Improving sample average approximation using distributional robustness *

E.J. Anderson† A.B. Philpott‡

October 5, 2019

Abstract

We consider stochastic optimization problems in which we aim to minimize the expected value of an objective function with respect to an unknown distribution of random parameters. We analyse the out-of-sample performance of solutions obtained by solving a distributionally robust version of the sample average approximation problem for unconstrained quadratic problems, and derive conditions under which these solutions are improved in comparison with those of the sample average approximation. We compare different mechanisms for constructing a robust solution: phi-divergence using both total variation and standard smooth ϕ functions; a CVaR-based risk measure; and a Wasserstein metric.

1 Introduction

In this paper we consider the general class of stochastic programming problems of the following form:

$$\text{SP:} \quad \min_{x \in X} \mathbb{E}_{\mathbb{P}}[c(x, \xi)].$$

*The authors would like to thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme Mathematics of Energy Systems when work on this paper was undertaken. This work was supported by EPSRC Grant Number EP/R014604/1, and the New Zealand Marsden Fund under contract UOA1520. The authors also acknowledge the contributions of discussions with Karen Willcox and Harrison Nguyen to this research.

†University of Sydney

‡University of Auckland

Here the decision variable x is constrained to lie in $X \subseteq \mathbb{R}^n$, and expectations are taken over the random variable $\xi(\omega)$, defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $\mathbb{R}^m$. We denote an optimal solution of SP by x^* and its optimal value by C^*. Given a sample $S = \{\xi_1, \xi_2, ..., \xi_N\}$, the problem SP can be approximated by the *sample average approximation problem*

$$\text{SAA:} \quad \min_{x \in X} \frac{1}{N} \sum_{i=1}^{N} c(x, \xi_i), \tag{1}$$

where we choose to suppress the dependence of ξ on ω when this is clear from the context. We write $\mathbb{E}_{\mathbb{P}_0}[c(x, S)]$ to denote the objective of (1), where the expectation uses the finite probability measure $\mathbb{P}_0$ that assigns mass $\frac{1}{N}$ to each $\xi_i \in S$.

Our focus in this paper is on *distributionally robust optimization* [26], in which the decision maker chooses x to solve

$$\min_{x \in X} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} \left[c(x, \xi) \right],$$

where $\mathcal{P}$ is a set of probability measures, from which a worst-case measure $\mathbb{Q}$ is chosen, and the expectation is taken over the random variable ξ with distribution $\mathbb{Q}$. The distributionally robust version of SAA is

$$\text{DRO:} \quad \min_{x \in X} \sup_{\mathbb{Q} \in \mathcal{P}_\delta} \mathbb{E}_{\mathbb{Q}} \left[c(x, S) \right], \tag{2}$$

where the objective function depends on both the sample S and the worst-case probability measure. In DRO $\mathcal{P}$ is defined to be a region $\mathcal{P}_\delta$ containing the sample distribution $\mathbb{P}_0$ and parametrized by δ, so that it increases in size with increasing δ. When $\delta = 0$ we have $\mathcal{P}_\delta = \{\mathbb{P}_0\}$ and DRO reverts to the SAA problem of minimizing the expectation under $\mathbb{P}_0$ of $c(x, \xi)$. When $\delta > 0$, the worst-case measure $\mathbb{Q} \in \mathcal{P}_\delta$ is chosen to evaluate the expectation.

There are many different parameterizations that we might use for $\mathcal{P}_\delta$, and thus a variety of different versions of the distributionally robust optimization. Early versions of these models ([19], [7]) choose a worst case result from a set of distributions $\mathcal{P}$ that are subject to constraints on their moments. The data-driven approach we have outlined in which $\mathcal{P}$ depends on a sample has been the focus of more recent work. There are many alternative approaches, for example, [5] constructs a confidence set for the first and second moments of $\mathbb{P}$ based on a sample, whereas [25] constructs $\mathcal{P}$ in terms of a likelihood function, and [2] chooses $\mathcal{P}$ to be the confidence region of a goodness-of-fit test.

A number of authors consider a DRO model where the set $\mathcal{P}_\delta$ is obtained from looking at distributions within a distance δ of the sample distribution

under some metric on the space of distributions. One choice is to use ϕ-divergence (such as the Kullback-Leibler divergence) to define the distance. Note though that a ϕ-divergence is typically not symmetric and may not satisfy the triangle inequality. So apart from some special cases such as total variation (which is an L_1 distance) this is not a metric, or semi-metric. Bayraksan and Love [1] give a tutorial discussion of the use of ϕ-divergence in this setting, and Shapiro [21] also discusses the different types of ϕ-divergence and their links with coherent risk measures. Gotoh, Kim and Lim 2018 [11] show that using ϕ-divergence leads to small changes in the mean compared with large changes in the variance when considering in-sample performance. Van Parys et al. [24] show that the Kullback-Leibler divergence (also called relative entropy) has optimal properties in terms of the asymptotic behavior for out-of-sample disappointment.

An alternative approach used by many authors is to define distances using the Wasserstein distance between probability measures. For example Pflug and Wozabal [17] apply this approach, where $\mathcal{P}_\delta$ is the set of distributions with a Wasserstein distance of less than δ to the sample distribution. The application here is to portfolio optimization, as is also the case for Wozabal [27]. The paper by Gao and Kleywegt [9] gives a comparison of the Wasserstein and ϕ-divergence approaches arguing for the better performance of the former and including some detailed comparisons on a newsvendor problem. An important consideration in the choice of approach is the computational burden involved in carrying out the inner maximization of DRO. The work by Esfahani and Kuhn [8] demonstrates how this can be done in the Wasserstein case for a wide variety of objective function forms.

It is well-known that when a sample is used to determine a decision variable, the resulting decision may perform relatively poorly on a new sample from the same distribution. The optimization can exploit particular features of the sample and delivers a decision that happens to do well on this set of values. This is related to *overfitting*, which has received a lot of attention in statistics and machine learning (see e.g. [20], [14], [12]). Here the coefficients of a model are estimated using a training set of data, and a model with many coefficients can choose these to match the training set very well. When applied to out-of-sample test data the model often performs worse than a simpler model with fewer coefficients. The solutions from sample average approximations with small sample sizes can also perform poorly out of sample (see e.g. [4],[6],[27]).

The most widely adopted machine-learning approach to overfitting is to add some form of regularization to the estimation problem. Examples are ridge regression [13] and LASSO [23]. These techniques are typically used to reduce the number of explanatory variables with nonzero weights, in the

expectation that the resulting models will generalize better when applied out of sample. There is also an interpretation [23], in which regularization (by shrinking the size of parameters) results in parameter estimates that have lower variance. It is well known (see e.g. [28]) that ridge regression and LASSO estimation problems have equivalent formulations that can be interpreted as robust optimization problems that specify uncertainty sets on the data values. Regression regularization can also be formulated as a distributionally robust optimization problem using a Wasserstein metric [3].

Our paper seeks to explain observed improvements in out-of-sample performance of DRO in a more general optimization context. For example, as shown in [4], solutions to financial optimization problems are very sensitive to sampling errors in estimated returns. Out-of-sample performance of solutions to such problems is often much better when a distributionally robust approach is used [5],[27]. However the nature of the improvement in out-of-sample performance varies. A particular set of data (corresponding to a single sample) may or may not give an improvement if a robust approach is used, but the variance of the out-of-sample outcomes when considered over multiple sets of data will be reduced. One might expect that it will be necessary to accept a higher average cost in order to achieve a reduction in variance. But in fact there are many cases where both the mean and the variance of the out-of-sample results are improved by using a DRO approach. For example Esfahani and Kuhn [8] carry out numerical experiments for a portfolio optimization problem (using synthetic data) and show that both mean and variance improve for a Wasserstein robustification (provided δ is not too large). Very similar results are found by Gotoh et al. [10] when using Kullback-Leibler divergence in an inventory problem and a logistic regression problem. Luo and Mehrotra [16] report improvements in mean out-of-sample behavior from using a Wasserstein approach for a logistic regression problem (with δ set by a cross-validation method). Nevertheless there is no guarantee that an improvement in out of sample mean is available: for example Gotoh et al. [10] show that in their setup a portfolio optimization problem never sees an improvement in mean.

Our aim in this paper is to characterize classes of problem SP for which solutions to DRO outperform solutions to SAA, when evaluated with the true distribution. To achieve this we simplify the analysis by restricting attention to unconstrained continuous optimization problems with strictly convex quadratic objective functions. Nevertheless, we believe that this is a significant first step towards understanding the behavior of distributional robust approaches applied to more general problems. The contributions of the paper are as follows:

1. We define the concept of incremental improvement, and derive an improvement lemma that provides general conditions under which solutions to DRO will outperform those of SAA.

2. We apply the improvement lemma to different formulations of DRO with quadratic objective function, using $\mathcal{P}_\delta$ derived from phi-divergence, coherent risk measures and Wasserstein formulations, and show how the conditions for incremental improvement in each case translate into conditions on the underlying probability distribution $\mathbb{P}$.

3. We present a number of simple examples with univariate objective functions for which analytical expressions of improvement can be derived. These examples provide a deeper understanding of the variation in out-of-sample outcomes that can result from different robustification approaches.

The paper is laid out as follows. The next section establishes our notation and terminology, formally defines the concept of incremental improvement, and establishes the improvement lemma. Section 3 shows how the improvement lemma specializes to quadratic problems. Section 4 then applies the improvement lemma to examples with $\mathcal{P}_\delta$ derived from ϕ-divergence, both for a smooth ϕ function and also when total variation is used. Section 5 repeats this analysis for a CVaR-based coherent risk measure, and section 6 applies the improvement lemma to a problem with the Wasserstein distance. In section 7 we conclude the paper with some general observations. The proofs of all the propositions in the paper are deferred to two appendices.

2 Improving SAA

Our interest is in the solution of the stochastic optimization problem SP using sample average approximation (1) and its distributionally robust version. We assume that $c(x, \xi(\omega))$ satisfies the following conditions:

1. $E_{\mathbb{P}}[c(x, \xi(\omega))]$ exists and has finite value for all $x \in X$;

2. $c(x, \xi(\omega))$ is differentiable in $x \in X$ at almost every $\omega \in \Omega$;

3. There exists a positive valued random variable $K(\omega)$ such that $\mathbb{E}_{\mathbb{P}}[K(\omega)] < \infty$, and for all $x, y \in X$, $|c(x, \xi(\omega)) - c(y, \xi(\omega))| < K(\omega) \|x - y\|$.

The last condition is needed for the interchange of expectation and gradient operators in (3) below.

We denote an optimal solution to SAA by $x_0(S)$. In general this may not be unique, but in nearly all our analysis in this paper we deal with SAA problems with a unique solution. For N large it can be shown (see [22]) that $x_0(S)$ will approach the solution set of SP. When $x_0(S)$ is unique we use $C_0(S)$ to denote the expected cost of $x_0(S)$ given the sample S. Thus

$$C_0(S) = \mathbb{E}_{\mathbb{P}}[c(x_0(S), \xi)].$$

Taking expectations over $\mathbb{P}$ amounts to looking at the out-of-sample performance of the solution $x_0(S)$ under the real distribution.

We write $\bar{c}(x) = \mathbb{E}_{\mathbb{P}}[c(x, \xi)]$. Given a sample S, we denote the gradient of $\bar{c}(x)$ evaluated at $x_0(S)$ by $\nabla\bar{c}(x_0(S))$. By Theorem 7.44 of [22] the above conditions on $c(x, \xi)$ imply

$$\nabla\bar{c}(x_0(S)) = [\nabla_x \mathbb{E}_{\mathbb{P}}[c(x, \xi)]]_{x_0(S)} = \mathbb{E}_{\mathbb{P}}[[\nabla_x c(x, \xi)]_{x_0(S)}]. \tag{3}$$

A distributionally robust version of SAA (DRO) generates a solution $x_\delta(S)$, that depends both on the sample S and a parameter $\delta > 0$ that controls the amount of robustness added to the SAA problem. A choice $\delta = 0$ will give $x_\delta(S) = x_0(S)$. Fundamentally we are interested in the quality of the solution as measured by $C_\delta(S) = \mathbb{E}_{\mathbb{P}}[c(x_\delta(S), \xi)]$ in comparison with the SAA alternative $C_0(S)$. Like $C_0(S)$, $C_\delta(S)$ is well defined only when $x_\delta(S)$ is unique, so when working with $C_\delta(S)$ we will make this assumption. Thus we will assume the existence of some tie breaking rule to determine a unique choice of $x_\delta(S)$. As we will show, it turns out that for many examples there is no need for a tie-breaking rule for $x_\delta(S)$, provided $x_0(S)$ is unique and δ is chosen sufficiently small. Since the solution quality depends on what sample is chosen, we are interested in the expectations of $C_0(S)$ and $C_\delta(S)$ over different samples that may occur, which we write using notation $\mathbb{E}_S$. This expectation can be derived using the underlying probability measure $\mathbb{P}$.

It is helpful to make the following definitions that apply when $x_0(S)$ and $x_\delta(S)$ are unique.

Definition The *expected value of the robust solution* (VRS(δ)) is

$$\text{VRS}(\delta) = \mathbb{E}_S[C_0(S) - C_\delta(S)].$$

Observe that in VRS(δ) the expectation is taken over the sampling distribution, accounting for the randomness driven by the choice of sample S as well as the random variable ξ. The value of VRS(0) is zero, and we will focus on circumstances in which VRS(δ) is positive for small positive δ, which means that $x_\delta(S)$ performs better out of sample than $x_0(S)$. We give this a formal definition.

Definition A given form of robustification applied to a problem SP *incrementally improves* SAA if $\text{VRS}(\delta) > 0$ for all $\delta > 0$ sufficiently small.

When considering robustification it is natural to try and quantify its effect and to seek a value of δ that yields the best possible improvement in out-of-sample performance. In general this is challenging to study analytically. Our approach is to quantify the improvement as δ increases incrementally from 0. As we show below this approach provides some analytical traction that gives a deeper theoretical understanding of some of the mechanisms that provide the improvement. In many examples we can provide conditions on the problem data that will give sufficient conditions for incremental improvement. Deriving conditions under which robustification fails to improve SAA for any $\delta > 0$ is much more challenging: the best result we have gives necessary conditions for incremental improvement.

When $\text{VRS}(\delta)$ is differentiable at $\delta = 0$, we can quantify incremental improvement using its derivative.

Definition The *marginal value of the robust solution* (MVRS) is

$$MVRS = \lim_{\delta \to 0} \frac{VRS(\delta)}{\delta}$$

where this limit exists.

It is easy to see that if MVRS is strictly positive then robustification incrementally improves SAA, but the size of MVRS is determined by an arbitrary decision on the way that the set $\mathcal{P}_\delta$ is parameterized. Moreover switching between δ and δ^2 can mean an MVRS that is zero, positive or undefined. Since MVRS is derived from changes in $\mathbb{E}_S[C_0(S) - C_\delta(S)]$, it is related to changes in the optimal solution $x_\delta(S)$ as δ increases from 0. We analyze these changes via the following definition.

Definition We say that problem DRO exhibits *linear variation* if for almost all samples S, DRO has a unique solution and there is some constant vector $\bar{y}(S)$ with

$$x_\delta(S) = x_0(S) + \bar{y}(S)\delta + O(\delta^2).$$

We now state the key result of this paper that establishes general necessary and sufficient conditions for incremental improvement of SAA using robustification.

Lemma 1 (Improvement Lemma.) *Suppose DRO exhibits linear variation. Then for almost all samples S there exists some vector $\bar{y}(S)$ with*

$$C_\delta(S) = C_0(S) + \nabla \bar{c}(x_0(S))^\top \bar{y}(S)\delta + O(\delta^2)$$

and $MVRS = -\mathbb{E}_S[\nabla \bar{c}(x_0(S))^\top \bar{y}(S)]$. *If the robustification incrementally improves SAA then*

$$\mathbb{E}_S[\nabla \bar{c}(x_0(S))]^\top \bar{y}(S)] \leq 0.$$

Conversely, if

$$\mathbb{E}_S[\nabla \bar{c}(x_0(S))]^\top \bar{y}(S)] < 0, \tag{4}$$

then the robustification incrementally improves SAA.

In the next section we show how the improvement lemma takes a specific form in cases with strictly convex quadratic objective functions.

3 Quadratic costs

Throughout the rest of this paper we will analyze a simple case where the cost function in SP is quadratic of the form

$$c(x, \xi) = \frac{1}{2} x^\top H x + v(\xi)^\top x + u(\xi), \tag{5}$$

where H is a symmetric positive definite matrix, and for simplicity we will take $X = \mathbb{R}^n$. We will also assume that $\{v(\xi(\omega)) : \omega \in \Omega\}$ has a density f with support having n dimensions. The "true" problem we seek to solve is therefore

$$\text{SQP:} \quad \min_{x \in X} \mathbb{E}_{\mathbb{P}}[\tfrac{1}{2} x^\top H x + v(\xi)^\top x + u(\xi)].$$

We will provide an exact analysis without resorting to the asymptotic case where $N \to \infty$. This will allow a direct comparison of different robustifications and also shows how characteristics of the underlying probability distribution determine the behavior of the robustification.

There are a number of applications in which this form of objective function arises and we mention three. First a problem of estimating a value x on the basis of noisy observations ξ may be posed as one of minimizing the expectation of a quadratic loss function, so the cost to be minimized is $E_{\mathbb{P}}[(x - \xi)^2]$. A second application occurs in production quantity optimization with quadratic costs and stochastic prices. The problem is to choose a (scalar) quantity x to manufacture, with quadratic costs $ax^2 + bx + c$ and a selling price ξ that is unknown in advance (but where we have a representative sample of previous prices). The cost to be minimized (or negative profit) is then $\mathbb{E}[ax^2 + bx + c - x\xi]$ which can be written in the required form (without the linear term bx) with a change of variables. A final application is in mean-variance portfolio optimization when the covariance matrix between

n individual stock returns is known, and short selling is allowed. For this problem the objective function to be minimized is

$$w^\top \Sigma w - \lambda \mathbb{E}\left[y^\top w\right]$$

where $w \in \mathbb{R}^n$ is the vector of portfolio allocations (and must satisfy the constraint that the component values sum to one), ξ is the vector of returns, and λ is given and represents the importance of the mean return versus the variance of returns. We assume that the covariance matrix Σ is known and the return distribution can be estimated from a sample of historical returns. Since we have an added constraint that $\sum_{i=1}^n w_i = 1$, we can rewrite the problem in the appropriate linear subspace by setting $w_n = 1 - \sum_{i=1}^{n-1} w_i$. This will introduce a linear term in w within the objective function and require a change of variables to recover a problem without the linear term.

Given a sample $S = \{\xi_1, \xi_2, \ldots, \xi_N\}$, the sample average approximation of SQP is

$$\text{SAA:} \quad \min_{x \in X} \left(\tfrac{1}{2} x^\top H x + \bar{v}_0(S)^\top x + \bar{u}_0(S)\right)$$

where $\bar{v}_0(S) = E_{\mathbb{P}_0}[v(\xi)]$ and $\bar{u}_0(S) = E_{\mathbb{P}_0}[u(\xi)]$ are the sample averages of $v(\xi)$ and $u(\xi)$. The solution to SQP is $x^* = -H^{-1}\bar{v}$ for $\bar{v} = E_{\mathbb{P}}[v(\xi)]$ and that of SAA is $x_0(S) = -H^{-1}\bar{v}_0(S)$. Observe that our positive definite quadratic assumption implies that the solution $x_0(S)$ is unique, and is an unbiased estimator of x^* since $\mathbb{E}_S[\bar{v}_0(S)] = \bar{v}$.

We remark that the matrix H in (5) is assumed throughout this paper to be deterministic, so that our formulation of SQP is not as general as it might be. If H depends on ξ then the solution to SQP is $x^* = -\bar{H}^{-1}\bar{v}$ and that of SAA is $x_0(S) = -\bar{H}(S)^{-1}\bar{v}_0(S)$ where $\bar{H} = \mathbb{E}_{\mathbb{P}}[H(\xi)]$ and $\bar{H}(S) = \mathbb{E}_{\mathbb{P}_0}[H(\xi)]$. In this case the estimator $-\bar{H}(S)^{-1}\bar{v}_0(S)$ will in general be biased. Our analysis below will focus on the estimator $x_0(S)$ and how it varies as SAA is robustified. We will derive conditions when H is deterministic that are sufficient for this variation to improve out-of-sample performance. The analysis is more complicated when H depends on ξ giving a bias in $x_0(S)$. Our intent in considering the simpler case is to identify the key drivers that improve out-of-sample performance in the unbiased case, as a first step to the more general case.

Given a random sample $S = \{\xi_1, \xi_2, \ldots, \xi_N\}$ with each element drawn independently from $\mathbb{P}$ and SAA solution $x_0(S) = -H^{-1}\bar{v}_0(S)$, we can order the elements of S so that

$$v(\xi_1)^\top x_0(S) \leq v(\xi_2)^\top x_0(S) \leq \ldots \leq v(\xi_N)^\top x_0(S).$$

We say that S is *strictly ordered by SAA* if

$$v(\xi_1)^\top x_0(S) < v(\xi_2)^\top x_0(S) < \ldots < v(\xi_N)^\top x_0(S).$$

Proposition 2 *If the random variable $v(\xi(\omega))$ has a density (with support of dimension n) then the set of samples $S = \{\xi_1, \xi_2, \ldots, \xi_N\}$ that are strictly ordered by SAA has probability measure* 1.

Given a sample S, the distributionally robust version of SAA is

$$\text{DRQP:} \quad \min_{x \in X} \left(\tfrac{1}{2} x^\top H x + \sup_{\mathbb{Q} \in \mathcal{P}_\delta} \mathbb{E}_{\mathbb{Q}} \left[v(\xi)^\top x + u(\xi) \right] \right),$$

where $\mathcal{P}_\delta$ is a set of probability distributions that are close to $\mathbb{P}_0$. The following proposition gives a formula for $C_\delta(S)$ that is defined when DRQP has a unique solution.

Proposition 3 *If DRQP has a unique solution $x_\delta(S)$ then*

$$\begin{aligned} C_\delta(S) = C_0(S) &- (\bar{v} - \bar{v}_0(S))^\top H^{-1} (\bar{v}_\delta(S) - \bar{v}_0(S)) \\ &+ \frac{1}{2} (\bar{v}_\delta(S) - \bar{v}_0(S))^\top H^{-1} (\bar{v}_\delta(S) - \bar{v}_0(S)). \end{aligned} \tag{6}$$

where $\bar{v}_\delta(S) = -H x_\delta(S)$.

Proposition 3 gives an exact value for VRS(δ) in terms of $\bar{v}_\delta(S)$. The expression (6) can be approximated to first order when DRQP exhibits linear variation so $x_\delta(S) = x_0(S) + \bar{y}(S)\delta + O(\delta^2)$. Then using

$$-H^{-1}(\bar{v}_\delta(S) - \bar{v}_0(S)) = x_\delta(S) - x_0(S),$$

and

$$\begin{aligned} \nabla \bar{c}(x_0(S)) &= \mathbb{E}_{\mathbb{P}}[[\nabla_x c(x, \xi)]_{x_0(S)}] \\ &= H x_0(S) + \bar{v} \\ &= \bar{v} - \bar{v}_0(S), \end{aligned}$$

we obtain the following form of the improvement lemma in this case.

Lemma 4 *Suppose DRQP exhibits linear variation. Then for almost all samples S there exists some vector $\bar{y}(S)$ with*

$$C_\delta(S) = C_0(S) - (\bar{v}_0(S) - \bar{v})^\top \bar{y}(S)\delta + O(\delta^2)$$

and $MVRS = \mathbb{E}_S[(\bar{v}_0(S) - \bar{v})^\top \bar{y}(S)]$. If the robustification incrementally improves SAA then

$$\mathbb{E}_S[(\bar{v}_0(S) - \bar{v})^\top \bar{y}(S)] \geq 0,$$

and if $\mathbb{E}_S[(\bar{v}_0(S) - \bar{v})^\top \bar{y}(S)] = 0$, *then*

$$\lim_{\delta \to 0} \mathbb{E}_S[(\bar{v}_0(S) - \bar{v})^\top (x_\delta(S) - x_0(S))]/\delta^2 \geq 0.$$

Conversely, if

$$\mathbb{E}_S[(\bar{v}_0(S) - \bar{v})^\top \bar{y}(S)] > 0, \tag{7}$$

then the robustification incrementally improves SAA.

To apply the inequality (7) in Lemma 4 we require a formula for the vector $\bar{y}(S)$, which depends on the sample and the particular form of robustification. In the following sections we will derive expressions for $\bar{y}(S)$ using three different versions of robustification. Observe that (7) will remain true for any positive scaling of $\bar{y}(S)$.

It is interesting to observe that the formulae for incremental improvement do not explicitly depend on the constant term $u(\xi)$, its expectation $\bar{u}$ or sample average $\bar{u}_0(S)$. Indeed the optimal solutions of SQP and SAA are independent of the constant terms, and so we can assume that $u(\xi) = 0$ when solving SQP and SAA. In what follows we will in general assume that $u(\xi) = 0$, and construct distributionally robust versions of SAA that do not include this constant term. It is important to realize however that the optimal solution $x_\delta(S)$ to DRQP will depend on the constant term, and so $\bar{y}(S)$ will implicitly account for the constant term. We will illustrate the difference this makes in the next section.

4 Phi divergence

Distributionally robust optimization using ϕ-divergence works with finite distributions, say $\nu_q = (q_1, q_2, ..., q_N)$ and $\nu_p = (p_1, p_2, ..., p_N)$, and defines

$$d_\phi(\nu_q, \nu_p) = \sum_{i=1}^{N} p_i \phi\left(\frac{q_i}{p_i}\right) \tag{8}$$

for ϕ a convex function defined on $[0, \infty)$ with $\phi(1) = 0$ (and achieving its minimum there). Given the sample distribution $\mathbb{P}_0$, we may define

$$\mathcal{P}_\delta = \{\mathbb{Q} : d_\phi(\mathbb{Q}, \mathbb{P}_0) \leq \delta\}.$$

Note that because (8) is not symmetric we obtain a different set $\mathcal{P}_\delta$ depending on whether $\mathbb{P}_0$ is chosen to be ν_p or ν_q in (8). We first study an example (total variation) where $\phi(t) = |t - 1|$ is non-smooth, and then consider general analytic functions ϕ.

4.1 Total variation

Given a sample $S = \{\xi_1, \xi_2, \ldots, \xi_N\}$, and $\phi(t) = |t-1|$, we define $\mathcal{P}_\delta \subseteq \{\mathbb{Q} : supp(\mathbb{Q}) = S\}$, by

$$\mathcal{P}_\delta = \{(q_1, q_2, \ldots, q_N) : \sum_{i=1}^{N} \left| q_i - \frac{1}{N} \right| \leq \delta\}.$$

The distributionally robust version of SAA (without the constant term) is

$$\text{DRQP:} \quad \min_{x \in X} \tfrac{1}{2} x^\top H x + Q_{\max}(x),$$

where

$$Q_{\max}(x) = \max_{(q_1, q_2, \ldots, q_N) \in \mathcal{P}_\delta} \sum_{i=1}^{N} q_i v(\xi_i)^\top x.$$

Recall $x_0(S)$ is the solution to SAA, and by Proposition 2 we have

$$v(\xi_1)^\top x_0(S) < v(\xi_2)^\top x_0(S) \ < ... < v(\xi_N)^\top x_0(S),$$

for all samples S apart from a set with probability 0. For the samples S that are strictly ordered by SAA we let $R(S) = v(\xi_N) - v(\xi_1)$.

Lemma 5 *DRQP exhibits linear variation with*

$$\bar{y}(S) = -\frac{1}{2} H^{-1} R(S).$$

For $\delta > 0$ sufficiently small

$$VRS(\delta) = \mathbb{E}_S[\frac{\delta}{2}(\bar{v} - \bar{v}_0(S))^\top H^{-1} R(S) - \frac{\delta^2}{8} R(S)^\top H^{-1} R(S)],$$

and

$$MVRS = \frac{1}{2} \mathbb{E}_S[(\bar{v} - \bar{v}_0(S))^\top H^{-1} R(S)]. \tag{9}$$

This robustification incrementally improves SAA if

$$\mathbb{E}_S[(\bar{v} - \bar{v}_0(S))^\top H^{-1} R(S)] > 0.$$

To illustrate the formulae in Lemma 5, consider a one-dimensional production optimization problem with prices given by $g(\xi)$ and costs $\frac{1}{2}x^2$, so $c(x, \xi) = \frac{1}{2}x^2 - g(\xi)x$. We assume that $g(\xi) > 0$ almost surely. We may take $S = \{\xi_1, \xi_2, \ldots, \xi_N\}$ ordered so that

$$g(\xi_1) \geq g(\xi_2) \geq \ldots \geq g(\xi_N).$$

Let $\bar{g}_0(S) = \frac{1}{N} \sum_{i=1}^{N} g(\xi_i)$. We can take $H = 1$ and $v(\xi) = -g(\xi)$ in our previous analysis and obtain the following result.

Proposition 6 *Suppose* $c(x,\xi) = \frac{1}{2}x^2 - g(\xi)x$ *and* $g(\xi) > 0$ *almost surely, then total variation robustification gives*

$$VRS(\delta) = (\delta/2)cov(\bar{g}_0(S), R(S)) - \left(\delta^2/8\right)\mathbb{E}_S[R(S)^2], \tag{10}$$

where $R(S) = g(\xi_1) - g(\xi_N)$. *If the distribution of prices* $g(\xi)$ *is symmetric about its mean then MVRS is zero and* $VRS(\delta) < 0$ *for all* δ. *If* $cov(\bar{g}_0(S), R(S)) > 0$ *then there is incremental improvement.*

Proposition 6 shows that robustification using total variation always makes the solution worse when the price distribution is symmetric. In contrast, when there is a skew in the distribution of outcomes we can expect to see $\text{cov}(\bar{g}_0(S), R(S)) \neq 0$. For small δ this is the dominant term and will determine whether or not there is incremental improvement.

We can observe that if the distribution of $g(\xi)$ has significant weight in the right tail, then both the mean and the range are large when there is a sample point that happens to be far out in the tail. This suggests that the range is positively correlated with the mean, and hence $\mathbb{E}_S[(\bar{g} - \bar{g}_0(S))R(S)] < 0$. A robust solution takes weight from a high outlier and moves it to the lowest value. On average these moves improve the solution.

To study the effect of skew in the distribution of $g(\xi)$, we will work with the random variable $W = g(\xi) - \bar{g}$ which has mean 0. Let W have density $f(w)$ and cumulative distribution function $F(w)$, and define $Q(z) = \int_z^\infty wf(w)dw$. The following result is established using order statistics to determine an exact expression for MVRS.

Proposition 7 *Suppose* $c(x,\xi) = \frac{1}{2}x^2 - g(\xi)x$ *and* $g(\xi) > 0$ *almost surely, then total variation robustification gives*

$$MVRS = \frac{1}{2}\int_{-\infty}^{\infty}\left(F(z)^{N-1} - (1-F(z))^{N-1}\right)Q(z)dz. \tag{11}$$

It is possible to precisely identify a set of distributions where a right skew will guarantee a positive value for MVRS independent of the size of the sample N. The condition we need compares densities on either side of w_0, which is defined as the median of W where $F(w_0) = 1/2$. Specifically we compare the density at $w = w_0 - \gamma$ for $\gamma > 0$ with the density at $F^{-1}(1-F(w))$ where this expression is simply $w_0 + \gamma$ in the case that W is symmetric.

Proposition 8 *If* $f(w) \geq f(F^{-1}(1-F(w)))$ *for all* $w < w_0$ *with strict inequality for some* w, *and* $g(\xi) > 0$ *almost surely, then total variation robustification incrementally improves SAA.*

We finish this section by discussing an example to illustrate the effect of the constant term $u(\xi)$ on incremental improvement.

Example 1 (Estimation in one dimension): We consider the estimation problem we mentioned earlier where the objective is $\mathbb{E}_{\mathbb{P}}[(x-\xi)^2]$. In one dimension we have

$$\text{SP: } \min_x \mathbb{E}_{\mathbb{P}}[x^2 - 2\xi x + \xi^2]$$

with optimal solution $x^* = \mathbb{E}_{\mathbb{P}}[\xi]$. The SAA problem is

$$\text{SAA: } \min_x \left(x^2 - 2x\bar{\xi}_0(S) + \frac{1}{N}\sum_{i=1}^N \xi_i^2 \right).$$

We can neglect the term $u(\xi) = \xi^2$ in SP to give the problem

$$\text{SP0: } \min_x \mathbb{E}_{\mathbb{P}}[x^2 - 2\xi x]$$

which has the same optimal solution as SP. The corresponding sample average approximation is

$$\text{SAA0: } \min_x \left(x^2 - 2x\bar{\xi}_0(S) \right).$$

If the distribution of ξ is symmetrical about its mean then Proposition 6 shows that robustification of SAA0 with total variation makes the solution worse.

Now consider robustification of SAA where $\mathcal{P}_\delta$ is defined by total variation. This gives

$$\min_x \sup_{(q_1,\ldots,q_N)\in\mathcal{P}_\delta} \left(x^2 - 2\sum_{i=1}^N q_i\xi_i x + \sum_{i=1}^N q_i\xi_i^2 \right),$$

with solution $x_\delta(S) = \sum_{i=1}^N q_i\xi_i$, for some $q \in \mathcal{P}_\delta$. The presence of the term $q_i\xi_i^2$ affects the solution of this problem. Consider a sample $S = \{\xi_1, \xi_2, \ldots, \xi_N\}$ where these are ordered. Given any x we denote $i_{\min} = \arg\min_i |\xi_i - x|$, and $i_{\max} = \arg\max_i |\xi_i - x|$ (the points closest and furthest from x respectively) The inner problem adds a weight of $\frac{\delta}{2}$ to $q_{i_{\max}}$ and subtracts that weight from $q_{i_{\min}}$. It is clear that when ξ has a continuous distribution $i_{\min}$ and $i_{\max}$ are uniquely determined at $x = x_0(S) = \bar{\xi}_0(S)$ for almost all samples S, and since they remain the same for small δ, we have

$$x_\delta(S) - x_0(S) = \frac{\delta}{2}(\xi_{i_{\max}} - \xi_{i_{\min}}).$$

Thus we have shown linear variation with $\bar{y}(S) = (\xi_{i_{\max}} - \xi_{i_{\min}})/2$. Suppose ξ has mean 0 then we can deduce from Lemma 4 (noting $v(\xi) = -2\xi$ here) that robustification of SAA with total variation gives incremental improvement if

$$\mathbb{E}_S[(\xi_{i_{\max}} - \xi_{i_{\min}})\bar{\xi}_0(S)] < 0.$$

We can illustrate the differences between robustifying SAA and SAA0 with a simple example. Suppose ξ has a uniform distribution on $[-\frac{1}{2}, \frac{1}{2}]$ so $F(\xi) = \xi + \frac{1}{2}$, and $f(\xi) = 1$. Consider a sample size of $N = 3$, giving $\xi_1 < \xi_2 < \xi_3$.

Here $x_0(S) = \bar{\xi}_0(S) = \frac{\xi_1+\xi_2+\xi_3}{3}$. It is not hard to show that in this case $\xi_{i_{\min}}$ is simply the middle point ξ_2. This is because the order $\xi_1 < \xi_2 < \xi_3$ implies both $\xi_0(S) - \xi_2 < \xi_0(S) - \xi_1$ and $\xi_0(S) - \xi_2 < \xi_3 - \xi_0(S)$. Thus

$$\xi_{i_{\max}} - \xi_{i_{\min}} = \begin{cases} \xi_3 - \xi_2 & \text{if} \quad \xi_2 < \frac{\xi_1+\xi_3}{2} \\ \xi_1 - \xi_2 & \text{if} \quad \xi_2 > \frac{\xi_1+\xi_3}{2}. \end{cases}$$

The joint density of ξ_1, ξ_2, ξ_3 is $f(\xi_1, \xi_2, \xi_3) = 6$ over the region $\{(\xi_1, \xi_2, \xi_3) \mid \xi_1 < \xi_2 < \xi_3,\ \xi_i \in [-\frac{1}{2}, \frac{1}{2}]\}$. This gives

$$\begin{aligned} \mathbb{E}_S[(\xi_{i_{\max}} - \xi_{i_{\min}})\bar{\xi}_0(S)] &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{x}^{\frac{1}{2}} \int_{2y-x}^{\frac{1}{2}} 6(z-y)(x+y+z)dzdydx \\ &\quad + \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{x}^{\frac{1}{2}} \int_{y}^{2y-x} 6(x-y)(x+y+z)dzdydx \\ &= -\frac{1}{2}, \end{aligned}$$

so robustification of SAA will give incremental improvement in this case (with a symmetric distribution) □

This example shows that it is possible to add a random term to SAA that has no effect on the optimal solution $x_0(S)$ to SAA, but will affect $x_\delta(S)$ when we robustify the problem with $\delta > 0$. We note that in this example of estimating $\mathbb{E}[\xi]$, robustification of SAA using total variation improves out-of-sample performance even for large δ. If $\delta = 1 - \frac{2}{N}$, then the worst-case distribution sets $q_i = 0$ except for the first and Nth order statistics (ξ_1 and ξ_N) that have $q_i = \frac{1}{2}$. When ξ has a uniform distribution, the estimate $\frac{\xi_1+\xi_N}{2}$ of the mean has a much lower variance than the sample mean $\bar{\xi}_0(S)$ (see [15]).

4.2 Smooth phi-divergence

We now consider the case where ϕ is an analytic strictly convex function with $\phi(1) = \phi'(1) = 0$, and $\phi''(1) > 0$. Given a sample S, we define $\mathcal{P}_\delta \subseteq \{\mathbb{Q} :$

$supp(\mathbb{Q}) = S\}$, by

$$\mathcal{P}_\delta = \{(q_1, q_2, \ldots, q_N) : \sum_{i=1}^{N} \phi\left(Nq_i\right) \leq N\delta^2\}.$$

Observe that we have chosen to parametrize $\mathcal{P}_\delta$ using δ^2 on the right-hand side of the inequality. Let us denote

$$V(S) = \frac{1}{N}\sum_{i=1}^{N} \left(v(z_i) - \bar{v}_0(S)\right)\left(v(z_i) - \bar{v}_0(S)\right)^\top. \tag{12}$$

We now have the following result.

Proposition 9 *For any analytic strictly convex ϕ, DRQP with ϕ-divergence robustification exhibits linear variation with*

$$\bar{y}(S) = \left(\frac{2}{\phi''(1)}\right)^{1/2} \frac{H^{-1}V(S)H^{-1}\bar{v}_0(S)}{\left(\bar{v}_0(S)^\top H^{-1}V(S)H^{-1}\bar{v}_0(S)\right)^{\frac{1}{2}}}.$$

Example 1 (Continued): We return to the problem

$$\text{SP0: } \min_x E_{\mathbb{P}}[x^2 - 2\xi x]$$

with corresponding sample average approximation

$$\text{SAA0: } \min_x \left(x^2 - 2x\bar{\xi}_0(S)\right)$$

given a sample $S = \{\xi_1, \xi_2, \ldots, \xi_N\}$, and $\bar{\xi}_0 = \frac{1}{N}\sum_{i=1}^{N}\xi_i$. Now consider a distributionally robust version

$$\text{DRO: } \min_x \sup_{\mathbb{Q}\in\mathcal{P}_\delta} \mathbb{E}_{\mathbb{Q}}[x^2 - 2\xi x],$$

where $\mathcal{P}_\delta$ is defined by modified χ^2 distance, with $\phi(t) = (t-1)^2$, and $d_\phi(\mathbb{Q}, \mathbb{P}_0) = N\sum_{i=1}^{N}\left(q_i - \frac{1}{N}\right)^2$. Applying Proposition 9 with $v(\xi) = -2\xi$, gives

$$\bar{y}(S) = -V(S)^{\frac{1}{2}},$$

where

$$V(S) = \frac{1}{N}\sum_{i=1}^{N}\left(\xi_i - \bar{\xi}_0\right)^2,$$

the standard deviation of the sample points. We can compare this with the solution to DRO for small δ which can be computed analytically (using Lemma 4 in [18] with $r = \frac{\delta}{\sqrt{N}}$):

$$\begin{aligned} x_\delta(S) &= \bar{\xi}_0 - \sum_i \frac{\xi_i(\xi_i - \bar{\xi}_0)}{\sqrt{NV(S)}} \frac{\delta}{\sqrt{N}} \\ &= \bar{\xi}_0 - V(S)^{\frac{1}{2}}\delta \end{aligned}$$

as required. □

Proposition 9 allows us to identify the condition for incremental improvement given in the Proposition below. Note that the condition is the same for any choice of analytic ϕ function, so that whether we use Kullback-Leibler or some other phi-divergence will not change the incremental improvement property.

Proposition 10 *Robustification with smooth phi-divergence gives incremental improvement in SAA for any analytic strictly convex ϕ if*

$$\mathbb{E}_S \left[\frac{(\bar{v}_0(S) - \bar{v})^\top H^{-1} V(S) H^{-1} \bar{v}_0(S)}{\left(\bar{v}_0^\top(S) H^{-1} V(S) H^{-1} \bar{v}_0(S)\right)^{\frac{1}{2}}} \right] > 0.$$

5 CVaR based robustness

Distributionally robust optimization can also be based on a coherent risk measure ρ where we solve

$$\min_{x \in X} \rho[c(x, \xi)]$$

which can be reformulated as

$$\min_{x \in X} \sup_{\mathbb{Q} \in \mathcal{P}_\delta} \mathbb{E}_{\mathbb{Q}}\left[c(x, \xi)\right]$$

for some convex set $\mathcal{P}_\delta$ of probability measures on the discrete set $\{c(x, \xi_i) : \xi_i \in S\}$. We shall focus on the particular risk measure

$$\rho[c(x, \xi)] = (1 - \delta)\mathbb{E}[c(x, S)] + \delta \mathrm{CVaR}_{1-\alpha}[c(x, S)].$$

Here $\mathcal{P}_\delta$ is a polyhedral set of probability measures that depend on δ. For example if $\alpha = \frac{1}{N}$, then $\mathcal{P}_\delta$ is the convex hull of the N points $(\frac{1-\delta}{N}, \frac{1-\delta}{N}, \ldots, \frac{1-\delta}{N}) + \delta e_i$, $i = 1, 2, \ldots, N$, where e_i is the i'th unit vector.

As in the previous section our analysis will be applied to $c(x,\xi) = \frac{1}{2}x^\top Hx + v(\xi)^\top x$, where H is positive definite and we ignore the constant term that arises from $u(\xi)$. Recall the formulation

$$\text{DRQP:} \quad \min_{x\in X} \left(\tfrac{1}{2}x^\top Hx + Q_{\max}(x)\right),$$

where

$$Q_{\max}(x) = \max_{(q_1,q_2,\ldots,q_N)\in\mathcal{P}_\delta} \sum_{i=1}^{N} q_i v(\xi_i)^\top x.$$

We obtain the following optimality conditions for this problem.

Lemma 11 *The solution to DRQP with CVaR robustification satisfies*

$$x_\delta(S) \in -H^{-1}\left((1-\delta)\bar{v}_0(S) + \delta G_{CVaR}(x_\delta(S))\right) \tag{13}$$

where $G_{CVaR}(x)$ is the subdifferential for $CVaR_{1-\alpha}\left[\{v(\xi_i)^\top x\}\right]$. When $CVaR_{1-\alpha}\left[\{v(\xi_i)^\top x\}\right]$ is differentiable at $x_\delta(S)$ with derivative $\bar{v}_{CVaR}(S)$ then

$$x_\delta(S) = -H^{-1}\left((1-\delta)\bar{v}_0(S) + \delta\bar{v}_{CVaR}(S)\right). \tag{14}$$

We can apply a similar analysis here to that used in the total variation phi-divergence section. Recall that SAA has a unique solution $x_0(S) = -H^{-1}\bar{v}_0$, and if $S = \{\xi_1, \xi_2, \ldots, \xi_N\}$ with

$$v(\xi_1)^\top x_0(S) \le v(\xi_2)^\top x_0(S) \le \ldots \le v(\xi_N)^\top x_0(S),$$

then Proposition 2 gives

$$v(\xi_1)^\top x_0(S) < v(\xi_2)^\top x_0(S) < \ldots < v(\xi_N)^\top x_0(S) \tag{15}$$

except for a set of samples with probability 0. For the samples satisfying (15), $\text{CVaR}_{1-\alpha}\left[\{v(\xi_i)^\top x\}\right]$ is differentiable at $x_0(S)$, with derivative

$$\bar{v}_{\text{CVaR}}(S) = \frac{1}{\alpha N}\sum_{i=1}^{m_\alpha} v(\xi_i) + (1 - \frac{m_\alpha}{\alpha N})v(\xi_{m_\alpha}), \tag{16}$$

where m_α is the unique integer with $\alpha \in (\frac{m_\alpha}{N}, \frac{m_\alpha+1}{N}]$.

Proposition 12 *Suppose $c(x,\xi) = \frac{1}{2}x^\top Hx + v(\xi)^\top x$. Then DRQP with CVaR robustification has linear variation with*

$$\bar{y}(S) = -H^{-1}(\bar{v}_{CVaR}(S) - \bar{v}_0(S)),$$

where $\bar{v}_{CVaR}(S)$ *is defined by (16). Also*

$$\begin{aligned} VRS(\delta) &= \mathbb{E}_S[\delta(\bar{v}-\bar{v}_0(S))^\top H^{-1}(\bar{v}_{CVaR}(S)-\bar{v}_0(S)) \\ &\quad -\frac{\delta^2}{2}(\bar{v}_{CVaR}(S)-\bar{v}_0(S))^\top H^{-1}(\bar{v}_{CVaR}(S)-\bar{v}_0(S))], \end{aligned}$$

giving

$$MVRS = \mathbb{E}_S[(\bar{v}-\bar{v}_0(S))^\top H^{-1}(\bar{v}_{CVaR}(S)-\bar{v}_0(S))]$$

with incremental improvement if this is positive.

If we consider $v(\xi_i)$ as the set of sample vectors, then this result expresses the MVRS value as the expected value over samples of a product involving the vector difference between the real mean and the sample mean, the inverse of H, and the difference between the high cost elements in the sample (that are represented in CVaR) and the sample mean.

From this result we can derive the following result for the one-dimensional case.

Proposition 13 *Suppose* $c(x,\xi) = \frac{1}{2}x^2 - g(\xi)x$*, where* $g(\xi)$ *has a density with mean* $\bar{g}$ *and variance* σ^2*, and let* $\bar{g}_0(S) = \frac{1}{N}\sum_{i=1}^N g(\xi_i)$ *and* $\alpha \in (0,1]$*. Then with CVaR robustification*

$$MVRS = \frac{\sigma^2}{N} + \mathbb{E}_S[(\bar{g}_0(S)-\bar{g})\,CVaR_{1-\alpha}[\{-sgn(\bar{g}_0(S))g(\xi_i)\}]].$$

In this scalar case it is possible to derive a more explicit form of MVRS if we know the distribution of $g(\xi)$, and we can assume that $\bar{g}_0(S)$ is always positive. We define $W = g(\xi) - \bar{g}$, having a density denoted $f(w)$ and cumulative distribution function $F(w)$. This gives the following result.

Proposition 14 *Suppose* $c(x,\xi) = \frac{1}{2}x^2 - g(\xi)x$ *where* $\bar{g}_0(S) > 0$*, and we solve DRQP with CVaR robustification where* $\alpha \in (0,1]$*. Then*

$$MVRS = \frac{\sigma^2}{N} - \int_{-\infty}^{\infty} Q(z)(1-F(z))^{N-1}\Lambda_\alpha(z)dz,$$

where $Q(z) = \int_z^\infty w f(w)dw$*, and*

$$\begin{aligned} \Lambda_\alpha(z) &= \frac{1}{\alpha N} + \frac{1}{\alpha N}(N-1)\frac{F(z)}{(1-F(z))} \\ &\quad + \frac{1}{\alpha N}\frac{(N-1)(N-2)}{2}\frac{F(z)^2}{(1-F(z))^2} \\ &\quad + \ldots + (1-\frac{m_\alpha}{\alpha N})\begin{pmatrix} N-1 \\ m_\alpha \end{pmatrix}\frac{F(z)^{m_\alpha}}{(1-F(z))^{m_\alpha}}, \end{aligned}$$

where m_α *is the unique integer satisfying* $\alpha \in (\frac{m_\alpha}{N}, \frac{m_\alpha+1}{N}]$*.*

There are some observations we can make in relation to the condition $\bar{g}_0(S) > 0$. This is included in order to ensure that $x_0(S) > 0$ and hence that it is the left rather than right tail of the $g(\xi)$ distribution that appears in the CVaR term. We can usually assume that $\bar{g}_0(S)$ is close to the mean of the $g(\xi)$ distribution for reasonable sample sizes. This is often enough to make the probability of $\bar{g}_0(S) < 0$ extremely small. In these cases we can take the expression for MVRS as a good approximation for the exact value. There are other cases in which $\bar{g}_0(S) < 0$ with probability close to 1. When this happens there are alternative formulae (which we will not give here) obtained through defining $W = \bar{g} - g(\xi)$.

We now study some examples of MVRS for the one-dimensional problem with $c(x, \xi) = \frac{1}{2}x^2 - g(\xi)x$. The formula in Proposition 14 shows that MVRS will be positive if the second term is small. There is a connection here to the skew in the distribution of $g(\xi)$. We consider an example with a large right-hand skew and show that MVRS is positive.

Example 2 (exponential distribution):

Suppose $g(\xi)$ is exponentially distributed on $[0, \infty)$, so $\bar{g} = 1, \sigma^2 = 1$. Then $f(w) = e^{-(w+1)}$ over the range $(-1, \infty)$. If we robustify with $\text{CVaR}_{1-\frac{1}{N}}(\xi)$ then $\alpha = \frac{1}{N}$, $\Lambda_\alpha(z) = 1$ and

$$\begin{aligned} \text{MVRS} &= \frac{1}{N} - \int_{-1}^{\infty} (1 - F(z))^{N-1} Q(z) dz \\ &= \frac{1}{N} - \int_{-1}^{\infty} (e^{-z-1})^N \, (z+1) \, dz. \end{aligned}$$

Now $\int_{-1}^{\infty} (e^{-z-1})^N \, (z+1) \, dz = \int_0^{\infty} e^{-Nw} w dw$ and integrating by parts shows this has the value $\frac{1}{N^2}$. Thus MVRS $= \frac{N-1}{N^2} > 0$ and the CVaR robustification is incrementally improving. □

It is not necessary to consider examples with a skew to end up with MVRS positive, and we now consider three symmetric examples to give a better understanding of the behavior of MVRS with CVaR robustification.

Example 3 (uniform distribution):

We take $g(\xi)$ to be uniform on $[0, 2a]$. Then $\bar{g} = a$ and we obtain F uniform on $[-a, a]$ so $\sigma^2 = \frac{a^2}{3}$, $f(w) = \frac{1}{2a}$, $F(w) = \frac{\xi + a}{2a}$, $Q(z) = \frac{1}{4a}(a^2 - z^2)$. Then

$$\begin{aligned} \text{MVRS} &= \frac{a^2}{3N} - \int_{-a}^{a} (1 - F(z))^{N-1} Q(z) dz \\ &= 2a^2 \left(\frac{1}{6N} - \frac{1}{(N+1)(N+2)} \right) \end{aligned}$$

which is positive when $N > 2$ showing that the CVaR robustification is incrementally improving in this case. □

Example 4 (normal distribution):

Consider a univariate example with a normal distribution where $g(\xi)$ is an $N(\mu, \sigma^2)$ random variable with μ large enough that we can ignore the possibility of negative sample values. Then F is an $N(0, \sigma^2)$ random variable, with $f(\xi) = \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-\xi^2}{2\sigma^2})$. Now

$$Q(z) = \int_z^\infty u \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-u^2}{2\sigma^2}) du = \frac{\sigma}{\sqrt{2\pi}} \exp(\frac{-z^2}{2\sigma^2}) = \sigma^2 f(z).$$

Thus we can approximate MVRS with

$$\text{MVRS}_\sim = \frac{\sigma^2}{N} - \sigma^2 \int_{-\infty}^{\infty} (1 - F(z))^{N-1} f(z) \Lambda_\alpha(z) dz.$$

The approximation arises from taking the $\text{sgn}(\bar{g}_0(S))$ term in Proposition 13 as always being 1. As μ gets larger the probability that this fails becomes vanishingly small. In Appendix 2, we show (Lemma 27) that

$$\int_{-\infty}^{\infty} (1 - F(z))^{N-1} f(z) \Lambda_\alpha(z) dz = \frac{1}{N},$$

which gives $\text{MVRS}_\sim = 0$. □

Example 5 (mixture of univariate normal distributions):

We consider a case where ξ is univariate and $g(\xi)$ is formed as a mixture of two normal distributions having the same mean (large enough to ensure that $\bar{g}_0 > 0$ with very high probability). Thus W has density $f(w) = (f_1(w) + f_2(w))/2$ where $f_i(w) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp(\frac{-w^2}{2\sigma_i^2})$. Then $\sigma^2 = \frac{(\sigma_1^2 + \sigma_2^2)}{2}$,

$$F(z) = \frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{z} \frac{1}{\sigma_1} \exp(\frac{-w^2}{2\sigma_1^2}) + \frac{1}{\sigma_2} \exp(\frac{-w^2}{2\sigma_2^2}) dw$$

and

$$\begin{aligned} Q(z) &= (1/2) \int_z^\infty w \left(f_1(w) + f_2(w)\right) dw \\ &= (1/2)(\sigma_1^2 f_1(z) + \sigma_2^2 f_2(z)). \end{aligned}$$

Taking $\alpha = 1/N$, we can approximate the value of MVRS (using the same

argument as in Example 4) by

$$\begin{aligned}\text{MVRS}_{\sim} &= \frac{(\sigma_1^2+\sigma_2^2)}{2N} - \frac{1}{2}\int_{-\infty}^{\infty}(\sigma_1^2 f_1(z)+\sigma_2^2 f_2(z))(1-(F_1(z)+F_2(z))/2)^{N-1}dz \\ &= \frac{(\sigma_1^2+\sigma_2^2)}{2N} - \frac{1}{2}\int_{-\infty}^{\infty}\left(\frac{\sigma_1}{\sqrt{2\pi}}\exp(\frac{-z^2}{2\sigma_1^2})+\frac{\sigma_2}{\sqrt{2\pi}}\exp(\frac{-z^2}{2\sigma_2^2})\right) \\ &\quad \times\left(1-\frac{1}{2\sqrt{2\pi}}\int_{-\infty}^{z}\left(\frac{1}{\sigma_1}\exp(\frac{-u^2}{2\sigma_1^2})+\frac{1}{\sigma_2}\exp(\frac{-u^2}{2\sigma_2^2})\right)du\right)^{N-1}dz.\end{aligned}$$

We can evaluate this numerically. For example, if $\sigma_1 = 1$, $\sigma_2 = 2$ and $N = 3$ we obtain MVRS $= -4.33 \times 10^{-2}$, and if $N = 5$, MVRS $= -7.50 \times 10^{-2}$.

We see that in comparison with a normal distribution, the heavy tails introduced by taking a mixture of normal distributions makes MVRS negative and the overall performance of this robustification worse. □

Comparison of the three cases we have considered, each symmetric about its mean, suggests that for a univariate problem with CVaR robustification the normal distribution effectively acts as a division point, with incremental improvement failing to hold if the distribution has heavier tails than the normal.

6 Wasserstein metric

Distributionally robust optimization using a Wasserstein metric chooses

$$\mathcal{P}_\delta = \{\mathbb{Q} : d_W(\mathbb{Q}, \mathbb{P}_0) \le \delta\},$$

where $d_W(\mathbb{Q}, \mathbb{P}_0)$ is the cost of a minimum cost transportation plan from one probability distribution to the other. Formally we have the Wasserstein distance from a distribution ν_1 on the set $M \subset \mathbb{R}^m$ to a distribution ν_2, also on the set M, defined as

$$d_W(\nu_1, \nu_2) = \min_{\gamma \in \Gamma(\nu_1,\nu_2)} \int_{M\times M} \|z_1 - z_2\| \; d\gamma(z_1, z_2) \tag{17}$$

where $\Gamma(\nu_1, \nu_2)$ is the set of all measures on the product space $M \times M$ with marginals ν_1 and ν_2. $\Gamma(\nu_1, \nu_2)$ can be thought of as a transportation plan with a density at (z_1, z_2) in $M \times M$ that represents the probability mass moved from point z_1 to point z_2. We will apply this robustification to DRQP assuming a Euclidean metric, and consider problems where the underlying set M is a closed and bounded convex set in $\mathbb{R}^m$ (so that when $m = 1$, M is an interval.)

In distributionally robust optimization the inner problem is to choose a distribution on M maximizing the expected cost subject to a bound on the Wasserstein distance to the sample distribution $\mathbb{P}_0$ (which has equal probabilities at each of the sample points $\xi_1, \xi_2, ..., \xi_N$). This gives the following inner problem:

$$\begin{array}{ll} \max_{\mathbb{Q}} & \mathbb{E}_{\mathbb{Q}}[c(x,z)] \\ \text{subject to} & d_W(\mathbb{Q}, \mathbb{P}_0) \leq \delta \end{array}$$

in which the expectation is taken over the random variable z in $\mathbb{R}^m$ with distribution $\mathbb{Q}$. More generally we will use z to indicate an element of the set M that contains ξ_i.

In the previous two sections the structure of the cost function with respect to the random variable ξ has not been critical; everything has been determined by the set of cost functions $c(x, \xi_i)$ evaluated at the sample points. With Wasserstein robustification we will consider moves in the sample points ξ_i and we need to pay much more attention to the behavior of $c(x, z)$ with respect to changes in z. Often we will take x fixed and it is convenient to write $c_x(z)$ for $c(x, z)$. We assume throughout that $c_x(z)$ is differentiable at all $z \in \mathbb{R}^m$.

Using (17), the inner maximization problem is equivalent to solving

$$\begin{array}{ll} \max_{\mathbb{Q},\gamma} & \mathbb{E}_{\mathbb{Q}}[c_x(z)] \\ \text{subject to} & \int_{M\times M} \|z - \xi\| \ d\gamma(z,\xi) \leq \delta, \\ & \gamma \in \Gamma(\mathbb{Q}, \mathbb{P}_0). \end{array}$$

Since $\gamma \in \Gamma(\mathbb{Q}, \mathbb{P}_0)$, the set of all measures on the product space $M \times M$ with marginals $\mathbb{Q}$ and $\mathbb{P}_0$, it has a discrete distribution as one of the marginals, and we may specify it through specifying the distribution that each of the sample points ξ_i matches to under γ. More precisely we can rewrite $\gamma \in \Gamma(\mathbb{Q}, \mathbb{P}_0)$ in terms of components γ_i that are measures on M with $\gamma_i = \gamma(\cdot, \xi_i)$. Since $\mathbb{P}_0$ has mass $1/N$ at ξ_i we have $\gamma_i(M) = 1/N$, and the probability measure $\mathbb{Q}$ is obtained from adding together the components from each sample point, $\mathbb{Q} = \sum_{i=1}^N \gamma_i$.

It is convenient to scale the individual components γ_i so that they are probability measures: $\mathbb{Q}_i = N\gamma_i$ (with the scaling of N applied so that total mass of $\mathbb{Q}_i$ is 1). We can then write the inner problem as

$$\begin{array}{ll} \bar{\text{P}}\text{: } \max_{\mathbb{Q}_i} & \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbb{Q}_i}[c_x(z_i)] \\ \text{subject to} & \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbb{Q}_i}[\|z_i - \xi_i\|] \leq \delta, \end{array}$$

where z_i is a random variable with distribution $\mathbb{Q}_i$.

We make use of a result of Gao and Kleywegt [9, Corollary 2, part iii].

Proposition 15 *(Gao and Kleywegt) If there is an optimal solution to $\bar{P}$, then there is an optimal solution with an index i_0 such that for every $i \neq i_0$, $\mathbb{Q}_i$ has weight 1 on a point $z_i^* \in \arg\max_{z\in M}\{c_x(z) - \lambda^* \|z - \xi_i\|\}$ where $\lambda^* \geq 0$ is the Lagrange multiplier for the constraint in $\bar{P}$, and $\mathbb{Q}_{i_0}$ has weight on at most two points in $\arg\max_{z\in M}\{c_x(z) - \lambda^* \|z - \xi_{i_0}\|\}$.*

In the case where c_x is strictly concave in z we can be more explicit about the solution of $\bar{\text{P}}$. In this case we can think about contour surfaces of $\|\nabla c_x(z_j)\|$ in M for varying values of δ. In the solution to $\bar{\text{P}}$ all the points outside such a surface are moved inwards to lie on that surface and points inside the surface are not moved.

Proposition 16 *When $c_x(z)$ is strictly concave then $\bar{P}$ has a solution in which each $\mathbb{Q}_i$ has support at a single point $z_i \in M$. If we write $\bar{J} = \{i : z_i \neq \xi_i\}$ for the points that move, then (a) $\nabla c_x(z_i) = \alpha_i(z_i - \xi_i)$ for some scalar α_i for $i \in \bar{J}$; (b) $\|\nabla c_x(z_i)\| = \|\nabla c_x(z_j)\|$ for $i \in \bar{J}$ and $j \in \bar{J}$; and (c) $\|\nabla c_x(z_i)\| \geq \|\nabla c_x(\xi_k)\|$ for $i \in \bar{J}$ and $k \notin \bar{J}$.*

When we do not have a strictly concave cost function c_x then the types of move that occur are in general more complex. For example in the case where c_x is convex the inner maximization sends weight at z_i to a point on the boundary of the region M. For small δ we will find that just one point is changed, with some weight left at its original position and a small part of the total weight moved to a point on the boundary of M. In general we will not want to have such a dependence on the boundary of M, since in many problems the probability of ξ being near the boundary of the region M is very small, and the choice of boundary will be somewhat arbitrary.

Our next result shows linear variation for the Wasserstein form of DRQP given that $c(x, z) = \frac{1}{2}x^\top Hx + v(z)^\top x$. Note that $\nabla c_x(z) = \sum_j x_j \nabla v_j(z)$ and $c_x(z)$ is strictly concave if each component of v is strictly concave. We write $J_v(z)$ for the Jacobian matrix for the vector function $v : \mathbb{R}^m \to \mathbb{R}^n$ so the ij th element of $J_v(z)$ is $\frac{\partial v_i}{dz_j}$. Thus the elements of the vector ∇v_i are on the i'th row of $J_v(z)$ which is an $n \times m$ matrix.

Proposition 17 *With the Wasserstein distance metric, if every component of $v(z)$ is strictly concave then DRQP exhibits linear variation with*

$$\bar{y}(S) = \frac{H^{-1}J_v(\xi^*(S))J_v(\xi^*(S))^\top H^{-1}\bar{v}_0(S)}{\|J_v(\xi^*(S))^\top H^{-1}\bar{v}_0(S)\|}$$

where $\xi^*(S) = \xi_{i_0}$ is a sample point with the largest gradient norm for the cost function evaluated at the SAA solution $x_0(S)$ (i.e. $i_0 = \arg\max_i \left\|\nabla_z(v(\xi_i)^\top H^{-1}\bar{v}_0(S))\right\|$).

Proposition 17 simplifies when x is a scalar, where we have $c(x,\xi) = \frac{1}{2}x^2 - g(\xi)x$. Then $H = 1$, and, writing $\bar{g}_0(S) = (1/N)\sum_{i=1}^N g(\xi_i)$,

$$\bar{y}(S) = -\frac{\nabla g(\xi^*(S))^\top \nabla g(\xi^*(S))\bar{g}_0(S)}{\|\bar{g}_0(S)\nabla g(\xi^*(S))\|} = -\|\nabla g(\xi^*(S))\|$$

provided we have $\bar{g}_0(S) > 0$.

Proposition 18 *(a) When $c(x,\xi) = \frac{1}{2}x^2 - g(\xi)x$ and g is a strictly convex and non-negative function of ξ then*

$$MVRS = \mathbb{E}_S[\,(\bar{g}_0(S) - \bar{g})\,\|\nabla g(\xi^*(S))\|]$$

where $\xi^*(S) = \arg\max_{\xi_i \in S}\{\|\nabla g(\xi_i)\|\}$.
(b) In the case that $c(x,\xi) = \frac{1}{2}x^2 - \xi^2 x$ and the ξ values are realizations of a random variable which is non-negative and has density and cdf given by f and F, then

$$\begin{aligned} MVRS &= 2(N-1)\int_0^\infty \left(\int_z^\infty uF(u)^{N-2}f(u)du\right) z^2 f(z)dz \\ &\quad + 2\int_0^\infty z^3 F(z)^{N-1}f(z)dz - 2N\left(\int_0^\infty u^2 f(u)du\right)\int_0^\infty zF(z)^{N-1}f(z)dz. \end{aligned}$$

Note that part (a) of this result is similar to the formula in the total variation case, but we have the sample range $R(S)$ replaced by the maximum of $\|\nabla g(\xi_i)\|$, $\xi_i \in S$. There is very similar behavior here to that we have seen in other cases. If there is a skew in the underlying distribution of ξ towards values with high values for $g(\xi)$, then we can expect to see samples where there is an outlier producing both a high value for $\bar{g}_0(S) - \bar{g}$ and also a high value for $\|\nabla g(\xi^*_S)\|$. This will give a positive correlation between the two and hence a positive value for MVRS. This is illustrated in the example below.

Example 6

We suppose that $c(x,\xi) = \frac{1}{2}x^2 - \xi^2 x$ and the underlying distribution of the random variable ξ is exponential with mean 1, so $f(\xi) = e^{-\xi}$, $F(\xi) = 1 - e^{-\xi}$. Thus

$$\begin{aligned} \text{MVRS} &= 2(N-1)\int_0^\infty \left(\int_z^\infty u(1-e^{-u})^{N-2}e^{-u}du\right) z^2 e^{-z}dz \\ &\quad + 2\int_0^\infty z^3(1-e^{-z})^{N-1}e^{-z}dz - 4N\int_0^\infty z(1-e^{-z})^{N-1}e^{-z}dz \end{aligned}$$

since $\int_0^\infty u^2 e^{-u}du = 2$. When $N = 5$ we can numerically evaluate the integrals and obtain MVRS = 2.497. □

7 Conclusions and discussion

The application of robustification to stochastic optimization problems to improve mean out-of-sample performance has been widely reported in the literature. Robustification has value from a risk reduction point of view, but it may also have value for a risk neutral decision maker. This paper contributes to our understanding of why this is the case.

Empirical evidence from many different studies has shown that a small amount of robustification can improve out-of-sample performance, so our analysis focuses on what we call incremental improvement, that is improvement in performance as the size of the distributional uncertainty set increases from zero. Incremental improvement arises from changes in the minimizing point. In many cases, namely those with linear variation, we can define a directional derivative of the minimizer that can be used to quantify incremental improvement, and evaluate the improvement in out-of-sample cost to first-order, expressed as the marginal value of robust solution (MVRS).

To illustrate the concepts, we quantify incremental improvement and MVRS for several examples, all with convex quadratic objective functions. MVRS depends on the form of the linear term in this objective function, the version of robustification applied, and the underlying "ground-truth"probability distribution. Our analysis shows that incremental improvement cannot be taken for granted and different robustification approaches applied to the same problem can give MVRS values having opposite signs. We also show by example how adding a random constant to the objective function of an optimization will change the optimal solution of the robustified problem while leaving the optimizers of the "true" problem and its sample-average approximation unchanged.

To understand the impact of small amounts of robustification, we can summarize the changes made on the SAA problem as follows.

1. For ϕ-divergence, weight is moved from points with low cost to points with high cost with the change in weight depending linearly on the cost values. For total variation, weight is removed from the point in the sample that gives the lowest cost and moved to the point in the sample that has the highest cost.

2. For CVaR robustification, weight is removed from all points in the sample and added to a small number of points in the sample that correspond to high costs.

3. For Wasserstein robustification, provided $c(x, \xi)$ is strictly concave in ξ, the sample point with the largest value for the norm of the gradient

with respect to ξ is moved incrementally to a higher cost position (the exact move depends on the function c).

These effects have a simple form in a univariate framework, when we have $c(x,\xi) = \frac{1}{2}x^2 - g(\xi)x$. With sample S, the sample average approximation solution $x_0(S)$ is equal to $\bar{g}_0(S)$. Since each of the different robustification approaches move weight to lower values of $g(\xi_i)$ (corresponding to higher costs) we have $x_\delta(S) < x_0(S)$. Though this introduces a bias in the value of $\mathbb{E}_S[x_\delta(S)]$ we can obtain improvement through shrinkage when there are larger moves to the left for samples with high values of $\bar{g}_0(S)$ (and hence high values for $x_0(S)$) than there are for samples with low values of $\bar{g}_0(S)$ (and hence low values for $x_0(S)$). Hence we get an advantage when the sample mean is positively correlated with the size of the change in optimal solution induced by the robustification.

When we consider the total variation form of the robustification it is only the tails that influence the change that is made, and $(x_0(S) - x_\delta(S))/\delta$ is simply half the range of values in the sample. Here any skew to the right in the distribution of $g(\xi)$ will induce a correlation that yields a positive value for MVRS. For more general ϕ-divergence we have similar behavior with skew to the right in $g(\xi)$ leading to incremental improvement from robustification. We note that MVRS is zero for symmetric distributions under ϕ-divergence robustification, which does not hold for the other two types of robustification.

For CVaR robustification the change in optimal solution, $x_0(S) - x_\delta(S)$, depends on the entire sample average since weight is removed from all the points in the sample, except those at the left hand end of $g(\xi_i)$. This produces the term σ^2/N that does not appear in the other robustifications that involve changes only to the points at the two extremes of the sample. The value of MVRS for CVaR robustification also depends on the left hand tail of the $g(\xi)$. Where that tail is long, the existence of a point in the sample that is far out in the left tail means that there will be a small sample average and also the CVaR robustification adds weight to a point far to the left. We end up with a negative correlation between $\bar{g}_0(S) - \bar{g}$ and $x_0(S) - x_\delta(S)$. This effect works in the opposite direction to the σ^2/N term.

Examples for CVaR robustification show that when the distribution is uniform over an interval, the σ^2/N term dominates and MVRS is positive; when the distribution is normal the effect from the left hand tail balances the positive term and MVRS is approximately zero; and when the distribution is a mixture of normals having a heavier left hand tail than the normal, then the tail behavior dominates and MVRS is negative. In loose terms we may think of the normal distribution as a kind of boundary between cases where MVRS for CVaR is positive or negative.

For the Wasserstein robustification and convex $g(\xi)$ the point where g has the highest gradient is moved. This will be a point towards the extremities of the ξ_i values (that in general occur in a multivariate space) - and hence is likely to be where $g(\xi_i)$ is large and so costs are low. In the special case of ξ scalar and $g(\xi) = \xi^2$ then it is the lowest cost point in the sample that is moved. Consistent with our discussion so far we have a positive value for MVRS when the distribution of ξ^2 has a positive skew.

From a practical perspective, our results give some guidance to risk-neutral decision makers facing a stochastic optimization problem. In practice, the true probability distribution of uncertain parameters will hardly ever be known, so MVRS cannot be computed as we have done in this paper. However, there are often circumstances when a decision maker has some knowledge of the underlying distribution that can be helpful in predicting how robustification will perform. Two characteristics of the distribution are particularly relevant: Is the distribution symmetric or skewed? And does the distribution have heavy tails?

We have shown how to quantify incremental improvement from robustification in univariate examples using the "ground-truth" probability distribution. With random variables that are multivariate or have unknown distributions, these analytical techniques are not applicable, and (assuming the DRO problem has linear variation) one has to resort to statistical estimation of MVRS from the data available.

A significant restriction in our analysis is the specific form of the objective function studied. First, we have assumed a strictly convex quadratic function. This ensures uniqueness of the true solution and that of the sample average approximation, which enables a simpler analysis of linear variation and incremental improvement. If the optimal solution is not unique then a more complicated set-valued variational analysis is required.

We have also assumed that the objective function has no stochastic constant term and the quadratic term in the objective function $x^\top Hx$ is not stochastic. As remarked above the stochastic constant term can alter the solution to the robustified problem. We have chosen to set this term to be zero for simplicity, but for any form of this term one could carry out a similar analysis to study the effect of robustification.

On the other hand if H is stochastic then the SAA solution will in general be biased. The advantage of our treatment (with deterministic H) is that it avoids confusion between bias and shrinkage. When there is a bias in the SAA solution it will be affected by the robustification, being either increased or decreased depending on circumstances.

References

[1] G. Bayraksan and D.K. Love. Data-driven stochastic programming using phi-divergences. In *The Operations Research Revolution*, pages 1–19. INFORMS, 2015.

[2] D. Bertsimas, V. Gupta, and N. Kallus. Robust sample average approximation. *Mathematical Programming*, 171(1-2):217–282, 2018.

[3] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627*, 2016.

[4] V. K. Chopra and W. T. Ziemba. The effect of errors in means, variances, and covariances on optimal portfolio choice. In *Handbook of the Fundamentals of Financial Decision Making: Part I*, pages 365–373. World Scientific, 2013.

[5] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[6] M. Drela. Pros & cons of airfoil optimization. In *Frontiers of Computational Fluid Dynamics 1998*, pages 363–381. World Scientific, 1998.

[7] J. Dupačová. The minimax approach to stochastic programming and an illustrative application. *Stochastics: An International Journal of Probability and Stochastic Processes*, 20(1):73–88, 1987.

[8] P.M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[9] R. Gao and A.J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[10] J-Y. Gotoh, M.J. Kim, and A.E.B. Lim. Calibration of distributionally robust empirical optimization models. *arXiv preprint arXiv:1711.06565*, 2017.

[11] J-Y. Gotoh, M.J. Kim, and A.E.B. Lim. Robust empirical optimization is almost the same as mean–variance optimization. *Operations Research Letters*, 46(4):448–452, 2018.

[12] D.M. Hawkins. The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12, 2004.

[13] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[14] S. Lawrence, C.L. Giles, and A.C. Tsoi. Lessons in neural network training: Overfitting may be harder than expected. In *AAAI/IAAI*, pages 540–545. Citeseer, 1997.

[15] E.H. Lloyd. Least-squares estimation of location and scale parameters using order statistics. *Biometrika*, 39(1/2):88–95, 1952.

[16] F. Luo and S. Mehrotra. Decomposition algorithm for distributionally robust optimization using Wasserstein metric. *arXiv preprint arXiv:1704.03920*, 2017.

[17] G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

[18] A.B. Philpott, V.L. de Matos, and L. Kapelevich. Distributionally robust SDDP. *Computational Management Science*, 15(3-4):431–454, 2018.

[19] H. Scarf. A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*, 1958.

[20] C. Schaffer. Overfitting avoidance as bias. *Machine Learning*, 10(2):153–178, 1993.

[21] A. Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

[22] A. Shapiro, A. Ruszczynski, and D. Dentcheva. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2014.

[23] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[24] B.P.G. Van Parys, P.M. Esfahani, and D. Kuhn. From data to decisions: Distributionally robust optimization is optimal. *arXiv preprint arXiv:1704.04118*, 2017.

[25] Z. Wang, P.W. Glynn, and Y. Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2):241–261, 2016.

[26] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

[27] D. Wozabal. Robustifying convex risk measures for linear portfolios: A nonparametric approach. *Operations Research*, 62(6):1302–1315, 2014.

[28] H. Xu, C. Caramanis, and S. Mannor. Robust regression and lasso. In *Advances in Neural Information Processing Systems*, pages 1801–1808, 2009.

Appendix 1: Proofs of propositions

Proof of Improvement Lemma

For the first part, for almost all S, we have

$$\begin{aligned} C_\delta(S) - C_0(S) &= \mathbb{E}_{\mathbb{P}}[c(x_\delta(S),\xi) - c(x_0(S),\xi)] \\ &= \mathbb{E}_{\mathbb{P}}[[\nabla_x c(x,\xi)]_{x_0(S)}]^\top (x_\delta(S) - x_0(S)) + O((x_\delta(S) - x_0(S))^2) \\ &= \nabla \bar{c}(x_0(S))^\top \bar{y}(S)\delta + O(\delta^2) \end{aligned}$$

by definition of $\nabla \bar{c}(x_0(S))$ and $\bar{y}(S)$, and since linear variation implies that $x_\delta(S) - x_0(S)$ is $O(\delta)$. It follows that

$$MVRS = -\mathbb{E}_S[\nabla \bar{c}(x_0(S))^\top \bar{y}(S)].$$

If $\mathbb{E}_S[C_\delta(S) - C_0(S)] < 0$ for $\delta > 0$ sufficiently small then we must have $\mathbb{E}_S[\nabla \bar{c}(x_0(S))^\top \bar{y}(S)] \leq 0$. Conversely if $\mathbb{E}_S[\nabla \bar{c}(x_0(S))^\top \bar{y}(S)] < 0$ then $\mathbb{E}_S[C_\delta(S) - C_0(S)] < 0$ for sufficiently small δ, so we get incremental improvement. □

Proof of Proposition 2

Consider $S = \{\xi_1, \xi_2, \ldots, \xi_N\}$ with

$$v(\xi_1)^\top x_0(S) \leq v(\xi_2)^\top x_0(S) \leq \ldots \leq v(\xi_N)^\top x_0(S).$$

Any sample S where $v(\xi_i)^\top x_0(S) = v(\xi_{i+1})^\top x_0(S)$, will have ξ_i satisfying

$$v(\xi_i)^\top H^{-1} \sum_{i=1}^{N} v(\xi_i) = v(\xi_{i+1})^\top H^{-1} \sum_{i=1}^{N} v(\xi_i)$$

which defines a linear subspace in $\mathbb{R}^n$. Since $v(\xi)$ has an n-dimensional density from which we sample independently, the probability measure of the set of samples satisfying $v(\xi_i)^\top x_0(S) = v(\xi_{i+1})^\top x_0(S)$ is 0. It follows that an ordered sample S satisfies

$$v(\xi_1)^\top x_0(S) < v(\xi_2)^\top x_0(S) < \ldots < v(\xi_N)^\top x_0(S)$$

with probability 1. □

Proof of Proposition 3

We have

$$\begin{aligned} C_0(S) &= \frac{1}{2} x_0(S)^\top H x_0(S) + \bar{v}^\top x_0(S) + \bar{u} \\ &= \frac{1}{2} v_0(S)^\top H^{-1} v_0(S) - \bar{v}^\top H^{-1} v_0(S) + \bar{u}. \end{aligned}$$

Suppose DRQP has a unique solution $x_\delta(S)$. Let $\bar{v}_\delta(S) = -Hx_\delta(S)$. Then

$$\begin{aligned} C_\delta(S) &= \frac{1}{2}x_\delta(S)^\top Hx_\delta(S) + \bar{v}^\top x_\delta(S) + \bar{u} \\ &= \frac{1}{2}\bar{v}_\delta^\top H^{-1}\bar{v}_\delta(S) - \bar{v}^\top H^{-1}\bar{v}_\delta(S) + \bar{u} \\ &= C_0(S) + (\bar{v}_0(S) - \bar{v})^\top H^{-1}(\bar{v}_\delta(S) - \bar{v}_0(S)) \\ &+ \frac{1}{2}(\bar{v}_\delta(S) - \bar{v}_0(S))^\top H^{-1}(\bar{v}_\delta(S) - \bar{v}_0(S)). \end{aligned}$$

(using the symmetry of H^{-1}) as required. □

Proof of Lemma 4

The improvement lemma gives the existence of a $\bar{y}(S)$ for almost every sample S, with

$$C_\delta(S) = C_0(S) + \nabla\bar{c}(x_0(S))^\top\bar{y}(S)\delta + O(\delta^2).$$

Substituting $\nabla\bar{c}(x_0(S)) = \bar{v} - \bar{v}_0(S)$ gives

$$C_\delta(S) = C_0(S) - (\bar{v}_0(S) - \bar{v})^\top\,\bar{y}(S)\delta + O(\delta^2),$$

and

$$MVRS = \mathbb{E}_S[(\bar{v}_0(S) - \bar{v})^\top\,\bar{y}(S)].$$

Substituting $-Hx_\delta(S)$ for $\bar{v}_\delta(S)$ and $-Hx_0(S)$ for $\bar{v}_0(S)$ in the formula for $C_\delta(S)$ in Proposition 3 yields

$$\begin{aligned} \mathbb{E}_S[C_\delta(S) - C_0(S)] \;=\; & \mathbb{E}_S[(\bar{v}_0(S) - \bar{v})^\top(x_0(S) - x_\delta(S)) \\ & + \frac{1}{2}(x_\delta(S) - x_0(S))^\top H(x_\delta(S) - x_0(S))]. \end{aligned} \tag{18}$$

The substitution $x_\delta(S) - x_0(S) = \bar{y}(S)\delta + O(\delta^2)$ gives

$$\mathbb{E}_S[C_\delta(S) - C_0(S)] = -\mathbb{E}_S[(\bar{v}_0(S) - \bar{v})^\top\bar{y}(S)]\delta + \frac{\delta^2}{2}\bar{y}(S)^\top H\bar{y}(S) + O(\delta^2).$$

In the limit of small δ the first term dominates and if $\mathbb{E}_S[(\bar{v}_0(S) - \bar{v})^\top\bar{y}(S)] > 0$ then the overall expression is negative for small δ and we obtain incremental improvement. Conversely if there is incremental improvement, we need $\mathbb{E}_S[(\bar{v}_0(S) - \bar{v})^\top\bar{y}(S)] \geq 0$.

If $\mathbb{E}_S[(\bar{v}_0(S) - \bar{v})^\top\bar{y}(S)] = 0$ then we need the second order terms in the right hand side of (18) to be at most zero. Since $\frac{1}{2}(x_\delta(S) - x_0(S))^\top H(x_\delta(S) -$

$x_0(S)) \geq 0$, we need the second order terms in $\mathbb{E}_S[(\bar{v}_0(S)-\bar{v})^\top(x_0(S)-x_\delta(S))]$ to be at most zero. It follows that

$$\lim_{\delta\to 0} \mathbb{E}_S[(\bar{v}_0(S)-\bar{v})^\top(x_\delta(S)-x_0(S))]/\delta^2 \geq 0$$

as required. □

Proof of Lemma 5

For an arbitrary x suppose we order the elements of a sample $S = \{\xi_1, \xi_2, \ldots, \xi_N\}$ so that

$$v(\xi_1)^\top x = \ldots = v(\xi_k)^\top x < v(\xi_{k+1})^\top x \leq \ldots \leq v(\xi_{l-1})^\top x < v(\xi_l)^\top x = \ldots = v(\xi_N)^\top x.$$

It is easy to see that

$$Q_{\max}(x) = \bar{v}_0(S)^\top x + (\delta/2)(\max_{\xi_i} v(\xi_i)^\top x - \min_{\xi_i} v(\xi_i)^\top x).$$

This is a convex function of x with subdifferential

$$\partial Q_{\max}(x) = \bar{v}_0(S) + (\delta/2)G(x)$$

where

$$G(x) = \text{conv}(\{v(\xi_l), ..., v(\xi_N)\}) + \text{conv}(\{-v(\xi_1), ..., -v(\xi_k)\}).$$

So an optimal solution $x_\delta(S)$ to DRQP satisfies

$$0 \in Hx_\delta(S) + \bar{v}_0(S) + (\delta/2)G(x_\delta(S)). \tag{19}$$

Since $G(x)$ is bounded, for any optimal solution $x_\delta(S)$ we have $\lim_{\delta\to 0} x_\delta(S) = x_0(S)$. Now by Proposition 2 we have all samples S apart from a set with probability 0 are strictly ordered by SAA, so

$$v(\xi_1)^\top x_0(S) < v(\xi_2)^\top x_0(S) \ < ... < v(\xi_N)^\top x_0(S), \tag{20}$$

and so for the samples that are strictly ordered by SAA, and for δ small enough we have (20) with $x_\delta(S)$ replacing $x_0(S)$. It follows that for all samples S strictly ordered by SAA, the relationship (19) becomes

$$Hx_\delta(S) + \bar{v}_0(S) + (\delta/2)(v(\xi_N) - v(\xi_1)) = 0,$$

so

$$x_\delta(S) = -H^{-1}(\bar{v}_0(S) + \delta R(S)/2),$$

where $R(S) = v(\xi_N) - v(\xi_1)$. Thus for δ small enough $x_\delta(S)$ is unique and

$$x_\delta(S) - x_0(S) = -\frac{1}{2}H^{-1}R(S)\delta,$$

so DRQP exhibits linear variation with $\bar{y}(S) = -\frac{1}{2}H^{-1}R(S)$. Furthermore

$$\begin{aligned} C_\delta(S) &= \frac{1}{2}\left(x_0(S) - \frac{1}{2}\delta H^{-1}R(S)\right)^\top H\left(x_0(S) - \frac{1}{2}\delta H^{-1}R(S)\right) + \bar{v}^\top\left(x_0(S) - \frac{1}{2}\delta H^{-1}R(S)\right) \\ &= C_0(S) - \frac{\delta}{2}x_0(S)^\top R(S) + \frac{\delta^2}{8}R(S)^\top H^{-1}R(S) - \frac{\delta}{2}\bar{v}^\top H^{-1}R(S) \\ &= C_0(S) - \frac{\delta}{2}(\bar{v} - \bar{v}_0(S))^\top H^{-1}R(S) + \frac{\delta^2}{8}R(S)^\top H^{-1}R(S). \end{aligned}$$

Thus we obtain

$$\mathrm{VRS}(\delta) = \mathbb{E}_S[\frac{\delta}{2}(\bar{v} - \bar{v}_0(S))^\top H^{-1}R(S) - \frac{\delta^2}{8}R(S)^\top H^{-1}R(S)].$$

So

$$\mathrm{MVRS} = \mathbb{E}_S[(1/2)(\bar{v} - \bar{v}_0(S))^\top H^{-1}R(S)]$$

with the remark on incremental improvement being immediate. □

Proof of Proposition 6

Since $H = 1$, Lemma 5 gives for sufficiently small $\delta > 0$,

$$\mathrm{VRS}(\delta) = \mathbb{E}_S[\frac{\delta}{2}(\bar{v} - \bar{v}_0(S))R(S) - \frac{\delta^2}{8}R(S)^2].$$

Since $g(\xi) > 0$ almost surely, we have $\bar{x}_0(S) > 0$ for almost all samples S, and so for sufficiently small δ

$$R(S) = v(\xi_N) - v(\xi_1) = g(\xi_1) - g(\xi_N).$$

Now defining $\bar{R} = E_S[R(S)]$, we get

$$\begin{aligned} \mathbb{E}_S[(\bar{v} - \bar{v}_0(S))R(S) &= \mathbb{E}_S[(\bar{g}_0(S) - \bar{g})(R(S) - \bar{R})] \\ &= \mathrm{cov}(\bar{g}_0(S), R(S)) \end{aligned}$$

so

$$\mathrm{VRS}(\delta) = \frac{\delta}{2}\mathrm{cov}(\bar{g}_0(S), R(S)) - \frac{\delta^2}{8}\mathbb{E}_S[R(S)^2]$$

and MVRS $=\mathrm{cov}(\bar{g}_0(S), R(S))$.

In the case that the distribution of prices $g(\xi)$ is symmetric about its mean then we can condition on R_S and observe that for any sample with outcomes $\{g(\xi_1), g(\xi_2), ..., g(\xi_N)\}$ there is another sample with outcomes $\{2\bar{g} - g(\xi_1), 2\bar{g} - g(\xi_2), ..., 2\bar{g} - g(\xi_N)\}$ which is equally likely, in which each outcome is replaced by an outcome at the same distance but on the opposite side of $\bar{g}$. This mirror sample has the same range but $(\bar{g} - \bar{g}_0(S))$ is reversed in sign since $\bar{g} - g(\xi_i)$ is replaced by $\bar{g} - (2\bar{g} - g(\xi_i)) = g(\xi_i) - \bar{g}$. From this we deduce that MVRS is zero, and thus $\mathbb{E}_S[C_\delta(S)] = \mathbb{E}_S[C_0(S)] + \left(\delta^2/8\right)\mathbb{E}_S[R(S)^2]$, giving $\text{VRS}(\delta) < 0$ for all $\delta > 0$. □

Proof of Proposition 7

We have

$$\begin{aligned} \text{MVRS} &= \frac{1}{2}\mathbb{E}[(\bar{g}_0(S) - \bar{g})R_S] \\ &= \frac{1}{2}\mathbb{E}\left[(z_N - z_1)\bar{z}\right]. \end{aligned}$$

where z_i is the i'th order statistic of $\{g(\xi_i) - \bar{g} : i = 1, ...N\}$. By Lemma 26 in Appendix 2,

$$\mathbb{E}[z_N\bar{z}] = \int_{-\infty}^{\infty} Q(z)F(z)^{N-1}dz$$

and

$$\mathbb{E}[z_1\bar{z}] = \int_{-\infty}^{\infty} Q(z)(1 - F(z))^{N-1}dz$$

which yields the result. □

Proof of Proposition 8

Let $\widetilde{f}$ be the density for $\widetilde{F}$, the symmetric distribution matching $f(w)$ for $w < w_0$. Thus $\widetilde{f}(w_0 + \gamma) = f(w_0 - \gamma)$ and $\widetilde{F}(w_0 + \gamma) = 1 - F(w_0 - \gamma)$, for $\gamma > 0$. Define $\tau(z) = F^{-1}(1 - F(2w_0 - z))$ for $z > w_0$ and $\tau(z) = z$ for $z \leq w_0$. Hence $F(z) = \widetilde{F}(\tau^{-1}(z))$, and so $f(z) = \widetilde{f}(\tau^{-1}(z))/\tau'(\tau^{-1}(z))$.

We know that F has mean 0 and hence

$$0 = \int_{-\infty}^{\infty} zf(z)dz = \int_{-\infty}^{\infty} \frac{z}{\tau'(\tau^{-1}(z))}\widetilde{f}(\tau^{-1}(z))dz = \int_{-\infty}^{\infty} \tau(w)\widetilde{f}(w)dw \quad (21)$$

using a change of variable $w = \tau^{-1}(z)$ so $\tau'(w)dw = dz$. We may write

$$\begin{aligned} \int_{-\infty}^{\infty} \tau(w)\widetilde{f}(w)dw &= \int_{-\infty}^{w_0} w\widetilde{f}(w)dw + \int_{w_0}^{\infty} \tau(w)\widetilde{f}(w)dw \\ &= \int_{w_0}^{\infty} (2w_0 - z)\widetilde{f}(z)dw + \int_{w_0}^{\infty} \tau(w)\widetilde{f}(w)dw \end{aligned}$$

using symmetry for $\widetilde{f}$. So

$$\int_{w_0}^{\infty} (\tau(z) + 2w_0 - z)\, \widetilde{f}(z)dz = 0. \tag{22}$$

We will use Proposition 7 and we begin by rewriting the required expression in terms of $\widetilde{F}$. From (11) we have

$$\begin{aligned}
\text{MVRS} &= \frac{1}{2}\int_{-\infty}^{\infty} \left(\widetilde{F}(\tau^{-1}(z))^{N-1} - (1-\widetilde{F}(\tau^{-1}(z)))^{N-1}\right) \times \left(\int_z^{\infty} u \frac{\widetilde{f}(\tau^{-1}(u))}{\tau'(\tau^{-1}(u))} du\right) dz \\
&= \frac{1}{2}\int_{-\infty}^{\infty} \left(\widetilde{F}(w)^{N-1} - (1-\widetilde{F}(w))^{N-1}\right) \times \left(\int_w^{\infty} \tau(z)\widetilde{f}(z)dz\right) \tau'(w)dw
\end{aligned}$$

using a change of variable $w = \tau^{-1}(z)$ and $z = \tau^{-1}(u)$. Since $\tau(w) = w$ for $w \leq w_0$, this expression can be written

$$\begin{aligned}
\text{MVRS} &= \frac{1}{2}\int_{-\infty}^{w_0} \left(\widetilde{F}(z)^{N-1} - (1-\widetilde{F}(z))^{N-1}\right) \left(\int_z^{w_0} u\widetilde{f}(u)dz + \int_{w_0}^{\infty} \tau(u)\widetilde{f}(u)du\right) dz \\
&\quad + \frac{1}{2}\int_{w_0}^{\infty} \left(\widetilde{F}(z)^{N-1} - (1-\widetilde{F}(z))^{N-1}\right) \left(\int_z^{\infty} \tau(u)\widetilde{f}(u)du\right) \tau'(z)dz.
\end{aligned}$$

Let $T(z) = \int_z^{\infty} \tau(u)\widetilde{f}(u)du \geq 0$ for $z \geq w_0$. From (22)

$$T(w_0) = \int_{w_0}^{\infty} \tau(z)\widetilde{f}(z)dz = \int_{w_0}^{\infty} (z - 2w_0)\, \widetilde{f}(z)dz > 0$$

since the skew in the distribution ensures that $w_0 < 0$ and hence $z - 2w_0 > 0$ for $z > w_0$. Now observe that $T(z)$ begins by increasing in z while $\tau(z) < 0$ and then decreases. It approaches the value zero, for z large, and hence $T(z) > 0$ for $z \geq w_0$. And thus $\left(\widetilde{F}(z)^{N-1} - (1-\widetilde{F}(z))^{N-1}\right)\left(\int_z^{\infty} \tau(u)\widetilde{f}(u)du\right) > 0$ for $z > w_0$.

Now from our assumption $f(w) \geq f(F^{-1}(1-F(w))$ and the definition of τ we obtain $f(2w_0 - z) \geq f(\tau(z))$. But as $F(\tau(z)) = 1 - F(2w_0 - z)$ we know that $f(\tau(z))\tau'(z) = f(2w_0 - z)$. And hence our assumption implies $\tau'(z) \geq 1$ with strict inequality for some range of values. Thus we have

$$\begin{aligned}
\text{MVRS} &> \frac{1}{2}\int_{-\infty}^{w_0} \left(\widetilde{F}(z)^{N-1} - (1-\widetilde{F}(z))^{N-1}\right) \left(\int_z^{w_0} u\widetilde{f}(u)dz + \int_{w_0}^{\infty} \tau(u)\widetilde{f}(u)du\right) dz \\
&\quad + \frac{1}{2}\int_{w_0}^{\infty} \left(\widetilde{F}(z)^{N-1} - (1-\widetilde{F}(z))^{N-1}\right) \left(\int_z^{\infty} \tau(u)\widetilde{f}(u)du\right) dz.
\end{aligned}$$

Now $\widetilde{F}(w)^{N-1}-(1-\widetilde{F}(w))^{N-1}$ is symmetric with a change of sign around z_0. We can use the same argument that established MVRS is zero for symmetric f to show the corresponding expression for $\widetilde{F}$ is zero after shifting to allow for the non zero mean:

$$\int_{-\infty}^{\infty}\left(\widetilde{F}(z)^{N-1}-(1-\widetilde{F}(z))^{N-1}\right)\left(\int_z^{\infty}(u-w_0)\widetilde{f}(u)du\right)dw=0.$$

We can subtract half this integral from the right hand side of the inequality to obtain

$$\begin{aligned}\text{MVRS} \;>\; & \frac{1}{2}\int_{-\infty}^{w_0}\left(\widetilde{F}(z)^{N-1}-(1-\widetilde{F}(z))^{N-1}\right)\left(\int_z^{w_0}w_0\widetilde{f}(u)du+A\right)dz\\ & +\frac{1}{2}\int_{w_0}^{\infty}\left(\widetilde{F}(w)^{N-1}-(1-\widetilde{F}(w))^{N-1}\right)\left(\int_z^{\infty}(\tau(u)-u+w_0)\widetilde{f}(u)du\right)dz\end{aligned}$$

where $A=\int_{w_0}^{\infty}(\tau(u)-u+w_0)\widetilde{f}(u)du$. We can use (22) to show that $A=-\frac{w_0}{2}$, but we don't use this fact. We want to split the second term in these integrals into a symmetric and non-symmetric part. We can write

$$\text{MVRS}>\frac{1}{2}\int_{-\infty}^{\infty}\left(\widetilde{F}(z)^{N-1}-(1-\widetilde{F}(z))^{N-1}\right)(U(z)+V(z))dz$$

where $U(z)=\int_z^{w_0}w_0\widetilde{f}(u)du+A$ for $z\le w_0$ and $U(z)=\int_{w_0}^{z}w_0\widetilde{f}(u)du+A$ for $z>w_0$. Note that U is symmetric around w_0 and is maximized at w_0 since $w_0<0$. Hence

$$\begin{aligned}&\left(\widetilde{F}(w_0-k)^{N-1}-(1-\widetilde{F}(w_0-k))^{N-1}\right)U(w_0-k)\\ =\;& -\left(\widetilde{F}(w_0+k)^{N-1}-(1-\widetilde{F}(w_0+k))^{N-1}\right)U(w_0+k),\end{aligned}$$

and so $\int_{-\infty}^{\infty}\left(\widetilde{F}(z)^{N-1}-(1-\widetilde{F}(z))^{N-1}\right)U(z)dz=0$.

Also $V(z)=0$ for $z\le w_0$ and for $z>w_0$ we have

$$\begin{aligned}V(z) &= \int_z^{\infty}(\tau(z)-u+w_0)\widetilde{f}(u)du-\left(\int_{w_0}^{z}w_0\widetilde{f}(u)du+A\right)\\ &= \int_{w_0}^{z}(-2w_0-\tau(u)+u)\widetilde{f}(u)du.\end{aligned}$$

So, from (22), $V(\infty)=0$. Now $\tau(u)+2z_0-u$ has derivative $\tau'(u)-1\ge 0$ and because the integral in (22) is zero, we can deduce that $\tau(u)+2w_0-u$ starts negative and becomes positive. Since

$$\frac{d}{dz}V(z)=(-2w_0-\tau(z)+z)\widetilde{f}(z),$$

we know that V starts by increasing and then decreases to zero. Moreover $V(w_0) = 0$. Hence it is always non-negative. Since $V(z)$ is zero for $z \leq w_0$ when $\widetilde{F}(z)^{N-1} - (1 - \widetilde{F}(z))^{N-1} < 0$, then

$$\int_{-\infty}^{\infty} \left(\widetilde{F}(z)^{N-1} - (1 - \widetilde{F}(z))^{N-1} \right) V(z) dz \geq 0,$$

and so we have established that MVRS > 0, as required. □

Proof of Proposition 9

We are interested in the solution $x_\delta(S)$ to DRQP where we solve the inner maximization:

$$\begin{array}{lll} \text{IP: } \max_q & \sum_{i=1}^N q_i v(\xi_i)^\top x & \\ \text{s.t.} & \sum_{i=1}^N \frac{1}{N} \phi\left(N q_i\right) \leq \delta^2, & [\lambda] \\ & \sum_{i=1}^N q_i = 1, & [\mu] \\ & q_i \geq 0. & \end{array}$$

The problem IP has Lagrangian

$$\mathcal{L} = -\sum_{i=1}^N q_i v(\xi_i)^\top x + \lambda(\delta, x)(\sum_{i=1}^N \frac{1}{N} \phi\left(N q_i\right) - \delta^2) + \mu(\sum_{i=1}^N q_i - 1)$$

which has first derivative

$$\frac{\partial \mathcal{L}}{\partial q_i} = -v(\xi_i)^\top x + \lambda(\delta, x) \phi'\left(N q_i\right) + \mu.$$

Let $q_i(\delta, x)$ denote the optimal solution for a given δ and x. We translate this into $r_i(\delta, x)$ by

$$q_i(\delta, x) = \frac{1}{N}(1 + r_i(\delta, x)),$$

where $\sum_{i=1}^N q_i(\delta, x) = 1$ implies

$$\sum_i r_i(\delta, x) = 0.$$

Then minimization of the Lagrangian implies

$$-v(\xi_i)^\top x + \lambda(\delta, x) \phi'\left(1 + r_i(\delta, x)\right) + \mu = 0, \quad i = 1, 2, \ldots, N. \tag{23}$$

We write

$$\phi(1 + w) = \frac{w^2}{2} \phi''(1) + \frac{w^2}{2} g(w)$$

with $g(w) \to 0$ as $w \to 0$, and so $\phi'(1+w) = w\phi''(1) + wg(w) + \frac{w^2}{2}g'(w)$ and since ϕ is analytic we have g and $g'(w)$ well defined. We let $g_0(w) = g(w) + \frac{w}{2}g'(w)$, and $g_0(w) \to 0$ as $w \to 0$. Then

$$\phi'(1+w) = w(\phi''(1) + g_0(w))$$

and (23) becomes

$$-v(\xi_i)^\top x + \lambda(\delta,x)r_i(\delta,x)\left(\phi''(1) + g_0(r_i(\delta,x))\right) + \mu = 0.$$

Then $\sum_i r_i(\delta,x) = 0$ implies

$$\mu = \bar{v}_0(S)^\top x - \frac{1}{N}\sum_i \lambda(\delta,x)r_i(\delta,x)g_0(r_i(\delta,x))$$

so for each $i = 1, 2, \ldots, N$,

$$\begin{array}{rcl} -v(\xi_i)^\top x + \lambda(\delta,x)r_i(\delta,x)\left(\phi''(1) + g_0(r_i(\delta,x))\right) & & \\ +\bar{v}_0(S)^\top x - \frac{1}{N}\sum_j \lambda(\delta,x)r_j(\delta,x)g_0(r_j(\delta,x)) & = & 0. \end{array} \tag{24}$$

Moreover from the constraint

$$\frac{1}{N}\sum_{i=1}^{N}\phi\left(1 + r_i(\delta,x)\right) = \delta^2$$

we get

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{r_i(\delta,x)^2}{2}(\phi''(1) + g(r_i(\delta,x)))\right) = \delta^2. \tag{25}$$

We will establish the proposition via several lemmas.

Lemma 19 *The solution $\lambda(\delta,x), r_i(\delta,x)$, $i = 1, 2, \ldots, N$ to (24) and (25) has components that are analytic functions of x and δ for all (x,δ) in a neighbourhood $(x_0(S), 0)$.*

Proof We establish the result using the implicit function theorem. We can rewrite (24) and (25) in the form $G(r_1, r_2, ..., r_N, \lambda, x, \delta) = 0$, where $G : \mathbb{R}^{N+2+n} \to \mathbb{R}^{N+1}$ has components

$$\begin{array}{rcll} G_i(r_1, r_2, ..., r_N, \lambda, x, \delta) & = & \left(\phi''(1) + g_0(r_i)\right)\lambda r_i - \frac{1}{N}\lambda\sum_j r_j g_0(r_j) & \\ & & -\left(v(\xi_i)^\top x - \bar{v}_0(S)^\top x\right), & i = 1, 2, ..., N, \\ G_{N+1}(r_1, r_2 ... r_N, \lambda, x, \delta) & = & \frac{1}{N}\sum_{i=1}^{N}\left(\frac{{r_i}^2}{2}(\phi''(1) + g(r_i))\right) - \delta^2 & \end{array}$$

We will apply the implicit function theorem to express $(r_1, r_2 ... r_N, \lambda)$ as functions of (x, δ). We require G to be continuously differentiable around the point $(x, \delta, \lambda, r) = (x_0(S), 0, \lambda(0, x_0(S)), 0)$, and the $(N+1) \times (N+1)$ Jacobian matrix with elements $J_{ij} = \partial G_i / dr_j$, $j = 1, 2, ...N$ and $J_{i,N+1} = \partial G_i / d\lambda$ to be non-singular at this point.

We have for $i = 1, 2, \ldots, N$,

$$\partial G_i / dr_j = \begin{cases} -\frac{1}{N}\lambda(g_0(r_j) + r_j g_0'(r_j)), & j \neq i \\ -\frac{1}{N}\lambda(g_0(r_i) + r_i g_0'(r_i)) + \lambda\left(\phi''(1) + g_0(r_i)\right) + \lambda r_i g_0'(r_i), & \text{otherwise,} \end{cases}$$

and

$$\partial G_i / d\lambda = \left(v(\xi_i) - \bar{v}_0(S)\right)^\top x\big) / \lambda,$$

using the fact that $G_i = 0$, $i = 1, 2, ..., N$.

Suppose that J does not have full rank so there is $w \neq 0$, with $Jw = 0$. Then Jw can be written

$$\begin{aligned} (Jw)_i &= -\frac{\lambda}{N}\sum_{j=1}^{N} w_j(r_j g_0'(r_j) + g_0(r_j)) \\ &\quad + \lambda w_i(\phi''(1) + g_0(r_i) + r_i g_0'(r_i)) \\ &\quad + \frac{w_{N+1}}{\lambda}\left(v(\xi_i) - \bar{v}_0(S)\right)^\top x\big) \\ (Jw)_{N+1} &= \sum_{j=1}^{N} w_j r_j\left(\phi''(1) + g_0(r_j)\right). \end{aligned}$$

We may sum the first N equations to obtain (from the definition of $\bar{v}_0(S)$)

$$-\lambda\sum_{j=1}^{N} w_j(r_j g_0'(r_j) + g_0(r_j)) + \lambda\sum_{i=1}^{N} w_i(\phi''(1) + g_0(r_i) + r_i g_0'(r_i)) = 0$$

which simplifies to

$$\sum_{i=1}^{N} w_i = 0. \tag{26}$$

Now consider the behavior of $g_0(r_i) + r_i g_0'(r_i)$. As $\delta \to 0$ this also approaches zero. We can write $(Jw)_i = 0$ as

$$w_i(\phi''(1) + g_0(r_i) + r_i g_0'(r_i)) = K_0 - \frac{w_{N+1}}{\lambda^2}(v(\xi_i) - \bar{v}_0(S))^\top x$$

where $K_0 = \frac{1}{N}\sum_{j=1}^{N} w_j(r_j g_0'(r_j) + g_0(r_j))$. And hence for δ small enough we have w_i approximately equal to $\frac{K_0}{\phi''(1)} - \frac{w_{N+1}}{\lambda^2\phi''(1)}(v(\xi_i) - \bar{v}_0(S))^\top x$.

We start by considering the case where $w_{N+1} \neq 0$. By considering $-w$ if necessary we can assume that $w_{N+1} < 0$. Then for small δ, w_i is approximately proportional to $(v(\xi_i) - \bar{v}_0(S))^\top x$, so we have established that for small δ, $v(\xi_i)^\top x < v(\xi_j)^\top x$ will imply $w_i < w_j$.

From equations (24)

$$\begin{aligned} & r_i(\delta, x)\left(\phi''(1) + g_0(r_i(\delta, x))\right) - \frac{1}{N}\sum_j r_j(\delta, x) g_0(r_j(\delta, x)) \\ = \quad & \frac{1}{\lambda(\delta, x)}(v(\xi_i) - \bar{v}_0(S))^\top x. \end{aligned} \tag{27}$$

From (26) and (27) we deduce

$$\begin{aligned} & \sum_i w_i r_i(\delta, x)\left(\phi''(1) + g_0(r_i(\delta, x))\right) - \sum_i \frac{w_i}{N}\sum_j r_j(\delta, x) g_0(r_j(\delta, x)) \\ = \quad & \sum_i \frac{w_i}{\lambda(\delta, x)}(v(\xi_i) - \bar{v}_0(S))^\top x, \end{aligned}$$

whence

$$\sum w_i r_i\left(\phi''(1) + g_0(r_i)\right) = \sum \frac{w_i}{\lambda(\delta, x)} v(\xi_i)^\top x.$$

Now $(Jw)_{N+1} = 0$ implies $\sum_i w_i v(\xi_i)^\top x = 0$. Since $\sum w_i = 0$ some of the w_i values are negative, and we may choose $\widehat{v}$ so that $w_i \leq 0$ for $v(\xi_i)^\top x \leq \widehat{v}$ and $w_i > 0$ for $v(\xi_i)^\top x > \widehat{v}$. Then since $\sum \widehat{v} w_j = 0$ we have a contradiction

$$0 = \sum_i w_i v(\xi_i)^\top x = \sum_i w_i (v(\xi_i)^\top x - \widehat{v}) > 0.$$

where the inequality arises because each term in the sum is non-negative and they cannot all be zero unless $v(\xi_i)$ are all the same or $w_i = 0$.

Next we consider the case where $w_{N+1} = 0$. Then $(Jw)_i = 0$ implies

$$w_i(\phi''(1) + g_0(r_i) + r_i g_0'(r_i)) = \frac{1}{N}\sum_{j=1}^N w_j(r_j g_0'(r_j) + g_0(r_j)).$$

Unless $w = 0$, we may scale the w_j values so that the right hand side is 1. But then $w_i = \frac{1}{\phi''(1) + g_0(r_i) + r_i g_0'(r_i)} > 0$ for δ chosen small enough, which contradicts (26).

Hence we have established that J has full rank for δ chosen small enough. Hence the analytic implicit function theorem implies that $\lambda(\delta, x), r_i(\delta, x)$ exist as analytic functions of x and δ in a neighbourhood of $(x_0(S), 0)$. $\square$

Lemma 20 *Suppose* $k = \sqrt{\frac{2}{\phi''(1)}}$. *Then*

$$r_i(\delta, x) = \delta k \frac{(v(\xi_i) - \bar{v}_0(S))^\top x}{(x^\top V(S) x)^{1/2}} + \delta^2 h_i(\delta, x) \tag{28}$$

where $h_i(\delta, x)$ *is bounded.*

Proof Rearranging (24) gives

$$r_i(\delta, x) = \frac{1}{\lambda(\delta, x)} \frac{v(\xi_i)^\top x - \bar{v}_0(S)^\top x}{\phi''(1) + g_0(r_i(\delta, x))} + \frac{\sum_j r_j(\delta, x) g_0(r_j(\delta, x))}{N(\phi''(1) + g_0(r_i(\delta, x)))}. \tag{29}$$

Let $\sigma(\delta, x) = r(\delta, x)\lambda(\delta, x)$, $\eta_i(\delta, x) = \frac{v(\xi_i)^\top x - \bar{v}_0(S)^\top x}{\phi''(1) + g_0(r_i(\delta, x))}$, and $G_{ij}(\delta, x) = \frac{g_0(r_j(\delta, x))}{N(\phi''(1) + g_0(r_i(\delta, x)))}$.

Then

$$\sigma_i(\delta, x) = \eta_i(\delta, x) + \sum_j G_{ij}(\delta, x) \sigma_j(\delta, x),$$

and $\lim_{\delta \to 0} G_{ij}(\delta, x) = 0$. In matrix form we obtain

$$\sigma(\delta, x) = (I - G(\delta, x))^{-1} \eta(\delta, x),$$

whence taking limits as $\delta \to 0$ yields

$$\lim_{\delta \to 0} r_i(\delta, x)\lambda(\delta, x) = \lim_{\delta \to 0} \eta_i(\delta, x) = \frac{\left(v(\xi_i)^\top x - \bar{v}_0(S)^\top x\right)}{\phi''(1)}. \tag{30}$$

Multiplying (25) by $\lambda(\delta, x)^2$ gives

$$\frac{1}{2N} \sum_{i=1}^{N} \left(\lambda(\delta, x) r_i(\delta, x)\right)^2 \left(\phi''(1) + g(r_i(\delta, x))\right) = \delta^2 \lambda(\delta, x)^2$$

which gives

$$\begin{aligned} \lim_{\delta \to 0} \delta^2 \lambda(\delta, x)^2 &= \frac{\phi''(1)}{2N} \sum_{i=1}^{N} \left(\frac{v(\xi_i)^\top x - \bar{v}_0(S)^\top x}{\phi''(1)} \right)^2 \\ &= \frac{x^\top V(S) x}{2\phi''(1)}, \end{aligned}$$

where

$$V(S) = \frac{1}{N} \sum_{i=1}^{N} \left(v(\xi_i) - \bar{v}_0(S)\right) \left(v(\xi_i) - \bar{v}_0(S)\right)^\top.$$

Thus

$$\lim_{\delta\to 0}\delta\lambda(\delta,x)=\left(\frac{x^\top V(S)x}{2\phi''(1)}\right)^{1/2}. \tag{31}$$

Since for almost all S we will have $\frac{x^\top V(S)x}{2\phi''(1)}>0$, (30) and (31) give

$$\lim_{\delta\to 0}\frac{r_i(\delta,x)}{\delta}=\frac{\left(v(\xi_i)^\top x-\bar{v}_0(S)^\top x\right)}{\phi''(1)}\left(\frac{x^\top V(S)x}{2\phi''(1)}\right)^{-1/2}$$

whereby applying Lemma 19 gives

$$r_i(\delta,x)=\delta k\frac{(v(\xi_i)-\bar{v}_0(S))^\top x}{(x^\top V(S)x)^{1/2}}+\delta^2 h_i(\delta,x)$$

for some bounded $h_i(\delta,x)$ as required. □

We now proceed to prove Proposition 9 The objective of the robust optimization DRQP is

$$\begin{aligned}&\frac{1}{2}x^\top Hx+\frac{1}{N}\sum_{i=1}^N(1+r_i(\delta,x))v(\xi_i)^\top x\\&=\frac{1}{2}x^\top Hx+\bar{v}_0(S)^\top x+\frac{1}{N}\sum_{i=1}^N r_i(\delta,x)v(\xi_i)^\top x.\end{aligned}$$

The first order conditions determining $x_\delta(S)$ are

$$Hx_\delta(S)+\bar{v}_0(S)+\nabla_x\left(\frac{1}{N}\sum_{i=1}^N r_i(\delta,x)v(\xi_i)^\top x\right)=0.$$

Taking derivatives with respect to δ we obtain

$$H\frac{d}{d\delta}x_\delta(S)+\frac{d}{d\delta}\nabla_x\left(\frac{1}{N}\sum_{i=1}^N r_i(\delta,x)v(\xi_i)^\top x\right)=0.$$

Now Lemma 20 gives

$$\begin{aligned}&\frac{\partial}{\partial x_j}\left(\frac{1}{N}\sum_{i=1}^N r_i(\delta,x)v(\xi_i)^\top x\right)\\=\;&\frac{\delta}{N}\left(\sum_{i=1}^N k\frac{\partial}{\partial x_j}\left(\frac{(v(\xi_i)-\bar{v}_0(S))^\top x}{(x^\top V(S)x)^{1/2}}\right)+\delta\frac{\partial}{\partial x_j}h_i(\delta,x)\right)v(\xi_i)^\top x\\&+\frac{\delta}{N}\sum_{i=1}^N\left(k\frac{(v(\xi_i)-\bar{v}_0(S))^\top x}{(x^\top V(S)x)^{1/2}}+\delta h_i(\delta,x)\right)v_j(\xi_i),\end{aligned}$$

and by Lemma 19 $h_i(\delta, x)$ and $\frac{\partial}{\partial x_j} h_i(\delta, x)$ are bounded as $\delta \to 0$, so it follows that

$$\begin{aligned}
& \lim_{\delta \to 0} \frac{d}{d\delta} \frac{\partial}{\partial x_j} \left(\frac{1}{N} \sum_{i=1}^{N} r_i(\delta, x) v(\xi_i)^\top x \right) \\
= \; & \frac{1}{N} \sum_{i=1}^{N} k \frac{\partial}{\partial x_j} \left(\frac{(v(\xi_i) - \bar{v}_0(S))^\top x}{(x^\top V(S) x)^{1/2}} \right) v(\xi_i)^\top x \\
& + \frac{1}{N} \sum_{i=1}^{N} \left(k \frac{(v(\xi_i) - \bar{v}_0(S))^\top x}{(x^\top V(S) x)^{1/2}} \right) v_j(\xi_i)
\end{aligned}$$

giving

$$\left[\frac{d}{d\delta} x_\delta(S) \right]_{\delta=0} = H^{-1} \zeta(x_0(S)),$$

where

$$\begin{aligned}
\zeta_j(x) \; = \; & \frac{k}{N} \sum_{i=1}^{N} \frac{\partial}{\partial x_j} \left(\frac{(v(\xi_i) - \bar{v}_0(S))^\top x}{(x^\top V(S) x)^{1/2}} \right) v(\xi_i)^\top x \\
& + \frac{k}{N} \sum_{i=1}^{N} \left(\frac{(v(\xi_i) - \bar{v}_0(S))^\top x}{(x^\top V(S) x)^{1/2}} \right) v_j(\xi_i).
\end{aligned}$$

We can simplify $\zeta_j(x)$ by noting

$$\nabla \frac{(v(\xi_i) - \bar{v}_0(S))^\top x}{(x^\top V(S) x)^{1/2}} = \frac{(v(\xi_i) - \bar{v}_0(S))}{(x^\top V(S) x)^{1/2}} - \frac{(v(\xi_i) - \bar{v}_0(S))^\top x}{(x^\top V(S) x)^{3/2}} V(S) x.$$

So at $\delta = 0$ we have

$$\begin{aligned}
\zeta(x) \; = \; & \frac{k}{N} \frac{1}{(x^\top V(S) x)^{1/2}} \sum_{i=1}^{N} \left(\left(v(\xi_i)^\top x \right) \left(v(\xi_i) - \bar{v}_0(S) \right) \right) \\
& - \frac{k}{N} \frac{1}{(x^\top V(S) x)^{1/2}} \sum_{i=1}^{N} \left(v(\xi_i)^\top x \frac{(v(\xi_i) - \bar{v}_0(S))^\top x}{x^\top V(S) x} V(S) x \right) \\
& + \frac{k}{N} \frac{1}{(x^\top V(S) x)^{1/2}} \sum_{i=1}^{N} v(\xi_i) (v(\xi_i) - \bar{v}_0(S))^\top x.
\end{aligned}$$

However from the definition of $V(S)$ we have

$$\frac{1}{N}\sum_{i=1}^{N}(v(\xi_i)^\top x)(v(\xi_i)-\bar{v}_0(S))^\top x$$

$$= \quad x^\top V(S)x + \frac{1}{N}\sum_{i=1}^{N} x^\top \bar{v}_0(S)\left(v(\xi_i)-\bar{v}_0(S)\right)^\top x = x^\top V(S)x.$$

So the term in the sum involving $x^\top V(S)x$ simplifies and we get

$$\zeta(x) = \frac{k}{(x^\top V(S)x)^{1/2}}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\left(v(\xi_i)^\top x\right)\left(v(\xi_i)-\bar{v}_0(S)\right)+v(\xi_i)(v(\xi_i)-\bar{v}_0(S))^\top x\right)\right)$$
$$- V(S)x$$
$$= \frac{k}{(x^\top V(S)x)^{1/2}}\frac{1}{N}\sum_{i=1}^{N} v(\xi_i)(v(\xi_i)-\bar{v}_0(S))^\top x,$$

where we have used the fact that $\frac{1}{N}\sum_{i=1}^{N}\left(v(\xi_i)-\bar{v}_0(S)\right)\left(\bar{v}_0(S)^\top x\right)=0$ and hence

$$V(S)x = \frac{1}{N}\sum_{i=1}^{N}\left(v(\xi_i)-\bar{v}_0(S)\right)v(\xi_i)^\top x.$$

Thus

$$\frac{d}{d\delta}x_\delta(S) = -H^{-1}\frac{k}{(x^\top V(S)x)^{1/2}}\frac{1}{N}\sum_{i=1}^{N} v(\xi_i)(v(\xi_i)-\bar{v}_0(S))^\top x.$$
$$= -H^{-1}\frac{k}{(x^\top V(S)x)^{1/2}}V(S)x$$

as $V(S)$ is symmetric. Thus in the limit as $\delta \to 0$, we get $x_\delta(S) \to -H^{-1}\bar{v}_0(S)$, and we obtain

$$\bar{y}(S) = k\frac{H^{-1}V(S)H^{-1}\bar{v}_0(S)}{\left(\bar{v}_0(S)^\top H^{-1}V(S)H^{-1}\bar{v}_0(S)\right)^{\frac{1}{2}}}$$

as required. □

Proof of Proposition 10

Immediately from the improvement lemma and Proposition 9 we know that robustification with smooth ϕ divergence incrementally improves SAA if

$$\mathbb{E}_S\left[\left(\bar{v}_0(S)-\bar{v}\right)^\top H^{-1}\frac{kV(S)H^{-1}\bar{v}_0(S)}{\left(\bar{v}_0^\top(S)H^{-1}V(S)H^{-1}\bar{v}_0(S)\right)^{\frac{1}{2}}}\right] > 0.$$

Since k is a positive constant the result follows immediately. □

Proof of Lemma 11

The translation equivariance of ρ yields

$$\rho[c(x,S)]=\frac{1}{2}x^\top Hx+(1-\delta)\frac{1}{N}\sum_{i=1}^{N}\left(v(\xi_i)^\top x\right)+\delta\text{CVaR}_{1-\alpha}[\{v(\xi_i)^\top x\}].$$

The first order conditions for RSAA(δ) become

$$0\in\partial\rho(c(x,S))=Hx+(1-\delta)\bar{v}_0+\delta G_{\text{CVaR}} \tag{32}$$

which gives (13) and (14) when the subgradient at the optimal solution is unique. □

Proof of Proposition 12

Consider all samples S satisfying (15). Suppose that $\alpha\in(\frac{m-1}{N},\frac{m}{N}]$ for some integer m. For a given x, we suppose that

$$v(\xi_1)^\top x\geq v(\xi_2)^\top x\geq ..v(\xi_k)^\top x=v(\xi_{k+1})^\top x=...=v(\xi_\ell)^\top x$$

with $v(\xi_\ell)^\top x>v(\xi_j)^\top x$, for all $j>\ell$, and $k\leq m\leq\ell$. When $k\neq\ell$ we have non-differentiability of CVaR at x and the subdifferential $\partial\text{CVaR}_{1-\alpha}\left[\{v(\xi_i)^\top x\}\right]$ is the set

$$G_{\text{CVaR}}(x)=\frac{1}{\alpha N}\sum_{i=1}^{k-1}v(\xi_i)+(1-\frac{k-1}{\alpha N})\text{conv}\{v(\xi_k),v(\xi_{k+1}),...,v(\xi_\ell)\}.$$

By Lemma 11

$$x_\delta(S)\in-H^{-1}\left((1-\delta)\bar{v}_0(S)+\delta G_{\text{CVaR}}(x_\delta(S))\right)$$

and since $G_{\text{CVaR}}(x_\delta(S)$ is a bounded set, we have $x_\delta(S)\to x_0(S)$ as $\delta\to 0$. Thus for all δ sufficiently small we must have

$$v(\xi_1)^\top x_\delta(S)<v(\xi_2)^\top x_\delta(S)<...<v(\xi_N)^\top x_\delta(S)$$

so $\text{CVaR}_{1-\alpha}\left[\{v(\xi_i)^\top x\}\right]$ is differentiable at $x_\delta(S)$, with derivative

$$\bar{v}_{\text{CVaR}}(S)=\frac{1}{\alpha N}\sum_{i=1}^{k-1}v(\xi_i)+(1-\frac{k-1}{\alpha N})v(\xi_m).$$

Lemma 11 then gives

$$x_\delta(S)=-H^{-1}\left((1-\delta)\bar{v}_0(S)+\delta\bar{v}_{\text{CVaR}}(S)\right)$$

and

$$x_\delta(S) - x_0(S) = -H^{-1}(\bar{v}_{\mathrm{CVaR}}(S) - \bar{v}_0(S))\delta,$$

whence DRQP has linear variation with

$$\bar{y}(S) = -H^{-1}(\bar{v}_{\mathrm{CVaR}}(S) - \bar{v}_0(S)).$$

If we write R for $(\bar{v}_{\mathrm{CVaR}}(S) - \bar{v}_0(S))$, then

$$\begin{aligned} C_\delta(S) &= \mathbb{E}_{\mathbb{P}}[c(x_\delta(S), \xi)] \\ &= \frac{1}{2}\left(x_0(S) - \delta H^{-1}R\right)^\top H\left(x_0(S) - \delta H^{-1}R\right) + \bar{v}^\top\left(x_0(S) - \delta H^{-1}R\right) \\ &= C_0(S) - \delta x_0(S)^\top R + \frac{\delta^2}{2}R^\top H^{-1}R - \delta\bar{v}^\top H^{-1}R \\ &= C_0(S) - \delta(\bar{v} - \bar{v}_0(S))^\top H^{-1}R + \frac{\delta^2}{2}R^\top H^{-1}R. \end{aligned}$$

Thus we obtain

$$\begin{aligned} \mathrm{VRS}(\delta) &= \mathbb{E}_S[\delta(\bar{v} - \bar{v}_0(S))^\top H^{-1}(\bar{v}_{\mathrm{CVaR}}(S) - \bar{v}_0(S)) \\ &\quad -\frac{\delta^2}{2}(\bar{v}_{\mathrm{CVaR}}(S) - \bar{v}_0(S))^\top H^{-1}(\bar{v}_{\mathrm{CVaR}}(S) - \bar{v}_0(S))], \end{aligned}$$

and

$$\mathrm{MVRS} = \mathbb{E}_S[(\bar{v} - \bar{v}_0)^\top H^{-1}(\bar{v}_{\mathrm{CVaR}}(S) - \bar{v}_0(S))]$$

as required. □

Proof of Proposition 13

We apply Proposition 12 with $H = 1$ and $v(\xi) = -g(\xi)$, so $\bar{v}_0(S) = -\bar{g}_0(S)$. Now $\bar{v}_{\mathrm{CVaR}}(S)$ is the derivative of $\mathrm{CVaR}_{1-\alpha}[\{v(\xi_i)^\top x\}]$ evaluated at $x_0(S) = \bar{g}_0(S)$. Thus

$$\bar{v}_{\mathrm{CVaR}}(S) = \mathrm{CVaR}_{1-\alpha}[\{-\mathrm{sgn}(\bar{g}_0(S))g(\xi_i)\}].$$

As in Proposition 12 there is a need for care when $x_0(S) = 0$ since at that point we have $\mathrm{CVaR}_{1-\alpha}[\{v(\xi_i)^\top x\}]$ non differentiable. The formulation here makes $\bar{v}_{\mathrm{CVaR}}(S) = 0$ in this case. But since the proposition statement involves an expectation over a continuous distribution we can see that $x_0(S) = 0$ with probability zero and our definition at this point will have no impact.

We obtain for all $\delta > 0$ sufficiently small

$$\begin{aligned} x_\delta(S) &= x_0(S) - \delta H^{-1}(\bar{v}_{\mathrm{CVaR}}(S) - \bar{v}_0) \\ &= x_0(S) - (\delta/2)(\mathrm{CVaR}_{1-\alpha}[\{-\mathrm{sgn}(\bar{g}_0(S))g(\xi_i)\}] + \bar{g}_0(S)) \\ &= x_0(S) - \frac{\delta}{2}(\bar{g}_0(S) + \mathrm{CVaR}_{1-\alpha}[\{-\mathrm{sgn}(\bar{g}_0(S))g(\xi_i)\}]) \end{aligned}$$

and

$$\begin{aligned}
\text{MVRS} &= \mathbb{E}_S[(\bar{v} - \bar{v}_0(S))^\top H^{-1}(\bar{v}_{\text{CVaR}}(S) - \bar{v}_0(S))] \\
&= \mathbb{E}_S[(-\bar{g} + \bar{g}_0(S))(\text{CVaR}_{1-\alpha}[\{-\text{sgn}(\bar{g}_0(S))g(\xi_i)\}] + \bar{g}_0(S))] \\
&= \mathbb{E}_S[(\bar{g}_0(S) - \bar{g})(\bar{g}_0(S) + \text{CVaR}_{1-\alpha}[\{-\text{sgn}(\bar{g}_0(S))g(\xi_i)\}])] \\
&= \frac{\sigma^2}{N} + \mathbb{E}_S[(\bar{g}_0(S) - \bar{g})\text{CVaR}_{1-\alpha}[\{-\text{sgn}(\bar{g}_0(S))g(\xi_i)\}],
\end{aligned}$$

as required. □

Proof of Proposition 14

We will use Proposition 13 and show that

$$-\mathbb{E}_S[(\bar{g}_0(S) - \bar{g})\text{CVaR}_{1-\alpha}[\{-g(\xi_i)\}]] = \int_{-\infty}^{\infty} Q(z)(1 - F(z))^{N-1}\Lambda_\alpha(z)dz. \tag{33}$$

First observe that

$$\mathbb{E}_S[(\bar{g}_0(S) - \bar{g})\bar{g}] = 0,$$

so

$$\begin{aligned}
-\mathbb{E}_S[(\bar{g}_0(S) - \bar{g})\text{CVaR}_{1-\alpha}[\{-g(\xi_i)\}]] &= -\mathbb{E}_S[(\bar{g}_0(S) - \bar{g})(\text{CVaR}_{1-\alpha}[\{-g(\xi_i)\}] + \bar{g})]] \\
&= -\mathbb{E}_S[\bar{w}\text{CVaR}_{1-\alpha}[\{-w_i\}]].
\end{aligned}$$

We have that $-\text{CVaR}_{1-\alpha}[\{-w_i\}]$ assigns probability 1 to the lowest $100\alpha\%$ outcomes of w_i, and takes the expectation. Thus, if $\alpha \in (\frac{m_\alpha}{N}, \frac{m_\alpha+1}{N}]$ then

$$-\text{CVaR}_{1-\alpha}[\{-w_i\}] = \frac{1}{\alpha N}z_1 + \frac{1}{\alpha N}z_2 + \ldots + (1 - \frac{m_\alpha}{\alpha N})z_m,$$

so

$$-\mathbb{E}_S\left[\bar{w}\text{CVaR}_{1-\alpha}[\{-w_i\}]\right] = \frac{1}{\alpha N}\mathbb{E}_S[\bar{w}z_1] + \frac{1}{\alpha N}\mathbb{E}_S[\bar{w}z_2] + \ldots + (1 - \frac{m_\alpha}{\alpha N})\mathbb{E}_S[\bar{w}z_m].$$

Since Lemma 26 gives

$$\mathbb{E}[\bar{w}z_j] = \frac{(N-1)!}{(N-j)!(j-1)!}\int_{-\infty}^{\infty} Q(z)(1 - F(z))^{N-j}F(z)^{j-1}dz$$

and

$$\begin{aligned}
\Lambda_\alpha(z) &= \frac{1}{\alpha N} + \frac{1}{\alpha N}(N-1)\frac{F(z)}{(1 - F(z))} \\
&\quad + \frac{1}{\alpha N}\frac{(N-1)(N-2)}{2}\frac{F(z)^2}{(1 - F(z))^2} \\
&\quad + \ldots + (1 - \frac{m_\alpha}{\alpha N})\begin{pmatrix} N-1 \\ m_\alpha \end{pmatrix}\frac{F(z)^{m_\alpha}}{(1 - F(z))^{m_\alpha}},
\end{aligned}$$

where m_α is the unique integer for which $\alpha \in (\frac{m_\alpha}{N}, \frac{m_\alpha+1}{N}]$, the identity (33) now follows. □

Proof of Proposition 16

We suppose that $c(x,\xi)$ is strictly concave in ξ. We first observe by Proposition 15 that this is enough to show that the solution to $\bar{\mathrm{P}}$ has each v_i supported on a single point (if v_i has weight p on z_{i1} and $(1-p)$ on z_{i2} then setting v_i to have weight 1 on $pz_{i1} + (1-p)z_{i2}$ increases the objective of $\bar{\mathrm{P}}$ and still satisfies the constraint). Thus $\bar{\mathrm{P}}$ becomes

$$\begin{array}{ll} \text{P1: } \max_{z_i} & \sum_{i=1}^N c_x(z_i) \\ \text{subject to} & \frac{1}{N}\sum_{i=1}^N \|z_i - \xi_i\| \le \delta. \end{array}$$

The Lagrangian of P1 is

$$\mathcal{L} = \sum_{i=1}^N \left(c_x(z_i) - \lambda \|z_i - \xi_i\|\right) + \lambda\delta$$

which is maximized at z_i. So

$$\nabla c_x(z_i) - \lambda \frac{z_i - \xi_i}{\|z_i - \xi_i\|} = 0$$

if $z_i \neq \xi_i$. This establishes (a) where $\alpha_i = \frac{\lambda}{\|z_i - \xi_i\|}$. To establish (b), notice that $\|\nabla c_x(z_i)\| = \lambda$ and so has the same value for each i where $z_i \neq \xi_i$.

In the case that $z_k = \xi_k$ we must have $\mathcal{L}$ is not increased when $z_k = \xi_k + \varepsilon \nabla c_x(\xi_k)$ for small $\varepsilon > 0$. Thus

$$\varepsilon \|\nabla c_x(\xi_k)\|^2 - \lambda\varepsilon \|\nabla c_x(\xi_k)\| \le 0,$$

giving $\|\nabla c_x(\xi_k)\| \le \lambda$. And hence for any choice of z_i with $z_i \neq \xi_i$, $\|\nabla c_x(\xi_k)\| \le \|\nabla c_x(z_i)\|$, as required. □

Lemma 21 *Let J_v be the $n \times m$ Jacobian matrix for $v(z)$ evaluated at some sample point z^*, and α a scalar constant. Then*

$$\frac{\partial}{\partial x_j}\left(v(z^* + \alpha \frac{J_v^\top x}{\|J_v^\top x\|})^\top x\right) = v_j(z^* + \alpha \frac{J_v^\top x}{\|J_v^\top x\|}).$$

Proof of Lemma 21

$$\begin{aligned} \frac{\partial}{dx_j} v_i(z^* + \alpha \frac{J_v^\top x}{\|J_v^\top x\|}) &= \sum_{k=1}^m \frac{\partial v_i}{dz_k}\frac{\partial}{dx_j}(z^* + \alpha \frac{J_v^\top x}{\|J_v^\top x\|})_k \\ &= \alpha \sum_{k=1}^m (J_v)_{ik} \frac{\partial}{dx_j}\frac{(J_v^\top x)_k}{\|J_v^\top x\|}. \end{aligned}$$

Now

$$\begin{aligned}
\frac{\partial}{dx_j}\frac{(J_v^\top x)_k}{\|J_v^\top x\|} &= \frac{\partial}{dx_j}\frac{(J_v^\top x)_k}{(x^\top J_v J_v^\top x)^{1/2}} = \frac{\partial}{dx_j}\frac{\sum_i x_i (J_v)_{ik}}{(x^\top J_v J_v^\top x)^{1/2}} \\
&= \frac{1}{\|J_v^\top x\|}\frac{\partial}{dx_j}\sum_i x_i (J_v)_{ik} + \sum_i x_i (J_v)_{ik}\frac{\partial}{dx_j}\frac{1}{(x^\top J_v J_v^\top x)^{1/2}} \\
&= \frac{1}{\|J_v^\top x\|}\left((J_v)_{jk} - \sum_i x_i (J_v)_{ik}\frac{\left(J_v J_v^\top x\right)_j}{(x^\top J_v J_v^\top x)}\right) \\
&= \frac{1}{\|J_v^\top x\|}\left((J_v)_{jk} - (J_v^\top x)_k\frac{\left(J_v J_v^\top x\right)_j}{(x^\top J_v J_v^\top x)}\right).
\end{aligned}$$

So

$$\begin{aligned}
\frac{\partial}{dx_j} v_i(z^* + \alpha\frac{J_v^\top x}{\|J_v^\top x\|}) &= \frac{\alpha}{\|J_v^\top x\|}\sum_{k=1}^m (J_v)_{ik}\left((J_v)_{jk} - (J_v^\top x)_k\frac{\left(J_v J_v^\top x\right)_j}{(x^\top J_v J_v^\top x)}\right) \\
&= \frac{\alpha}{\|J_v^\top x\|}\left(\left(J_v J_v^\top\right)_{ij} - (J_v J_v^\top x)_i\frac{\left(J_v J_v^\top x\right)_j}{(x^\top J_v J_v^\top x)}\right).
\end{aligned}$$

Hence

$$\begin{aligned}
\sum_j x_j\frac{\partial}{dx_j} v_i(z^* + \alpha\frac{J_v^\top x}{\|J_v^\top x\|}) &= \sum_j x_j\frac{\alpha}{\|J_v^\top x\|}\left(\left(J_v J_v^\top\right)_{ij} - (J_v J_v^\top x)_j\frac{\left(J_v J_v^\top x\right)_i}{(x^\top J_v J_v^\top x)}\right) \\
&= \frac{\alpha}{\|J_v^\top x\|}\left(\left(J_v J_v^\top x\right)_i - (x^\top J_v J_v^\top x)\frac{\left(J_v J_v^\top x\right)_i}{(x^\top J_v J_v^\top x)}\right) = 0,
\end{aligned}$$

which yields the result. □

Proof of Proposition 17

Recall

$$\text{DRQP:} \quad \min_{x\in X}\left(\tfrac{1}{2}x^\top Hx + \sup_{\mathbb{Q}\in\mathcal{P}_\delta}\mathbb{E}_{\mathbb{Q}}\left[v^\top x\right]\right)$$

so

$$\nabla c_x(z_i) = \sum_j x_j \nabla v_j(z_i).$$

We require the linear variation property for almost all samples S. Since we assume strict concavity for v_j we know that ∇v_j takes a range of values and almost everywhere the sample $S = \{\xi_1, \xi_2, ..., \xi_N)$ has each point with a

different value for $\left\|\sum_j x_j \nabla v_j(\xi_i)\right\|$, so we can make this assumption. Then for small δ we will move just one point. We deduce this from Proposition 16 part (b), since for small δ it is impossible for two different points to end up with the same value for $\|\nabla c_x(z_i)\|$ without moving a combined distance more than δ. Moreover part (c) of Proposition 16 shows that the point that is moved is $\xi^*(S) = \xi_{i_0}$, the sample point with the highest gradient norm for the cost function, so $i_0 = \arg\max_i \left\|\sum_j x_j \nabla v_j(\xi_i)\right\|$ (which is well-defined under our assumption). For brevity we write ξ^* for $\xi^*(S)$. The solution we obtain after robustification, given a distance limit δ, moves ξ^* to the point $\xi^* + N\delta \frac{\sum_j x_j \nabla v_j(\xi^*)}{\left\|\sum_j x_j \nabla v_j(\xi^*)\right\|}$, where the term $N\delta$ arises from the way that we define the Wasserstein distance, and the fact that we move in the z-gradient direction of the cost function $c_x(z)$ follows from part (a) of Proposition 16.

After the robustifying move, and substituting $J_v(\xi^*)^\top x$ for $\sum_j x_j \nabla v_j(\xi^*)$, the term $v_k(\xi^*)$ is replaced by

$$v_k\left(\xi^* + N\delta \frac{J_v(\xi^*)^\top x}{\|J_v(\xi^*)^\top x\|}\right).$$

The objective function of DRQP is therefore

$$\frac{1}{2}x^\top H x + \frac{1}{N} v\left(\xi^* + N\delta \frac{J_v(\xi^*)^\top x}{\|J_v(\xi^*)^\top x\|}\right)^\top x + \frac{1}{N}\sum_{j \neq i_0} v(\xi_j)^\top x.$$

The first order conditions determining $x_\delta(S)$ are hence

$$Hx + \frac{1}{N}\nabla_x \left(v(\xi^* + N\delta \frac{J_v(\xi^*)^\top x}{\|J_v(\xi^*)^\top x\|})^\top x\right) + \frac{1}{N}\sum_{j \neq i_0} v(\xi_j) = 0.$$

Now applying Lemma 21 with $\alpha = N\delta$, we get that $x_\delta(S)$ satisfies

$$Hx + \frac{1}{N} v(\xi^* + N\delta \frac{J_v(\xi^*)^\top x}{\|J_v(\xi^*)^\top x\|}) + \frac{1}{N}\sum_{j \neq i_0} v(\xi_j) = 0,$$

where

$$\frac{1}{N} v_k(\xi^* + N\delta \frac{J_v(\xi^*)^\top x}{\|J_v(\xi^*)^\top x\|}) = \frac{1}{N} v_k(\xi^*) + \frac{\delta}{\|J_v(\xi^*)^\top x\|} \nabla v_k(\xi^*)^\top J_v(\xi^*)^\top x + O(\delta^2).$$

So we have first order conditions

$$Hx + \delta \frac{J_v(\xi^*) J_v(\xi^*)^\top x}{\|J_v(\xi^*)^\top x\|} + \bar{v}_0(S) = O(\delta^2).$$

We have $x_0(S) = -H^{-1}\bar{v}_0(S)$, so

$$H(x - x_0(S)) = -\delta \frac{J_v(\xi^*)J_v(\xi^*)^\top x}{\|J_v(\xi^*)^\top x\|} + O(\delta^2),$$

giving $x = x_0(S) + O(\delta)$, whence

$$\frac{J_v(\xi^*)J_v(\xi^*)^\top x}{\|J_v(\xi^*)^\top x\|} = \frac{J_v(\xi^*)J_v(\xi^*)^\top x_0(S)}{\|J_v(\xi^*)^\top x_0(S)\|} + O(\delta)$$

and

$$H(x - x_0(S)) = -\delta \frac{J_v(\xi^*)J_v(\xi^*)^\top x_0(S)}{\|J_v(\xi^*)^\top x_0(S)\|} + O(\delta^2). \tag{34}$$

From (34) and its definition it follows that

$$\bar{y}(S) = -H^{-1} \frac{J_v(\xi^*)J_v(\xi^*)^\top x_0(S)}{\|J_v(\xi^*)^\top x_0(S)\|}.$$

Substituting for $x_0(S)$ gives the expression we require. □

Proof of Proposition 18

(a) In the univariate example $J_v(\xi^*) = -\nabla g(\xi^*(S))$ so from Proposition 17 we have

$$\bar{y}(S) = -\left\|\nabla g(\xi^*(s))\right\|$$

and $v(\xi) = -g(\xi)$, so Lemma 4 gives

$$\text{MVRS} = \mathbb{E}_S\left[(\bar{g}_0(S) - \bar{g})\left\|\nabla g(\xi^*(s))\right\|\right].$$

(b) In the case that $g(\xi) = \xi^2$ and ξ is non-negative then $\xi^*(S)$ is the largest ξ_i in S, which we write as the order statistic ξ_N. Then since $\nabla g(\xi) = 2\xi$ we have

$$\begin{aligned} \text{MVRS} &= \mathbb{E}_S\left[2((1/N)\sum_{i=1}^N \xi_i^2 - \mathbb{E}[\xi^2])\xi_N\right] \\ &= 2\mathbb{E}_S[\frac{\xi_N}{N}\sum_{i=1}^N \xi_i^2] - 2\mathbb{E}[\xi^2]\mathbb{E}[\xi_N]. \end{aligned}$$

Writing ξ_i for the order statistics we have, for $i < N$, (essentially this is the result of Lemma 23 with $j = N$)

$$\begin{aligned} &\mathbb{E}_S\left(\xi_N\xi_i^2\right) \\ = \;& \frac{N!}{(i-1)!(N-i-1)!}\int_0^\infty \int_{x_a}^\infty x_a^2 x_b F(x_a)^{i-1} f(x_a) f(x_b)\left(F(x_b) - F(x_a)\right)^{N-i-1} dx_b dx_a. \end{aligned}$$

But

$$\sum_{i=1}^{N-1} \frac{N!}{(i-1)!(N-i-1)!} F(x_a)^{i-1} \left(F(x_b) - F(x_a)\right)^{N-i-1} = N\left(N-1\right) F_b^{N-2},$$

so

$$\sum_{i=1}^{N-1} \mathbb{E}_S\left(\xi_N \xi_i^2\right) = N(N-1) \int_0^\infty \int_{x_a}^\infty x_a^2 x_b F(x_b)^{N-2} f(x_a) f(x_b) dx_b dx_a.$$

Now ξ_N has distribution $F(z)^N$ so has density $NF(z)^{N-1} f(z)$. Thus

$$\mathbb{E}_S\left(\xi_N\right) = N \int_0^\infty z F(z)^{N-1} f(z) dz,$$

$$\mathbb{E}_S\left(\xi_N^3\right) = N \int_0^\infty z^3 F(z)^{N-1} f(z) dz.$$

We have

$$\begin{aligned}
\text{MVRS} &= \frac{2}{N} \sum_{i=1}^{N-1} \mathbb{E}_S[\xi_N \xi_i^2] + \frac{2}{N} \mathbb{E}_S\left(\xi_N^3\right) - 2\mathbb{E}_S[\xi^2]\mathbb{E}_S[\xi_N] \\
&= 2(N-1) \int_0^\infty \left(\int_z^\infty u F(u)^{N-2} f(u) du \right) z^2 f(z) dz \\
&\quad + 2 \int_0^\infty z^3 F(z)^{N-1} f(z) dz \\
&\quad - 2N \left(\int_0^\infty u^2 f(u) du \right) \int_0^\infty z F(z)^{N-1} f(z) dz
\end{aligned}$$

as required. □

Appendix 2: Identities for order statistics

In this appendix we derive some identities for order statistics from samples of a random variable W with mean 0 and cumulative distribution function F and density f. We let

$$P_W(z) = \int_{-\infty}^z u f(u) du, \quad Q_W(z) = \int_z^\infty u f(u) du,$$

where we usually drop the explicit dependence on the distribution W. Thus $P(z) + Q(z) = 0$, and $P(\infty) = Q(-\infty) = 0$. Suppose $\{w_1, w_2, \ldots, w_N\}$ is a

random sample of W, with order statistics $z_1 \leq z_2 \leq \ldots \leq z_N$. The sample mean is $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$.

Lemma 22 $\mathbb{E}[z_i^2] = \frac{N!}{(N-i)!(i-1)!}\int_{-\infty}^{\infty} z^2 F(z)^{i-1} f(z)(1-F(z))^{N-i}dz.$

Proof. Consider the event $A_i = \{z_i \in (x_a, x_a+\varepsilon)\}$. Then

$$\begin{aligned}
\mathbb{P}(A_i) &= \mathbb{P}(z_i \in (x_a, x_a+\varepsilon)) \\
&= \mathbb{P}\left(\begin{array}{c} i-1 \text{ of the } w_i \text{ in } (-\infty, x_a), \\ \text{one } w_i \text{ in } (x_a, x_a+\varepsilon),\ N-i \text{ of } w_i > x_a+\varepsilon. \end{array}\right) \\
&= \frac{N(N-1)(N-2)\ldots(N-i+1)}{(i-1)!} \\
&\quad \times F(x_a)^{i-1}\left(F(x_a+\varepsilon)-F(x_a)\right)\left(1-F(x_a+\varepsilon)\right)^{N-i} \\
&= \frac{N!}{(N-i)!(i-1)!} f(x_a)F(x_a)^{i-1}(1-F(x_a))^{N-i}\varepsilon + o(\varepsilon).
\end{aligned}$$

Thus

$$\mathbb{E}[z_i^2] = \frac{N!}{(N-i)!i!}\int_{-\infty}^{\infty} z^2 F(z)^{i-1} f(z)(1-F(z))^{N-i}dz.$$

■

Lemma 23 *If $i < j$ then*

$$\mathbb{E}[z_i z_j] = \frac{N(N-1)...(N-j+1)}{(i-1)!(j-i-1)!}\int_{-\infty}^{\infty}\int_{x_a}^{\infty} x_a x_b B(x_a, x_b)dx_b dx_a \qquad (35)$$

where

$$B(x_a, x_b) = F(x_a)^{i-1} f(x_a) f(x_b)\left(F(x_b)-F(x_a)\right)^{j-i-1}\left(1-F(x_b)\right)^{N-j}.$$

Proof. The joint distribution of z_i and z_j can be expressed in terms of the event $A_{ij} = \{z_i \in (x_a, x_a+\varepsilon) \text{ and } z_j \in (x_b, x_b+\varepsilon')\}$.

$$\begin{aligned}
\mathbb{P}(A_{ij}) &= \mathbb{P}(z_i \in (x_a, x_a+\varepsilon) \text{ and } z_i \in (x_b, x_b+\varepsilon')) \\
&= \mathbb{P}\left(\begin{array}{c} (i-1)\ w_i \text{ in } (-\infty, x_a), \text{ one } w_i \text{ in } (x_a, x_a+\varepsilon), \\ j-i-1 \text{ of the } w_i \text{ in } (x_a+\varepsilon, x_b), \\ \text{one } w_i \text{ in } (x_b, x_b+\varepsilon'), \text{ rest of the } w_i > x_b+\varepsilon'. \end{array}\right) \\
&= \left(\begin{array}{c} N \\ i-1 \end{array}\right)(N-i+1)\left(\begin{array}{c} N-i \\ j-i-1 \end{array}\right)(N-j+1) \\
&\quad \times F(x_a)^{i-1}\left(F(x_a+\varepsilon)-F(x_a)\right)\left(F(x_b)-F(x_a+\varepsilon)\right)^{j-i-1} \\
&\quad \times \left(F(x_b+\varepsilon')-F(x_b)\right)\left(1-F(x_b+\varepsilon')\right)^{N-j}.
\end{aligned}$$

But

$$\begin{aligned} & \binom{N}{i-1}(N-i+1)\binom{N-i}{j-i-1}(N-j+1) \\ = \; & \frac{N(N-1)...(N-j+1)}{(i-1)!(j-i-1)!}, \end{aligned}$$

and

$$\begin{aligned} & F(x_a)^{i-1}\left(F(x_a+\varepsilon)-F(x_a)\right)\left(F(x_b)-F(x_a+\varepsilon)\right)^{j-i-1} \\ & \times\left(F(x_b+\varepsilon')-F(x_b)\right)\left(1-F(x_b+\varepsilon')\right)^{N-j} \\ = \; & B(x_a,x_b)\varepsilon\varepsilon'+o(\varepsilon\varepsilon'). \end{aligned}$$

Thus

$$\mathbb{E}[z_i z_j] = \frac{N(N-1)...(N-j+1)}{(i-1)!(j-i-1)!}\int_{-\infty}^{\infty}\int_{x_a}^{\infty} x_a x_b B(x_a,x_b)dx_b dx_a$$

as required. ■

Lemma 24

$$\sum_{j=i+1}^{N}\mathbb{E}[z_i z_j] = \frac{N!}{(i-1)!(N-i-1)!}\int_{-\infty}^{\infty} zF(z)^{i-1}(1-F(z))^{N-i-1}f(z)Q(z)dz. \tag{36}$$

Proof.

$$\begin{aligned} & \sum_{j=i+1}^{N}\frac{N(N-1)\ldots(N-j+1)}{(i-1)!(j-i-1)!}\left(F(x_b)-F(x_a)\right)^{j-i-1}\left(1-F(x_b)\right)^{N-j} \\ = \; & \frac{N(N-1)\ldots(N-i+1)\,(N-i)}{(i-1)!} \\ & \times\sum_{k=0}^{N-i-1}\frac{(N-i-1)...(N-i-k)}{k!}\left(F(x_b)-F(x_a)\right)^{k}\left(1-F(x_b)\right)^{N-1-i-k} \\ = \; & \frac{N!}{(i-1)!(N-i-1)!}(1-F(x_a))^{N-i-1}. \end{aligned}$$

Substituting in (35) and substituting for $Q(z)$ yields (36). ■

Lemma 25

$$\sum_{i=1}^{j-1} \mathbb{E}[z_i z_j] = \frac{N!}{(j-2)!(N-j)!} \int_{-\infty}^{\infty} P(z) z f(z) (1-F(z))^{N-j} F(z)^{j-2} dz.$$

Proof. Observe that $\sum_{i=1}^{j-1} \mathbb{E}[z_i z_j]$ is the same as $\sum_{i=N-j+2}^{N} \mathbb{E}[w_i w_{N-j+1}]$ when $w = -z$. Now using Lemma 24 we have

$$\sum_{i=N-j+2}^{N} \mathbb{E}[w_{N-j+1} w_i] = \frac{N!}{(N-j)!(j-2)!} \int_{-\infty}^{\infty} z F_W(z)^{N-j} (1-F_W(z))^{j-2} f_W(z) Q_W(z) dz$$

where we use a subscript W to show that the relevant quantity is with regard to w not z. Since $F_W(z) = 1 - F(-z)$ and $Q_W(z) = \int_z^{\infty} u f(-u) du$ we can change variables $v = -z$ and obtain

$$\sum_{i=N-j+2}^{N} \mathbb{E}[w_{N-j+1} w_i] = \frac{N!}{(N-j)!(j-2)!} \int_{-\infty}^{\infty} -v (1-F(v))^{N-j} F(v)^{j-2} f(v) \int_{-v}^{\infty} u f(-u) du dv.$$

Finally changing variables $t = -u$ gives $\int_{-v}^{\infty} u f(-u) du = \int_{v}^{-\infty} t f(t) dt = -P(v)$ and we recover the expression we require. ■

Lemma 26

$$\mathbb{E}[z_j \bar{z}] = \frac{(N-1)!}{(N-j)!(j-1)!} \int_{-\infty}^{\infty} Q(z) (1-F(z))^{N-j} F(z)^{j-1} dz. \tag{37}$$

Proof. Applying Lemmas 22, 24, and 25, we obtain

$$\begin{aligned}
\mathbb{E}[z_j \bar{z}] &= \frac{1}{N} \left(\sum_{i=1}^{j-1} \mathbb{E}[z_i z_j] + \mathbb{E}[z_j^2] + \sum_{i=j+1}^{N} \mathbb{E}[z_j z_i] \right) \\
&= \frac{(N-1)!}{(j-2)!(N-j)!} \int_{-\infty}^{\infty} P(z) z f(z) (1-F(z))^{N-j} F(z)^{j-2} dz \\
&\quad + \frac{(N-1)!}{(N-j)!(j-1)!} \int_{-\infty}^{\infty} z^2 F(z)^{j-1} f(z) (1-F(z))^{N-j} dz \\
&\quad + \frac{(N-1)!}{(j-1)!(N-j-1)!} \int_{-\infty}^{\infty} z F(z)^{j-1} (1-F(z))^{N-j-1} f(z) Q(z) dz \\
&= \frac{(N-1)!}{(N-j)!(j-1)!} A
\end{aligned}$$

where

$$\begin{aligned} A &= \int_{-\infty}^{\infty} (j-1)P(z)zf(z)(1-F(z))^{N-j}F(z)^{j-2}dz \\ &\quad + \int_{-\infty}^{\infty} z^2F(z)^{j-1}f(z)(1-F(z))^{N-j}dz \\ &\quad + \int_{-\infty}^{\infty} (N-j)zF(z)^{j-1}(1-F(z))^{N-j-1}f(z)Q(z)dz. \end{aligned}$$

Integrating the third term of A by parts gives

$$\begin{aligned} &\left[-(1-F(z))^{N-j}Q(z)zF(z)^{j-1}\right]_{-\infty}^{\infty} \\ &+ \int_{-\infty}^{\infty} (1-F(z))^{N-j}\frac{d}{dz}\left(Q(z)zF(z)^{j-1}\right)dz \\ = &\int_{-\infty}^{\infty} (1-F(z))^{N-j}Q(z)F(z)^{j-1}dz \\ &- \int_{-\infty}^{\infty} (1-F(z))^{N-j}\left[z^2f(z)F(z)^{j-1}\right]dz \\ &+ \int_{-\infty}^{\infty} (1-F(z))^{N-j}\left[Q(z)z(j-1)F(z)^{j-2}f(z)\right]dz \end{aligned}$$

which cancels with the first two terms of A (using the fact that $P(z)+Q(z) = 0$) to yield

$$A = \int_{-\infty}^{\infty} (1-F(z))^{N-j}Q(z)F(z)^{j-1}dz,$$

which demonstrates (37) as required. ■

Lemma 27 *Suppose z has density f and cumulative distribution function F. For all $\alpha \in (0,1]$,*

$$\int_{-\infty}^{\infty} f(z)(1-F(z))^{N-1}\Lambda_\alpha(z)dz = \frac{1}{N}. \tag{38}$$

Proof. First observe that if $\alpha < \frac{1}{N}$, then $\Lambda_\alpha(z) = 1$ and

$$\int_{-\infty}^{\infty} f(z)(1-F(z))^{N-1}dz = \left[-\frac{1}{N}(1-F(z))^N\right]_{-\infty}^{\infty} = \frac{1}{N}.$$

We next show (38) for every $\alpha = \frac{m}{N}$, $m = 1, 2, \ldots, N$. In this case

$$\Lambda_\alpha(z) = \frac{1}{m}\left(1 + (N-1)\frac{F(z)}{(1-F(z))} + \ldots + \left(\begin{array}{c} N-1 \\ m-1 \end{array}\right)\frac{F(z)^{m-1}}{(1-F(z))^{m-1}}\right).$$

Now

$$\begin{aligned}
&\int_{-\infty}^{\infty} f(z)(1-F(z))^{N-1} \left(\begin{array}{c} N-1 \\ m-1 \end{array} \right) \frac{F(z)^{m-1}}{(1-F(z))^{m-1}} dz \\
= \; & \frac{(N-1)!}{(N-m)!(m-1)!} \int_{-\infty}^{\infty} (1-F(z))^{N-m} F(z)^{m-1} f(z) dz \\
= \; & \frac{1}{N} \left(\int_0^1 \frac{N!}{(N-m)!(m-1)!} u^{m-1}(1-u)^{N-m} du \right) = \frac{1}{N}
\end{aligned}$$

where the final equality follows from observing that the integrand is the density of a beta distribution and hence integrates to 1.

So

$$\int_{-\infty}^{\infty} f(z)(1-F(z))^{N-1} \Lambda_\alpha(z) dz = \frac{1}{m}(\frac{1}{N} + \frac{1}{N} + \ldots + \frac{1}{N})$$

where the sum is over m terms. This yields the result for $\alpha = \frac{m}{N}$, $m = 1, 2, \ldots, N$.

Now suppose $\alpha \in (\frac{m}{N}, \frac{m+1}{N}]$, $m = 1, 2, \ldots, N-1$. Then

$$\Lambda_\alpha(z) = \Lambda_{\frac{m}{N}}(z) + (1 - \frac{m}{\alpha N}) \left(\begin{array}{c} N-1 \\ m \end{array} \right) \frac{F(z)^m}{(1-F(z))^m}$$

which is linear in $\frac{1}{\alpha} \in [\frac{N}{m+1}, \frac{N}{m})$, so $\int_{-\infty}^{\infty} f(z)(1-F(z))^{N-1} \Lambda_\alpha(z) dz$ is also linear in $\frac{1}{\alpha}$ in this range. Since we have established that

$$\int_{-\infty}^{\infty} f(z)(1-F(z))^{N-1} \Lambda_\alpha(z) dz = \frac{1}{N}$$

for $\alpha = \frac{m}{N}$ and $\alpha = \frac{m+1}{N}$, and for each z, $\Lambda_\alpha(z)$ is continuous at $\alpha = \frac{m}{N}$, the identity must hold throughout this range which gives the result. ■