ *is continuous on* $\mathcal{X} \times \mathbb{R}^d$

(ii) For each $x \in \mathcal{X}$, F_x *is strictly increasing and continuous in some neighborhood of* $F_x^{-1}(\beta)$

(iii) $\mathbb{E}\left[\, Y|Y \in \mathcal{R}^c \,\right] \in \operatorname{int}\left(\mathcal{R}^c\right)$

(iv) $\mathcal{X}$ *is compact.*

Then, with probability 1, for n large enough $\tilde{\rho}_{n,\beta} \equiv \hat{\rho}_{N(n),\beta}$.

Proof. Note that if

$$z > \max\left\{ f\left(x, \frac{1}{N(n)(\omega)-n} \sum_{i\in 1\ldots,N(n)(\omega):\ Y_i(\omega)\in\mathcal{R}^c} Y_i(\omega)\right), f\left(x, \mathbb{E}\left[\ Y|Y\in\mathcal{R}^c\ \right]\right)\right\}$$

then

$$\begin{aligned}\hat{F}_{n,x}(z)(\omega) &= \frac{N(n)(\omega)-n}{N(n)(\omega)} \\ &\quad + \frac{1}{N(n)(\omega)}|\{1\le i\le N(n)(\omega)\mid f(x,Y_i(\omega))\le z \text{ and } Y_i(\omega)\in\mathcal{R}\}| \\ &= \tilde{F}_{N(n),x}(z)(\omega).\end{aligned}$$

So if we have

$$\hat{F}^{-1}_{n,x}(\beta)(\omega) > \max\left\{ f\left(x, \frac{1}{N(n)(\omega)-n} \sum_{i\in 1\ldots,N(n)(\omega):\ Y_i(\omega)\in\mathcal{R}^c} Y_i(\omega)\right), f\left(x, \mathbb{E}\left[\ Y|Y\in\mathcal{R}^c\ \right]\right)\right\} \tag{A.19}$$

then this implies that $\hat{F}^{-1}_{n,x}(u)(\omega) = \tilde{F}^{-1}_{N(n),x}(u)(\omega)$ for all $u \ge \beta$, which in turn implies $\hat{\rho}_{\beta,n}(x)(\omega) = \tilde{\rho}_{\beta,N(n)}(x)(\omega)$. Hence, it is enough to show that with probability 1, for sufficiently large n, the inequality (A.19) holds for all $x \in \mathcal{X}$.

Since $\mathbb{E}\left[\ Y|Y\in\mathcal{R}^c\ \right] \in \text{int}\left(\mathcal{R}^c\right)$ we have that

$$f(x, \mathbb{E}\left[Y|Y\in\mathcal{R}^c\right]) < F_x^{-1}(\beta) \qquad \text{for all } x\in\mathcal{X}$$

and since $\mathcal{X}$ is compact there exists $\delta > 0$ such that

$$\sup_{x\in\mathcal{X}} \left(F_x^{-1}(\beta) - f\left(x, \mathbb{E}\left[\ Y|Y\in\mathcal{R}^c\ \right]\right)\right) > \delta. \tag{A.20}$$

The continuity of $f(x,y)$ and again the compactness of $\mathcal{X}$ implies that there exists $\gamma > 0$ such that

$$|y - \mathbb{E}\left[\ Y|Y\in\mathcal{R}^c\ \right]| < \gamma \implies \sup_{x\in\mathcal{X}} |f(x,y) - f(x, \mathbb{E}\left[\ Y|Y\in\mathcal{R}^c\right])\ | < \frac{\delta}{2}$$

Thus, by Corollary 4.3, with probability 1, for n large enough

$$\left| f\left(x, \frac{1}{N(n)-n} \sum_{i\in 1\ldots,N(n):\ Y_i\in\mathcal{R}^c} Y_i\right) - f\left(x, \mathbb{E}\left[\ Y|Y\in\mathcal{R}^c\right]\right) \right| < \frac{\delta}{2} \tag{A.21}$$

Also, by Theorem 4.1, given $N(n) > n$, for n large enough

$$\sup_{x \in \mathcal{X}} \left| F_x^{-1}(\beta) - \tilde{F}_{N(n),x}^{-1}(\beta) \right| < \frac{\delta}{2}, \tag{A.22}$$

which implies for all $x \in \mathcal{X}$

$$\begin{aligned}
\tilde{F}_{N(n),x}^{-1}(\beta) - f\left(x, \frac{1}{N(n)-n} \sum_{i \in 1\ldots,N(n):\ Y_i \in \mathcal{R}^c} Y_i\right) & \\
\geq \left(F_x^{-1}(\beta) - \frac{\delta}{2}\right) - \left(f\left(x, \mathbb{E}\left[\, Y | Y \in \mathcal{R}^c \,\right]\right) + \frac{\delta}{2}\right) & \quad \text{by (A.21) and (A.22)} \\
= \underbrace{\left(F_x^{-1}(\beta) - f\left(x, \mathbb{E}\left[\, Y | Y \in \mathcal{R}^c \,\right]\right)\right)}_{> \delta \text{ by (A.20)}} - \delta & \\
> 0. &
\end{aligned}$$

Similarly with probability 1 for n large enough we have $\tilde{F}_{N(n),x}^{-1}(\beta) > f\left(x, \mathbb{E}\left[\, Y | Y \in \mathcal{R}^c \,\right]\right)$ for all $x \in \mathcal{X}$. Therefore the inequality (A.19) holds with probability 1 for sufficiently large n as required. □

5 Risk regions for the portfolio selection problem

In this section we characterize exactly the risk region of the portfolio selection problem when the asset returns are elliptically distributed. In Section 5.1 we formulate the basic problem and, to provide some intuition, we find the risk region by brute force for an arbitrary discrete distribution. In Section 5.2 we define elliptical distributions and give the non-risk region for the unconstrained problem, and finally in Section 5.3 we characterize the non-risk region when portfolios are constrained to a a convex set.

5.1 Problem statement and brute force aggregation

In the portfolio selection problem, one aims to choose a portfolio of financial assets with uncertain returns. For $i = 1, \ldots, d$, let x_i denote the amount to invest in asset i, and Y_i the random return of asset i. The loss function in this

problem is the negative total return, that is $f(x, Y) = \sum_{i=1}^{d} -x_i Y_i = -x^T Y$. The optimization problem will typically try to balance the expected profit against the risk in some way, and so our problem is usually of one of the following forms:

$$\begin{aligned}
&\text{(i)} \underset{x \in \mathcal{X}}{\text{minimize}}\ \rho_\beta(-x^T Y) \\
&\qquad\qquad \text{subject to } \mathbb{E}\left[x^T Y\right] \geq t \\
&\text{(ii)} \underset{x \in \mathcal{X}}{\text{maximize}}\ \mathbb{E}\left[x^T Y\right] \\
&\qquad\qquad \text{subject to } \rho_\beta(-x^T Y) \leq s \\
&\text{(iii)} \underset{x \in \mathcal{X}}{\text{minimize}}\ \rho_\beta(-x^T Y) + v\mathbb{E}\left[-x^T Y\right]
\end{aligned}$$

where $v \geq 0$ and $\mathcal{X} \subset \mathbb{R}^d$ represents the set of valid portfolios. The set $\mathcal{X}$ of feasible portfolios may encompass constraints like no short-selling ($x \geq 0$), total investment ($\sum_{i=1}^{d} x_i = 1$) and quotas on certain stocks or combinations of stocks ($x \leq c$).

For a given portfolio $x \in \mathcal{X}$, the corresponding risk region is the half-space of points where loss is greater than or equal to the β-quantile:

$$\mathcal{R}_x = \{y \in \mathbb{R}^d : \ -x^T y \geq F_x^{-1}(\beta)\}$$

For a discrete distribution of returns, finding the β-quantile of the loss associated to a particular portfolio is a case of ordering all scenarios according to their loss and selecting the appropriate order statistic. In Figure A.3 we have illustrated a scenario set of returns for two hypothetical assets sampled i.i.d. from a multivariate Normal distribution. The line in this figure separates all those scenarios with loss below the β-quantile from those with loss above for the portfolio $x = (\frac{1}{2}, \frac{1}{2})$.

Recall that the risk region associated to a set of feasible decisions is the union of all risk regions for decisions in that set. Thus, we can find all scenarios in the risk region by calculating the β-quantile for all feasible portfolios. On the left hand side of Figure A.4, for the same scenario set in Figure

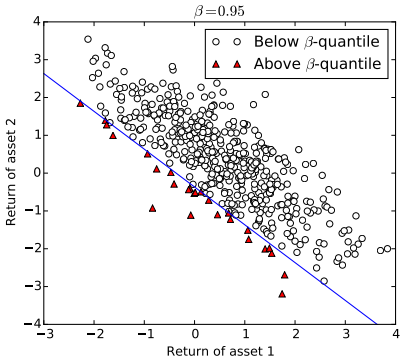


Fig. A.3: Scenarios with loss above and below β-quantile for one portfolio

A.3, we have identified the risk scenarios for the set of feasible portfolios $\mathcal{X} = \{(x_1, x_2) \in \mathbb{R}^2 : \ x_1, x_2 \geq 0, \ x_1 + x_2 = 1\}$.

Corollary 2.13 states that if the aggregation condition holds, then all the mass in the non-risk region can be aggregated into its conditional expectation without affecting the value of the expectation of the loss or any tail risk measure. For a discrete distribution, we noted in Section 2.2 that the aggregation condition always holds for the risk region. On the right-hand side of Figure A.4 is illustrated the same scenario set where all non-risk scenarios have been aggregated into a single point. Note that the β-quantile lines have not changed after aggregation. By aggregating all the non-risk scenarios into a single point we substantially reduce the computational cost of solving the corresponding portfolio selection problem.

The following corollary gives sufficient conditions for the risk region to satisfy the aggregation condition for continuous distributions.

Corollary 5.1. *Suppose that $\mathcal{Y} = \mathbb{R}^d$ and there exist $x_1, x_2 \in \mathcal{X}$ which are linearly independent. Then, for any $\mathcal{R} \supseteq \mathcal{R}_{\mathcal{X}}$, $\mathcal{R}$ satisfies the aggregation condition. Moreover, if $\mathcal{R}$ is convex, Y is continuous and $\mathcal{X}$ is compact, then aggregation sampling with respect to $\mathcal{R}$ is consistent in the sense of Theorem 4.4.*

Proof. For the first part of this result, it is enough to show that $\mathcal{R}_{\mathcal{X}}$ satisfies

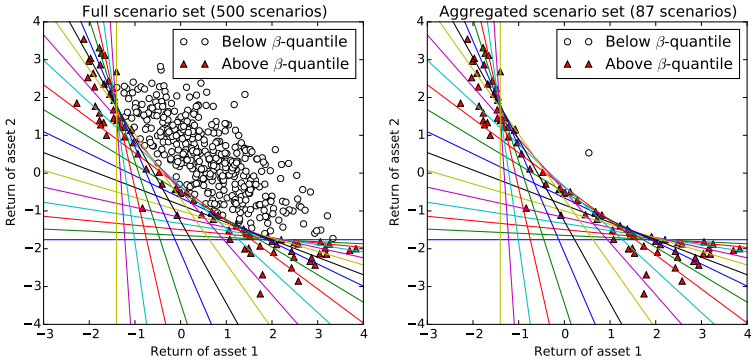


Fig. A.4: Scenario set separated into risk and non-risk scenarios: full scenario set (left) and aggregated scenario set (right)

the aggregation condition. We prove this by showing that all the conditions of Proposition 2.10 hold. Note that $x \mapsto -x^T y$ is continuous so condition (ii) holds immediately.

For all $x \in \mathcal{X}$ the interior of the corresponding risk region and non-risk region are open half-spaces:

$$\operatorname{int}(\mathcal{R}_x) = \{y \in \mathbb{R}^d : -x^T y > F_x^{-1}(\beta)\}$$
$$\operatorname{int}(\mathcal{R}_x^c) = \{y \in \mathbb{R}^d : -x^T y < F_x^{-1}(\beta)\}$$

Fix $\bar{x} \in \mathcal{X}$. Then either $\bar{x}$ is linearly independent to x_1 or it is linearly independent to x_2. Assume it is linearly independent to x_1. Now, $\operatorname{int}(\mathcal{R}_{\bar{x}})$ and $\operatorname{int}(\mathcal{R}_{x_1})$ are non-parallel half-spaces and so both $\operatorname{int}(\mathcal{R}_{\bar{x}} \cap \mathcal{R}_{x_1})$ and $\operatorname{int}(\mathcal{R}_{x_1} \setminus \mathcal{R}_{\bar{x}}) = \operatorname{int}(\mathcal{R}_{x_1}) \cap \operatorname{int}(\mathcal{R}_{\bar{x}}^c)$ are non-empty so condition (iii) is satisfied.

Since $\mathcal{R}_{x_1}$ and $\mathcal{R}_{x_2}$ are non-parallel half-spaces, their union $\mathcal{R}_{x_1} \cup \mathcal{R}_{x_2}$ is connected. Similarly, for any $x \in \mathcal{X}$, we must have $\mathcal{R}_x$ being non-parallel with either $\mathcal{R}_{x_1}$ or $\mathcal{R}_{x_2}$ and so $\mathcal{R}_x \cup \mathcal{R}_{x_1} \cup \mathcal{R}_{x_2}$ must also be connected. Hence, $\mathcal{R}_{\mathcal{X}} = \bigcup_{x \in \mathcal{X}} (\mathcal{R}_x \cup \mathcal{R}_{x_1} \cup \mathcal{R}_{x_2})$ is connected so condition (i) is also satisfied.

It now remains to show that aggregation sampling is consistent in the sense of Theorem 4.4. Conditions (i) and (iv) of this theorem hold trivially.

Condition (iii) also holds immediately since $\mathcal{R}$ is convex, so it only remains to verify condition (ii). Since Y is continuous and has support $\mathcal{Y} = \mathbb{R}^d$, Y has a density f such that $f(y) > 0$ for all $y \in \mathbb{R}^d$. Hence, for all $x \in \mathcal{X}$, the function F_x is continuous and increasing everywhere. □

In the illustrative example above we used brute force to test whether or not a point belonged to the risk region. This approach requires the calculation of the β-quantile for all feasible decisions, which is roughly equivalent to the computational cost required to enumerate the value of the tail-risk measure for all feasible decisions. To benefit from risk regions, we instead need a convenient method to test whether or not a point belongs to it.

5.2 Non-risk region for elliptically distributed returns

By exploiting the structure of a parametric distribution, it may be possible to characterize its associated risk region in a more convenient manner. In this section we do this for elliptically distributed returns.

Elliptical distributions are a general class of distributions which include among others the multivariate Normal and multivariate t-distributions. See [28] for a full overview of the subject.

Definition 5.2 (Spherical and Elliptical Distributions)**.** *Let X be a random vector in $\mathbb{R}^d$, then X is said to be* spherical *if its distribution is invariant under orthonormal transformations; that is, if*

$$X \sim UX \qquad \text{for all } U \in \mathbb{R}^{d\times d} \text{ orthonormal}.$$

Let Y be a random vector in $\mathbb{R}^d$, then Y is said to be elliptical *if it can be written $Y = PX + \mu$ where $P \in \mathbb{R}^{d\times d}$ is non-singular, $\mu \in \mathbb{R}^d$, and X is random vector with spherical distribution. We will denote this $Y \sim \text{Elliptical}(X, P, \mu)$.*

We will assume throughout that Y is continuous and $\mathcal{Y} = \mathbb{R}^d$ so that we can apply Corollary 5.1. An important property of elliptical distributions is that for any random vector with such a distribution, we can characterize

exactly the distribution of any linear combination of the components of the vector. That is, for an an elliptical distribution $Y \sim \text{Elliptical}(X, P, \mu)$ in $\mathbb{R}^d$ and $x \in \mathbb{R}^d$ we have

$$x^T Y \sim \|Px\| X_1 + x^T \mu. \tag{A.23}$$

where X_1 is the first component of the random vector X, and $\|\cdot\|$ denotes the standard Euclidean norm. This property allows us to solve some portfolio selection problems for elliptical distributions where the risk measure is positive homogeneous and translation invariant via quadratic programming or interior point algorithms. Such risk measures include the β-VaR, β-CVaR and all coherent risk measures [20]. For more details, and a proof of (A.23) see [29]. By (A.23) the β-quantile of the loss of a portfolio is as follows:

$$F_x^{-1}(\beta) = \|Px\| F_{X_1}^{-1}(\beta) - x^T \mu.$$

Therefore, using (A.3) the non-risk region for $Y \sim \text{Elliptical}(X, P, \mu)$, is the following:

$$\{y \in \mathbb{R}^d : -x^T y \leq \|Px\| F_{X_1}^{-1}(\beta) - x^T \mu \qquad \forall x \in \mathcal{X}\} \tag{A.24}$$

If we take $\mathcal{X} = \mathbb{R}^d$, then it can be shown that the set (A.24) is in fact just an ellipsoid (see Proposition (B.1)):

$$\mathcal{R}^c_{\mathbb{R}^d} = \{y \in \mathbb{R}^d : (y - \mu)^T \Sigma^{-1} (y - \mu) \leq F_{X_1}^{-1}(\beta)^2\}. \tag{A.25}$$

where $\Sigma = P^T P$. Note that by (A.5) the set $\mathcal{R}_{\mathcal{X}} \subset \mathcal{R}_{\mathbb{R}^d}$ and so $\mathcal{R}_{\mathbb{R}^d}$ always satisfies the aggregation condition. Unlike (A.24) this characterization in (A.25) allows us to easily test whether or not an arbitrary point is in the risk region.

As discussed in Section 3 on scenario generation, the greater the probability of the non-risk region, the greater the benefit of our methodology over regular sampling. To gauge the utility of our methodology we calculate the probability of the region (A.25) for the Normal distribution. If $Y \sim \mathcal{N}(\mu, \Sigma)$

this can be calculated exactly:

$$\mathbb{P}\left((Y-\mu)^T\Sigma^{-1}(Y-\mu) \leq \Phi^{-1}(\beta)^2\right) = \mathbb{P}\left(X^T X \leq \Phi^{-1}(\beta)^2\right)$$
$$\text{where } Y = PX + \mu$$
$$= \mathbb{P}\left(\chi_d^2 \leq \Phi^{-1}(\beta)^2\right),$$

where Φ is the distribution function of the standard Normal distribution. That is, the probability of the non-risk region is invariant to the mean and covariance and can be calculated from a χ_d^2 distribution function. In Figure A.5 we have plotted how the probability of the non-risk region varies with the value of β and the dimension. It shows that as the dimension increases, the probability of the non-risk region converges to zero. This convergence is so quick that for even relatively small dimensions and high values of β, the probability of the ellipsoid is tiny. This means that the potential benefit of aggregating scenarios using this region for reasonably sized problems would be negligible. However, as we show in the next subsection, by using the constraints of our problem we can significantly increase this probability.

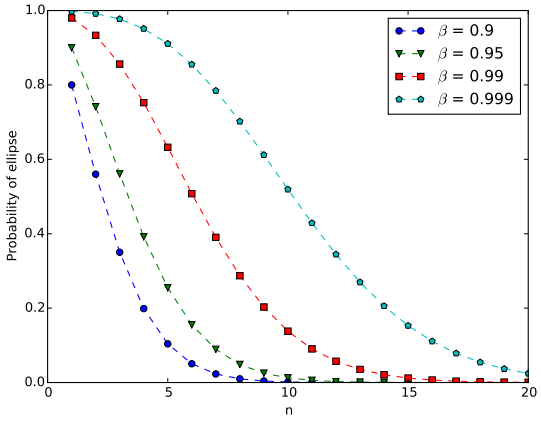


Fig. A.5: Plot of how mass of ellipse varies with dimension

5.3 Non-risk region with convex constraints

We now treat the more general case where the portfolios are constrained to a convex set. As well as convexity we also require the related concepts of cone and conic hull.

Definition 5.3 (Cones and Conic Hull)**.** *A set $K \subset \mathbb{R}^d$ is a cone if for all $x \in K$ and $\lambda \geq 0$ we have $\lambda x \in K$. A cone is convex if for all $x_1, x_2 \in K$ and $\lambda_1, \lambda_2 \geq 0$ we have $\lambda_1 x_1 + \lambda_2 x_2 \in K$. The conic hull of a set $\mathcal{A} \subset \mathbb{R}^d$ is the smallest convex cone containing $\mathcal{A}$, and is denoted* $\text{conic}(\mathcal{A})$.

The characterization of this region also makes use of the concept of a projection onto a convex set which we recall now.

Definition 5.4 (Projection)**.** *Let $C \subset \mathbb{R}^d$ be a closed convex set. Then for any point $y \in \mathbb{R}^d$, we define the projection of y onto C to be the unique point $p_C(y) \in C$ such that*

$$\inf_{x \in C} \|x - y\| = \|p_C(y) - y\|$$

By a slight abuse of notation, for a set $\mathcal{A} \subset \mathbb{R}^d$ and a matrix $T \in \mathbb{R}^{d \times d}$, we write $T(\mathcal{A}) := \{Ty : y \in \mathcal{A}\}$. Now, letting $K = \text{conic}(\mathcal{X})$, Corollary B.5 in Appendix B applied to the set (A.24) gives us the non-risk region:

$$P^T \left(\{\tilde{y} : \|p_{K'}(\tilde{y} - \mu)\| \leq F_{X_1}^{-1}(\beta)\} \right) \tag{A.26}$$

where $K' = PK$. Like (A.25) the characterization (B.6) allows us easily to check whether or not a point lies in the risk region.

We now repeat our calculations of the probability of the non-risk region assuming now that $\mathcal{X} = \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1,\ x \geq 0\}$. The probability of the non-risk region is no longer invariant to the parameters of the Normal distribution, so for simplicity we take $\mu = 0$ and $\Sigma = I_n$. In this case we have $P = I_d$ and so $K' = K = \mathbb{R}^d_+$. Also, $p_K(y) = y_+$ where $y_+ = \max\{0, y\}$, hence

$$\mathcal{R}^c_{\mathcal{X}} = \{y \in \mathbb{R}^d : \|y_+\| \leq \Phi^{-1}(\beta)\}.$$

The probability of this region cannot be calculated analytically, and so we estimate it by Monte Carlo simulation. As Figure A.6 shows, the probability of the region decays at a much slower rate as the dimension increases. This underlines the importance of making use of our constraints for finding sets which satisfy the aggregation condition.

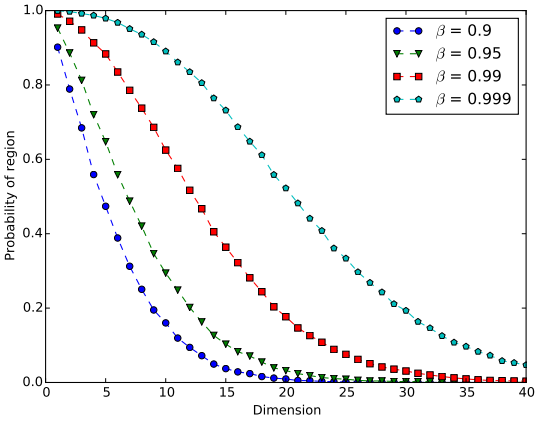


Fig. A.6: Plot of how probability of non-risk region varies when $K = \mathbb{R}^d_+$, and $Y \sim \mathcal{N}(0, I)$

6 Numerical tests

In this section, we test the performance of our aggregation sampling algorithm from Section 3 on the portfolio selection problem with elliptical distributions, using the non-risk region found in Section 5. The purpose of this test is to compare the performance of our methodology against "standard" scenario generation methods, which spread their scenarios evenly across the support of a distribution. For simplicity, we do this by comparing the performance of aggregation sampling method against that of basic sampling. Although we could run tests comparing aggregation sampling against more sophisticated scenario generation methods (such as sampling with variance reduction techniques), as mentioned in Section 3.2, it is often possible to combine our methods with these techniques, in which case we would have to use

the enhanced version of our method to ensure a fair comparison.

6.1 Experimental Set-up

For our numerical tests we use the following problem:

$$\begin{aligned} \underset{x \geq 0}{\text{minimize}} \ & \beta\text{-CVaR}(-x^T Y) \\ \text{subject to } & \mathbb{E}\left[x^T Y\right] \geq t. \end{aligned} \tag{P}$$

We will in particular assume that the asset returns follow a Normal distribution, that is $Y \sim \mathcal{N}(\mu, \Sigma)$. We construct our Normal distributions from monthly return data between January 2007 and February 2015 from randomly selected companies in the FTSE 100 index.

To ensure our non-risk has non-neglible probability we have imposed positivity constraints on our portfolios. Note that by (B.6), the aggregation sampling algorithm will require the calculation of projections onto the finitely generated cones $K' = P\mathbb{R}^d_+$ where $\Sigma = P^T P$. For details on how this is done see [30].

This problem has been constructed so that we can easily calculate exactly the optimality gap of any candidate portfolio $x \geq 0$. The following formula is easily verified by recalling that for continuous probability distributions, the β-CVaR is just the conditional expectation of the random variable above the β-quantile (see [6] for instance):

$$\beta\text{-CVaR}(-x^T Y) = (1-\beta)\mu^T x + \sqrt{x^T \Sigma x} \int_{\Phi^{-1}(\beta)}^{\infty} z \, d\Phi(z) \tag{A.27}$$

where Φ denotes the distribution function of the standard Normal distribution. The problem (P) can therefore be solved exactly using an interior point algorithm.

The application of aggregation sampling to the problem (P) is valid as all the conditions of Corollary 5.1 hold. We are interested in the quality and

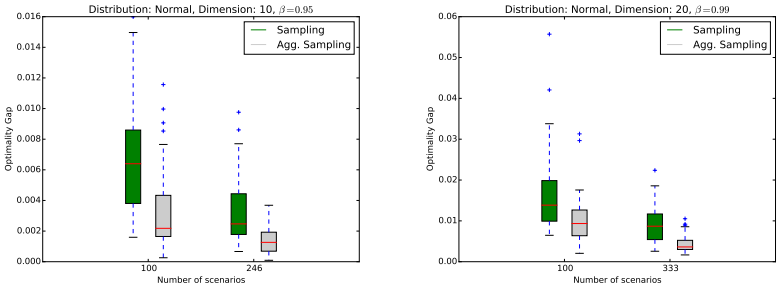
stability of the solutions that are yielded by our method as compared to sampling. To this end, in each experiment we construct 50 scenario sets using sampling and aggregation sampling, solve the resulting problems, and calculate the optimality gaps for the solutions that these yield. We then estimate the probability of the non-risk region, and repeat the stability test for scenario sets of the expected effective sample size of aggregation sampling with respect to the first sample size.

6.2 Results

In Figure A.7 are presented the results of these stability tests for two different problems. In the first problem we have $d = 10$ and $\beta = 0.95$ and the probability of the non-risk region is estimated to be 0.594; in the second problem we have $d = 20$ and $\beta = 0.99$ for which the probability of the non-risk region is estimated to be 0.700. Note that for both of these experiments the probability of the non-risk region is much larger than for the case where asset returns are independently distributed as in Figure A.6. For the first problem the expected effective sample size of aggregation sampling with 100 risk scenarios, as given by (A.13), is $100 + \frac{0.594}{0.406}100 \approx 246$. Similarly, the expected effective sample size of aggregation sampling for the second problem is $100 + \frac{0.7}{0.3}100 \approx 333$. In both cases, the performance of aggregation sampling for 100 risk scenarios is on a par with that of sampling for the much larger expected effective sample size, in terms of both the quality of solutions and their stability.

7 Conclusions

In this paper we have demonstrated that in stochastic programs which use a tail risk measure, a significant portion of the support of the random variables in the problem do not participate in the calculation of that tail risk measure, whatever feasible decision is used. As a consequence, for scenario-based



(a) Probability of risk region: 0.406 **(b)** Probability of risk region: 0.300

Fig. A.7: Optimality gap for 50 scenarios sets constructed via sampling and aggregation sampling

problems, if we concentrate our scenarios in the region of the distribution which is important to the problem, the risk region, we can represent the uncertainty in our problem in a more parsimonious way, thus reducing the computational burden of solving it.

We have proposed and analyzed two specific methods of scenario generation using risk regions: aggregation sampling and aggregation reduction. Both of these methods were shown to be more effective as the probability of the non-risk region increases: in essence the higher this probability the more redundancy there is in the original distribution. Therefore, our methodology becomes more valuable as as our problem becomes more constrained, and as the level of the tail-risk increases since these changes cause the probability of the non-risk region to decrease.

However, the application of this work relies on the ability to characterize the risk region in a way which makes it convenient to test whether or not a point belongs to it. This is difficult as it depends on the cost function, the distribution of uncertain parameters, and the set of feasible decisions. An exact characterization of the risk region may not be possible for most problems, but it may be possible to find conservative regions which contain the true risk region.

For some problems the issue might be that the non-risk region has negligible probability or is even empty. Indeed we observed for the portfolio selection problem that the probability of the non-risk region quickly tended to zero as the dimension of our problem increases. A potential strategy for overcoming this problem, and more generally for improving the effectiveness of our methodology, would be the addition of artificial constraints to the problem to enlarge the non-risk region. However, even if a non-risk region has small mass, for large and difficult problems, for example those involving integer variables or with non-linear recourse problems, the reduction in computation time gained from aggregation may be significant.

In the case of the portfolio selection problem we were able to characterize the risk region in a convenient form when the distributions of asset returns are elliptical, and demonstrated the gain from aggregation sampling for simple test problems. In the paper [31] we demonstrate that our methodology may be applied to more difficult and realistic portfolio selection problems such as those involving integer variables, and for which the asset returns are no longer elliptically distributed. In the same paper we also some of the technical issues involved in applying the method, such as finding the conic hull of the feasible region, and methods of projecting points onto this. We also investigate the use of artificial constraints as a way of making our methodology more effective.

Acknowledgments

We would like to thank Burak Buke for bringing our attention to the connection between aggregation sampling and renewal-reward processes.

A Continuity of Distribution and Quantile Functions

Throughout we use the following set-up: $\mathcal{X} \subset \mathbb{R}^k$ a decision space, Y a random vector with support $\mathcal{Y} \subset \mathbb{R}^d$ defined on a probability space $(\Omega, \mathcal{B}, \mathbb{P})$, and a cost function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The quantity is $f(x, Y)$ is assumed to be measurable for all $x \in \mathcal{X}$. In this appendix we prove a series of technical results related to the continuity of the distribution and quantile functions for $f(x, Y)$. These are required for the proofs in Section 4.

The following is an elementary result from the stochastic optimization literature concerning the continuity of an expectation function.

Proposition A.1. *Suppose for $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and a given $\bar{x} \in \mathcal{X}$ the following holds:*

(i) $x \mapsto g(x, Y)$ is continuous at $\bar{x}$ with probability 1.

(ii) There exists a neighborhood W of $\bar{x}$ and integrable $h : \mathcal{Y} \to \mathbb{R}$ such that for all $x \in W$ we have $g(x, Y) \leq h(Y)$ with probability 1.

Then, $x \mapsto \mathbb{E}\left[g(x, Y)\right]$ is continuous at $\bar{x}$.

Proof. Let $(x_k)_{k=1}^{\infty}$ be some sequence in $\mathcal{X}$ such that $x_k \to \bar{x}$ as $k \to \infty$. Without loss of generality $x_k \in W$ for all $k \in \mathbb{N}$. By assumption (i), almost surely we have $g(x_k, Y) \to f(\bar{x}, Y)$ as $k \to \infty$. Using assumption (ii) we can apply the Lebesgue theorem of dominated convergence so that:

$$\begin{aligned}\lim_{k\to\infty} \mathbb{E}\left[g(x_k, Y)\right] &= \mathbb{E}\left[\lim_{k\to\infty} g(x_k, Y)\right] \\ &= \mathbb{E}\left[g(\bar{x}, Y)\right]\end{aligned}$$

and hence $x \mapsto \mathbb{E}\left[g(x, Y)\right]$ is continuous at $\bar{x}$. □

Recall that we use the following notation for simplicity of exposition:

$$F_x(z) := \mathbb{P}\left(f(x, Y) \leq z\right)$$
$$F_x^{-1}(\beta) := \inf\{z \in \mathbb{R} : \ F_x(z) \geq \beta\}$$

The continuity of the distribution function immediately follows from the above proposition.

Corollary A.2. *Suppose for a given $\bar{x} \in \mathcal{X}$ that $x \mapsto f(x, Y)$ is continuous with probability 1 at $\bar{x}$, and for $z \in \mathbb{R}$ the distribution function $F_{\bar{x}}$ is continuous at z. Then, $x \mapsto F_x(z)$ is continuous at $\bar{x}$.*

Proof. Let $g(x, Y) = \mathbb{1}_{\{f(x,Y) \leq z\}}$ so that $F_x(z) = \mathbb{E}\left[g(x, Y)\right]$. The function $g(x, Y)$ is clearly dominated by the integrable function $h(Y) = 1$. It is therefore enough to show that $x \mapsto g(x, Y)$ is almost surely continuous at $\bar{x}$ as the result will then follow from Proposition A.1.

Since $F_{\bar{x}}$ is continuous at z, we must have $\mathbb{P}\left(f(\bar{x}, Y) = z\right) = 0$. Almost surely, we have that for $\omega \in \Omega$ that $x \mapsto f(x, Y(\omega))$ is continuous at $\bar{x}$. Let's first assume that $f(\bar{x}, Y(\omega)) > z$. In this case, there exist some neighborhood V of $\bar{x}$ such that $x \in V \Rightarrow f(x, Y(\omega)) > z$, which in turn implies $|g(x, Y) - g(\bar{x}, Y)| = 0$. Hence $x \mapsto g(x, Y(\omega))$ is continuous at $\bar{x}$. The same argument holds if $f(\bar{x}, Y(\omega)) < z$. Hence, with probability 1, $x \mapsto g(x, Y)$ is continuous at $\bar{x}$. □

Continuity of the quantile function follows from the continuity of the distribution function but requires that the distribution function is strictly increasing at the required quantile.

Proposition A.3. *Suppose for some $\bar{x} \in \mathcal{X}$, and $z = F_{\bar{x}}^{-1}(\beta)$ that the conditions of Corollary A.2 hold, and in addition that $F_{\bar{x}}$ is strictly increasing at $F_{\bar{x}}^{-1}(\beta)$, that is for all $\epsilon > 0$*

$$F_{\bar{x}}\left(F_{\bar{x}}^{-1}(\beta) - \epsilon\right) < \beta < F_{\bar{x}}\left(F_{\bar{x}}^{-1}(\beta) + \epsilon\right).$$

Then $x \mapsto F_x^{-1}(\beta)$ is continuous at $\bar{x}$.

Proof. Assume $x \mapsto F_x^{-1}(\beta)$ is not continuous at $\bar{x}$. This means there exists $\epsilon > 0$ such that for all neighborhoods W of $\bar{x}$

$$\text{there exists } x' \in W \text{ such that } \left|F_{\bar{x}}^{-1}(\beta) - F_{x'}^{-1}(\beta)\right| > \epsilon.$$

Now set,

$$\begin{aligned}\gamma &:= \min\{\beta - F_{\bar{x}}\left(F_{\bar{x}}^{-1}(\beta) - \epsilon\right), F_{\bar{x}}\left(F_{\bar{x}}^{-1}(\beta) + \epsilon\right) - \beta\} \\ &> 0 \qquad \text{since } F_{\bar{x}} \text{ strictly increasing at } F_{\bar{x}}^{-1}(\beta).\end{aligned}$$

By the continuity of $x \mapsto F_x\left(F_{\bar{x}}^{-1}(\beta)\right)$ at $\bar{x}$ there exists W a neighborhood of $\bar{x}$, such that:

$$x \in W \Longrightarrow \left|F_x\left(F_{\bar{x}}^{-1}(\beta)\right) - F_{\bar{x}}\left(F_{\bar{x}}^{-1}(\beta)\right)\right| < \gamma. \tag{A.28}$$

But for the x' identified above That is,

$$\begin{aligned}F_{x'}^{-1}(\beta) &< F_{\bar{x}}^{-1}(\beta) - \epsilon \\ \text{or } F_{x'}^{-1}(\beta) &> F_{\bar{x}}^{-1}(\beta) + \epsilon\end{aligned}$$

and so given that $F_{\bar{x}}$ is non-decreasing, and by the definition of γ we must have:

$$\left|F_{\bar{x}}\left(F_{\bar{x}}^{-1}(\beta)\right) - F_{\bar{x}}\left(F_{x'}^{-1}(\beta)\right)\right| \geq \gamma$$

which contradicts (A.28). □

Recall, that for a sequence of i.i.d. random vectors $Y_1, Y_2, \ldots$ with the same distribution as Y, we define the sampled distribution function as follows:

$$F_{n,x}(z) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{f(x,Y_i)\leq z\}}.$$

The final result concerns the continuity of the sampled distribution function.

Lemma A.4. *Suppose for $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and $\bar{x} \in \mathcal{X}$ the conditions from Proposition A.1 hold. Then for all $\epsilon > 0$ there exists a neighborhood W, of $\bar{x}$, such that with probability 1*

$$\limsup_{n\to\infty} \sup_{x\in W\cap\mathcal{X}} \left| \frac{1}{n}\sum_{i=1}^{n} g(x, Y_i) - \frac{1}{n}\sum_{i=1}^{n} g(\bar{x}, Y_i) \right| < \epsilon$$

In particular, if $x \mapsto f(x, Y)$ is continuous at $\bar{x}$ with probability 1 and $F_{\bar{x}}$ is continuous at $z \in \mathbb{R}$ then for all $\epsilon > 0$ there exists a neighborhood W, of $\bar{x}$ such that with probability 1

$$\limsup_{n\to\infty} \sup_{x\in W\cap\mathcal{X}} |F_{n,x}(z) - F_{n,\bar{x}}(z)\,| < \epsilon. \tag{A.29}$$

Proof. Fix $\bar{x} \in \mathcal{X}$, and $\epsilon > 0$. Let $(\gamma_k)_{k=1}^{\infty}$ be any sequence of positive numbers converging to zero and define

$$V_k := \{x \in \mathcal{X} : \|x - \bar{x}\| \le \gamma_k\},$$
$$\delta_k(Y) := \sup_{x\in V_k} |g(x, Y) - g(\bar{x}, Y)\,|.$$

Note first that the quantity $\delta_k(Y)$ is Lebesgue measurable (see [26, Theorem 7.37] for instance). By assumption (1) the mapping $x \mapsto g(x, Y)$ is continuous at $\bar{x}$ with probability 1, hence $\delta_k(Y) \to 0$ almost surely as $k \to \infty$. Now, since $|g(x, Y)| \le h(Y)$ we must have $|\delta_k(Y)| \le 2h(Y)$, therefore, by the Lebesgue dominated convergence theorem, we have that

$$\lim_{k\to\infty} \mathbb{E}\left[\delta_k(Y)\right] = \mathbb{E}\left[\lim_{k\to\infty} \delta_k(Y)\right] = 0. \tag{A.30}$$

Note also that

$$\sup_{x\in V_k} \left| \frac{1}{n}\sum_{i=1}^{n} g(x, Y_i) - \frac{1}{n}\sum_{i=1}^{n} g(\bar{x}, Y_i) \right| \le \frac{1}{n}\sum_{i=1}^{n} \sup_{x\in V_k} |g(x, Y_i) - g(\bar{x}, Y_i)|$$

and so

$$\sup_{x\in V_k} \left| \frac{1}{n}\sum_{i=1}^{n} g(x, Y_i) - \frac{1}{n}\sum_{i=1}^{n} g(\bar{x}, Y_i) \right| \le \frac{1}{n}\sum_{i=1}^{n} \delta_k(Y_i).$$

Since the sequence of random vectors $Y_1, Y_2, \ldots$ is i.i.d. we have by the strong law of large numbers that the right-hand side of (A.31) converges with prob-

ability 1 to $\mathbb{E}\left[\delta_k(Y)\right]$ as $n \to \infty$. Hence, with probability 1

$$\limsup_{n\to\infty} \sup_{x\in V_k} \left| \frac{1}{n}\sum_{i=1}^{n} g(x, Y_i) - \frac{1}{n}\sum_{i=1}^{n} g(\bar{x}, Y_i) \right| \leq \mathbb{E}\left[\delta_k(Y)\right]. \tag{A.31}$$

By (A.30) we can pick $k \in \mathbb{N}$ such that $\mathbb{E}\left[\delta_k(Y)\right] < \epsilon$ and so setting $W = V_k$ we have by (A.31) with probability 1

$$\limsup_{n\to\infty} \sup_{x\in W\cap\mathcal{X}} \left| \frac{1}{n}\sum_{i=1}^{n} g(x, Y_i) - \frac{1}{n}\sum_{i=1}^{n} g(\bar{x}, Y_i) \right| < \epsilon.$$

The result (A.29) follows immediately as the special case $g(x, Y) = \mathbb{1}_{\{f(x,Y)\leq z\}}$. □

B Convex cone results

The results in this appendix relate to the characterization of the non-risk region for the portfolio selection problem with elliptically distributed returns. This first result allows for an exact characterization of this region for the unconstrained portfolio selection problem.

Proposition B.1. *Suppose $\alpha > 0, \mu \in \mathbb{R}^d$ and $P \in \mathbb{R}^{d\times d}$. Then, for all $y \in \mathbb{R}^d$:*

$$\left(y^T - \mu\right)\Sigma^{-1}\left(y - \mu\right) \leq \alpha^2 \iff x^T\left(y - \mu\right) \leq \|Px\|\ \alpha \qquad \forall x \in \mathbb{R}^d, \tag{A.32}$$

where $\Sigma = P^T P$.

Proof. Assume without loss of generality that $\mu = 0$. So we have to prove:

$$y^T\Sigma^{-1}y \leq \alpha^2 \iff x^T y \leq \sqrt{x^T\Sigma x}\ \alpha\ \forall x \in \mathbb{R}^d. \tag{A.33}$$

We first prove the forward implication. We do this by proving the converse, that is, we suppose for some $y \in \mathbb{R}^d$ that there exists $\tilde{x} \in \mathbb{R}^d$ such that $\tilde{x}^T y > \|P\tilde{x}\|\ \alpha$. First, set $y_0 = \frac{\Sigma\tilde{x}\alpha}{\|P\tilde{x}\|} = \frac{\Sigma\tilde{x}\alpha}{\sqrt{\tilde{x}^T\Sigma\tilde{x}}}$. Now,

$$y_0\Sigma^{-1}y_0 = \frac{\tilde{x}^T\Sigma^T\Sigma^{-1}\Sigma\tilde{x}\alpha^2}{\tilde{x}^T\Sigma\tilde{x}} = \alpha^2$$

and

$$\begin{aligned}\tilde{x}^T y_0 &= \tilde{x}^T \frac{\Sigma \tilde{x} \alpha}{\sqrt{\tilde{x}^T \Sigma \tilde{x}}} \\ &= \sqrt{\tilde{x}^T \Sigma \tilde{x}}\ \alpha \\ &= \|P\tilde{x}\|\ \alpha\end{aligned}$$

That is, y_0 satisfies the inequalities of this Proposition with equality. Note that we also have,

$$(y - y_0)^T \Sigma^{-1} (y - y_0) > 0 \qquad \text{since } \Sigma^{-1} \text{ is positive definite.}$$

Expanding and rearranging this expression we have,

$$\begin{aligned}& y^T\Sigma^{-1}y - 2y_0^T\Sigma^{-1}y + y_0\Sigma^{-1}\tilde{y_0} > 0 \\ \Leftrightarrow \quad & y^T\Sigma^{-1}y - 2\frac{\alpha}{\sqrt{\tilde{x}^T\Sigma\tilde{x}}}\tilde{x}^T\Sigma\Sigma^{-1}y + \alpha^2 > 0 \\ \Leftrightarrow \quad & y^T\Sigma^{-1}y - 2\frac{\alpha}{\sqrt{\tilde{x}^T\Sigma\tilde{x}}}\tilde{x}^T y + \alpha^2 > 0 \\ \Leftrightarrow \quad & y^T\Sigma^{-1}y - 2\frac{\alpha}{\sqrt{\tilde{x}^T\Sigma\tilde{x}}}\tilde{x}^T y + \alpha^2 > 0 \\ \Rightarrow \quad & y\Sigma^{-1}y > \alpha^2 \qquad \text{since } \tilde{x}^T y > \|Px\|\ \alpha,\end{aligned}$$

as required.

We now prove the backwards implication. We again do this by proving the converse, in this case, that if $y^T\Sigma^{-1}y > \alpha^2$ then there exists $\tilde{x} \in \mathbb{R}^d \setminus \{0\}$ such that $\tilde{x}^T y > \sqrt{x^T\Sigma x}\alpha$.

Let $\tilde{x} = \Sigma^{-1}y$. Now,

$$\begin{aligned}\tilde{x}^T y &= y^T\Sigma^{-1}y \\ &= \underbrace{\sqrt{y^T\Sigma^{-1}y}}_{=\sqrt{\tilde{x}^T\Sigma\tilde{x}}} \underbrace{\sqrt{y^T\Sigma^{-1}y}}_{>\alpha} \\ &> \sqrt{\tilde{x}^T\Sigma\tilde{x}}\ \alpha \\ &= \|Px\|\ \alpha,\end{aligned}$$

as required. □

The following two propositions give properties about projections onto convex cones which are required in the proof of the main results of this appendix.

Proposition B.2. *Suppose $K \subset \mathbb{R}^d$ is a convex cone, then, for all $y \in \mathbb{R}^d$:*

$$p_K(y)^T \left(y - p_K(y)\right) = 0$$

Proof. First note that we must have $p_K(y)^T y \geq 0$. If this is not the the case then

$$\begin{aligned} \|y - p_K(y)\|^2 &= \|p_K(y)\|^2 - 2p_K(y)^T y + \|y\|^2 \\ &> \|y\|^2 = \|y - 0\|^2 \end{aligned}$$

which contradicts the definition of $p_K(y)$ since $0 \in K$. Now assume that $p_K(y)^T \left(y - p_K(y)\right) \neq 0$, and set $\tilde{x} = \frac{p_K(y)^T y}{\|p_K(y)\|^2} p_K(y) \in K$. Now,

$$\begin{aligned} p_K(y)^T(\tilde{x} - y) &= p_K^T y - p_K^T y \\ &= 0. \end{aligned}$$

By assumption $p_k^T y \neq \|p_K(y))\|^2$, and so $\tilde{x} \neq p_K(y)$, hence

$$\begin{aligned} \|p_K(y) - y\|^2 &= \|(p_K(y) - \tilde{x}) + (\tilde{x} - y)\|^2 \\ &= \|(p_K(y) - \tilde{x})\|^2 - 2\underbrace{(p_K(y) - \tilde{x})^T(\tilde{x} - y)}_{=0} + \|(\tilde{x} - y)\|^2 \\ &> \|(\tilde{x} - y)\|^2 \end{aligned}$$

which, again, contradictions the definition of $p_K(y)$ since $\tilde{x} \in K$. □

Proposition B.3. *Let $K \subset \mathbb{R}^d$ be a convex cone and $x \in K$. Then for any $y \in \mathbb{R}^d$*

$$x^T y \leq x^T p_K(y).$$

Proof. The result holds trivially if $y \in K$ so we assume $y \notin K$. Assume there exists $\tilde{x} \in K$ such that $\tilde{x}^T y > \tilde{x}^T p_K(y)$. For all $0 \leq \lambda \leq 1$ we have

$\lambda x + (1-\lambda)p_K(y) \in K$. Now,

$$\begin{aligned}
&\|(\lambda\tilde{x} + (1-\lambda)p_K(y)) - y\|^2 - \|y - p_K(y)\|^2 \\
&\qquad = \|\lambda(\tilde{x} - p_K(y)) + (p_K(y) - y)\|^2 - \|y - p_K(y)\|^2 \\
&\qquad = \lambda^2 \|\tilde{x} - p_K(y)\|^2 + 2\lambda(\tilde{x} - p_K(y))^T(p_K(y) - y) \\
&\qquad = \lambda^2 \|\tilde{x} - p_K(y)\|^2 - 2\lambda \underbrace{\tilde{x}^T(y - p_K(y))}_{>0 \text{ by assumption}}.
\end{aligned}$$

That is, for $0 < \lambda < \frac{\tilde{x}^T(y-p_K(y))}{2\|p_K(y)-\tilde{x}\|}$ we have $\|\lambda\tilde{x} + (1-\lambda)p_K(y) - y\| < \|y - p_K(y)\|$ which contradicts the definition of $p_K(y)$. □

The next two results generalize Proposition B.1 to the case where $x \in \mathbb{R}^d$ is restricted to a convex cone. The first describes the region in the case where $P = I$, and the second generalizes the result to any non-singular matrix. In particular, it is Corollary B.5 that allows us to characterize the maximal non-risk region of portfolio selection problem for a convex feasible region.

Theorem B.4. *Let $\mathcal{X} \subset \mathbb{R}^d$ be convex, and let*

$$\mathcal{A} := \{y : x^T y \leq \|x\| \ \alpha \ \forall x \in \mathcal{X}\}$$

and

$$\mathcal{B} := \{y : \|p_K(y)\| \leq \alpha\}$$

where $K = \text{conic}(\mathcal{X})$. Then, $\mathcal{A} = \mathcal{B}$.

Proof. $(\mathcal{B} \subseteq \mathcal{A})$
Suppose $y \in \mathcal{B}$ and let $x \in \mathcal{X}$, then $x \in K$ and so

$$\begin{aligned}
x^T y &\leq x^T p_K(y) && \text{by Proposition B.3} \\
&\leq \|x\| \ \|p_K(y)\| && \text{by the Cauchy-Schwartz inequality} \\
&\leq \|x\| \ \alpha && \text{since } y \in \mathcal{B}.
\end{aligned}$$

Hence $y \in \mathcal{A}$.
$(\mathcal{A} \subseteq \mathcal{B})$

Suppose $y \notin \mathcal{B}$ and set $x = p_K(y) \in K$. Now,

$$\begin{aligned} x^T y &= p_K(y)^T y \\ &= p_K(y)^T p_K(y) + p_K(y)^T (y - p_K(y)) \\ &= p_K(y)^T p_K(y) \qquad \text{by Proposition B.2} \\ &\geq \|x\| \ \alpha \qquad \text{since } y \notin \mathcal{B}. \end{aligned}$$

Since $\mathcal{X}$ is convex we have $x = \lambda \bar{x}$ for some $\bar{x} \in \mathcal{X}$ and so we must also have $\bar{x}^T y > \|\bar{x}\| \alpha$, hence $y \notin \mathcal{A}$. □

The projection of a point onto a cone, used in the characterization above, is illustrated in Figure A.8.

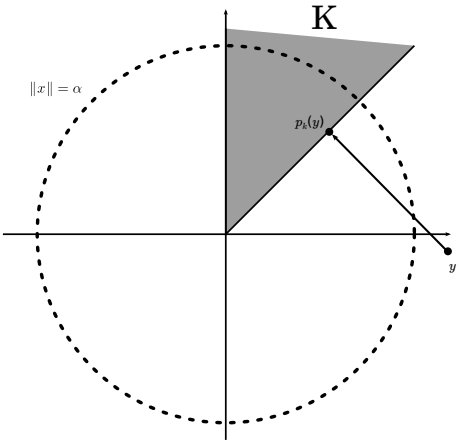


Fig. A.8: Projection onto a convex cone

Corollary B.5. *Suppose K is a convex cone, and $P \in \mathbb{R}^{d \times d}$ is a non-singular matrix. Let,*

$$\mathcal{A} := \{y : x^T y \leq \|Px\| \ \alpha \ \forall x \in K\}$$

and

$$\mathcal{B} := P^T \left(\{\tilde{y} : \|p_{K'}(\tilde{y})\| \leq \alpha\}\right)$$

where $K' = PK$. Then, $\mathcal{A} = \mathcal{B}$.

Proof.

$$
\begin{aligned}
\mathcal{B} &= P^T\left(\{\tilde{y} : \|p_{K'}(\tilde{y})\| \leq \alpha\}\right) \\
&= P^T\left(\{\tilde{y} : \tilde{x}^T\tilde{y} \leq \|\tilde{x}\| \ \alpha \ \forall \tilde{x} \in K'\}\right) \qquad \text{by Theorem B.4} \\
&= \{y : \tilde{x}^T(P^T)^{-1}y \leq \sqrt{\tilde{x}^T\tilde{x}}\alpha \ \forall \tilde{x} \in K'\} \\
&= \{y : x^T P^T(P^T)^{-1}y \leq \|Px\| \ \alpha \ \forall x \in K\} \\
&= \{y : x^T y \leq \|Px\| \ \alpha \ \forall x \in K\} \\
&= \mathcal{A}
\end{aligned}
$$

□

References

[1] A. J. King and S. W. Wallace, *Modeling with Stochastic Programming*, ser. Springer Series in Operations Research and Financial Engineering. Springer, 2012.

[2] P. Kall and S. W. Wallace, *Stochastic Programming*. Chichester: John Wiley & Sons, 1994.

[3] J. R. Birge and F. Louveaux, *Introduction to stochastic programming*. New York: Springer-Verlag, 1997.

[4] M. Kaut and S. W. Wallace, "Evaluation of scenario-generation methods for stochastic programming," *Pacific Journal of Optimization*, vol. 3, no. 2, pp. 257–271, 2007.

[5] P. Jorion, *Value at Risk: The New Benchmark for Controlling Market Risk*. Irwin Professional, 1996.

[6] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *The Journal of Risk*, vol. 2, no. 3, pp. 21–41, 2000.

[7] A. J. King and R. T. Rockafellar, "Asymptotic theory for solutions in statistical estimation and stochastic programming," *Math. Oper. Res.*, vol. 18, no. 1, pp. 148–162, 1993.

[8] A. Shapiro, "Monte Carlo sampling methods," in *Stochastic Programming*, ser. Handbooks in Operations Research and Management Science, A. Ruszczyński and A. Shapiro, Eds. Amsterdam: Elsevier Science B.V., 2003, vol. 10, ch. 6, pp. 353–425.

[9] J. Linderoth, A. Shapiro, and S. Wright, "The empirical behavior of sampling methods for stochastic programming," *Annals of Operations Research*, vol. 142, no. 1, pp. 215–241, 2006.

[10] W. Mak, D. Morton, and R. Wood, "Monte Carlo bounding techniques for determining solution quality in stochastic programs," *Operations Research Letters*, vol. 24, pp. 47–56, 1999.

[11] G. C. Pflug, "Scenario tree generation for multiperiod financial optimization by optimal discretization," *Mathematical Programming*, vol. 89, no. 2, pp. 251–271, 2001.

[12] H. Heitsch and W. Römisch, "Scenario tree reduction for multistage stochastic programs," *Computational Management Science*, vol. 6, no. 2, pp. 117–133, 2009.

[13] K. Høyland and S. W. Wallace, "Generating scenario trees for multistage decision problems," *Management Science*, vol. 47, no. 2, pp. 295–307, 2001.

[14] H. Markowitz, "Portfolio selection," *Journal of Finance*, vol. 7, pp. 77–91, 1952.

[15] K. Høyland, M. Kaut, and S. W. Wallace, "A heuristic for moment-matching scenario generation," *Computational Optimization and Applications*, vol. 24, no. 2–3, pp. 169–185, 2003.

[16] M. Kaut and S. W. Wallace, "Shape-based scenario generation using copulas," *Computational Management Science*, vol. 8, no. 1–2, pp. 181–199, 2011.

[17] R. García-Bertrand and R. Mínguez, "Iterative scenario based reduction technique for stochastic optimization using conditional value-at-risk," *Optimization and Engineering*, vol. Online First, 2012.

[18] D. Tasche, "Expected shortfall and beyond," *Journal of Banking & Finance*, vol. 26, no. 7, pp. 1519–1533, 2002.

[19] P. Billingsley, *Probability and Measure*, 3rd ed. New York, NY: Wiley, 1995.

[20] P. Artzner, F. Delbaen, J. Eber, and D. Heath, "Coherent measures of risk," *Mathematical Finance*, vol. 9, no. 3, pp. 203–228, 1999.

[21] W. Ogryczak and A. Ruszczyński, "Dual stochastic dominance and related mean-risk models," *SIAM J. Optim.*, vol. 13, no. 1, pp. 60–78 (electronic), 2002.

[22] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *Journal of Banking & Finance*, vol. 26, no. 7, pp. 1443–1471, 2002.

[23] C. Acerbi and D. Tasche, "On the coherence of expected shortfall," *Journal of Banking & Finance*, vol. 26, no. 7, pp. 1487–1503, 2002.

[24] J. L. Higle, "Variance reduction and objective function evaluation in stochastic linear programs," *INFORMS Journal on Computing*, vol. 10, no. 2, pp. 236–247, 1998.

[25] R. Serfling, *Approximation Theorems of Mathematical Statistics*, ser. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 1980. [Online]. Available: https://books.google.co.uk/books?id=eIXGaQP6qLsC

[26] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, ser. MPS-SIAM Series on Optimization. Philadelphia: SIAM, 2009, vol. 9.

[27] G. Grimmett and D. Stirzaker, *Probability and Random Processes*, ser. Probability and Random Processes. OUP Oxford, 2001. [Online]. Available: https://books.google.co.uk/books?id=G3ig-0M4wSIC

[28] K.-T. Fang, S. Kotz, and K. W. Ng, *Symmetric Multivariate and Related Distributions (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 11 1989.

[29] B. Kaynar, Ş. I. Birbil, and J. Frenk, "Application of a general risk management model to portfolio optimization problems with elliptical distributed returns for risk neutral and risk averse decision makers," Erasmus Research Institute of Management, Tech. Rep., 2007.

[30] M. Ujvári, "On the projection onto a finitely generated cone," *Acta Cybernetica*, 2007.

[31] J. Fairbrother, A. Turner, and S. W. Wallace, "Scenario generation for portfolio selection problems with tail risk measure," ArXiv e-print 1511.04935, Nov. 2015. [Online]. Available: http://arxiv.org/abs/1511.04935

Paper B

Scenario Generation for Portfolio Selection with Tail Risk Measures

Jamie Fairbrother, Amanda Turner, Stein W. Wallace

1 Introduction

In the portfolio selection problem one must decide how to invest in a collection of financial instruments with uncertain returns which in some way balances one's expected profit of the investment against its risk. In the typical set-up the uncertain returns are modeled by random variables, the total return of a portfolio is some linear combination of these, and riskiness is measured by a real-valued function of the total return which should in some way penalize potential large losses. This approach was first proposed by Markowitz [1] who used variance as a risk measure.

The use of variance as a measure of risk is problematic for a few reasons. The foremost of these is perhaps that variance penalizes large profits as well as large losses. As a consequence, in the case where the returns of financial assets are not normally distributed, using the variance can lead to patently bad decisions; for instance, a portfolio can be chosen in favor of one which always has higher returns (see [2] for an example of this). This particular issue can be overcome by using a "downside" risk measure, that is one which only depends on losses greater than the mean, or some other specified threshold, for example the semi-variance [3, Chapter 9], mean regret [4], or value-at-risk [5]. More recently, much research has been given to coherent risk measures, a concept introduced in [6]. These are risk measures which have sensible properties such as subadditivity, which in particular ensures that a risk measure incentivizes diversification of a portfolio. Using a coherent risk measure in a portfolio selection problem should avoid flawed decisions, such as the one cited in the case of variance.

In this work, we are interested in portfolio selection problems involving *tail risk measures*. These can be thought of as risk measures which only depend on the upper tail of a distribution above some specified quantile. A canonical example of a tail risk measure is the value-at-risk (VaR) [5]. The β-VaR is defined to be the β-quantile of a random variable. In portfolio selec-

tion problems this has the appealing interpretation as the amount of capital required to cover up to $\beta \times 100\%$ of potential losses. Thus, tail risk measures, in particular those which dominate the β-VaR, are useful as they can give us some idea of the amount capital at risk in the worst $(1-\beta) \times 100\%$ of potential losses. Like variance, the value-at-risk is also problematic as it is not a coherent measure of risk. Specifically, it is not subadditive (see [7] for example). Moreover, β-VaR leads to difficult and intractable problems when used in an optimization context. The conditional value-at-risk (CVaR), sometimes referred to as the expected shortfall, is another tail risk measure and can be roughly thought as the conditional expectation of a random variable above the β-VaR. It is both coherent [8], and more tractable in an optimization setting [9].

However, the use of risk measures, even coherent ones such as β-CVaR, is still problematic in portfolio selection problems where the asset returns are modeled with continuous probability distributions. This is because the evaluation of many risk measures for arbitrary continuously distributed returns would involve the evaluation of multidimensional integrals, and this becomes computationally infeasible when our problems involve many assets. On the other hand, the evaluation of such an integral reduces to a summation if the returns have a discrete distribution.

Scenario generation is the construction of a finite discrete distribution to be used in a stochastic optimization problem. This may involve fitting a parametric model to asset returns and then discretizing this distribution, or directly modeling them with a discrete distribution, for example via moment-matching [10]. In either case, standard scenario generation methods struggle to adequately represent the uncertainty in problems using tail risk measures. This is because the value of a tail risk measure, by definition, only depends on a small subset of the support of a random variable, and typical scenario generation methods will spread their scenarios evenly across the whole support of the distribution. This means that the region on which the value of the

tail risk depends, is represented by relatively few scenarios. Hence, unless there are a very large number of scenarios, the value of of tail risk measure is very unstable (see [11] for example).

The natural remedy to this problem is to represent the regions of the distribution on which the tail risk measure depends with more scenarios. Intuition would tell us that these correspond to the "tails" of the distribution. However, for a multivariate distribution there is no canonical definition of the tails. If by tails, we simply mean the region where at least one of the components exceeds a large value, then the probability of this region quickly converges to one with the problem dimension, and thus prioritizing scenarios in this region will be of little benefit. Finding the relevant tails of the distribution is a non-trivial problem.

In the paper [12] we addressed the problem of scenario generation for general stochastic programs using tail risk measures, and for this we defined the concept of a β-risk region. In portfolio selection, to each valid portfolio there is a distribution of losses (or returns). The β-risk region consists of all potential asset returns which lead to a loss in the β-tail for some portfolio. We have shown that the value of a tail risk measure in effect only depends on the distribution of returns in the risk region. Although characterizing this region in a convenient way is generally not possible, we have been able to do this for the portfolio selection problem when the asset returns are elliptically distributed. We have also proposed a sampling approach to scenario generation using these risk regions which prioritizes the generation of scenarios in the risk region. We demonstrated for simple examples that this methodology can produce scenario sets which yield better and more stable solutions than does basic sampling. In Sections 2 and 4 we review respectively the requisite theory of risk regions, and how risk regions can be used in scenario generation.

In this paper we address issues related to the application of this methodology to realistic portfolio selection problems. Firstly, we deal with how

problem constraints are used to calculate the risk region. In [12] we showed that for elliptically distributed returns, the risk region depends on the conic hull of the feasible region but we only did calculations for the case where this is the positive quadrant, that is, we only use the constraint of no short-selling. In practice, one may wish to impose other constraints on portfolios, such as quotas on the amount one can invest an asset or industry. In Section 3 we describe how the conic hull of the feasible region can be calculated from linear constraints, and how this is used to test whether or not a point lies in the risk region.

The effectiveness of the presented methodology depends directly on the probability of the risk region. In effect, the smaller the probability of the risk region, the more redundant scenarios we can discard. In [12] we observed that the probability of this region tends to one as the problem dimension increases. In Section 5 we make some more general observations on how this probability varies with distribution, and in particular observe that for distributions with heavy tails and positive correlations, characteristics of typical stock return data, the probability of the risk region is low.

In practice it may not be appropriate to model asset returns with elliptical distributions as these are likely to exhibit non-elliptical features such as skewness [13]. Moreover, when the asset returns have an elliptical distribution, the portfolio selection problem may be solvable by more efficient methods [14]. In Section 6 we test our methodology for a variety of distributions constructed from real data. We calculate the probability of the risk region for a range of constraints, and test the performance of our methodology for scenario generation and scenario reduction. We demonstrate here that when asset returns are near-elliptical in distribution, we can approximate its risk region with the risk region of an elliptical distribution to good effect.

Finally, in Section 7 we demonstrate for a difficult case study problem how our methodology can be applied. In particular we demonstrate how the addition of artificial constraints to the problem can be used to find better

solutions.

2 Portfolio selection and risk regions

In this section we recall the requisite concepts and results from our previous paper [12]. In particular we define the risk region for the portfolio selection problem and give a convenient characterization of this when asset returns have elliptical distributions.

2.1 Tail risk measures and risk regions

As mentioned above, a risk measure is a function of a real-valued random variable representing a loss. For $0 < \beta \leq 1$, a β-tail risk measure can be thought of as a function of a random variable which depends only on the upper $(1-\beta)$-tail of the distribution. The precise definition uses the *generalized inverse distribution function* or *quantile function*.

Definition 2.1 (Quantile function and β-tail risk measure)**.** *Suppose Z is a random variable with distribution function F_Z. Then the generalized inverse distribution function, or* quantile function *is defined as follows:*

$$\begin{aligned} F_Z^{-1} : (0,1] &\to \mathbb{R} \cup \{\infty\} \\ \beta &\mapsto \inf\{z \in \mathbb{R} : F_Z(x) \geq \beta\} \end{aligned}$$

Now a β-tail risk measure is any function of a random variable, $\rho_\beta(Z)$, which depends only on the quantile function of a random variable above β.

Example 2.2 (Value at risk (VaR))**.** *Let Z be a random variable, and $0 < \beta < 1$. Then, the $\beta-$VaR for Z is defined to be the β-quantile of Z:*

$$\beta\text{-VaR}(Z) := F_Z^{-1}(\beta)$$

.

Example 2.3 (Conditional value at risk (CVaR))**.** *Let Z be a random variable, and $0 < \beta < 1$. Then, the β-*CVaR *can be thought roughly as the conditional expectation of a random variable above its β-quantile. The following alternative characterization of β-*CVaR *[8] shows directly that it is a β-tail risk measure.*

$$\beta\text{-CVaR}(Z) = \int_{\beta}^{1} F_Z^{-1}(u)\, du$$

The observation that we exploit for this work is that very different random variables will have the same β-tail risk measure as long as their β-tails are the same. Such a situation is illustrated in Figure B.1 for two discrete random variables.

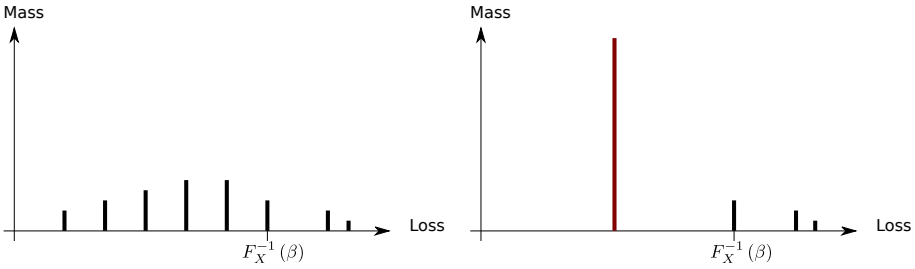


Fig. B.1: Two very different random variables with identical β-tails

In this paper we are interested in portfolio selection problems which use β-tail risk measures. We use the following basic set-up: we have a set of financial assets indexed by $i = 1, \ldots, d$, by x_i we denote how much we invest in asset i, and by Y_i we denote the random future return of asset i. The profit associated to a particular investment decision $x = (x_1, \ldots, x_d)$ and return $Y = (Y_1, \ldots, Y_d)$ is $x^T Y = \sum_{i=1}^{d} x_i Y_i$. The loss associated to an investment decision is thus $-x^T Y$, and so for a given β-tail risk measure ρ_β we would like an investment with small risk $\rho_\beta(-x^T Y)$. The aim of a portfolio selection problem is to choose a decision which balances choosing a portfolio with high expected profit against choosing one with small risk. This typically corresponds to solving a problem of one of the following forms:

$$\text{(i)} \quad \underset{x\in\mathcal{X}}{\text{minimize}}\ \rho_\beta(-x^TY) \tag{P1}$$
$$\text{subject to } \mathbb{E}\left[x^TY\right] \geq t,$$
$$\text{(ii)} \quad \underset{x\in\mathcal{X}}{\text{maximize}}\ \mathbb{E}\left[x^TY\right] \tag{P2}$$
$$\text{subject to } \rho_\beta(-x^TY) \leq s,$$
$$\text{(iii)} \quad \underset{x\in\mathcal{X}}{\text{minimize}}\ \lambda\rho_\beta(-x^TY) + (1-\lambda)\mathbb{E}\left[-x^TY\right], \tag{P3}$$

where $0 \leq \lambda \leq 1$ and $\mathcal{X} \subset \mathbb{R}^d$ represents the set of valid portfolios. This feasibility region will typically encompass a constraint which specifies the amount of capital to be invested, and may include others which, for example the exclusion of short-selling, or a limit on the amount that can be invested in certain industries.

In [12] we introduced the concept of a risk region for a stochastic program using a tail-risk measure. We define this now for the portfolio selection problem.

Definition 2.4 (Risk region)**.** *The β-risk region associated associated with the random vector Y and the feasible region $\mathcal{X} \subseteq \mathbb{R}^d$ is as follows:*

$$\mathcal{R}_{Y,\mathcal{X}}(\beta) := \bigcup_{x\in\mathcal{X}} \{y \in \mathbb{R}^d :\ -x^Ty \geq F^{-1}_{-x^TY}(\beta)\}. \tag{B.1}$$

The risk region consists precisely of those outcomes of Y which have a loss in the β-tail of the loss distribution for *some* feasible portfolio. We refer to the complement of the risk region as the non-risk region and this consists of outcomes which never lead to a loss in the β-tail; it can be written as follows:

$$\mathcal{R}_{Y,\mathcal{X}}(\beta)^c = \bigcap_{x\in\mathcal{X}} \{y \in \mathbb{R}^d : -x^Ty < F^{-1}_{-x^TY}(\beta)\}. \tag{B.2}$$

Note that since this set is the intersection of half-spaces, it is convex.

For a discrete distribution of returns, we can easily calculate the β-quantile of the loss for a particular portfolio by calculating the loss for all scenarios and ordering them based on this. In Figure B.2 we illustrate a scenario set of returns for two hypothetical assets, along with the line separating those scenarios with loss above and below the β-quantile for the portfolio $x = (\frac{1}{2}, \frac{1}{2})$.

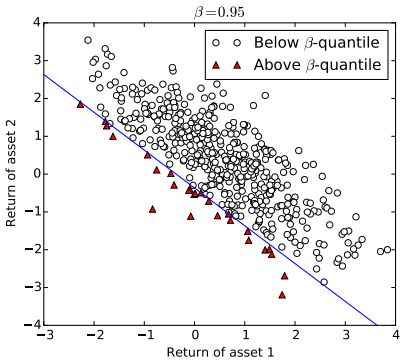


Fig. B.2: Scenarios with loss above and below β-quantile for one portfolio

The risk region is the union over all feasible portfolios of the half spaces of points with returns above the β-quantile. We can find this region by brute force, and this is illustrated on the left-hand side of Figure B.3. Also illustrated in this figure is the set of returns where all the mass in the non-risk region has been aggregated into a single scenario. We call the latter the *aggregated scenario set*. As is also shown in the figure, the β-quantile lines do not change after aggregation and so neither does the corresponding value of any β-tail risk measure. Aggregating scenarios leads to more concise optimization problems which are easier to solve.

The transformed random vector where all mass in a region has been concentrated into its conditional expectation plays a special role in this work. We call this the *aggregated random vector*.

Definition 2.5 (Aggregated Random Vector)**.** *For some set* $\mathcal{R} \supseteq \mathcal{R}_{Y,\mathcal{X}}$ *the* ag-

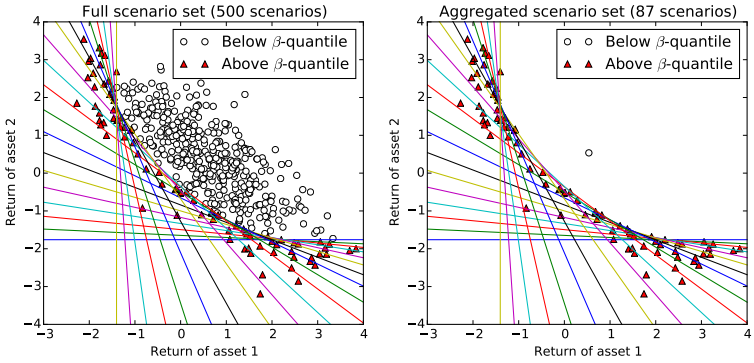


Fig. B.3: Return points with loss below the β-quantile for all non-negative portfolios (left) and aggregated scenario set (right)

gregated random vector *is defined as follows:*

$$\psi_{\mathcal{R}}(Y) := \begin{cases} Y & \text{if } Y \in \mathcal{R}, \\ \mathbb{E}\left[\, Y | Y \in \mathcal{R}^c \,\right] & \text{otherwise.} \end{cases}$$

In [12] we showed that under mild conditions the value of a tail risk measure is completely determined by the the distribution of the random vector Y in the risk region. That is, the values of the tail risk measure of any two random vectors with identical distributions in the risk region will be the same for all feasible decisions.

Theorem 2.6. *Let $\mathcal{R} \supseteq \mathcal{R}_{Y,\mathcal{X}}(\beta)$ be such that for all $x \in \mathcal{X}$ the following condition holds:*

$$\mathbb{P}\left(Y \in \{y : z' < -x^T y \leq F^{-1}_{-x^T Y}(\beta)\} \cap \mathcal{R}\right) > 0 \qquad \forall\, z' < F^{-1}_{-x^T Y}(\beta)\,. \tag{B.3}$$

If $\tilde{Y}$ is a random vector for which the following holds:

$$\mathbb{P}\left(Y \in \mathcal{A}\right) = \mathbb{P}\left(\tilde{Y} \in \mathcal{A}\right) \qquad \text{for any } \mathcal{A} \subseteq \mathcal{R}, \tag{B.4}$$

then $\rho_\beta\left(\left(f(x,Y)\right)\right) = \rho_\beta\left(f(x,\tilde{Y})\right)$ for all $x \in \mathcal{X}$, for any β-tail risk measure ρ_β.

The technical condition (B.3) precludes certain degenerate cases. If $\mathcal{R}$ is convex, we have that $\mathbb{E}\left[Y | Y \in \mathcal{R}^c\right] \in \mathcal{R}^c$ in which case the aggregated random vector defined above has, by definition, the same distribution as Y in

the risk region. The aggregated random vector has the additional property of preserving the overall expected return of the original random vector. The following corollary taken from [12] summarizes this result and provides sufficient conditions so that (B.3) holds.

Corollary 2.7. *Suppose $\mathcal{R}_{Y,\mathcal{X}}(\beta) \subseteq \mathcal{R} \subset \mathbb{R}^d$, Y is a continuous random vector with support $\mathcal{Y} = \mathbb{R}^d$, and $\mathcal{X}$ contains at least two linearly independent elements. Then Y satisfies (B.3). In addition, if $\mathcal{R}$ is convex then $\tilde{Y} = \psi_{\mathcal{R}}(Y)$ satisfies condition (B.4) and so for all $x \in \mathcal{X}$ we have:*

$$\rho_\beta\left(-x^T Y\right) = \rho_\beta\left(-x^T \tilde{Y}\right),$$
$$\mathbb{E}\left[x^T Y\right] = \mathbb{E}\left[x^T \tilde{Y}\right].$$

2.2 Risk regions for elliptical distributions

In order to exploit risk regions for scenario generation one has to be able to characterize these in a way which allows one to conveniently test whether or not a point belongs to it. In our previous paper, we were able to do this in the case where the asset returns have *elliptical distributions*. Elliptical distributions are a general class of distributions which include, among others, multivariate Normal and multivariate t-distributions. See [15] for a full overview of the subject.

Definition 2.8 (Elliptical Distribution)**.** *Let $X = (X_1, \ldots, X_d)$ be a random vector in $\mathbb{R}^d$, then X is said to be* spherical, *if*

$$X \sim UX \qquad \text{for all orthonormal matrices } U$$

where $\sim$ means the two operands have the same distribution function.

Let Y be a random vector in $\mathbb{R}^d$, then Y is said to be elliptical *if it can be written $Y = PX + \mu$ where $P \in \mathbb{R}^{d\times d}$ is non-singular, $\mu \in \mathbb{R}^d$, and X is random vector with spherical distribution. Such an elliptical distribution will be denoted* Elliptical (X, μ, P).

This definition says that a random vector with a spherical distribution is rotation invariant, and that an elliptical distribution is an affine transformation of a spherical distribution. Elliptical distributions are convenient in the context of portfolio selection as we can write down exactly the distribution of loss of a portfolio:

$$-x^T Y \sim \|Px\| X_1 - x^T \mu,$$

and so, the β-quantile of the loss $-x^T Y$ is as follows:

$$F^{-1}_{-x^T Y}(\beta) = \|Px\| F^{-1}_{X_1}(\beta) - x^T \mu.$$

For elliptically distributed returns, we can thus rewrite the risk region in (B.1) as follows:

$$\mathcal{R}_{Y,\mathcal{X}}(\beta) := \bigcup_{x \in \mathcal{X}} \{y \in \mathbb{R}^d : -x^T y \geq \|Px\| F^{-1}_{X_1}(\beta) - x^T \mu\}. \tag{B.5}$$

In this form it is still difficult to check whether a given point $\tilde{y} \in \mathbb{R}^d$ belongs to it. In [12] we provided a more convenient characterization of the risk region for elliptical returns. This characterization makes use of the conic hull of the feasible region $\mathcal{X} \subset \mathbb{R}^d$.

Definition 2.9 (Convex cones and conic hull)**.** *A set $K \subset \mathbb{R}^d$ is a cone if for all $x \in K$ and $\lambda \geq 0$ we have $\lambda x \in K$. A cone is convex if for all $x_1, x_2 \in K$ and $\lambda_1, \lambda_2 \geq 0$ we have $\lambda_1 x_1 + \lambda_2 x_2 \in K$. The conic hull of a set $\mathcal{A} \subset \mathbb{R}^d$ is the smallest convex cone containing $\mathcal{A}$, and is denoted* conic $(\mathcal{A})$.

For example, suppose that our feasible region consists of portfolios with non-negative investments (i.e. no short-selling) and whose total investment is normalized to one, that is:

$$\mathcal{X} = \{x \in \mathbb{R}^d : \sum_{i=1}^{d} x_i = 1, \;\; x_i \geq 0 \text{ for each } i = 1, \ldots, d\},$$

then the conic hull of this is the positive quadrant, that is conic $(\mathcal{X}) = \mathbb{R}^d_+$. The alternative characterization also makes use of projections.

Definition 2.10 (Projection). *Let $C \subset \mathbb{R}^d$ be a closed convex set, then for any point $y \in \mathbb{R}^d$ we define its projection onto C to be the unique point $p_C(y) \in C$ such that*

$$\inf_{x \in C} \|x - y\| = \|p_C(y) - y\|.$$

We are now ready to give a characterization of our risk region for which we use the following convenient abuse of notation: for a set $\mathcal{A} \subset \mathbb{R}^d$ and a matrix $T \in \mathbb{R}^{d \times d}$, we write $T(\mathcal{A}) := \{Ty : y \in \mathcal{A}\}$. The following result was proved in [12].

Theorem 2.11. *Suppose $Y \sim \text{Elliptical}(X, P, \mu)$, $\mathcal{X} \subseteq \mathbb{R}^d$ is convex and let $K = \text{conic}(\mathcal{X})$. Then the risk region can be characterized exactly as follows:*

$$\mathcal{R}_{Y,\mathcal{X}}(\beta) = P^T \left(\{\tilde{y} \in \mathbb{R}^d : \|p_{K'}(\tilde{y} - \mu)\| \geq F_{X_1}^{-1}(\beta)\} \right), \tag{B.6}$$

where $K' = PK$ is a linear transformation of the conic hull K.

When $K = \mathbb{R}^d$ the complement of the region in (B.6) has a convenient geometric description; it is an open ellipsoid.

3 Projections and the conic hull

The characterization of the risk region for portfolio selection problems given in (B.6) relies on one being able to calculate the conic hull of the set of feasible portfolios, and also the ability to project points onto a transformation of this. In Section 3.1 we show how one can find the conic hull of the feasible region for typical constraints of a portfolio selection problem. This conic hull is a *finitely generated cone*. In 3.2 we show how one can project points onto this type of cone.

3.1 Conic hull of feasible region

In portfolio problems, the feasible region is usually defined by linear constraints, that is $\mathcal{X} = \{x \in \mathbb{R}^d : Ax \leq b\}$, where $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$. That

is, the feasible region is the intersection of a finite number of half-spaces. It is a well-known fact that any such intersection can be written as the convex hull of a finite number of points plus the conical combination of some more points (see Theorem 1.2 in [16] for example). That is, there exists $x_1, \ldots, x_k \in \mathbb{R}^d$ and $y_1, \ldots, y_l \in \mathbb{R}^d$ such that

$$\mathcal{X} = \{\sum_{i=1}^{k} \lambda_i x_i + \sum_{j=1}^{l} \nu_j y_j : \lambda, \nu \geq 0, \ \sum_{i=1}^{k} \lambda_i = 1\}. \tag{B.7}$$

The conic hull of this region is the following *finitely generated cone*:

$$\text{conic}\,(\mathcal{X}) = \{\sum_{i=1}^{k} \lambda_i x_i + \sum_{j=1}^{l} \nu_j y_j : \lambda, \nu \geq 0\}.$$

To express the intersection of half-spaces in the form (B.7), we could use *Chernikova's algorithm* (also known as the double description method) [17], [18]. Every finitely generated cone can also be written as a *polyhedral cone*, that is, of the form $\{x \in \mathbb{R}^d : Dx \geq 0\}$, and vice versa (see [16, Chapter 1]). Chernikova's algorithm again provides a concrete method for going between these two different representations. Although these two representations are mathematically equivalent, as we shall see, they are algorithmically different.

We will suppose the constraints for our portfolio selection problem have the following form:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d : \begin{array}{rl} \mathbf{1}^T x = c & \\ a_i^T x \leq b_i & \text{for } i = 1, \ldots, m, \\ x \geq 0, & \end{array} \right\} \tag{B.8}$$

where $\mathbf{1}$ is column vector of ones and $c > 0$. The first of these constraints specifies the total of amount of capital to be invested, the inequalities represent other constraints such as quotas on the amount one can invest in a specific company or industry. In this case, we can describe immediately the conic hull as a polyhedral cone.

Proposition 3.1. *Let $\mathcal{X}$ be the set defined in (B.8) and let*

$$\mathcal{Y} = \left\{ x \in \mathbb{R}^n : \left(\frac{b_i}{c}\mathbf{1} - a_i\right)^T x \geq 0 \text{ for } i = 1, \ldots, m,\ x \geq 0 \right\}$$

then conic $(\mathcal{X}) = \mathcal{Y}$.

Proof. Given that $\mathcal{X}$ is convex, to show that conic $(\mathcal{X}) = \mathcal{Y}$, it suffices to show that

$$x \in \mathcal{Y} \setminus \{0\} \iff \exists\ \lambda > 0 \text{ such that } \lambda x \in \mathcal{X}.$$

We demonstrate first the forward implication. Suppose $x \in \mathcal{Y} \setminus \{0\}$. Then, given that $x > 0$, we must have $v := \mathbf{1}^T x > 0$. Then, setting $\lambda = \frac{c}{v}$, we have

$$\mathbf{1}^T(\lambda x) = v\frac{c}{v} = c.$$

Since $\mathcal{Y}$ is a cone, we have $\lambda x \in \mathcal{Y}$, hence

$$\begin{aligned} & (\frac{b_i}{c}\mathbf{1} - a_i)^T \frac{c}{v} x \ \geq\ 0 \\ \therefore \quad & \frac{c}{v} a_i^T x \ \leq\ \frac{b_i}{c}\frac{c}{v}\underbrace{\mathbf{1}^T x}_{=v} \\ \therefore \quad & a_i^T(\frac{c}{v}x) \ \leq\ b_i \end{aligned}$$

and so $\lambda x \in$ conic $(\mathcal{X})$.

We now prove the backwards implication. Suppose $x \in$ conic $(\mathcal{X}) \setminus \{0\}$. Then there exists $\lambda > 0$ such that $\lambda x \in \mathcal{X}$, that is

$$\begin{aligned} \mathbf{1}^T \lambda x &= c \\ a_i^T \lambda x &\leq b_i \end{aligned}$$

Therefore,

$$\begin{aligned} & \frac{a_i^T \lambda x}{\mathbf{1}^T \lambda x} \ \leq\ \frac{b_i}{c} \\ \text{and so} \quad & \left(\frac{b_i}{c}\mathbf{1} - a_i\right)^T x \ \geq\ 0. \end{aligned}$$

Hence $x \in \mathcal{Y}$ as required. □

Figure B.4 shows how simple constraints in $\mathbb{R}^2$ affect the conic hull of the feasible region given the total investment and positivity constraints.

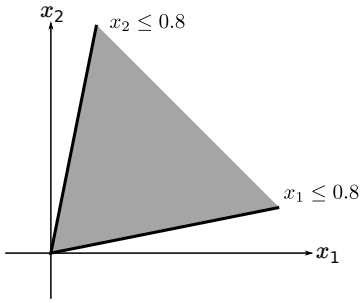


Fig. B.4: Conic hull from simple quota constraints given $x_1 + x_2 = 1$ and $x_1, x_2 \geq 0$

3.2 Projection onto a finitely generated cone

First, suppose that we can represent the conic hull of the feasible region $\mathcal{X} \subset \mathbb{R}^d$ as a finitely generated cone $K = \{Ay : y \geq 0\}$ where $A \in \mathbb{R}^{k \times d}$. By definition, the projection of a point $x_0 \in \mathbb{R}^d$ can be found by solving the following quadratic program:

$$\underset{y \geq 0}{\text{minimize}} \quad \|Ay - x_0\|_2^2 \tag{B.9}$$

In particular, if y^* is the optimal solution then $p_K(x_0) = Ay^*$. By formulating the KKT conditions [19, Chapter 5] of this problem, it can be seen that this problem is equivalent to solving the following linear complementarity problem (LCP):

$$\begin{aligned} &\text{Find } y, z \in \mathbb{R}^d \text{ such that} \\ &\qquad z - A^T A y = -A^T x_0 \\ &\qquad z^T y = 0 \\ &\qquad y, z \geqslant 0. \end{aligned}$$

If (y, z) is a solution to the above problem, then the required projection is $p_K(x_0) = Ay$. LCPs can be solved by more specialized algorithms than standard quadratic programs such as Lemke's algorithm [20].

Now, suppose instead we have a polyhedral characterization of the conic

hull, that is a cone of the form:

$$K = \{x \in \mathbb{R}^d : Bx \geq 0\}. \tag{B.10}$$

The projection of a point $x_0 \in \mathbb{R}^d$ onto the polyhedral cone in (B.10) is the solution of the following quadratic program:

$$\begin{aligned} &\underset{x}{\text{minimize}} \ \|x - x_0\|_2^2 \\ &\text{subject to } Bx \geq 0. \end{aligned}$$

Although the former problem in (B.9) can be solved using specialized algorithms, we will in practice use both approaches. For conic hulls with a small number of extremal rays, for example $K = \mathbb{R}^d_+$ we will use use the former method. As we add more constraints to the problem, we have found from experience that the number of extremal rays can exponentially increase, which for the former approach leads to cumbersomely large LCP problems. In this case we will use the polyhedral representation for projection.

4 Scenario generation

In this Section we present how risk regions can be exploited for the purposes of scenario generation. In Section 4.1 we propose two specific methods which work essentially by prioritizing the construction of scenarios in the risk region. These methods are specifically adapted to asset returns which have elliptical distributions, and so in Section 4.2 we discuss their usage for non-elliptic distributions. Finally, in Section 4.3 we discuss how the effectiveness of our methodology can be improved through the addition of artificial constraints to our problem.

4.1 Aggregation sampling and reduction

In this section we will assume that asset returns have elliptical distributions from which we can sample. In [12] we proposed two methods to exploit risk

regions. The first of these allows the user to specify the final number of scenarios in advance. The algorithm, which we call *aggregation sampling*, samples scenarios, aggregating all samples in the non-risk region and keeping all in the risk region, until we have the required number of risk scenarios. This is described in Algorithm 1.

Let $q = \mathbb{P}\left(Y \in \mathcal{R}^c_{Y,\mathcal{X}}\right)$ be the probability of the non-risk region, and n be the number of risk scenarios required. Define $N(n)$ to be the effective sample size from aggregation sampling, that is, the number of draws until the algorithm terminates[1]. The quantity $N(n)$ is a random variable:

$$N(n) \sim n + \mathcal{NB}(n, q),$$

where $\mathcal{NB}(N, q)$ denotes a *negative binomial* random variable. Recall that a negative binomial random variable $\mathcal{NB}(n, q)$ is the number of failures in a sequence of Bernoulli trials with probability of success q until n successes have occurred. The expected effective sample size of aggregation sampling is thus as follows:

$$\mathbb{E}\left[N(n)\right] = n + n\frac{q}{1-q}$$

The expected effective sample size can be thought of as the sample size required for basic sampling to produce the same number of scenarios in the risk region. The difference between the desired number of risk scenarios, and expected effective sample size is proportional to the ratio $\frac{q}{1-q}$. In particular, as the probability of the non-risk region approaches one, this gain tends to infinity.

The converse to aggregation sampling is sampling a set of a given size n and then aggregating all scenarios in the risk region of the underlying distribution. We call this *aggregation reduction*. This can be viewed as a sequence of n Bernoulli trials, where success and failure are defined in the same way as described above. The number of scenarios in the reduced sample, $R(n)$ is

[1] For simplicity of exposition we discount the event that the while-loop of the algorithm terminates with $n_{\mathcal{R}^c} = 0$ which occurs with probability q^n

input : $N_{\mathcal{R}}$ number of required risk scenarios, β level of tail risk measure, K conic hull of feasible region, (X, P, μ) elliptical distribution

output: $\{(y_s, p_s)\}_{s=1}^{N_{\mathcal{R}}+1}$ scenario set

$n_{\mathcal{R}^c} \leftarrow 0$, $n_{\mathcal{R}} \leftarrow 0$, $y_{\mathcal{R}^c} \leftarrow 0$;

$K' \leftarrow PK$;

while $n_{\mathcal{R}} < N_{\mathcal{R}}$ **do**

  Sample new point y;

  $y_{\text{trans}} \leftarrow P^{-T}(y - \mu)$;

  if $\| p_{K'}(y_{trans}) \| \leq F_{X_1}^{-1}(\beta)$ **then**

    $y_{\mathcal{R}^c} \leftarrow \frac{1}{n_{\mathcal{R}^c}+1}(n_{\mathcal{R}^c} y_{\mathcal{R}^c} + y)$;

    $n_{\mathcal{R}^c} \leftarrow n_{\mathcal{R}^c} + 1$;

  end

  else

    $y_{n_{\mathcal{R}}} \leftarrow y$;

    $n_{\mathcal{R}} \leftarrow n_{\mathcal{R}} + 1$;

  end

end

foreach i *in* $1, \ldots, N_{\mathcal{R}}$ **do** $p_i \leftarrow \frac{1}{n_{\mathcal{R}^c}+n_{\mathcal{R}}}$;

;

if $n_{\mathcal{R}}^c \geq 1$ **then**

  $y_{N_{\mathcal{R}}+1} \leftarrow y_{\mathcal{R}^c}$, $p_{N_{\mathcal{R}}+1} \leftarrow \frac{n_{\mathcal{R}^c}}{n_{\mathcal{R}^c}+n_{\mathcal{R}}}$;

end

else

  Sample new point y;

  $y_{N_{\mathcal{R}}+1} \leftarrow y$, $p_{N_{\mathcal{R}}+1} \leftarrow \frac{1}{N_{\mathcal{R}}+1}$;

end

Algorithm 1: Aggregation sampling algorithm for risk region of an elliptical distribution

as follows:

$$R(n) \sim n - \mathcal{B}(n, q) + 1$$

where $\mathcal{B}(n, q)$ denotes a binomial random variable. The expected reduction in scenarios in aggregation reduction is thus $nq - 1$.

The reason that aggregation sampling and aggregation reduction work is that, for large samples, they are equivalent to sampling from the aggregated random vector, and Corollary 2.7 tells us that this random vector has the correct tail risk measure and expectation. See [12, Section 4] for a full proof of the consistency of these algorithms.

In the above methods, for every sampled point we must be able to test whether the magnitude of its projection onto a cone is above or below a given threshold. As explained in Section 3.2, the projection of a point onto a finite cone involves solving a small LCP or quadratic program and so for large sample sizes and high dimensions this will become computationally expensive. However, given that each sample is independent of every other, this algorithm is naturally parallelizable. It may also be possible to make the algorithm more efficient if for any point $y \in \mathbb{R}^d$, a way could be found of testing the condition $\|p_K(y)\| \leq \alpha$ directly without calculating the full projection $p_K(y)$. For example, the quadratic program used to calculate the projection could be solved only to an accuracy sufficient to test this condition.

4.2 Approximation of risk regions

The above algorithm is specifically adapted to risk regions of elliptically distributed returns. However, the utility of using our scenario generation methodology with only elliptical distributions is limited. Firstly, it may be unrealistic to model returns with elliptical distributions. Real financial returns may exhibit properties which elliptical distributions do not, such as skewness. Secondly, using elliptical distributions may allow one to formulate the optimization problem in a more convenient way. For example, when

using the β-CVaR as a tail-risk measure, we have the following relation:

$$\beta\text{-CVaR}\left(-x^T Y\right) = -\|Px\|\,\beta\text{-CVaR}\left(X_1\right) - x^T\mu$$

which may mean we can solve the optimization problems with interior point or quadratic programming algorithms. See [14] for more details.

In this work, we put forward the idea that risk regions of elliptical distributions can be used for aggregation sampling and reduction, for distributions which are near-elliptical. We propose to use the risk region of an elliptical distribution as a surrogate for the risk region of a near-elliptical distribution.

The danger of approximating the risk region is that if for a particular decision, the β-quantile, is attained inside the surrogate non-risk region (that is the surrogate risk region is too small), then the value of the tail-risk measure may be distorted. On the other hand, if the surrogate risk region contains the true risk region (that is, the surrogate risk region is too large) then Corollary 2.7 guarantees that the associated aggregated random vector has the correct tail risk measure. However, we should be cautious about constructing a surrogate risk region which is excessively large. If this is the case then the probability of the surrogate may be very large, which means there will be little benefit in aggregation.

Through a careful probabilistic analysis of the distribution of returns for valid portfolios, one may be able to construct a surrogate risk region which tightly covers the true risk region. If this is not possible, one way to mitigate against the danger of using a surrogate risk region which is too small would be to represent the non-risk region with several points rather than a single point. For example, instead of aggregating all sampled points in the non-risk region, one could apply a clustering algorithm to these such as k-means. For simplicity, we will only test the basic aggregation methods which represents the non-risk region with a single point. For the non-elliptical distributions used in this paper we are able to rely on heuristic rules to construct our surrogates.

The first class of non-elliptical distributions we use in this paper are known as multivariate Skew-t distributions [21]. This class of distributions generalizes the elliptical multivariate t-distributions through the inclusion of an extra set of parameters which regulate the skewness. In this case we approximate the risk region with the risk region of the corresponding t-distribution.

The second class we use are discrete distributions constructed using the moment matching algorithm of [10]. These distributions have been applied previously to financial problems [11]. This algorithm constructs scenario sets with a specified correlation matrix and whose marginals have specified first four moments. This algorithm works by first taking a sample from a multivariate Normal distribution, and then iteratively applying transformations to this until the difference between its marginal moments and correlation matrix are sufficiently close to their target values. Since the algorithm is initialized with a sample from an elliptical distribution, the final distribution is near elliptical and we approximate the risk region for these distributions with the risk region of a multivariate normal distribution with the same mean and covariance structure.

4.3 Ghost constraints

We noted above that the performance of our methodology improves as the probability of the non-risk region decreases. In particular, the expected effective sample size in aggregation sampling increases as the probability of the non-risk risk region increases. Now, by its definition (B.2) the non-risk regions shrinks as the problem becomes more constrained. This suggests that it may be helpful to add constraints to our problem which shrink the set of feasible portfolios, but which are not themselves active, in the sense that their presence does not affect the set of optimal solutions. We will refer to a constraint added to a problem to boost the performance of our methodology, loosely, as a *ghost constraint*.

Finding non-active constraints to add to our problem is non-trivial as it relies on some knowledge of the optimal solution set. We will resort to heuristic rules to choose ghost constraints. For example, one could constrain our set of feasible portfolios to some neighborhood of a good quality solution.

Verifying whether or not a ghost constraint is active is difficult in general for stochastic programs. For a deterministic objective function which is convex and for which all constraints are convex (and the optimal solution is unique) a constraint $\{x : g(x) \leq 0\}$ is active if and only if it is binding at the optimal solution x^*, that is $g(x^*) = 0$. For a stochastic program, we are typically solving a scenario-based approximation and so a constraint which is not binding with respect to the scenario-based approximation may be binding with respect to the true problem and vice versa.

A rigorous test of whether a ghost constraint is active in the sense above is beyond the scope of this paper. We simply promote the idea here that ghost constraints may be a useful way of finding better solutions. We demonstrate the usage of ghost constraints on a difficult problem in Section 7.

5 Probability of the non-risk region

The benefit of aggregation sampling and reduction depends on the probability of the non-risk region. As was observed in [12] the probability of the non-risk region tends to decrease as the problem dimension increases, but increases as we tighten our problem constraints, and as we increase β, the level of the tail risk measure. In this section we make some empirical observations on how this probability varies with heaviness of the tails, and the correlations of the distribution.

The first observation is that in the presence of positivity constraints, the probability of the risk region improves as the the correlation between random variables increases. This can be seen in Figure B.5 which plots the probability of the risk region as a function of correlation for some two-dimensional

distributions. An intuitive explanation for this type of behavior is that in the case of positive correlations there is much more overlap in the risk regions of the individual portfolios.

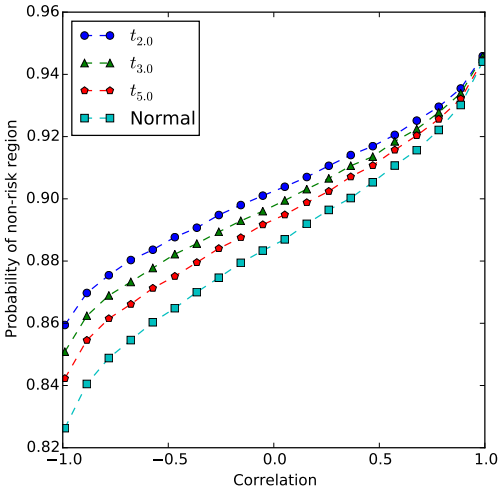


Fig. B.5: Correlation vs. Probability of non-risk region for some 2-dimensional elliptical distributions, positivity constraints and $\beta = 0.95$

The extent to which probabilities vary with correlation seems to be much greater in higher dimensions. In Figure B.6 we have plotted for Normal returns and a range of dimensions, the probabilities of the non-risk region for a particular type of correlation matrix: $\Lambda(\rho) \in \mathbb{R}^{d\times d}$ where $\Lambda(\rho)_{ij} = \rho$ for $i \neq j$ and $\rho > 0$. In the case of $\rho = 0$, the probability decays very quickly to zero as the dimension increases, whereas as when ρ is close to one, the probability of the non-risk region approaches β for all dimensions.

Our next observation is that the probability of the non-risk region seems to increase as the tails of the distributions become heavier. In Figure B.7 are plotted the probabilities of risk regions for some spherical distributions and a range of dimensions. Note that multivariate t-distributions have heavier tails than the multivariate Normal distribution, but the tails get lighter as

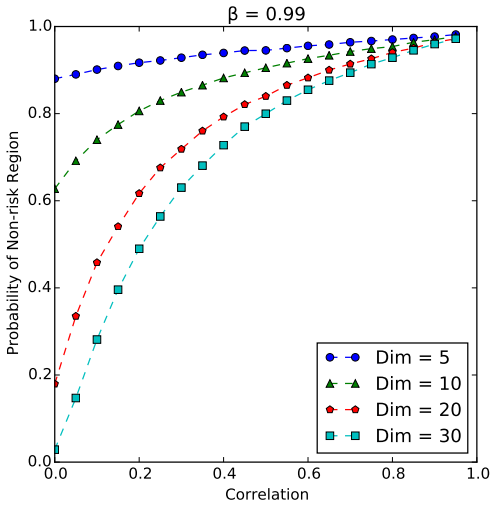


Fig. B.6: Probability of non-risk region for a range of correlation matrices and dimensions for Normal returns

the degrees of freedom parameter increases. This phenomenon can also be observed in Figure B.5.

The observations made in this section suggest that that the application of our methodology will be particularly effective when applied to real stock data tend to be positively correlated and have heavy tails.

6 Numerical tests

In this Section we test the numerical performance of our methodology for realistic distributions. There are three parts to these tests: the calculation of the probability of the non-risk region for a range of distributions and constraints, the performance of aggregation sampling, and the performance of aggregation reduction. In Section 6.1 we describe our experimental set-up, in particular we justify the distributions constructed for these experiments. The remaining three sections detail the individual experiments and their results.

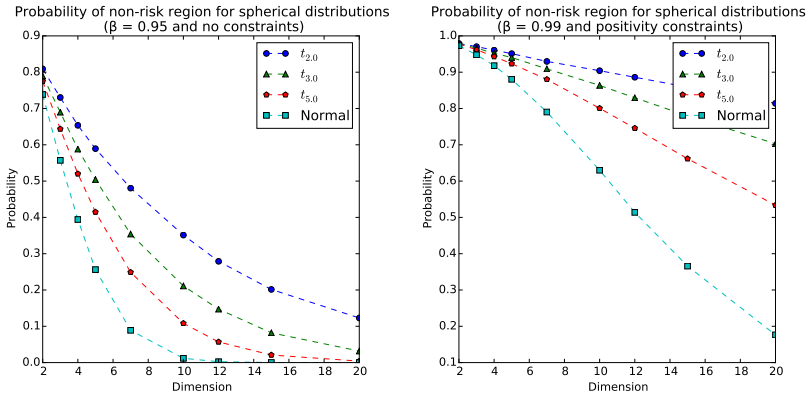


Fig. B.7: Probability of non-risk regions for different spherical distributions and dimensions

6.1 Experimental set-up

For robustness we will use several randomly constructed distributions for each family of distribution and each dimension we are testing. We construct these by fitting parametric distributions to real data. We use real data rather than arbitrarily generating problem parameters for two reasons. Firstly, generating parameters which correspond to well-defined distributions can be problematic. For example, for the moment matching algorithm, there may not exist a distribution which has a given set of target moments (see [22] and [23] for instance). Secondly, as was observed in Section 5, the probability of the non-risk region can vary wildly, and so it is most meaningful to test the performance of our methodology for distributions which are realistic for portfolio selection problems.

We construct our distributions from monthly return data from between January 2007 and February 2015 for 90 companies in the FTSE 100 index. For each dimension in the test, we randomly sample five sets of companies of that size, and for each of these sets fit Normal, t distributions and Skew-t distributions to the associated return data. Figure B.8 shows for two stocks the contours of the fitted density functions overlaying the historical return

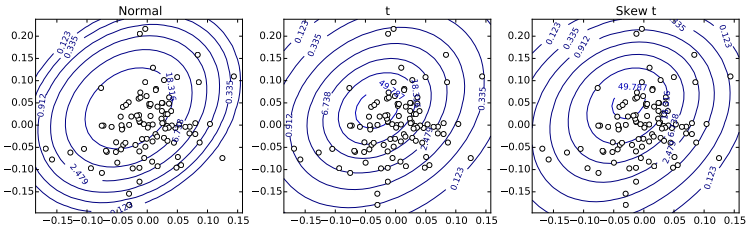


Fig. B.8: Contour plots of distributions fitted to financial return data for two assets

data. For the t distributions we fix the degrees of freedom parameter to 4.0. This is so that we can more easily compare the effect of heavier tails on the results of our tests. We allow the corresponding parameter for the skew-t distributions to be fitted from the data.

These three distributions are fitted to the data through maximum likelihood estimation, weighing each observation equally; our aim here is not to construct distributions which accurately capture the uncertainty of future returns, but to simply construct distributions which are realistic for this type of problem. We also use scenario sets constructed using the moment-matching algorithm. For each random set of companies, we calculate all the required marginal moments and correlations from their historical returns, and use these as input to the moment-matching algorithm. To allow us to compare results, the same constructed distributions are used across the three sets of numerical tests.

Throughout this section we use the β-CVaR as our tail risk measure. This is not only because the β-CVaR leads to tractable scenario-based optimization problems, but also for elliptically distributed returns we can evaluate the β-CVaR exactly which provides us with a means to evaluate the true performance of the solutions yielded by the approximate scenario-based problems. In addition, to ensure that the non-risk region does not have negligible probability, we will assume that we always have positivity constraints on our investments (i.e. no short selling).

6.2 Probability of non-risk region with quota constraints

We first estimate the probability of the non-risk region for a range of distributions, dimensions and constraints. We calculate these probabilities only for the Normal and t distributions as skew-t distributions and moment matching scenario sets use surrogate risk regions based on these distributions. The main purpose of this is to provide intuition about under what circumstances the methodology is effective: there is little to be gained from aggregating scenarios in a non-risk region of negligible probability.

For each distribution we sample 2000 scenarios and calculate the proportion of points in the non-risk region for different levels of β and constraints. In particular, for each dimension we calculate for $\beta = 0.95$, and $\beta = 0.99$, and for a range of *quotas*. The feasible region corresponding to quota $0 < q < 1$ is $\{x \in \mathbb{R}^d : 0 \leq x_i \leq q \text{ for } i = 1, \ldots, d, \ \sum_{i=1}^{d} x_i = 1\}$. Quotas are quite a natural constraint to use in the portfolio selection problem as they ensure that a portfolio is not overexposed to one asset. The quotas may also be viewed as ghost constraints to be used in cases where the probability of the non-risk region with only positivity constraints is too small.

In Figure B.9 for each each dimension we have tested we have plotted the results of one trial. The full results can be found in Appendix A. The first important observation from these is that the proportion of scenarios in the non-risk region, as compared to the uncorrelated case in Figure B.6, is surprisingly high; even for $\beta = 0.95$ and dimension 30, this proportion is non-negligible. As expected, the proportion of scenarios in the non-risk region increases as we tighten our quota. However, for higher dimensions the quotas need to be a lot tighter to make a significant difference. The plots also provide further evidence that the t distribution has non-risk regions with higher probabilities than the lighter-tailed Normal distribution. In Figure B.9, the non-risk region for the t-distributions has probability around 0.05 to 0.1 higher for dimensions 5 and 10, and around 0.1 to 0.2 higher for dimensions

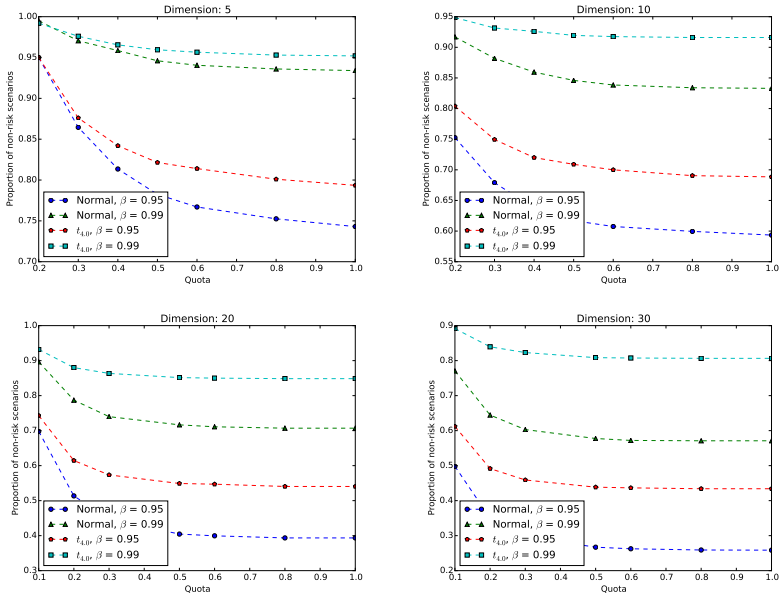


Fig. B.9: Proportions of non-risk scenarios

20 and 30.

6.3 Aggregation sampling

In this section we compare the quality of solutions yielded by sampling and aggregation sampling by observing the optimality gaps of the solutions that these scenario generation methods yield. For this, we use the following version of the portfolio selection problem.

$$\begin{aligned} &\underset{x \geq 0}{\text{minimize}} \ \beta\text{-CVaR}\left(-x^T Y\right) \\ &\text{such that } x^T \mu \geq \tau, \\ &\sum_{i=1}^{d} x_i = 1, \\ &0 \leq x \leq u, \end{aligned}$$

where μ is the mean of the input distribution (rather than scenario set), τ is the target return and u is a vector of asset quotas. The primary reason for using this formulation over those in (P2) and (P3) is that given a distribution of asset returns it is easy to select an appropriate expected target return τ. For simplicity, in our tests we set $\tau = \frac{1}{n}\sum_{i=1}^{n}\mu_i$ which ensures that the constraint is feasible but not trivially satisfied. Notice that in the above formulation we use the deterministic constraint, $x^T\mu \geq \tau$ rather than $\mathbb{E}\left[x^T Y\right] \geq \tau$. This is because the latter constraint depends on the scenario set. Therefore, the solution from a scenario-based approximation might be infeasible with respect to the original problem, which makes measuring solution quality problematic.

In this experiment, we test the performance of the aggregation sampling algorithm for three families of distributions: the Normal distribution, the t-distribution and the skew-t distribution.

For each distribution and problem dimension we run five trials using our constructed distributions (as described in Section 6.1). Each trial consists of generating 50 scenario sets via sampling and aggregation sampling, solving the corresponding scenario-based problem for each of these sets, and calculating the optimality gap for each solution which is yielded. For each scenario generation method we then calculate the mean and standard deviation (S.D.) of the optimality gap. For the skew-t distributions, although we are able to evaluate the objective function value for any candidate solution, to find the true optimal solution value (or one close to it), we resort to solving the problem for a very large sampled set of size 200000.

The full results for this experiment can be found in Appendix B. In Figure B.10 we have plotted for one trial the raw results for dimensions 10 and 30. We observe that there is a consistent improvement in solution quality in using aggregation sampling over basic sampling. In addition the solution values are much more stable. The improvement in solution quality and stability is particularly big for the t-distributions. This is because the probability of the non-risk region is greater for heavier-tailed distributions as observed in

Section 5. Aggregation sampling even lead to consistently better solutions for the skew-t distributions where we are approximating the risk region with a risk region for a t-distribution.

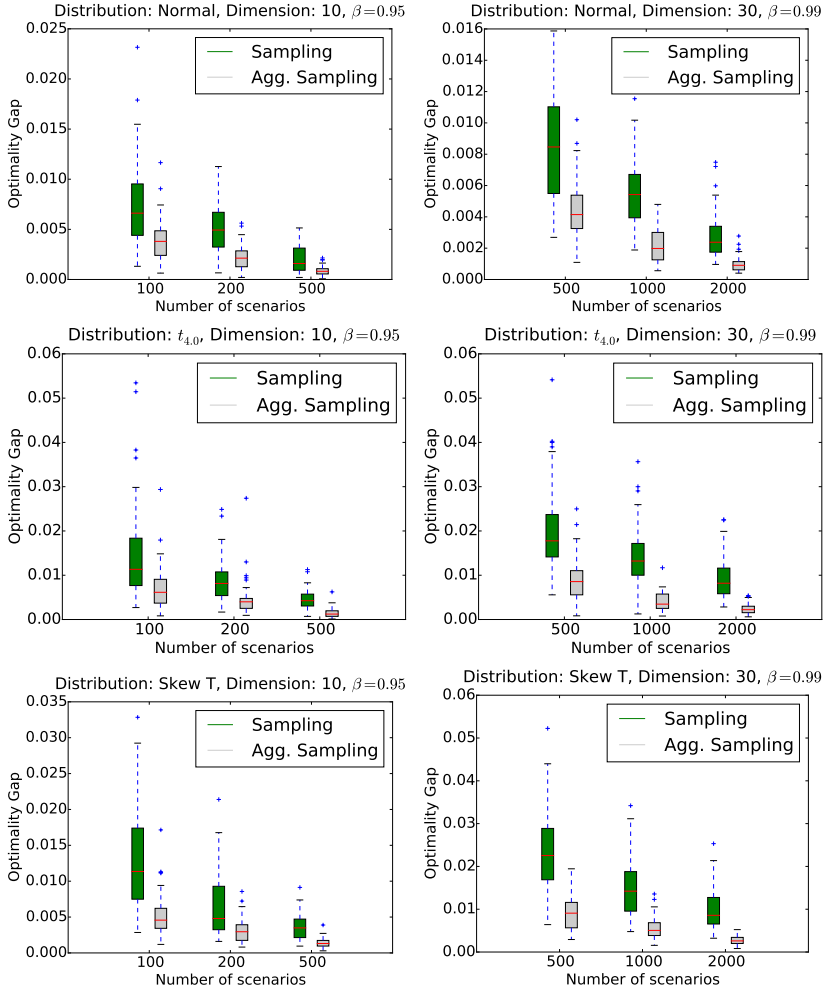


Fig. B.10: Stability test comparing performance of sampling and aggregation sampling

6.4 Aggregation reduction

The aim of these tests is to quantify the error induced through the use of aggregation reduction. In particular, we calculate the error induced in the optimal solution value. For a given scenario set, we aggregate the non-risk scenarios, solve the problem with respect to this reduced set, and calculate the optimality gap of this solution with respect to the original scenario set.

For these tests we use the same problem as in Section 6.3 and run tests for Normal, t and moment matching distributions. As explained in Section 4, we use the risk region of a Normal distribution to approximate the risk region for moment-matched scenario sets. For each family of distributions and problem dimension we again run five trials for different instances of the distribution. In each trial for different initial scenario set sizes, $n = 100,\ 200,\ 500$, and two different levels of tail risk measure $\beta = 0.95,\ 0.99$, we calculate the reduction error for 30 different scenario sets and report the mean error.

The full results can be found in Appendix C. These show that the reduction error is generally very small, in fact for almost all problems using $\beta = 0.95$, there is no error induced. For $\beta = 0.99$, and the smallest scenario set size $n = 100$, there is a small amount of error (< 0.01) for the Normal distributions, slightly larger errors for the heavier-tailed t distribution (< 0.02), and the largest errors (0.1-0.5) occur for reduced moment-matching scenario sets whose risk regions have been approximated with that of a Normal distribution. However, as the scenario set size is increased, all errors are reduced, and for the largest scenario set size $n = 500$, there is no error induced for almost all problems.

Comparing the reduction errors with the corresponding non-risk region probabilities in Appendix A, we see that the larger errors generally occur for the higher dimensional distributions whose non-risk region has a larger probability. This is to be expected as the larger the non-risk region the more scenarios that are aggregated. In Table B.37 in Appendix C we have also in-

cluded the proportions of reduced scenarios for moment matching scenario sets for which we approximated the risk region with that of a Normal distribution. The proportions of reduced scenarios in this case are generally slightly higher than that of the corresponding Normal distributions. This might suggest that the surrogates for the risk region are slightly too small, but this could equally be explained by the fact that moment matching scenario sets generally have heavier tails than the corresponding Normal distribution, which, as we observed in Section 5, also leads to non-risk regions of higher probabilities. In either case, the larger errors which are induced by reducing small moment-matching scenario sets could be explained by these increased probabilities.

7 Case study

In this section we demonstrate how our methodology can be applied to difficult problems which may occur in practice. For this, we use problems which are high-dimensional, have non-elliptical return distributions, and use integer variables. Note that the use of integer variables precludes the use of interior point algorithms to solve this problem. For a fixed computational budget we will compare the performance of sampling and aggregation sampling through estimation of the optimality gap. We also demonstrate how ghost constraints can be used to improve the quality of solutions while highlighting the possible pitfalls of this approach.

In these tests we use the following problem:

$$\begin{aligned}
&\underset{x,z}{\text{minimize}}\ \beta\text{-CVaR}\left(-x^T Y\right)\\
&\text{such that } x^T \mu \geq \tau,\\
&\qquad x_i \leq z_i \text{ for each } i = 1,\ldots,d,\\
&\qquad \sum_{i=1}^{d} x_i = 1,\\
&\qquad \sum_{i=1}^{d} z_i = M,\\
&\qquad 0 \leq x \leq u,\\
&\qquad z_i \in \{0,1\} \text{ for each } i = 1,\ldots,d.
\end{aligned}$$

This problem is similar to that used in Section 6.3 except that we now use integer variables to bound the number of assets in which we may invest. The extra constraints involving integer variables may change the conic hull of feasible solutions, however the method presented in Section 3.1 for calculating conic hulls of feasible regions cannot handle these. We therefore ignore these constraints when constructing a risk region to use for aggregation sampling. This is acceptable as the resulting conic hull will contain the true conic hull. Again, the random vector Y used in these tests is constructed by fitting Skew-t distributions to return data for companies from the FTSE100 stock index.

In each experiment we find candidate solutions for the above problem by solving large scenario-based approximations: we find one candidate solution for a scenario set constructed by basic sampling, and another for a scenario set constructed by aggregation sampling. The optimality gap for each of these solutions is then estimated by employing the *multiple replication procedure* of [24], which involves solving several n_g (smaller) problems for n_g independent scenario sets constructed by sampling and aggregation sampling as appropriate. Specifically, for $k = 1,\ldots,n_g$ denote by Y^k the empirical random vector corresponding to the k-th scenario set, and z_k^* the correspond-

ing optimal solution value. For a candidate solution $\tilde{x} \in \mathbb{R}^d$, and for each scenario set $k = 1, \ldots, n_g$ (of size n) the following is a conservative estimate of the optimality gap:

$$G_n^k = \beta\text{-CVaR}\left(-\tilde{x}^T Y^k\right) - z_k^*.$$

Now, for $0 < \alpha < 1$ an α confidence interval for the optimality gap is

$$(0, \bar{G}_n + \epsilon_{n_g,\alpha}), \tag{B.11}$$

where

$$\begin{aligned}
\bar{G}_n &= \frac{1}{n_g} \sum_{j=1}^{n_g} G_n^j, \\
S_{n_g}^2 &= \frac{1}{n_g - 1} \sum_{j=1}^{n_g} \left(G_n^j - \bar{G}_n\right), \\
\epsilon_{n_g,\alpha} &= t_{n_g-1,\alpha} \frac{S_{n_g}}{\sqrt{n_g}}
\end{aligned}$$

and $t_{n_g-1,\alpha}$ is the α-quantile of the (univariate) t-distribution with $n_g - 1$ degrees of freedom. Note that other procedures for estimating the optimality gap exist which only require the solution of one or two problems [25], [26].

Given the potential dangers in approximating the risk region, and misspecifying ghost constraints, it is important to verify the quality of a solution by the calculation of its corresponding *out-of-sample* value (out value) [27]. That is, we calculate the β-CVaR for our candidate solutions with respect to a large independently sampled scenario set. A bias in our aggregation sampling method may be indicated by a significantly higher out-of-sample value compared to that of sampling. Similarly, the introduction of ghost constraints which are too tight will lead to no improvement in, or a potentially worse out-of-sample value of the new candidate solution. Finally, to aid us in interpreting the results, we include an estimate of the probability of the risk region.

We set our computational budget so that our problems can be solved relatively quickly (a few seconds) on a personal computer. If our problem

is of dimension d and we have n scenarios, the number of floating point operations required to calculate β-CVaR for a particular solution is of order $O(d \times n)$. We therefore limit the number of scenarios in our problems so that $d \times n \leq 10000$ to ensure it can be solved quickly. For the estimation of the optimality gap we solve five problems which use half the number of scenarios used to calculate the candidate solution. For both problems we use $\beta = 0.99$, $\tau = 0.01$ and $\alpha = 0.95$. Our aim is to find a solution for which the upper limit of the 95% confidence interval $G_n + \epsilon_{n_g,\alpha}$ on its optimality is less than 0.015.

Case 1: $d = 30$ We begin with a problem of moderate dimension. Using our rules we use a scenario set size of $n = \frac{10000}{30} \approx 3300$ to find our candidate solutions, and for estimating the optimality gap we use a scenario set size of $\frac{n}{2} = \frac{3300}{2} = 1650$. The results are shown in Table B.1.

Sampling			Agg. Sampling			
Gap ($\bar{G}_n$)	Error ($\epsilon_{n_g,\alpha}$)	Out value	Gap ($\bar{G}_n$)	Error ($\epsilon_{n_g,\alpha}$)	Out value	Risk region prob.
0.018	0.005	0.140	0.004	0.002	0.140	0.157

Table B.1: Estimated optimality gaps for $n = 30$ with 95% confidence level

The out-of-sample values reveal that the quality of candidate solutions are about the same, however the estimation of the optimality gap using aggregation sampling gives us greater assurance that our solution is near optimal. Using (C.4) and Table B.1, the upper limit of the confidence interval on the optimality gap for aggregation sampling is $0.004 + 0.002 = 0.006$ meets our target of being less than 0.015.

Case 2: $d = 50$ We now increase the dimension of the problem substantially. Our rules for scenario set size now prescribe the use of $n = \frac{10000}{50} = 2000$ for calculating our candidate solutions, and a maximum scenario set size of $\frac{n}{2} = \frac{2000}{2} = 1000$ for estimating the optimality gap. The results are shown in

Table B.2.

Sampling			Agg. Sampling			
Gap ($\bar{G}_n$)	Error ($\epsilon_{n_g,\alpha}$)	Out value	Gap ($\bar{G}_n$)	Error ($\epsilon_{n_g,\alpha}$)	Out value	Risk region prob.
0.062	0.013	0.191	0.020	0.004	0.164	0.211

Table B.2: Estimated optimality gaps for $n = 50$ with 95% confidence level

This time both the out-of-sample value and optimality gap yielded by aggregation sampling are smaller than those yielded by basic sampling. Despite this however, the quality of the solutions appear to be much lower than those in Case 1. Using (C.4) and Table B.2, the upper limit of the confidence interval on the optimality gap for aggregation sampling is $0.02 + 0.004 = 0.024$ so does not meet our target.

In an attempt to improve our solution to the previous problem we now add ghost constraints to our problem. As noted in Section 6.2, it is only when constraints become very restrictive that these make a difference to the probability of the risk region and so in the first instance we will add constraints which are tight. We use the constraints $x_i \leq \tilde{x}_i + 0.05$ for $i = 1, \ldots, 50$ where $\tilde{x}$ denotes the candidate solution from aggregation sampling for our previous trial. The results for this trial are shown in Table B.3.

Sampling			Agg. Sampling			
Gap ($\bar{G}_n$)	Error ($\epsilon_{n_g,\alpha}$)	Out value	Gap ($\bar{G}_n$)	Error ($\epsilon_{n_g,\alpha}$)	Out value	Risk region prob.
0.018	0.010	0.168	0.011	0.005	0.165	0.119

Table B.3: Estimated optimality gaps for $n = 50$ with 95% confidence level where tight quota constraints have been added to the problem

The proportion of samples in the risk region has roughly halved and so our scenarios are concentrated on a much smaller region of the support. We see this time that the optimality gap of our solutions is much reduced for aggregation sampling, so much so that the error is now within our desired

tolerance. However, inspecting the out-of-sample value we see that it has not improved, despite the estimated optimality gap being much lower compared to the previous experiment. In addition, some of the ghost constraints we have added are binding. This strongly indicates that the added constraints may be too tight (they are active), in which case the optimality gaps are not valid with respect to the original problem.

We now try slightly looser ghost constraints: $x_i \leq \tilde{x}_i + 0.1$ where $\tilde{x}$ is still the candidate solution from our first trial without ghost constraints. The results are shown in Table B.4.

Sampling			Agg. Sampling			
Gap ($\bar{G}_n$)	Error ($\epsilon_{n_g,\alpha}$)	Out value	Gap ($\bar{G}_n$)	Error ($\epsilon_{n_g,\alpha}$)	Out value	Risk region prob.
0.034	0.009	0.169	0.009	0.004	0.162	0.162

Table B.4: Estimated optimality gaps for $n = 50$ with 95% confidence level where loose quota constraints have been added to the problem

The out-of-sample value in this trial is significantly improved compared to the previous one and there are fewer binding ghost constraints in our solution. The upper limit of the estimate of optimality gap is now within the desired tolerance. Again, this estimate only applies to the original problem (the problem without ghost constraints) if the ghost constraints are guaranteed to be non-active.

It is generally difficult to guarantee that ghost constraints are non-active, but nevertheless, the example above demonstrates that their inclusion, with careful calibration can lead to significantly improved solutions. In the example above, we only used simple quotas for our ghost constraints and it may be possible to further improve the out-of-sample value by adding lower bounds to investments as well as upper bounds. Additionally, when constructing our ghost constraints we used the same amount of slack for each variable and so one could try varying this for different assets.

8 Conclusions

In the paper [12] we proposed a general approach to scenario generation using risk regions for stochastic programs with tail risk measure. As proof-of-concept we demonstrated how this applied for portfolio selection problems for elliptically distributed returns. In this work, we have presented how this methodology may be used for more realistic portfolio selection problems, and studied under what conditions it is effective.

To find the risk region for our problem, we must be able to describe the conic hull of the feasible region and be able to project points onto this. In the presence of positivity constraints, we were able to describe exactly this conic hull from general linear constraints on our portfolio, and identified that the projection of a point onto a cone requires the solution of a small quadratic program, or a linear complementarity problem. The solution of these small programs becomes a significant bottleneck for high dimensions in our methodology and so one possible avenue of future research would be to investigate how this calculation could be done more efficiently. For example, instead of calculating the whole projection, one could calculate the projection only to an accuracy sufficient to test if a point belongs in the risk region.

The efficacy of using risk regions for scenario generation depends upon the probability of the risk region: the greater the probability of the non-risk region, the more scenarios that can be aggregated. It follows directly from the definition of risk regions that this probability decreases as the problem becomes more constrained and as the level of the tail risk β increases. In our case study we exploited the former property through the addition of non-binding or ghost constraints to our problem. A more systematic way of selecting ghost constraints, and finding some way to guarantee they are non-active are thus important directions of research. In our numerical experiments we observed that the probability of the risk region decreases for

heavier tailed distributions, and in the presence positivity constraints, the probability decreases as the correlations between asset returns increases. It is desirable to develop theory which explain these phenomena.

Finally, we tested the performance of our methodology for solving realistic problems where the return distributions were fitted from real financial return data. Aggregation sampling generally outperformed basic sampling in terms of solution quality and stability. We also showed that aggregation reduction induces almost no error in the solution for reasonably sized scenario sets. These results not only held for elliptical distributions, but also non-elliptical distributions for which we have approximated the risk regions. However, in a small number of cases, the mis-specification of these surrogate risk regions lead to worse results. Thus, research needs to done to determine how one can choose such more reliable surrogate risk regions for non-elliptical regions.

A Reduction proportion tables

The following tables list the estimated probabilities of the non-risk region for a variety of distributions constructed from real data. See Section 6.2 for details. Each table corresponds to a family of distributions at a given dimension, and each row gives the proportions for a given set of companies. In addition, the distributions corresponding the i-th row of each table of dimension d have been fitted using the same set of companies.

$\beta = 0.95$							$\beta = 0.99$						
$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.4$	$x \leq 0.3$	$x \leq 0.2$	$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.4$	$x \leq 0.3$	$x \leq 0.2$
0.743	0.752	0.767	0.782	0.814	0.865	0.950	0.934	0.936	0.941	0.946	0.959	0.971	0.995
0.738	0.744	0.760	0.777	0.809	0.855	0.949	0.922	0.925	0.928	0.934	0.949	0.965	0.992
0.767	0.775	0.793	0.807	0.832	0.872	0.948	0.930	0.932	0.937	0.943	0.953	0.969	0.990
0.763	0.771	0.784	0.801	0.830	0.880	0.951	0.931	0.934	0.944	0.949	0.957	0.973	0.987
0.755	0.763	0.777	0.798	0.829	0.883	0.955	0.927	0.929	0.935	0.940	0.951	0.966	0.991

Table B.5: Proportion of reduced scenarios for Normal distributed returns and $d = 5$

$\beta = 0.95$							$\beta = 0.99$						
$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.4$	$x \leq 0.3$	$x \leq 0.2$	$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.4$	$x \leq 0.3$	$x \leq 0.2$
0.594	0.600	0.608	0.617	0.637	0.679	0.752	0.833	0.834	0.839	0.846	0.860	0.882	0.917
0.617	0.621	0.632	0.647	0.669	0.703	0.777	0.851	0.852	0.856	0.860	0.868	0.879	0.914
0.506	0.509	0.523	0.534	0.560	0.606	0.689	0.779	0.780	0.787	0.791	0.806	0.837	0.889
0.564	0.566	0.573	0.590	0.615	0.658	0.748	0.827	0.828	0.835	0.846	0.857	0.877	0.921
0.537	0.540	0.552	0.566	0.586	0.624	0.727	0.820	0.822	0.825	0.832	0.843	0.870	0.912

Table B.6: Proportion of reduced scenarios for Normal distributed returns and $d = 10$

$\beta = 0.95$							$\beta = 0.99$						
$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.3$	$x \leq 0.2$	$x \leq 0.1$	$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.3$	$x \leq 0.2$	$x \leq 0.1$
0.394	0.394	0.400	0.405	0.447	0.513	0.698	0.707	0.707	0.711	0.717	0.740	0.787	0.896
0.325	0.326	0.332	0.342	0.392	0.457	0.635	0.653	0.653	0.655	0.662	0.696	0.740	0.851
0.344	0.344	0.348	0.354	0.389	0.460	0.668	0.648	0.648	0.653	0.656	0.683	0.743	0.870
0.384	0.385	0.390	0.401	0.440	0.507	0.708	0.695	0.695	0.698	0.704	0.740	0.782	0.896
0.417	0.418	0.424	0.432	0.479	0.540	0.738	0.727	0.727	0.730	0.735	0.764	0.813	0.906

Table B.7: Proportion of reduced scenarios for Normal distributed returns and $d = 20$

$\beta = 0.95$							$\beta = 0.99$						
$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.3$	$x \leq 0.2$	$x \leq 0.1$	$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.3$	$x \leq 0.2$	$x \leq 0.1$
0.259	0.259	0.263	0.267	0.297	0.350	0.498	0.571	0.571	0.572	0.578	0.603	0.644	0.770
0.264	0.266	0.269	0.272	0.299	0.347	0.511	0.587	0.587	0.589	0.591	0.616	0.661	0.790
0.282	0.282	0.286	0.291	0.321	0.378	0.533	0.599	0.599	0.602	0.607	0.631	0.681	0.785
0.247	0.247	0.251	0.257	0.281	0.333	0.502	0.555	0.555	0.556	0.558	0.586	0.630	0.769
0.293	0.293	0.296	0.301	0.324	0.374	0.548	0.583	0.583	0.584	0.587	0.613	0.665	0.802

Table B.8: Proportion of reduced scenarios for Normal distributed returns and $d = 30$

$\beta = 0.95$							$\beta = 0.99$						
$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.4$	$x \leq 0.3$	$x \leq 0.2$	$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.4$	$x \leq 0.3$	$x \leq 0.2$
0.793	0.801	0.814	0.822	0.842	0.876	0.950	0.952	0.953	0.957	0.960	0.966	0.976	0.992
0.775	0.782	0.796	0.812	0.837	0.877	0.946	0.949	0.950	0.954	0.956	0.961	0.972	0.988
0.808	0.815	0.829	0.841	0.859	0.898	0.953	0.958	0.960	0.962	0.964	0.969	0.980	0.992
0.799	0.808	0.819	0.828	0.855	0.882	0.950	0.949	0.951	0.954	0.957	0.965	0.977	0.990
0.793	0.799	0.809	0.822	0.848	0.887	0.951	0.960	0.960	0.963	0.965	0.969	0.976	0.991

Table B.9: Proportion of reduced scenarios for $t_{4.0}$ distributed returns and $d = 5$

$\beta = 0.95$							$\beta = 0.99$						
$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.4$	$x \leq 0.3$	$x \leq 0.2$	$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.4$	$x \leq 0.3$	$x \leq 0.2$
0.689	0.691	0.700	0.709	0.720	0.750	0.804	0.916	0.916	0.917	0.919	0.926	0.931	0.949
0.711	0.713	0.719	0.730	0.742	0.769	0.829	0.923	0.924	0.925	0.926	0.930	0.940	0.956
0.616	0.617	0.630	0.640	0.656	0.677	0.754	0.895	0.896	0.898	0.900	0.905	0.915	0.935
0.642	0.642	0.647	0.657	0.672	0.703	0.783	0.896	0.896	0.900	0.904	0.913	0.925	0.941
0.652	0.655	0.666	0.675	0.690	0.723	0.785	0.905	0.905	0.907	0.908	0.913	0.924	0.944

Table B.10: Proportion of reduced scenarios for $t_{4.0}$ distributed returns and $d = 10$

$\beta = 0.95$							$\beta = 0.99$						
$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.3$	$x \leq 0.2$	$x \leq 0.1$	$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.3$	$x \leq 0.2$	$x \leq 0.1$
0.540	0.540	0.547	0.549	0.574	0.615	0.743	0.849	0.849	0.850	0.852	0.864	0.880	0.932
0.461	0.463	0.467	0.475	0.509	0.560	0.703	0.835	0.836	0.840	0.844	0.858	0.870	0.919
0.506	0.507	0.510	0.515	0.551	0.595	0.753	0.839	0.839	0.839	0.840	0.855	0.874	0.931
0.511	0.511	0.514	0.519	0.562	0.612	0.753	0.860	0.860	0.862	0.865	0.876	0.894	0.939
0.567	0.568	0.572	0.576	0.609	0.657	0.797	0.866	0.867	0.867	0.870	0.881	0.901	0.952

Table B.11: Proportion of reduced scenarios for $t_{4.0}$ distributed returns and $d = 20$

$\beta = 0.95$							$\beta = 0.99$						
$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.3$	$x \leq 0.2$	$x \leq 0.1$	$x \leq 1.0$	$x \leq 0.8$	$x \leq 0.6$	$x \leq 0.5$	$x \leq 0.3$	$x \leq 0.2$	$x \leq 0.1$
0.434	0.434	0.436	0.439	0.459	0.491	0.612	0.806	0.806	0.807	0.808	0.823	0.840	0.891
0.466	0.466	0.468	0.469	0.495	0.532	0.649	0.821	0.821	0.823	0.824	0.838	0.853	0.897
0.443	0.443	0.445	0.448	0.474	0.512	0.637	0.821	0.822	0.822	0.824	0.834	0.854	0.898
0.444	0.445	0.448	0.454	0.470	0.513	0.635	0.812	0.813	0.814	0.814	0.823	0.841	0.889
0.417	0.417	0.419	0.421	0.444	0.487	0.617	0.808	0.808	0.810	0.811	0.823	0.844	0.891

Table B.12: Proportion of reduced scenarios for $t_{4.0}$ distributed returns and $d = 30$

B Aggregation sampling tables

The following tables list the relative reduction in the mean and standard deviation of optimality gaps for aggregation sampling compared with sampling for a variety of distributions. See Section 6.3 for more details.

n = 100		n = 200		n = 500	
Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.
2.747	2.542	3.226	3.321	3.697	2.871
3.905	4.427	3.226	3.323	3.646	4.439
3.803	2.993	4.889	3.538	4.567	3.927
3.376	3.040	3.402	2.517	5.182	4.357
3.240	3.257	3.432	2.246	4.807	4.708

Table B.13: Comparison for $d = 5$, $\beta = 0.95$, and Normal returns

n = 100		n = 200		n = 500	
Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.
1.989	1.876	2.670	2.422	2.460	2.495
2.018	2.494	2.711	2.227	3.126	2.864
1.559	1.652	1.736	1.230	2.727	2.678
1.869	2.089	2.275	2.181	2.551	2.731
1.996	2.085	2.285	2.061	2.466	2.828

Table B.14: Comparison for $d = 10$, $\beta = 0.95$, and Normal returns

n = 500		n = 1000		n = 2000	
Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.
2.357	2.124	2.890	3.039	3.026	2.809
2.504	3.054	2.750	2.839	2.873	2.689
2.308	1.963	2.546	2.854	2.803	2.791
2.341	2.699	2.948	3.369	2.592	2.367
2.802	2.657	3.421	2.494	3.725	3.547

Table B.15: Comparison for $d = 20$, $\beta = 0.99$, and Normal returns

n = 500		n = 1000		n = 2000	
Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.
1.943	1.842	2.161	2.148	2.901	2.846
1.779	2.195	2.197	2.067	2.590	2.483
1.990	2.227	2.246	2.033	2.405	2.514
2.019	2.012	2.076	2.057	2.010	1.891
1.866	1.769	2.457	1.921	2.853	3.138

Table B.16: Comparison for $d = 30$, $\beta = 0.99$, and Normal returns

n = 100		n = 200		n = 500	
Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.
2.857	2.661	2.762	1.981	3.500	3.709
3.407	3.431	3.692	3.416	5.572	6.167
4.335	3.062	3.872	4.195	3.244	3.149
4.280	3.748	4.636	6.732	4.974	6.593
2.578	1.773	3.664	3.500	4.019	4.160

Table B.17: Comparison for $d = 5$, $\beta = 0.95$, and $t_{4.0}$ returns

n = 100		n = 200		n = 500	
Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.
1.899	2.091	2.169	1.805	2.939	2.599
2.078	1.910	2.358	2.229	2.982	2.340
1.996	2.923	2.639	3.126	2.088	1.727
2.658	2.958	2.436	2.222	2.357	2.312
2.080	2.171	1.980	1.232	2.957	2.114

Table B.18: Comparison for $d = 10$, $\beta = 0.95$, and $t_{4.0}$ returns

n = 500		n = 1000		n = 2000	
Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.
4.142	5.028	4.215	4.383	5.571	5.221
3.039	3.843	4.096	4.346	4.857	6.084
3.378	3.831	4.020	4.267	5.007	5.617
3.722	4.886	3.744	3.247	4.339	5.336
3.616	3.524	4.999	3.739	5.116	6.277

Table B.19: Comparison for $d = 20$, $\beta = 0.99$, and $t_{4.0}$ returns

n = 500		n = 1000		n = 2000	
Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.
3.035	3.068	2.950	2.547	3.741	4.042
2.359	1.983	3.513	5.068	3.384	3.029
3.507	4.356	2.977	3.966	3.686	4.915
2.950	3.005	3.079	1.964	3.936	4.240
2.228	2.043	3.549	3.227	3.950	4.267

Table B.20: Comparison for $d = 30$, $\beta = 0.99$, and $t_{4.0}$ returns

n = 100		n = 200		n = 500	
Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.
1.917	1.601	2.766	3.020	3.352	2.644
1.887	1.857	2.748	2.416	3.414	3.290
3.171	3.489	4.433	3.427	3.949	3.774
2.620	3.170	3.038	3.518	2.872	3.178
2.391	2.408	2.027	1.891	3.466	3.434

Table B.21: Comparison for $d = 5$, $\beta = 0.95$, and Skew T returns

n = 100		n = 200		n = 500	
Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.
1.839	2.189	2.215	1.925	2.977	2.650
1.631	2.021	2.203	2.087	2.150	2.554
1.962	1.671	1.872	1.187	3.172	3.513
1.627	1.868	1.661	2.136	1.775	1.439
2.502	2.417	2.152	2.577	2.647	2.580

Table B.22: Comparison for $d = 10$, $\beta = 0.95$, and Skew T returns

n = 500		n = 1000		n = 2000	
Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.
4.646	5.803	4.921	4.384	5.843	6.268
4.639	4.025	6.296	5.028	6.513	7.438
3.355	3.840	3.655	3.163	3.305	3.359
3.317	2.257	3.448	3.623	4.794	4.732
3.395	3.365	3.164	3.145	4.351	4.306

Table B.23: Comparison for $d = 20$, $\beta = 0.99$, and Skew T returns

n = 500		n = 1000		n = 2000	
Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.	Mean Imp.	S.D. Imp.
2.631	3.659	3.364	4.298	4.000	4.099
2.285	2.809	2.667	3.201	3.482	2.882
3.266	4.545	3.617	4.340	3.791	3.138
2.923	3.334	3.750	3.796	4.304	5.492
2.486	2.289	2.658	2.754	3.659	4.918

Table B.24: Comparison for $d = 30$, $\beta = 0.99$, and Skew T returns

C Reduction error tables

The following tables list the mean error induced by aggregating scenarios in the non-risk region for a variety of distributions. See Section 6.4 for details.

$\beta = 0.95$			$\beta = 0.99$		
$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
0.000	0.000	0.000	0.008	0.002	0.000
0.000	0.000	0.000	0.009	0.001	0.000
0.000	0.000	0.000	0.005	0.002	0.000
0.000	0.000	0.000	0.007	0.001	0.000
0.000	0.000	-0.000	0.007	0.001	0.000

Table B.25: Reduction error induced for d=5 Normal returns

$\beta = 0.95$			$\beta = 0.99$		
$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
0.000	0.000	0.000	0.006	0.000	0.000
0.000	0.000	0.000	0.006	0.001	0.000
0.000	-0.000	0.000	0.006	0.001	0.000
0.000	0.000	-0.000	0.004	0.000	0.000
0.000	-0.000	-0.000	0.005	0.000	0.000

Table B.26: Reduction error induced for d=10 Normal returns

$\beta = 0.95$			$\beta = 0.99$		
$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
0.000	0.000	-0.000	0.003	0.000	0.000
0.000	-0.000	-0.000	0.002	0.000	0.000
0.000	-0.000	-0.000	0.002	0.000	0.000
0.000	0.000	0.000	0.002	0.000	-0.000
0.000	0.000	0.000	0.003	0.000	0.000

Table B.27: Reduction error induced for d=20 Normal returns

$\beta = 0.95$			$\beta = 0.99$		
$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
0.000	-0.000	0.000	0.001	0.000	0.000
0.000	-0.000	0.000	0.002	0.000	0.000
-0.000	0.000	-0.000	0.002	0.000	-0.000
-0.000	-0.000	0.000	0.001	0.000	-0.000
0.000	0.000	0.000	0.001	0.000	0.000

Table B.28: Reduction error induced for d=30 Normal returns

$\beta = 0.95$			$\beta = 0.99$		
$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
0.001	0.000	0.000	0.014	0.005	0.000
0.000	0.000	0.000	0.012	0.003	0.000
0.000	0.000	0.000	0.017	0.002	0.001
0.000	0.000	0.000	0.015	0.005	0.000
0.001	0.000	0.000	0.015	0.004	0.001

Table B.29: Reduction error induced for d=5 $t_{4.0}$ returns

$\beta = 0.95$			$\beta = 0.99$		
$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
0.000	0.000	0.000	0.015	0.003	0.000
0.000	0.000	0.000	0.015	0.004	0.000
0.000	-0.000	0.000	0.012	0.002	-0.000
0.000	0.000	-0.000	0.017	0.004	0.000
0.000	-0.000	-0.000	0.020	0.003	0.000

Table B.30: Reduction error induced for d=10 $t_{4.0}$ returns

$\beta = 0.95$			$\beta = 0.99$		
$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
0.000	0.000	-0.000	0.013	0.003	0.000
0.000	0.000	-0.000	0.017	0.000	0.000
0.000	0.000	-0.000	0.016	0.003	0.000
0.000	0.000	0.000	0.012	0.002	0.000
0.000	-0.000	0.000	0.016	0.002	0.000

Table B.31: Reduction error induced for d=20 $t_{4.0}$ returns

$\beta = 0.95$			$\beta = 0.99$		
$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
0.000	-0.000	0.000	0.015	0.001	0.000
0.000	0.000	0.000	0.015	0.004	0.000
0.000	-0.000	0.000	0.013	0.002	0.000
0.000	-0.000	-0.000	0.015	0.004	0.000
0.000	0.000	0.000	0.016	0.004	0.000

Table B.32: Reduction error induced for d=30 $t_{4.0}$ returns

$\beta = 0.95$			$\beta = 0.99$		
$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
0.000	-0.000	0.000	0.001	0.000	-0.000
0.000	0.000	0.000	0.002	0.000	0.000
0.000	0.000	0.000	0.000	-0.000	0.000
-0.000	0.000	0.000	0.000	0.000	-0.000
0.000	0.000	0.000	0.001	0.001	0.000

Table B.33: Reduction error induced for d=5 Moment Matching returns

$\beta = 0.95$			$\beta = 0.99$		
$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
-0.000	-0.000	-0.000	0.003	0.000	0.000
-0.000	-0.000	0.000	0.001	-0.000	-0.000
0.000	0.000	0.000	0.509	0.001	0.001
-0.000	0.000	-0.000	0.003	0.000	0.000
0.000	0.000	0.000	0.002	0.000	0.000

Table B.34: Reduction error induced for d=10 Moment Matching returns

$\beta = 0.95$			$\beta = 0.99$		
$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
0.000	0.000	0.000	0.089	0.000	0.000
-0.000	0.000	0.000	0.407	0.003	0.000
0.000	-0.000	-0.000	0.003	0.001	0.000
0.000	0.000	0.000	0.231	0.001	0.000
0.000	0.000	0.000	0.000	0.000	0.000

Table B.35: Reduction error induced for d=20 Moment Matching returns

$\beta = 0.95$			$\beta = 0.99$		
$n = 100$	$n = 200$	$n = 500$	$n = 100$	$n = 200$	$n = 500$
0.000	0.000	0.000	0.205	0.000	0.000
0.000	0.000	0.000	0.111	0.001	0.000
0.000	0.000	0.000	0.206	0.001	0.000
0.000	0.000	0.000	0.218	0.001	0.000
0.000	0.000	0.000	0.071	0.001	0.000

Table B.36: Reduction error induced for d=30 Moment Matching returns

$d = 5$		$d = 10$		$d = 20$		$d = 30$	
$\beta = 0.95$	$\beta = 0.99$	$\beta = 0.95$	$\beta = 0.99$	$\beta = 0.95$	$\beta = 0.99$	$\beta = 0.95$	$\beta = 0.99$
0.786	0.919	0.638	0.840	0.481	0.734	0.380	0.646
0.743	0.900	0.623	0.827	0.477	0.741	0.365	0.647
0.761	0.905	0.660	0.869	0.445	0.729	0.381	0.650
0.770	0.907	0.625	0.847	0.455	0.716	0.366	0.655
0.747	0.917	0.640	0.860	0.446	0.712	0.333	0.616

Table B.37: Proportions of scenarios reduced for moment matching scenario sets

References

[1] H. Markowitz, "Portfolio selection," *Journal of Finance*, vol. 7, pp. 77–91, 1952.

[2] M. R. Young, "A minimax portfolio selection rule with linear programming solution," *Management Science*, vol. 44, no. 5, pp. 673–683, 1998. [Online]. Available: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.44.5.673

[3] H. Markowitz, *Portfolio selection: Efficient diversification of investment*. New Haven: Yale University Press, 1959.

[4] R. Dembo and D. Rosen, "The practice of portfolio replication. A practical overview of forward and inverse problems," *Annals of Operations Research*, vol. 85, pp. 267–284, 1999.

[5] P. Jorion, *Value at Risk: The New Benchmark for Controlling Market Risk*. Irwin Professional, 1996.

[6] P. Artzner, F. Delbaen, J. Eber, and D. Heath, "Coherent measures of risk," *Mathematical Finance*, vol. 9, no. 3, pp. 203–228, 1999.

[7] D. Tasche, "Expected shortfall and beyond," *Journal of Banking & Finance*, vol. 26, no. 7, pp. 1519–1533, 2002.

[8] C. Acerbi and D. Tasche, "On the coherence of expected shortfall," *Journal of Banking & Finance*, vol. 26, no. 7, pp. 1487–1503, 2002.

[9] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *The Journal of Risk*, vol. 2, no. 3, pp. 21–41, 2000.

[10] K. Høyland, M. Kaut, and S. W. Wallace, "A heuristic for moment-matching scenario generation," *Computational Optimization and Applications*, vol. 24, no. 2–3, pp. 169–185, 2003.

[11] M. Kaut, S. W. Wallace, H. Vladimirou, and S. Zenios, "Stability analysis of portfolio management with conditional value-at-risk," *Quantitative Finance*, vol. 7, no. 4, pp. 397–409, 2007.

[12] J. Fairbrother, A. Turner, and S. W. Wallace, "Scenario generation for stochastic programs with tail risk measures," ArXiv e-print 1511.03074, Nov. 2015. [Online]. Available: http://arxiv.org/abs/1511.03074

[13] M. Kaut and S. W. Wallace, "Shape-based scenario generation using copulas," *Computational Management Science*, vol. 8, no. 1–2, pp. 181–199, 2011.

[14] B. Kaynar, Ş. I. Birbil, and J. Frenk, "Application of a general risk management model to portfolio optimization problems with elliptical distributed returns for risk neutral and risk averse decision makers," Erasmus Research Institute of Management, Tech. Rep., 2007.

[15] K.-T. Fang, S. Kotz, and K. W. Ng, *Symmetric Multivariate and Related Distributions (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 11 1989.

[16] G. M. Ziegler, *Lectures on Polytopes (Graduate Texts in Mathematics)*. Springer, 4 2008.

[17] N. K. Chernikova, "Algorithm for finding a general formula for the non-negative solutions of a system of linear inequalities," *U.S.S.R. Computational Mathematics and Mathematical Physics*, vol. 5, no. 2, pp. 228–233, 1965.

[18] H. Le Verge, "A note on Chernikova's algorithm," IRISA, Rennes, France, Tech. Rep. 635, 1992.

[19] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004. [Online]. Available: http://stanford.edu/~boyd/cvxbook/

[20] R. W. Cottle, J.-S. Pang, and R. E. Stone, *The linear complementarity problem*. Siam, 1992, vol. 60.

[21] A. Azzalini and A. Capitanio, "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, pp. 367–389, 2003.

[22] C. Klaassen, P. Mokveld, and B. V. Es, "Squared skewness minus kurtosis bounded by 186/125 for unimodal distributions," *Statistics & Probability Letters*, vol. 50, no. 2, pp. 131–135, 2000.

[23] N. Johnson and C. Rogers, "The moment problem for unimodal distributions," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 433–439, 1951.

[24] W. Mak, D. Morton, and R. Wood, "Monte Carlo bounding techniques for determining solution quality in stochastic programs," *Operations Research Letters*, vol. 24, pp. 47–56, 1999.

[25] G. Bayraksan and D. P. Morton, "Assessing solution quality in stochastic programs," *Mathematical Programming*, vol. 108, no. 2–3, pp. 495–514, sep 2006.

[26] R. Stockbridge and G. Bayraksan, "A probability metrics approach for reducing the bias of optimality gap estimators in two-stage stochastic linear programming," *Mathematical Programming*, vol. 142, no. 1–2, pp. 107–131, 2013.

[27] M. Kaut and S. W. Wallace, "Evaluation of scenario-generation methods for stochastic programming," *Pacific Journal of Optimization*, vol. 3, no. 2, pp. 257–271, 2007.

Paper C

Scenario Generation for Newsvendor Problems

Jamie Fairbrother

1 Introduction

Stochastic programming is a tool for making decisions under uncertainty. In a stochastic program uncertain quantities are modeled by random variables. One has a *loss function* associating to each combination of decision and realization of the random variables a loss, and the aim is to minimize the expectation, or some other risk measure, of this loss function. The power of stochastic programming is that it allows one to model explicitly the costs of future decisions based on the outcomes of random variables.

This flexibility comes at a cost: stochastic programs are typically analytically tractable. Moreover, when the underlying random variables have continuous random variables these problems are also numerically intractable as the calculation of the expected loss function usually involves the multidimensional integration of a loss function which may be only implicitly defined.

Scenario generation is the construction of a finite discrete random variable to use within a stochastic program. For this class of random variables, the calculation of the expected cost function reduces to a summation. In the case of stochastic linear programs, the resultant problem is a (large) linear program. Furthermore, there are many algorithms which exploit the structure of this type of problem to allow a solution more efficient than standard linear programming techniques, for example Bender's decomposition [1].

Scenario generation may involve the discretization of a continuous distribution or the direct construction of a discrete random vector. The simplest way to discretize a random vector is to represent it with a large sample. The resultant problem is known as the sample average approximation. Other discretization approaches such as that used in [2] attempt to construct a discretization which minimizes the distance between the approximation and the true distribution with respect to some probability metric. Property-matching approaches attempt to construct discrete distributions with desired statistical properties. This approach, first proposed in [3], works on the principle that

the solution to a stochastic program only depends on certain properties of a distribution. However, it is not usually clear *a priori* which properties are important for a particular problem and so when using such methods this must be investigated.

A draw-back to the above standard approaches above is that they do not explicitly exploit the structure of the underlying problem. A problem-driven approach to scenario generation, may allow one to represent the uncertainty in a more parsimonious way. In our previous papers [4, 5] we proposed a scenario generation approach to problems involving tail risk measures. For these problems, we identified that a (potentially large) region of the support of the random vector did not contribute to the evaluation of the tail risk measure and so could be represented with a single scenario. We demonstrated that by concentrating the construction of scenarios in the rest of the support one could find better solutions with fewer scenarios. The drawback of this method is that it relies on one being able to characterize the aforementioned region in a convenient manner. This is difficult because this region depends not only on the problem constraints but also the distribution of the random variables.

In this paper we propose another problem-driven approach to scenario generation. This approach assumes one can partition the support of the distribution into *active* and *inactive components*. The value of the expected cost function on the inactive components depends only on their conditional expectation (or some other statistic restricted to the inactive component). The inactive components can therefore be represented with a single scenario (respectively, a very few). Unlike the approach in [4], this partition is determined only by the loss function and is independent of the distribution of the random variables.

As proof of concept, we apply this approach to simple recourse problems. These are a class of stochastic programs which aim to minimize the deviation between the availability of a set of resources and the stochastic demands for

each of them. As such, simple recourse problems are useful in modeling inventory problems.

This paper is organized as follows: in Section 2 give the general set-up for this work and recall some required results from the stochastic programming literature; in Section 3 we demonstrate our approach on a basic newsvendor problem; in Section 4 we generalize our methodology, and provide a probabilistic analysis of our approach; in Section 5 we show how the approach of the previous section can be applied to simple recourse problems; in Section 6 we demonstrate the performance of our method in a small numerical test; finally, in Section 6 we make summarize this work and suggest some avenues of future research.

2 Preliminaries

In this section, we give the general set-up for this work and present some prerequisite theory from the stochastic programming literature that will be required for the analysis and testing of our methodology in the later sections of this paper.

Let $\tilde{\xi}$ be a random vector with support $\Xi \subset \mathbb{R}^d$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, $\mathcal{X} \subset \mathbb{R}^k$ a set of feasible decisions and $\mathrm{f} : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}$ a function for which $\tilde{\xi} \mapsto \mathrm{f}\left(x, \tilde{\xi}\right)$ is measurable (and integrable) for all $x \in \mathcal{X}$. We refer to $\mathrm{f}\left(x, \tilde{\xi}\right)$ throughout as the *loss function*, and the problem we consider is the minimization of the expectation of this function:

$$\underset{x \in \mathcal{X}}{\text{minimize}}\ \mathbb{E}\left[\mathrm{f}\left(x, \tilde{\xi}\right)\right]. \tag{C.1}$$

We will assume for all $x \in \mathcal{X}$ that $\mathbb{E}\left[\mathrm{f}(x, \tilde{\xi})^2\right] < \infty$ which allows us to use the central limit theorem (CLT).

In Section 2.1 we introduce the *Wasserstein distance*, a metric used to bound the error induced by approximating a distribution in a stochastic program. In Section 2.2 we show how one can estimate the optimality gap of a feasible solution for a stochastic program via sampling.

2.1 Wasserstein distance

The discretization of a continuous random vector to solve a stochastic program leads to another stochastic program that is an approximation of the original. The error induced by this approximation is most meaningfully quantified by the optimality gap of the solution that the approximate problem yields.

Definition 2.1 (Approximation error)**.** *The* approximation error *induced by using the random vector $\check{\xi}$ in the place of $\tilde{\xi}$ with respect to the problem* (C.1) *is as follows:*

$$\mathrm{e}(\tilde{\xi}, \check{\xi}) = \sup_{x_0 \in \operatorname*{argmin}_{x \in \mathcal{X}} \mathrm{F}_{\check{\xi}}(x)} \{\min_{x \in \mathcal{X}} \mathrm{F}_{\tilde{\xi}}(x) - \mathrm{F}_{\tilde{\xi}}(x_0)\}$$

A convenient way to bound the approximation error is to use the sup-distance between the true and approximate expected cost functions. The following elementary lemma is taken from [2].

Lemma 2.2.

$$\mathrm{e}(\tilde{\xi}, \check{\xi}) \leq 2\, ||\mathrm{F}_{\tilde{\xi}} - \mathrm{F}_{\check{\xi}}||_\infty$$

Therefore, to reduce the approximation error it suffices to minimize the sup-distance between the objective functions. This sup-distance can be bounded in turn by the *Wasserstein distance* between the true and approximate random vectors.

Definition 2.3 (Wasserstein distance)**.** *The Wasserstein distance (with respect to the 1-norm) between two random vectors $\tilde{\xi}$ and $\check{\xi}$ on $\mathbb{R}^m$ is defined as follows:*

$$\inf \{\mathbb{E}\left[\|Y_1 - Y_2\|_1\right] : \textit{for all } Y_1 \sim \tilde{\xi},\ Y_2 \sim \check{\xi} \textit{ defined on the same probability space}\}$$

The Wasserstein distance is related to the sup-distance by the Kantorovich-Rubinstein Theorem. Before stating this we first recall the definition of a Lipschitz function.

Definition 2.4 (Lipschitz). *For a function* $\mathrm{g} : \Xi \subset \mathbb{R}^m \to \mathbb{R}$*, its Lipshitz constant is defined as follows:*

$$L(\mathrm{g}) = \inf\{L : |\mathrm{g}(u) - \mathrm{g}(v)| \leq L\|u - v\| \text{ for all } u, v \in \mathbb{R}^m\} \tag{C.2}$$

The function g is said to be Lipschitz if $L(g) < \infty$*.*

Theorem 2.5 (Kantorovich-Rubinstein). *For random vectors* $\tilde{\xi}$ *and* $\check{\xi}$ *on* $\mathbb{R}^n$*, the Wasserstein distance (with respect to the 1-norm) can be written as follows:*

$$d_W(\tilde{\xi}, \check{\xi}) = \sup\{\mathbb{E}\left[\mathrm{g}\left(\tilde{\xi}\right)\right] - \mathbb{E}\left[\mathrm{g}\left(\check{\xi}\right)\right] : L_1(g) \leq 1\}.$$

For a proof of this theorem see [6, Chapter 1].

Suppose now that $\bar{L} > 0$ is a Lipschitz constant for our loss function, uniform across all decisions $x \in \mathcal{X}$, that is

$$|\mathrm{f}(x, \xi_1) - \mathrm{f}(x, \xi_2)| \leq \bar{L} d_W(\xi_1, \xi_2) \qquad \text{for all } x \in X, \text{ and } \xi_1, \xi_2 \in \Xi.$$

Then, $\xi \mapsto \frac{1}{\bar{L}}\mathrm{f}(x, \xi)$ is Lipschitz with Lipschitz less than or equal to 1 for all $x \in \mathcal{X}$ and so applying the Kantorovich-Rubinstein Theorem we have

$$\begin{aligned} ||\mathrm{F}_{\tilde{\xi}} - \mathrm{F}_{\check{\xi}}||_\infty &= \sup_{x \in X}\{\mathbb{E}_{\tilde{\xi}}\left[\mathrm{f}(x, \tilde{\xi})\right] - \mathbb{E}\left[\mathrm{f}(x, \check{\xi})\right]\} \\ &\leq \bar{L} d_W(\tilde{\xi}, \check{\xi}). \end{aligned} \tag{C.3}$$

Since the Wasserstein distance bounds the approximation error, some scenario generation and reduction algorithms have been designed so as to minimize it, see for example [2, 7]. In this work the Wasserstein distance is just used to analyze the performance of our methodology.

2.2 Estimation of the optimality gap

The bound on the approximation error described above is typically too conservative to be used in practice. Instead we resort to a statistical method to measure the quality of a solution.

Suppose we have a feasible solution $x_0 \in \mathcal{X}$ to the problem (C.1). Recall that the optimality gap of a feasible solution x_0 is defined as follows:

$$G = \mathbb{E}\left[f(x_0, \tilde{\xi})\right] - z^*$$

where $z^* = \min_{x \in \mathcal{X}} \mathbb{E}_{\tilde{\xi}}\left[f(x, \tilde{\xi})\right]$. A confidence interval for G can be constructed by solving the problem for multiple sampled scenario sets. The method presented here is sometimes called the *multiple replication procedure* (MRP) and originates from [8].

Let $\tilde{\xi}_{i1}, \ldots, \tilde{\xi}_{in}$ for $1 \leq i \leq n_g$, be independent identically distributed (i.i.d.) batches of random vectors and define $z^*_{ni} = \min_{x \in \mathcal{X}} \frac{1}{n} \sum_{j=1}^{n} f(x, \tilde{\xi}_{ij})$. Note that the elements $\tilde{\xi}_{ij}$ for $j = 1, \ldots, n$ within a batch do not need to be i.i.d. as it is sufficient that a batch yields an unbiased estimator for the expected cost function. Define G^i_n to be the estimated optimality gap for the i-th batch of scenarios:

$$G^i_n = z^*_{ni} - \frac{1}{n} \sum_{j=1}^{n} f(x_0, \tilde{\xi}_{ij}).$$

For $0 < \alpha < 1$, a conservative α confidence interval for the optimality gap G is

$$(0, \bar{G}_n + \epsilon_{n_g, \alpha}), \tag{C.4}$$

where

$$\begin{aligned}
\bar{G}_n &= \frac{1}{n_g} \sum_{j=1}^{n_g} G^j_n, \\
S^2_{n_g} &= \frac{1}{n_g - 1} \sum_{j=1}^{n_g} \left(G^j_n - \bar{G}_n\right), \\
\epsilon_{n_g, \alpha} &= t_{n_g - 1, \alpha} \frac{S_{n_g}}{\sqrt{n_g}}
\end{aligned}$$

and $t_{n_g - 1, \alpha}$ is the α-quantile of the (univariate) t-distribution with $n_g - 1$ degrees of freedom.

The main drawback of the above method for estimating a confidence interval for the optimality gap is that it involves solving multiple problems.

Other procedures have been proposed which require only one or two replications [9], [10].

3 The univariate newsvendor problem

In this section we sketch our approach to scenario generation for the univariate newsvendor problem. The more general approach is presented in Section 4.

In the newsvendor problem one must choose a quantity of stock to satisfy an uncertain demand. The problem can be formulated as a stochastic program:

$$\underset{l \le x \le u}{\text{minimize}} \; \mathbb{E}_{\tilde{\xi}} \left[h(x - \tilde{\xi})_+ + R(\tilde{\xi} - x)_+ \right], \tag{C.5}$$

where x is the decision of how much stock to order, $\tilde{\xi}$ is random variable representing demand, h is the unit storage cost of unsold product, and R is the unit rejection cost of surplus demand. Crucially, note that we have also assumed bounds l and u on the amount of stock we can order. These bounds may come from the context of the problem (e.g. a lower bound representing a minimum order size, and an upper bound for a budget restriction), or they may just define an interval inside which one is sure the optimal solution resides.

The above problem can be solved exactly without recourse to discretization. The optimal solution to the above problem is as follows:

$$x^* = \begin{cases} G^{-1}\left(\frac{R}{R+h}\right) & \text{if } l \le F^{-1}\left(\frac{R}{R+h}\right) \le u \\ l & \text{if } F^{-1}\left(\frac{R}{R+h}\right) < l \\ u & \text{otherwise,} \end{cases} \tag{C.6}$$

where G^{-1} denotes the generalized inverse distribution function of $\tilde{\xi}$. For illustrative purposes we will suppose that for the problem (C.5) we need to to discretize $\tilde{\xi}$.

Given $l \leqslant x \leqslant u$, we can rewrite the objective function as follows:

$$
\begin{aligned}
&h\mathbb{E}\left[\left(x-\tilde{\xi}\right)_{+}\right]+R\mathbb{E}\left[\left(\tilde{\xi}-x\right)_{+}\right] \\
&\quad=h\mathbb{E}\left[x-\tilde{\xi}|\tilde{\xi}\leqslant x\right]\mathbb{P}\left(\tilde{\xi}\leqslant x\right)+R\mathbb{E}\left[\tilde{\xi}-x|\tilde{\xi}\geqslant x\right]\mathbb{P}\left(\tilde{\xi}\geq x\right) \\
&\quad=h\mathbb{E}\left[x-\tilde{\xi}|\tilde{\xi}<l\right]\mathbb{P}\left(\tilde{\xi}<l\right)+h\mathbb{E}\left[x-\tilde{\xi}|l\leqslant\tilde{\xi}\leqslant x\right]\mathbb{P}\left(l\leqslant\tilde{\xi}\leqslant x\right) \\
&\qquad\quad+R\mathbb{E}\left[\tilde{\xi}-x|\tilde{\xi}>u\right]\mathbb{P}\left(\tilde{\xi}>u\right)+R\mathbb{E}\left[\tilde{\xi}-x|x\leqslant\tilde{\xi}\leqslant u\right]\mathbb{P}\left(x\leqslant\tilde{\xi}\leqslant u\right) \\
&\quad=h\mathbb{P}\left(\tilde{\xi}<l\right)\left(\mathbb{E}\left[\tilde{\xi}|\tilde{\xi}<l\right]-x\right)+R\mathbb{P}\left(\tilde{\xi}>u\right)\left(x-\mathbb{E}\left[\tilde{\xi}|\tilde{\xi}>u\right]\right) \\
&\qquad\quad+h\mathbb{E}\left[x-\tilde{\xi}|l\leqslant\tilde{\xi}\leqslant x\right]\mathbb{P}\left(l\leq\tilde{\xi}\leqslant x\right) \\
&\qquad\qquad+R\mathbb{E}\left[\tilde{\xi}-x|x\leqslant\tilde{\xi}\leqslant u\right]\mathbb{P}\left(x\leq\tilde{\xi}\leqslant u\right)
\end{aligned}
$$

The final lines show that in order to approximate the expected loss function correctly for $l \leqslant x \leqslant u$, with respect to the distribution of $\tilde{\xi}$ below l, only the probability of this event and its conditional expectation are important. Similarly, for the part of the distribution of $\tilde{\xi}$ above u, only the this probability and the conditional expectation are important.

Now, for a discrete approximation $\breve{\xi}$ of the random vector $\tilde{\xi}$, the conditional expectation of the event $\{\breve{\xi} \leqslant l\}$ and its probability can be set correctly with just one scenario in this part of the distribution: $\left(\mathbb{E}\left[\tilde{\xi}|\tilde{\xi}\leqslant l\right], \mathbb{P}\left(\tilde{\xi}\leqslant l\right)\right)$. Similarly, the conditional expectation and probability of $\{\breve{\xi} > u\}$ can be set correctly with a single scenario: $\left(\mathbb{E}\left[\tilde{\xi}|\tilde{\xi}>u\right], \mathbb{P}\left(\tilde{\xi}>u\right)\right)$. This suggests the following approach to scenario generation for this problem: use the single scenarios above for the lower and upper tails of the distribution, and discretize the body of the distribution using standard methods, normalizing the probabilities of the these scenarios appropriately. We test this approach where the main body of distribution is discretized via (rejection) sampling and call this method *newsvendor sampling*.

We have tested newsvendor sampling for the problem (C.5) where demand follows a scaled Beta distribution $\tilde{\xi} \sim 5\ \mathrm{Beta}(0.5, 0.5)$, $h = 0.5$, $R = 5.0$ and we use the bounds $l = 1.0$ and $u = 4.0$. Note that this distribution was chosen in particular because it has a lot of mass in its tails. In Figure C.1 we

have plotted examples of approximate expected loss functions for sampling and newsvendor sampling for this problem. Whereas the approximation for sampling for 10 scenarios is quite bad, the approximation for newsvendor sampling is able to match the true objective function well.

We compare the performance of newsvendor sampling against basic sampling by measuring the optimality gap of solutions that each method yields. For a range of scenario set sizes, we construct 30 scenario sets via sampling and newsvendor sampling, solve the corresponding stochastic program and calculate the optimality gap of the solutions with respect to the true problem. The results are shown in box plots in Figure C.2. This clearly demonstrates that newsvendor sampling performs much better than sampling in terms of the quality of solution and the stability.

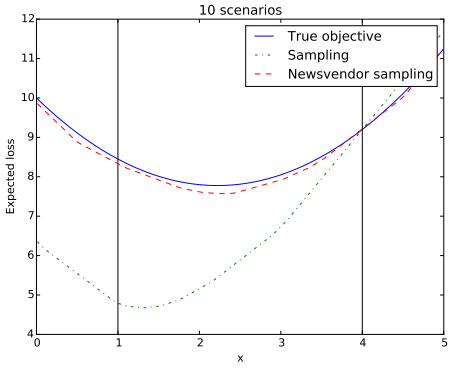


Fig. C.1: Comparison of sampling and newsvendor sampling approximations for univariate newsvendor problem

4 General case

In this section we generalize the above approach to scenario generation. In Section 4.1 we present a decomposition of the loss function which is required for newsvendor sampling, and in Section 4.2 we show how this is exploited, and give a simple probabilistic analysis of our approach.

4.1 Inactive Components

For some measurable set $I \subset \mathbb{R}^d$ denote by $\mathbb{1}_I$ its indicator function:

$$\begin{aligned} \mathbb{1}_I : &\mathbb{R}^d \to \{0,1\} \\ &\xi \mapsto \begin{cases} 1 & \text{if } \xi \in I \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Our approach to scenario generation in this work relies on there existing measurable $I \subset \mathbb{R}^d$ such that our loss function can be decomposed as follows:

$$f(x,\xi) = \mathbb{1}_I(\xi)\ \langle \mathrm{a}(x), \mathrm{b}(\xi) \rangle + \mathbb{1}_{I^c}(\xi)\ \mathrm{g}(x,\xi)$$

where $\mathrm{a} : \mathcal{X} \to \mathbb{R}^m$, $\mathrm{b} : \mathbb{R}^d \to \mathbb{R}^m$ and $\mathrm{g} : \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}$, and $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product. Given this decomposition, the expected loss function can now be written in following way:

$$\mathbb{E}\left[f(x,\tilde{\xi})\right] = \langle a(x), \mathbb{E}\left[\mathbb{1}_I(\tilde{\xi})\ b\left(\tilde{\xi}\right)\right] \rangle + \mathbb{E}\left[\mathbb{1}_{I^c}(\tilde{\xi})\ g(x,\tilde{\xi})\right].$$

If we are trying to approximate the random vector $\tilde{\xi}$ with another in order to well approximate the expected loss function, in the region I we only need to approximate the distribution so that the expectation $\mathbb{E}\left[\mathbb{1}_I(\tilde{\xi})\ b(\tilde{\xi})\right]$ is correct. The distribution of $\tilde{\xi}$ in I does not in any other way affect the value of the

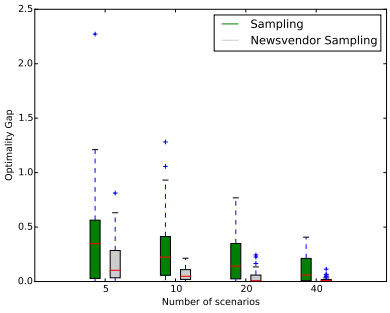


Fig. C.2: Comparison of performance of newsvendor sampling and sampling

expected loss function. This motivates the term *inactive component* for such a region.

Generalising the above, we now suppose our loss function has disjoint inactive components $I_1, \dots I_p \subset \Xi$, that is

$$\mathrm{f}\left(x, \xi\right) = \mathbb{1}_{I_i}(\xi)\ \left\langle \mathrm{a}_i\left(x\right), \mathrm{b}_i\left(\xi\right)\right\rangle + \mathbb{1}_{I_i^c}(\xi)\ \mathrm{g}_i\left(x, \tilde{\xi}\right) \qquad \text{for } i = 1, \dots, p.$$

Now, by fixing x and comparing the values of the above expressions for different $\tilde{\xi}$ we see that

$$\mathrm{f}\left(x, \xi\right) = \sum_{i=1}^{p} \mathbb{1}_{I_1}(\xi)\ \left\langle \mathrm{a}_i\left(x\right), \mathrm{b}_i\left(\xi\right)\right\rangle + \mathbb{1}_{\left(\bigcup_{i=1}^{p} I_i\right)^c}(\xi)\ \mathrm{g}_i\left(x, \xi\right)$$

and so

$$\mathbb{E}\left[\mathrm{f}\left(x, \tilde{\xi}\right)\right] = \sum_{i=1}^{p} \left\langle \mathrm{a}_i\left(x\right), \mathbb{E}\left[\mathbb{1}_{I_1}(\xi)\ \mathrm{b}_i\left(\xi\right)\right]\right\rangle + \mathbb{E}\left[\mathbb{1}_{\left(\bigcup_{i=1}^{p} I_i\right)^c}(\xi)\ \mathrm{g}_i\left(x, \xi\right)\right]$$

and so again, to approximate the expected loss function, in the inactive components I_i, only the expectations $\mathbb{E}\left[\mathbb{1}_{I_1}(\xi)\ \mathrm{b}_i\left(\xi\right)\right]$ need be correct. We refer to the complement of the union of inactive components $\left(\bigcup_{i=1}^{p} I_i\right)^c$ as the *active region*.

As an example, for the newsvendor problem given in Section 1, the loss function is decomposed as follows:

$$\begin{aligned}\mathrm{f}\left(x, \xi\right) = \mathbb{1}_{(-\infty, l)}(\xi)\ \left\langle \begin{pmatrix} hx \\ -h \end{pmatrix}, \begin{pmatrix} 1 \\ \xi \end{pmatrix}\right\rangle + \mathbb{1}_{(u, \infty)}(\xi)\ \left\langle \begin{pmatrix} R \\ -Rx \end{pmatrix}, \begin{pmatrix} \xi \\ 1 \end{pmatrix}\right\rangle \\ + \mathbb{1}_{[l, u]}(\xi)\ \left(\mathbb{1}_{[l, x]}(\xi)\ h(x - \xi) + \mathbb{1}_{(x, u]}(\xi)\ R(\xi - x)\right). \qquad \text{(C.7)}\end{aligned}$$

Therefore there are two inactive components for this problem, $(-\infty, l)$ and (u, ∞), and the active region is $[l, u]$.

Decomposition for the newsvendor problem was straight-forward because because x and $\tilde{\xi}$ were scalar quantities. Decomposition is similarly easy when the loss function can be separated into decomposable functions as follows:

$$\begin{aligned}\mathrm{f}\left(x, \xi\right) &= \mathrm{f}_1\left(x, \xi\right) + \mathrm{f}_2\left(x, \xi\right) \\ \text{where } \mathrm{f}_i\left(x, \xi\right) &= \sum_{j=1}^{n_i} \mathbb{1}_{I_{1j}}(\xi)\ \left\langle a_{1j}(x), b_{1j}(\xi)\right\rangle + \mathbb{1}_{\left(\bigcup_{i=1}^{n_1} I_{1j}\right)^c}(\xi)\ g_1(x, \xi). \qquad \text{(C.8)}\end{aligned}$$

Noticing that $\langle a_{1j}(x), b_{1j}(\tilde{\xi})\rangle + \langle a_{2k}(x), b_{2k}(\tilde{\xi})\rangle$ can be rewritten as $\langle \begin{pmatrix} a_{1j}(x) \\ a_{2k}(x) \end{pmatrix}, \begin{pmatrix} b_{1j}(\tilde{\xi}) \\ b_{2k}(\tilde{\xi}) \end{pmatrix} \rangle$ we can write:

$$
\begin{aligned}
\mathrm{f}(x,\xi) = & \sum_{k=1}^{n_2}\sum_{j=1}^{n_1} \mathbb{1}_{I_{1j}\cap I_{2k}}(\xi)\ \langle \begin{pmatrix} a_{1j}(x) \\ a_{2k}(x) \end{pmatrix}, \begin{pmatrix} b_{1j}(\xi) \\ b_{2k}(\xi) \end{pmatrix} \rangle \\
& + \mathbb{1}_{\left(\cup_{j,k}(I_{1j}\cap I_{2k})\right)^c}(\xi)\ \Bigg(\sum_{k=1}^{n_2}\sum_{j=1}^{n_1} \mathbb{1}_{I_{1j}^c\cap I_{2k}}(\xi)\ \left(g_1(x,\xi) + \langle a_{2k}(x), b_{2k}(\xi)\rangle\right) \\
& + \sum_{k=1}^{n_2}\sum_{j=1}^{n_1} \mathbb{1}_{I_{1j}\cap I_{2k}^c}(\xi)\ \left(\langle a_{1k}(x), b_{1k}(\xi)\rangle + g_2(x,\xi)\right) \\
& + \sum_{k=1}^{n_2}\sum_{j=1}^{n_1} \mathbb{1}_{I_{1j}^c\cap I_{2k}^c}(\xi)\ \left(g_1(x,\xi) + g_2(x,\xi)\right) \Bigg)
\end{aligned}
$$

Therefore, the set of inactive components for $\mathrm{f}(x,\xi)$ is $\{I_{1j}\cap I_{2k} :\ 1\le j\le n_1,\ 1\le k\le n_2\ \}$. More generally, suppose $\mathrm{f}(x,\xi) = \sum_{i=1}^{N}\mathrm{f}_i(x,\xi)$ where each $\mathrm{f}_i(x,\xi)$ is defined as in (C.8). Setting $J = \prod_{k=1}^{N}[1,n_k]$, we can write:

$$
\begin{aligned}
\mathrm{f}(x,\xi) = & \sum_{(j_1,\ldots,j_N)\in J} \mathbb{1}_{\cap_{k=1}^{N} I_{kj_k}}(\xi)\ \langle \begin{pmatrix} a_{1j_1}(x) \\ \vdots \\ a_{Nj_N}(x) \end{pmatrix}, \begin{pmatrix} b_{1j_1}(\xi) \\ \vdots \\ b_{Nj_N}(\xi) \end{pmatrix} \rangle \\
& + \mathbb{1}_{\left(\cup_{(j_1,\ldots,j_N)\in J}\cap_{k=1}^{N} I_{kj_k}\right)^c}(\xi)\ \left(\ldots\right)
\end{aligned}
$$

where the ellipsis covers all other possible intersections of I_{ij} and I_{ij}^c each of which involves at least one instance of the function $\mathrm{g}_i(x,\xi)$. The set of inactive components in this case is thus

$$
\{\cap_{k=1}^{N} I_{kj_k} : (j_1,\ldots,j_k)\in\prod_{k=1}^{N}[1,n_k]\}. \tag{C.9}
$$

4.2 Scenario generation

We suppose we have the following decomposition of the cost function:

$$
\mathrm{f}(x,\xi) = \sum_{i=1}^{p} \mathbb{1}_{I_i}(\xi)\ \langle \mathrm{a}_i(x), \mathrm{b}_i(\xi)\rangle + \mathbb{1}_A(\xi)\ \mathrm{g}(x,\xi)
$$

where $A = \left(\bigcup_{i=1}^{p} I_i\right)^c$, and that we are trying to approximate continuous random vector $\tilde{\xi}$ with some scenario set in order to approximate the expected cost function well. Our general approach to scenario generation will be to represent the inactive components I_i by a number of scenarios $(\xi_{ik}, p_{ik})_{k=1}^{n_i}$ sufficient so that:

$$\sum_{k=1}^{n_i} p_{ik} b_i(\xi_{ik}) = \mathbb{E}\left[\mathbb{1}_{I_i} b_i(\tilde{\xi})\right],$$
$$\sum_{k=1}^{n_i} p_{ik} = \mathbb{P}\left(\tilde{\xi} \in I_i\right).$$

In the case where $\mathrm{b}_i(\xi)$ is an affine function, and I_i is convex, it is enough to represent the region I_i with a single scenario:

$$\xi_i = \mathbb{E}\left[\tilde{\xi} | \tilde{\xi} \in I_i\right], \; p_i = \mathbb{P}\left(\tilde{\xi} \in I_i\right). \tag{C.10}$$

We assume here that the expectations $\mathbb{E}\left[\mathrm{b}_i\left(\tilde{\xi}\right) | \tilde{\xi} \in I_i\right]$ and probabilities can be calculated accurately, for example by numerical integration or Monte Carlo simulation. The distribution in the active region can be approximated by some other method. The generalization of newsvendor sampling in Section 3 is to represent each inactive component by the single scenario (C.10), and to construct scenarios in the active region through rejection sampling, again normalizing the probabilities of the scenarios in the active region appropriately. Figure C.3 shows a scenario set constructed by rejection sampling for a simple recourse problem. See Section 5 for more details.

Probabilistic Analysis Let $\tilde{\xi}$ and $\breve{\xi}$ be random vectors which have measure probability measures $\tilde{\mu}$ and $\breve{\mu}$ respectively. Denote by $\tilde{\xi}_A$ the random vector $\tilde{\xi}$ conditioned on being in the active region A which we assume to be non-negligible. That is, $\tilde{\xi}_A$ is the random vector with measure $\tilde{\mu}_A$ which is defined a follows:

$$\tilde{\mu}_A(B) = \frac{1}{\tilde{\mu}(A)} \tilde{\mu}(A \cap B), \qquad \text{for all measurable } B \subset \mathbb{R}^n.$$

The random vector $\breve{\xi}_A$ is defined analogously.

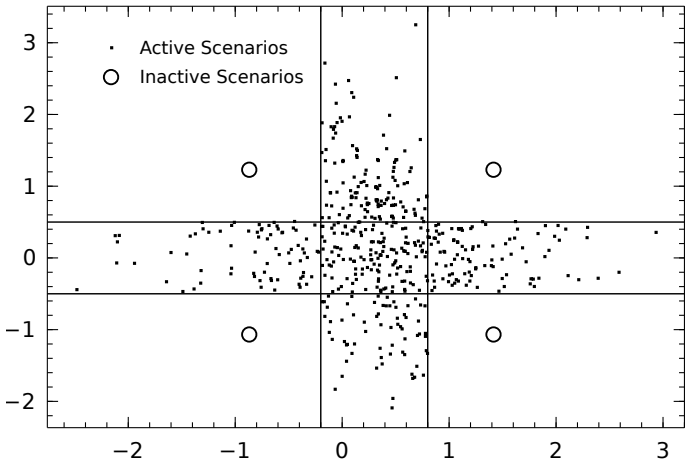


Fig. C.3: Example of scenario generation approach for a two-dimensional newsvendor problem

Now, suppose that for each $i = 1, \ldots, p$ we have

$$\mathbb{E}\left[\mathbb{1}_{I_i} b_i(\tilde{\xi})\right] = \mathbb{E}\left[\mathbb{1}_{I_i} b_i(\breve{\xi})\right]$$
$$\mathbb{P}\left(\tilde{\xi} \in I_i\right) = \mathbb{P}\left(\breve{\xi} \in I_i\right)$$

then $\mathbb{P}\left(\tilde{\xi} \in A\right) = \mathbb{P}\left(\breve{\xi} \in A\right)$ and

$$\begin{aligned}
||\mathrm{F}_{\tilde{\xi}} - \mathrm{F}_{\breve{\xi}}||_\infty &= \left|\mathbb{E}_{\tilde{\xi}}\left[\mathrm{f}\left(x, \tilde{\xi}\right)\right] - \mathbb{E}_{\breve{\xi}}\left[\mathrm{f}\left(x, \breve{\xi}\right)\right]\right| \\
&= \left|\mathbb{E}_{\tilde{\xi}}\left[\mathbb{1}_A(\tilde{\xi})\ \mathrm{g}\left(x, \tilde{\xi}\right)\right] - \mathbb{E}_{\breve{\xi}}\left[\mathbb{1}_A(\breve{\xi})\ \mathrm{g}\left(x, \breve{\xi}\right)\right]\right| \\
&= \left|\mathbb{P}\left(\tilde{\xi} \in A\right)\ \mathbb{E}\left[\mathrm{g}\left(x, \tilde{\xi}_A\right)\right] - \mathbb{P}\left(\breve{\xi} \in A\right)\mathbb{E}\left[\mathrm{g}\left(x, \breve{\xi}_A\right)\right]\right| \\
&= \mathbb{P}\left(\tilde{\xi} \in A\right)\left|\ \mathbb{E}\left[\mathrm{g}\left(x, \tilde{\xi}_A\right)\right] - \mathbb{E}\left[\mathrm{g}\left(x, \breve{\xi}_A\right)\right]\right| \\
&\leq \mathbb{P}\left(\tilde{\xi} \in A\right)\ \bar{L}_A\, d_W(\tilde{\xi}_A, \breve{\xi}_A)
\end{aligned} \tag{C.11}$$

where $\bar{L}_A$ is a uniform Lipschitz constant for $\xi \mapsto g(x, \xi)$ for $\xi \in A$ and over $x \in \mathcal{X}$. The final inequality follows by Theorem 2.5.

Note that the inequality in (C.11) can be expected to be significantly tighter than that in (C.3). Firstly, for a fixed scenario set size, we can expect $d_W(\tilde{\xi}_A, \breve{\xi}_A)$ to be smaller than $d_W(\tilde{\xi}, \breve{\xi})$ since we will be spreading more

scenarios over a smaller region of the distribution. Secondly, the bound has been scaled down by $\mathbb{P}\left(\tilde{\xi} \in A\right)$. We can therefore expect our methodology to be more effective the larger the combined probability of the inactive regions.

5 Simple recourse problems

We now characterize the inactive components of a class of stochastic linear programs known as *simple recourse problems*. These have the following form:

$$\begin{aligned} &\text{minimize } c^T x + \mathbb{E}\left[\mathrm{f}\left(x, \tilde{\xi}\right)\right] \\ &\text{subject to } Ax \leq b \\ &\qquad x \geq 0 \\ &\text{where } f(x, \tilde{\xi}) = \min_{y_+, y_- \geq 0} \{q^T y_+ + r^T y_- : Tx + Iy_+ - Iy_- = \tilde{\xi}\} \end{aligned}$$

and $c \in \mathbb{R}^k$, $q, r \in \mathbb{R}^d$ and $T \in \mathbb{R}^{k \times d}$. Denoting the rows of T by T_i for $i = 1, \ldots, d$, the loss function can be decomposed as follows:

$$\begin{aligned} \mathrm{f}\left(x, \tilde{\xi}\right) &= \sum_{i=1}^{m} \mathrm{f}_i\left(x, \tilde{\xi}\right) \\ \text{where } \mathrm{f}_i\left(x, \tilde{\xi}\right) &= q_i\left(\tilde{\xi}_i - T_i x\right)_+ + r_i\left(T_i x - \tilde{\xi}_i\right)_+ . \end{aligned}$$

Assuming that for each $i = 1, \ldots, m$ we have the constraints $l_i \leq T_i x \leq u_i$, the above summands can be decomposed further, in a similar way as in the basic newsvendor problem in (C.7):

$$\begin{aligned} \mathrm{f}_i\left(x, \xi\right) = {} & q_i \mathbb{1}_{\{\xi:\ \xi_i > u_i\}}(\xi)\ \left(\xi_i - T_i x\right) + r_i \mathbb{1}_{\{\xi:\ \xi_i < l_i\}}(\xi)\ \left(T_i x - \xi_i\right) \\ & + \mathbb{1}_{\{\xi:\ l_i \leq \xi_i \leq u_i\}}(\xi)\ \left(q_i \mathbb{1}_{\{\xi:\ \xi_i \geq T_i x\}}(\xi)\ \left(\xi_i - T_i x\right) + r_i \mathbb{1}_{\{\xi:\ \xi_i \leq T_i x\}}(\xi)\ \left(T_i x - \xi_i\right)\right). \end{aligned}$$

and so the inactive components of $\mathrm{f}_i\left(x, \xi\right)$ with respect to the i-th component are $\{\xi : \xi_i < l_i\}$ and $\{\xi : \xi_i > u_i\}$. By (C.9) the inactive components for

this problem are therefore all possible intersections of these two sets over $i = 1, \ldots, d$.

General constraints Suppose that our feasible region is given by $\mathcal{X} = \{x \in \mathbb{R}^n : Ax \leq b\}$. It is possible to extract simple bounds $l_i \leq T_i x \leq u_i$ from this feasible region by solving a series of simple linear programs. Specifically, the lower bounds are given by $ł_i = \min\{T_i x : Ax \leq b\}$ and the upper bounds are given by $u_i = \max\{T_i x : Ax \leq b\}$. Rather than solving each of these problems independently, a more efficient way would be to express $\mathcal{X}$ as a convex-conical combination. Classical results from polyhedral geometry [11, Chapter 1] state that the finite intersection of half-spaces can be expressed as follows:

$$\{Vt + Yu : t \geq 0,\ u \geq 0,\ \mathbb{1}^T t = 1\}$$

where $V \in \mathbb{R}^{k \times m_1}$, $Y \in \mathbb{R}^{k \times m_2}$, $t \in \mathbb{R}^{m_1}$ and $u \in \mathbb{R}^{m_2}$. Many efficient algorithms exist for calculating this representation, for example the generalized Chernikova algorithm [12]. Using the convention that $V = \mathbf{0}$ if there are no convex hull generators, and $Y = \mathbf{0}$ if there are no cone generators, the simple bounds can now be calculated as follows:

$$l_i = \begin{cases} \min_{j=1,\ldots,m_1} T_i V_j & \text{if } \min_{j=1,\ldots,m_2} T_i Y_j \geq 0 \\ -\infty & \text{otherwise} \end{cases}$$

$$u_i = \begin{cases} \max_{j=1,\ldots,m_1} T_i V_j & \text{if } \max_{j=1,\ldots,m_2} T_i Y_j \leq 0 \\ +\infty & \text{otherwise} \end{cases}$$

where $V_i \in \mathbb{R}^k$ for $i = 1, \ldots, m_1$ and $Y_i \in \mathbb{R}^k$ for $i = 1, \ldots, m_2$ denote the columns of V and Y respectively.

The extraction of simple box constraints from polyhedral constraints is illustrated in Figure C.4.

Probabilities of inactive regions In Section 4.2 we noted that the performance of our methodology improves as the probability of the inactive com-

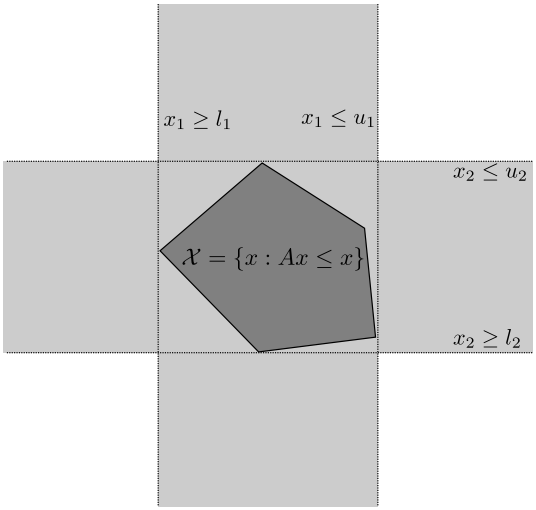


Fig. C.4: The extraction of simple constraints from general polyhedral constraints.

ponents increases. In this section we calculate the probability of inactive components for some basic distributions.

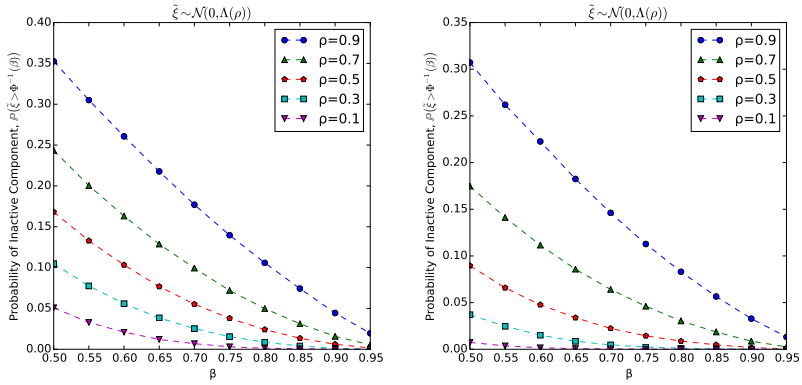
Suppose first that $\tilde{\xi}$ is uniformly distributed on $[0,1]^d$ and that $0 < l < u < 1$. If the inactive components for the problem are all possible intersections of the sets $\{\xi \in \mathbb{R}^d : \xi_i < l\}$ and $\{\xi \in \mathbb{R}^d : \xi_i > u\}$ over $i = 1, \ldots, d$ then the total probability of the inactive components is as follows:

$$\begin{aligned}\sum_{j=0}^{d} \binom{d}{j} \mathbb{P}\left(\tilde{\xi}_1 < l\right)^j \mathbb{P}\left(\tilde{\xi}_1 > u\right)^{d-j} &= \sum_{j=0}^{d} \binom{d}{j} l^j (1-u)^{d-j} \\ &= (l + (1-u))^d.\end{aligned}$$

Since we have $l + (1-u) < 1$ this probability will rapidly diminish to zero as the dimension of the random vector increases.

However, in the case of strong correlations this probability decreases much more slowly. In Figure C.5 we have plotted for the multivariate Normal distribution the probabilities of the inactive component $\{\xi \in \mathbb{R}^d : \xi > \Phi^{-1}(\beta)\}$ where Φ^{-1} denotes the inverse distribution function of a standard Normal distribution and the inequality applies element-wise. These calculations are

done for a particular type of correlation matrix $\Lambda(\rho) \in \mathbb{R}^{m \times m}$ where $\Lambda(\rho)_{ij} = \rho$ for $i \neq j$ and $\rho > 0$.



(a) $d = 5$ **(b)** $d = 10$

Fig. C.5: Probability of an inactive component for the multivariate Normal distribution

The figure shows that the stronger the correlations, the higher the probability of the inactive component. The probability for moderate correlations ($\rho \geq 0.5$) is much higher than was the case for the uniform distributions. For example for $d = 5$, when $\tilde{\xi}$ is uniformly distribution on $[0,1]^d$ and $u = 0.7$ we have $\mathbb{P}\left(\tilde{\xi} > u\right) = (0.3)^5 \approx 0.0025$, whereas when $\tilde{\xi} \sim \mathcal{N}\left(\mathbf{0}, \Lambda(0.5)\right)$ and we use the corresponding upper bound $u = \Phi^{-1}(0.7)$ we have $\mathbb{P}\left(\tilde{\xi} > u\right) \approx 0.05$.

Note that for Normal distributions, by symmetry the inactive component $\{\xi \in \mathbb{R}^d : \xi < -\Phi^{-1}(\beta)\}$ will have the same probability as $\{\xi \in \mathbb{R}^d : \xi > \Phi^{-1}(\beta)\}$. Note also that strong negative correlations, or a mixture of strong positive and negative correlations will yield other inactive components with high probability.

6 Numerical Test

In this section, we compare the performance of our newsvendor sampling method against basic sampling. In particular, for a fixed computational bud-

get we will compare the performance of sampling and newsvendor sampling through estimation of the optimality gap using the MRP outlined in Section 2.2.

For this experiment we use a multi-product newsvendor problem with budget constraint:

$$\begin{aligned} \underset{x}{\text{maximize}} \ & \sum_{i=1}^{d} \left(h_i \mathbb{E}\left[\left(x_i - \tilde{\xi}_i \right)_+ \right] + R_i \mathbb{E}\left[\left(\tilde{\xi}_i - x \right)_+ \right] \right) \\ \text{subject to} \ & \sum_{i=1}^{d} x_i \leq \tau \\ & l \leq x \leq u. \end{aligned}$$

The experiment is carried out for a 5-dimensional problem. The parameters in this test have been constructed in such a way that the problem is sufficiently unstable that sampling doesn't perform well. This is done by selecting rejection penalties R_i that are a lot bigger than the holding cost h_i, which ensures the solution is in the upper tail of the distribution (see the solution of the univariate newsvendor problem in (C.6)). In this way, the solution to the sample average approximation will be unstable as only a small number of scenarios will fall in this region.

The distribution and constraints of the problem are chosen so that the total probability of the inactive components is large. The covariance matrix has been constructed to have high strong positive correlations and we have used relatively tight simple bounds $l \leq x \leq u$. None of the simple bounds $l \leq x \leq u$ are binding, that is, their presence does not change the set of optimal solutions to the problem. In this way, they can be viewed as "ghost" constraints that are only included to boost the performance of the methodology. The use of artificial constraints to boost the performance of a scenario generation methodology was prevalent in the paper [5]. The full set of problem data is given in Appendix A.

We compare the quality of solutions yielded by sampling and newsvendor sampling for scenario sets consisting of 100 scenarios. For the estimation of

the optimality gap we use five replications of scenario sets of size 50, and estimate the error using $\alpha = 0.95$. We repeat this across five trials and the results are shown in Table C.1.

Sampling		Newsvendor Sampling	
Gap ($\bar{G}_n$)	Error ($\epsilon_{n_g,\alpha}$)	Gap ($\bar{G}_n$)	Error ($\epsilon_{n_g,\alpha}$)
1.002	0.491	0.358	0.219
2.123	1.352	0.325	0.196
0.745	0.335	0.517	0.301
0.937	0.641	0.331	0.295
1.155	0.439	0.551	0.210

Table C.1: Estimated optimality gaps from sampling and newsvendor sampling. Inactive probability: 0.76

The results show that newsvendor sampling consistently produces solutions whose optimality gaps are much smaller than that of basic sampling.

7 Discussion and Future Work

In this paper, we have proposed a methodology of scenario generation which exploits the partition of the support of a random vector into active and inactive components. The inactive components are represented by a single (or very few) scenarios in such a way that the expectation of the loss function restricted to these components is exact. The rest of the scenarios are spread over the active region. This methodology was demonstrated in particular on simple recourse problems whose separability makes the partition into active and inactive components straight-forward. However, many methods already exist for solving simple recourse problems which limits the utility of our approach. The most important avenue of future research is therefore to find decompositions of loss functions for more difficult problems.

A simple probabilistic analysis of our method suggested that it would

yield scenario sets whose approximation error would be much lower than scenario generation methods which spread scenarios evenly across the support of the distribution. In particular, the effectiveness of the methodology improves as the probability of the inactive components increases. In the case of the simple recourse problem, the inactive components grew larger as our problem becomes more constrained. It is reasonable to expect this property to hold for other problems as well: the fewer feasible decisions available, the less the loss function varies, and so the fewer scenarios required to accurately calculate the expected loss. By adding artificial constraints (which do affect the set of optimal solutions) one could improve the performance of our methodology. Further work is required to determine how one can reliably construct such constraints.

We suggested a concrete scenario generation method for our methodology which we called newsvendor sampling. In this method the scenarios in the active region are constructed via rejection sampling. Another possibility would be to use this methodology as a scenario reduction technique, aggregating all scenarios in each inactive component into a single point.

A Numerical Test Problem Data

For the numerical test in Section 6 we use a 5-dimensional multi-product newsvendor problem. The exact parameters are detailed below.

Distribution The random vector $\tilde{\xi}$ is modeled by a multivariate t distribution.

$$\tilde{\xi} \sim t_{3.0}\left(\mu, \Sigma\right)$$

where

$$\mu = (1.0, 1.0, 1.0, 1.0, 1.0),\ \Sigma = \begin{pmatrix} 0.51 & 1.18 & 0.56 & 0.57 & 0.88 \\ 1.18 & 2.99 & 1.43 & 1.22 & 2.31 \\ 0.56 & 1.43 & 1.36 & 0.70 & 1.12 \\ 0.57 & 1.22 & 0.70 & 0.93 & 0.92 \\ 0.88 & 2.31 & 1.12 & 0.92 & 1.82 \end{pmatrix}.$$

Problem data We use constant values for the holding and rejection costs.

$$h = (2.5, 2.5, 2.5, 2.5, 2.5),\ R = (17.5, 17.5, 17.5, 17.5, 17.5).$$

We choose lower and upper bounds for the decision based on the quantiles of the marginal distributions:

$$l_i = G^{-1}_{\tilde{\xi}_i}(0.8),\ u_i = G^{-1}_{\tilde{\xi}_i}(0.95),$$

where $G^{-1}_{\tilde{\xi}_i}$ denotes the distribution function of the marginal random variable $\tilde{\xi}_i$, which gives

$$l = [2.7, 3.69, 3.14, 2.94, 3.32],\ u = (3.68, 6.07, 4.74, 4.27, 5.18).$$

Finally, the budget is set as follows:

$$\tau = \frac{1}{2}\sum_{i}^{5}(l_i + u_i) = 19.86.$$

References

[1] R. Van Slyke and R. J.-B. Wets, "L-shaped linear programs with applications to optimal control and stochastic programming," *SIAM Journal of Applied Mathematics*, vol. 17, pp. 638–663, 1969.

[2] G. C. Pflug, "Scenario tree generation for multiperiod financial optimization by optimal discretization," *Mathematical Programming*, vol. 89, no. 2, pp. 251–271, 2001.

[3] K. Høyland and S. W. Wallace, "Generating scenario trees for multistage decision problems," *Management Science*, vol. 47, no. 2, pp. 295–307, 2001.

[4] J. Fairbrother, A. Turner, and S. W. Wallace, "Scenario generation for stochastic programs with tail risk measures," ArXiv e-print 1511.03074, Nov. 2015. [Online]. Available: http://arxiv.org/abs/1511.03074

[5] ——, "Scenario generation for portfolio selection problems with tail risk measure," ArXiv e-print 1511.04935, Nov. 2015. [Online]. Available: http://arxiv.org/abs/1511.04935

[6] C. Villani, *Topics in Optimal Transportation*, ser. Graduate studies in mathematics. American Mathematical Society, 2003. [Online]. Available: https://books.google.be/books?id=GqRXYFxe0l0C

[7] J. Dupačová, N. Gröwe-Kuska, and W. Römisch, "Scenario reduction in stochastic programming: An approach using probability metrics," *Mathematical Programming*, vol. 95, no. 3, pp. 493–511, 2003.

[8] W. Mak, D. Morton, and R. Wood, "Monte Carlo bounding techniques for determining solution quality in stochastic programs," *Operations Research Letters*, vol. 24, pp. 47–56, 1999.

[9] G. Bayraksan and D. P. Morton, "Assessing solution quality in stochastic programs," *Mathematical Programming*, vol. 108, no. 2–3, pp. 495–514, sep 2006.

[10] R. Stockbridge and G. Bayraksan, "A probability metrics approach for reducing the bias of optimality gap estimators in two-stage stochastic linear programming," *Mathematical Programming*, vol. 142, no. 1–2, pp. 107–131, 2013.

[11] G. M. Ziegler, *Lectures on Polytopes (Graduate Texts in Mathematics)*. Springer, 4 2008.

[12] H. Le Verge, "A note on Chernikova's algorithm," IRISA, Rennes, France, Tech. Rep. 635, 1992.

Thesis Conclusions

This thesis has been concerned with the development of problem-driven scenario generation for stochastic programs, that is, methods of scenario generation which use the underlying structure of a problem to provide a more parsimonious, and thus more tractable representation of uncertain problem parameters. As was noted in the thesis summary in the introduction, such approaches in the literature have been rare and somewhat heuristic in nature. The aim of this research was therefore to develop scenario generation methods which were *mathematically adapted* to a specific problem.

Two approaches to problem-driven scenario generation were proposed in this thesis. The first of these approaches was adapted to stochastic programs which use tail-risk measures, that is risk measures which depend only on the upper tail of a distribution function. This was the subject of the first two papers of this thesis. The second approach, and the subject of the third and final paper of this thesis, exploited a special type of decomposition of the loss function, was demonstrated in particular for simple recourse problems.

We now present a condensed summary of the achievements and limitations of each of the papers, and finally discuss more broadly the contribution of this thesis and possible extensions.

Paper A Paper A of this thesis introduced the risk region methodology for stochastic programs with a tail risk measure. The risk region of such a problem was defined, loosely, to be the set of all possible outcomes of the un-

derlying random vector which lead to a loss in the upper tail of the loss distribution for some feasible decision. We showed that under mild conditions this region completely determines the value of a tail risk measure. In particular, all the mass in the complement of this region, the non-risk region, could be aggregated into a single point without affecting the value of the tail risk measure.

This observation motivated the proposal of two scenario generation methods: aggregation sampling and aggregation reduction. In the former algorithm, one samples scenarios scenarios sequentially, keeping any which lie in the risk region and aggregating any which lie in the non-risk region until one has sampled a specified number of scenarios in the risk region. In the latter algorithm, one samples a specified number of scenarios and then aggregates all scenarios in the non-risk region. We demonstrated that both of these methods were equivalent to sampling from the random vector where all the mass in the non-risk region has been concentrated into its conditional expectation.

The effectiveness of both of these methods were shown to improve as the probability of the non-risk region increases. As a consequence, our method performs better for higher levels of tail risk, and for problems which are more constrained, as both of these changes increase the size of the non-risk region. On the other hand, we observed that as the problem dimension increases, the probability of non-risk region tends to zero, rendering our method useless in high dimensions.

Finally, we gave a convenient characterization of the risk region for portfolio selection problems when the asset returns have elliptical distributions, and presented a simple numerical test which demonstrated the improvement in solution quality and stability of aggregation sampling over basic sampling.

The main limitation of this methodology is the convenient of characterization of the risk region. This is difficult as it depends on the distribution of the random vector, the loss function and the problem constraints. For more dif-

ficult problems we suspect that an exact represent of the risk region may not be possible, However, we may be able to find conservative risk regions, that is, regions which contain the true risk region. For the purposes of the above algorithms, it is valid to use a conservative risk region rather than the true risk region. The characterization of risk regions and conservative risk regions for problems other than portfolio selection is the most important direction of future research of this work.

Another limitation of this methodology is that aggregating all scenarios in the non-risk region does not in general preserve the expectation of the loss function. If expectation is used in the stochastic program then one might represent the non-risk region with a few points rather than a single one. For instance, one could adapt the above algorithms to cluster into a specified number of points, any scenarios sampled from the non-risk region. The ideal proportion of points used to represent the non-risk region, and how points in the non-risk region are clustered are other important directions for future research.

Paper B Paper B of this thesis concerns the application of the risk region approach to more realistic portfolio selection problems. The paper first addressed some technical issues which concern testing whether or not a point lies in the risk region. Next, we studied the empirical behavior of the probability of the non-risk region. Here, it was found that this probability increases as the tails of the distributions become heavier and more positively correlated. This is good news for the application of our methodology to real portfolio selection problems, as historical stock return data exhibit both of these characteristics.

The scenario generation methods were then tested for a wide range of distributions constructed from real data. We also tested here the use of approximate risk regions for non-elliptical distributions. There is a danger in using an approximate risk region: if the approximate risk region is too small,

the value of the tail risk measure may be distorted. Our results indicated that our methodology performs consistently well, even when using the approximate risk regions.

Finally, we demonstrated the use of ghost constraints for a difficult case study problem. As noted above, the performance of our methodology improves as the problem becomes more constrained as this increases the size of the non-risk region. A ghost constraint is an artificial constraint added to the problem simply to improve the performance of our methodology. A ghost constraint should be as tight as possible without removing any optimal solutions. However, this is necessarily difficult as the construction of such a constraint requires some knowledge of the optimal solution to the problem we want to solve. In our case study we resorted to the following heuristic: we constrain our feasible decision to some neighborhood of an optimal solution to a sampled problem. Out-of-sample testing was then used to verify that judicious usage of these constraints does indeed improve the solution.

There are several major directions in which this work can be extended. Firstly, it would be useful to prove results concerning the behavior of the probability of the non-risk region with respect to how heavy are the tails of the distribution, and with respect to its correlation structure. Next, to allow us to apply this methodology to more distributions, we need a better way of constructing an approximate risk region, and of diagnosing potential problems with these. Finally, the development of a systematic method of constructing ghost constraints would allow us to extract better solutions from a problem with less trial and error.

Paper C The final paper of this thesis proposed a different approach to scenario generation which exploits a special decomposition of the loss function. This decomposition induces a partition of the support of the problem random vector into inactive components and an active region. Each of the inactive components of the problem can be typically represented by a single scenario.

We then proposed the following simple approach to scenario generation: the scenarios for the inactive components are calculated by numerical integration or simulation, and a specified number of scenarios to represent active region are constructed via rejection sampling. We called this approach *newsvendor sampling*, as the newsvendor problem served as our primary example in the paper.

Like the risk region approach, the performance of newsvendor sampling improves as the probability of the inactive components increases. Since the inactive regions grow as the problem becomes more constrained, this approach is again constraint-driven. Although not studied in the paper, this means that one could again add ghost constraints to a problem to improve the performance of our methodology.

Newsvendor sampling also has the same dimensionality problems as with the risk region approach: the higher the dimension of the random vector, the smaller the probability of the inactive components. The severity of this effect again depends on the underlying distribution of the problem random vector. In the presence of strong correlations in particular mitigates against this effect.

Unlike the risk region approach, the form of the inactive components does not depend on the distribution[1], which vastly simplifies the test of whether or not a point belongs to an inactive component. In the risk region approach, one has to solve a small quadratic program for this test, whereas for the newsvendor approach, one only has to verify whether some simple inequalities hold.

The approach of this paper was only described in detail and tested for simple recourse problems. The next step in this thread of research would be to study how the loss functions of other problems can be decomposed in order to use this approach. As with the risk region approach, this method could be made more useful if we had a more systematic way of constructing

[1]The probability of the inactive components do depend on the distribution however.

ghost constraints.

Final Remarks The methodologies developed in this thesis work in following way: they partition the support of the random vector associated to a stochastic program into an active region which should be represented by many points, and inactive region which can be represented by a single or very few points. Moreover, the inactive regions grew as the problems became more constrained. In effect, the fewer decisions one can make, the more redundancy in the distribution, and so the more effective our methods. This observation means that if artificial constraints are added to the problem, their performance improves. The most important thread of future research is therefore the development of a systematic way of constructing such ghost constraints to a stochastic program.

The explicit scenario generation methods developed were based on sampling. Sampling is flexible, easy to implement and has desirable asymptotic properties. However, the essence of the methodologies was the partitioning of the support into active and inactive regions; the actual method of scenario construction in these regions could be more refined. For example, as we mentioned in Paper A, scenarios in the non-risk region could constructed through a clustering algorithm. However, one has to be careful in how the scenarios are constructed: the active regions constructed in this thesis were generally non-convex, and so scenarios constructed via, for example, the k-means clustering method, would not necessarily be in the active region.

For both methodologies of this thesis, their exact detailed application (in particular, the convenient characterizations of the active and inactive regions) were only explicitly given for restricted simple families of problems. As mentioned above, the next obvious step is to extend the presented analyses to more problem classes. However, exact analyses may not be possible for more complicated problems.

For problems where only an approximate analysis of the active regions is

available, an advanced sampling method such as stratified sampling or importance sampling may be more appropriate. These flexible sampling techniques allow one to prioritize the generation of samples in certain regions of a distribution, and the usual asymptotic theory can still be applied to them. This means that misspecifications of active regions would not (asymptotically) be a problem, unlike, for example, with the misspecification of a risk region in the aggregation sampling algorithm. However, given that the somewhat arbitrary shape of the active region, the development of such schemes is likely to be a non-trivial task.