Submitted to *INFORMS Journal on Optimization*
manuscript MS-0001-1922.65

Improving sample average approximation using distributional robustness

E.J. Anderson
University of Sydney and Imperial College, London, edward.anderson@sydney.edu.au

A.B. Philpott
University of Auckland, a.philpott@auckland.ac.nz

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.

Key words: distributional robustness, sample average approximation, phi-divergence, Wasserstein
History: This paper was first submitted on June 22, 2019.

1. Introduction

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

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

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*

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

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* (Wiesemann et al. 2014), 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 $\mathbb{Q}$. In applications we seldom have enough information to specify $\mathbb{P}$, so the set $\mathcal{P}$ of distributions is chosen because we seek a solution that performs well irrespective of the choice of distribution.

If one has a sample drawn from $\mathbb{P}$ then this can be used to construct a suitable set $\mathcal{P}$. The distributionally robust version of SAA is then

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

where the objective function depends on the sample S through the worst-case probability measure chosen from 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 (Scarf 1958, Dupačová 1987) 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, Delage and Ye (2010) construct a confidence set for the first and second moments of $\mathbb{P}$ based on a sample, whereas Wang et al. (2016) construct $\mathcal{P}$ in terms of a likelihood function, and Bertsimas et al. (2018) choose $\mathcal{P}$ 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 (2015) give a tutorial discussion of the use of ϕ-divergence in this setting, and Shapiro (2017) also discusses the different types of ϕ-divergence and their links with coherent risk measures. Gotoh et al. (2018) 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. (2017) 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 (2007) 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 2014). The paper by Gao and Kleywegt (2016) 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. Esfahani and Kuhn (2018) demonstrate 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. Schaffer 1993, Lawrence et al. 1997, Hawkins 2004). 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. Chopra and Ziemba 2013, Drela 1998,Wozabal 2014).

The most widely adopted machine-learning approach to overfitting is to add some form of regularization to the estimation problem. Examples are ridge regression (Hoerl and Kennard 1970) and LASSO (Tibshirani 1996). The literature concerning improved estimation performance from techniques related to regularization is very extensive and we will not attempt to review it here. However it is well known (see e.g. Tibshirani 1996) that regularization (by shrinking the size of parameters) results in parameter estimates with lower variance that often outweighs the modest increase in bias. In our context, 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 (see e.g. Xu et al. 2009). Regression regularization can also be formulated as a distributionally robust optimization problem using a Wasserstein metric (Blanchet et al. 2016).

There is substantial evidence that distributionally robust optimization can improve expected out-of-sample performance on a wider range of models than those for estimation. For example,

as shown by Chopra and Ziemba (2013), 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 (Delage and Ye 2010, Wozabal 2014). 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 (2018) 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. (2017) when using Kullback-Leibler divergence in an inventory problem and a logistic regression problem. Luo and Mehrotra (2017) 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. (2017) show that in their setup a portfolio optimization problem never sees an improvement in mean.

The paper by Gotoh et al. (2017) is closest to our analysis and also considers the behaviour of robust optimization for small values of δ. Their analysis is for a general convex cost function and for smooth ϕ-divergence measures. They use the notation $x^*(\delta)$ for the limit of $x_\delta(S)$ as S gets large and discuss both the way that $x_\delta(S)$ approaches $x^*(\delta)$ in distribution and also the relationship of $x^*(\delta)$ to the true optimal solution $x^*(0)$. Based on this asymptotic analysis for large sample sizes they give an explicit expansion for the expected value of the out-of-sample objective function for small δ. However the expressions involved become very complex (for example involving derivatives with respect to δ of the expectation of the Jacobian of the convex conjugate of ϕ evaluated at the objective $c(x^*(\delta), \xi)$).

Our paper considers specific features of optimization models in order to demonstrate that, even with the simplest problems, improvements in expected out-of-sample performance can depend on the DRO method used and the underlying distributions. Large values of δ result in solutions to DRO that are conservative, and as observed in the literature these will not perform well in expectation when evaluated with $\mathbb{P}$. On the other hand small values of $\delta > 0$ can give improvements in expected out-of-sample performance in comparison with SAA ($\delta = 0$). We shall therefore focus on the change in expected out-of-sample performance of the solution to DRO as δ increases from 0. We call this *incremental improvement.* The factors that affect incremental improvement are the

form of robustification, the form of $c(x,\xi)$ and X, and the true probability distribution of the random variable $\xi(\omega)$. Since robustification can either increase or reduce bias in the solution to SAA, we will restrict attention to problems in which the solution (i.e. the minimizing x) of SAA is unbiased. This means robustification will always make a solution more biased, and any observed improvements in out-of-sample performance can then be attributed to other factors. To ensure that the solution to SAA is unbiased for all possible probability distributions on ξ we assume that $X = \mathbb{R}^n$, and take $c(x,\xi)$ to be a positive definite quadratic function of x with a deterministic Hessian matrix.

The main contributions of the paper are as follows.

1. We formally define the concept of incremental improvement for DRO;

2. We analyse incremental improvement for distributionally robust quadratic programs using $\mathcal{P}_\delta$ derived from phi-divergence, coherent risk measures and Wasserstein formulations;

3. We present a number of simple one-dimensional examples with univariate objective functions for which analytical expressions defining incremental improvement can be derived. These examples show that incremental improvement cannot be taken for granted. The outcome depends on the form of robustification and the underlying probability distribution. For example, we provide some problem instances where incremental improvement will be obtained if we robustify with phi-divergence, but will not be if we use a CVaR approach.

The paper is laid out as follows. The next section establishes our notation and terminology, and formally defines the concept of incremental improvement. Section 3 then analyses quadratic examples with $\mathcal{P}_\delta$ derived from ϕ-divergence, both for a smooth ϕ function and also when total variation is used. Section 4 and section 5 repeat this analysis for a CVaR-based coherent risk measure, and Wasserstein distance respectively. In section 6 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:

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

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

(c) 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 Shapiro et al. 2014) 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 (Shapiro et al. 2014) 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.

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 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 some examples we can provide conditions on the problem data that will give sufficient 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.

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 If for almost all samples S, DRO has a unique solution and there is some constant vector $\bar{y}(S)$ with

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

then we say that problem DRO exhibits *linear variation with direction* $\bar{y}(S)$.

We now state a general result that will be used to establish both necessary and sufficient conditions for incremental improvement of SAA using robustification.

LEMMA 1. *Suppose DRO exhibits linear variation with direction* $\bar{y}(S)$*. Then for almost all samples* S

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

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.

Lemma 1 is a simple consequence of the first-order variation in $x_\delta(S)$ at $\delta = 0$. To verify linear variation and evaluate $\nabla \bar{c}(x_0(S))^\top \bar{y}(S)$ for specific instances of $c(x,\xi)$ and different forms of robustification, we require some more assumptions on the form of $c(x,\xi)$ and X. The following assumptions will be assumed to hold throughout the rest of the paper.

ASSUMPTION 2. *$c(x,\xi) = \frac{1}{2}x^\top Hx + v(\xi)^\top x + u(\xi)$ for some deterministic positive definite matrix H;*

ASSUMPTION 3. *$X = \mathbb{R}^n$;*

Under Assumptions 2 and 3, the "true" problem we seek to solve is

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

The objective function of SQP is $\bar{c}(x) = \frac{1}{2}x^\top Hx + \bar{v}^\top x + \bar{u}$, where $\bar{v} = \mathbb{E}_{\mathbb{P}}[v(\xi)]$ and $\bar{u} = \mathbb{E}_{\mathbb{P}}[u(\xi)]$. The gradient $\nabla \bar{c}(x) = Hx + \bar{v}$, and the unique solution to SQP is $x^* = -H^{-1}\bar{v}$.

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

$$\text{SAA:}\ \ \min_{x\in\mathbb{R}^n} \left(\tfrac{1}{2}x^\top Hx + \bar{v}_0(S)^\top x + \bar{u}_0(S)\right)$$

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 unique solution to SAA is $x_0(S) = -H^{-1}\bar{v}_0(S)$. Since $\mathbb{E}_{\mathbb{S}}[\bar{v}_0(S)] = \bar{v}$, it is easy to see that $x_0(S)$ is unbiased for any probability distribution on ξ. On the other hand, 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. Similarly, if X is a proper subset of $\mathbb{R}^n$ then $x_0(S)$ is generally biased, even if Assumption 2 holds.

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

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

where $\mathcal{P}_\delta$ is a set of probability distributions that are close to $\mathbb{P}_0$. Recall that $\frac{1}{2}x^\top Hx$ is strictly convex and $\sup_{\mathbb{Q}\in\mathcal{P}_\delta} \mathbb{E}_{\mathbb{Q}}\left[v(\xi)^\top x + u(\xi)\right]$ is a convex function of x, so DRQP has a unique solution, denoted $x_\delta(S)$.

LEMMA 2. *Suppose DRQP exhibits linear variation with direction* $\bar{y}(S)$. *Then for almost all samples* S

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

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

$$\mathbb{E}_S[(\bar{v}_0(S) - \bar{v})^\top \bar{y}(S)] > 0, \tag{5}$$

then the robustification incrementally improves SAA.

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.

To apply the inequality (5) in Lemma 2 we require a formula for the vector $\bar{y}(S)$ that defines the direction of linear variation. This 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 (5) will remain true for any positive scaling of $\bar{y}(S)$.

In what follows, we apply Lemma 2 to models that robustify SAA using ϕ-divergence, conditional value at risk, and a Wasserstein metric. In the first two cases (and for part of the Wasserstein discussion) the distribution $\mathbb{Q} \in \mathcal{P}_\delta$ will be confined to be a finite distribution with weights $q_1, q_2, \ldots, q_N$ on the sample points in S. This gives

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

where

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

To determine a direction of linear variation in these models it is convenient to make the following assumption that we will require for all instances of DRQP that use such a set $\mathcal{P}_\delta$.

ASSUMPTION 4. *$\{v(\xi(\omega)) : \omega \in \Omega\}$ has a density f with support having n dimensions.*

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

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

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

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

PROPOSITION 1. *Under Assumption 4 the set of samples $S = \{\xi_1, \xi_2, \ldots, \xi_N\}$ that are strictly ordered by SAA has probability measure* 1.

3. 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

$$d_\phi(\nu_q, \nu_p) = \sum_{i=1}^{N} p_i \phi\left(\frac{q_i}{p_i}\right) \tag{6}$$

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 (6) 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 (6). We first study an example (total variation) where $\phi(t) = |t-1|$ is non-smooth, and then consider general analytic functions ϕ.

3.1. 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

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

Recall $x_0(S)$ is the solution to SAA, and by Proposition 1 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 3. *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

$$MVRS = \frac{1}{2}\mathbb{E}_S[(\bar{v} - \bar{v}_0(S))^\top H^{-1}R(S)]. \tag{7}$$

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

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

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

$$MVRS = \frac{1}{2}\int_{-\infty}^{\infty} \left(F(z)^{N-1} - (1-F(z))^{N-1}\right) Q(z)dz. \tag{9}$$

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 4. *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 2 shows that robustification of SAA0 with total variation makes the solution worse.

Now consider robustification of SAA (including the constant term) 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 $x_\delta(S)$. 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 $q_{i_{\min}}$. Under Assumption 4 $i_{\min}$ 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 2 (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 $\bar{\xi}_0(S)-\xi_2<\bar{\xi}_0(S)-\xi_1$ and $\bar{\xi}_0(S)-\xi_2<\xi_3-\bar{\xi}_0(S)$. Thus

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

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

$$\begin{aligned}\mathbb{E}_S[(\xi_{i_{\max}}-\xi_{i_{\min}})\bar{\xi}_0(S)] &= \int_{-\frac{1}{2}}^{\frac{1}{2}}\int_x^{\frac{1}{2}}\int_{2y-x}^{\frac{1}{2}} 6(z-y)(\frac{x+y+z}{3})dzdydx\\ &\quad+\int_{-\frac{1}{2}}^{\frac{1}{2}}\int_x^{\frac{1}{2}}\int_y^{2y-x} 6(x-y)(\frac{x+y+z}{3})dzdydx\\ &= -\frac{1}{6},\end{aligned}$$

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 $q_1 = q_N = \frac{1}{2}$. 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 (Lloyd 1952)).

3.2. 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 $\mathcal{P}_\delta \subseteq \{\mathbb{Q} : supp(\mathbb{Q}) \subseteq S\}$, by

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

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

$$V(S) = \frac{1}{N} \sum_{i=1}^{N} (v(\xi_i) - \bar{v}_0(S)) (v(\xi_i) - \bar{v}_0(S))^\top. \tag{10}$$

We now have the following result.

PROPOSITION 5. *For any analytic strictly convex ϕ, DRQP with ϕ-divergence robustification exhibits linear variation with*

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

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 $\bar{\xi}_0(S) = \frac{1}{N}\sum_{i=1}^{N} \xi_i$. 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 5 with $v(\xi) = -2\xi$, gives

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

where

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

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 (Philpott et al. 2018) with $r = \frac{\delta}{\sqrt{N}}$):

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

so there are no $O(\delta^2)$ terms in this case. □

Proposition 5 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 6. *Robustification with smooth phi-divergence gives incremental improvement in SAA for any analytic strictly convex ϕ if*

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

We can apply this to the scalar case where $c(x,\xi) = x^2 - g(\xi)x$. Let $\sigma(S) = \left(\frac{1}{N}\sum_i \left(g(\xi_i) - \overline{g}_0(S)\right)^2\right)^{1/2}$ be the standard deviation of the $g(\xi_i)$ values in the sample. Then we obtain incremental improvement if $\mathbb{E}_S\left[\sigma(S)(\overline{g}_0(S) - \overline{g})\right] > 0$. Notice that for any symmetric distribution $\mathbb{E}_S\left[\sigma(S)(\overline{g} - \overline{g}_0(S))\right] = 0$ and MVRS will be zero.

4. 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

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

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 4. *The solution to DRQP with CVaR robustification satisfies*

$$x_\delta(S) \in -H^{-1} \left((1-\delta)\bar{v}_0(S) + \delta G_{CVaR}(x_\delta(S)) \right) \tag{11}$$

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

$$x_\delta(S) = -H^{-1} \left((1-\delta)\bar{v}_0(S) + \delta \bar{v}_{CVaR}(S) \right). \tag{12}$$

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 we order $S = \{\xi_1, \xi_2, \ldots, \xi_N\}$ so that

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

then under Assumption 4 Proposition 1 gives

$$v(\xi_1)^\top x_0(S) > v(\xi_2)^\top x_0(S) > \ldots > v(\xi_N)^\top x_0(S) \tag{13}$$

except for a set of samples with probability 0. It is convenient in this section to arrange $v(\xi_i)^\top x_0(S)$ in decreasing order. For the samples satisfying (13), $\text{CVaR}_{1-\alpha}[\{v(\xi_i)^\top x\}]$ is differentiable at $x_0(S)$, with derivative

$$\bar{v}_{\text{CVaR}}(S) = \frac{1}{\alpha N} \sum_{i=1}^{m_\alpha} v(\xi_i) + (1 - \frac{m_\alpha}{\alpha N}) v(\xi_{m_\alpha}), \tag{14}$$

where $m_\alpha = \lceil \alpha N \rceil - 1$.

PROPOSITION 7. *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 (14). Also*

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

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 8. *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 9. *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*

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

where $m_\alpha = \lceil \alpha N \rceil - 1$.

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 9 shows that MVRS will be positive if the second term is small. There is a connection here to 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

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

The approximation arises from taking the $\text{sgn}(\bar{g}_0(S))$ term in Proposition 8 as always being 1. As μ gets larger the probability that this fails becomes vanishingly small. In Appendix 2, we show (Lemma 13) that

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

which gives $\widetilde{\text{MVRS}} = 0$. □

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

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

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

We see that in comparison with a normal distribution, the heavy tails introduced by taking a mixture of normal distributions makes $\widetilde{\text{MVRS}}$ 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.

5. Wasserstein metric

Distributionally robust optimization using a Wasserstein metric chooses

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

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

$$d_W(\nu_1, \nu_2) = \min_{\gamma \in \Gamma(\nu_1, \nu_2)} \int_{M \times M} \|z_1 - z_2\| \; d\gamma(z_1, z_2) \tag{15}$$

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:

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

in which the expectation is taken over the random variable ξ in $\mathbb{R}^m$ with distribution $\mathbb{Q}$.

The simplest version of DRQP in the Wasserstein setting requires $\mathbb{Q}$ to have the same support as $\mathbb{P}_0$. As this bears a close relation to the previous two sections we will discuss this first. We write $\{q_1, q_2, \ldots, q_N\}$ for the probability assigned by $\mathbb{Q}$ to points in S which gives a transportation problem that has optimal value

$$d_W(\mathbb{Q}, \mathbb{P}_0) = \left\{ \begin{array}{lll} \min_{w_{ij}} & \sum_{i=1}^N \sum_{j=1}^N w_{ij} \|\xi_i - \xi_j\| & \\ \text{subject to} & \sum_{j=1}^N w_{ij} = \frac{1}{N}, & i = 1, 2, \ldots, N, \\ & \sum_{i=1}^N w_{ij} = q_j, & j = 1, 2, \ldots, N, \\ & w_{ij} \geq 0. & \end{array} \right.$$

In DRQP, recall

$$Q_{\max}(x) = \begin{cases} \max & \sum_{i=1}^{N} q_i v\left(\xi_i\right)^\top x \\ \text{subject to} & d_W(\mathbb{Q}, \mathbb{P}_0) \le \delta \end{cases}$$

which is equivalent to setting $q_j = \sum_{i=1}^{N} w_{ij}, \quad j = 1, 2, \ldots, N,$ where w_{ij} is the mass transferred from point ξ_i to point ξ_j and solves

$$\begin{array}{ll} \text{WP: max} & \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} v\left(\xi_j\right)^\top x \\ \text{subject to} & \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} \left\|\xi_i - \xi_j\right\| \le \delta, \\ & \sum_{j=1}^{N} w_{ij} = \frac{1}{N}, \qquad\qquad i = 1, \ldots, N \\ & w_{ij} \ge 0. \end{array}$$

Since WP is a linear program, it has an optimal basic solution with $N+1$ basic variables. For small δ the optimal solution will thus have w_{ii} basic for each i, and choose a single $i \neq j$ and $w_{ij} = \delta / \left\|\xi_i - \xi_j\right\|$ so that w_{ij} shifts probability from point ξ_i to point ξ_j so as to maximize $\frac{v\left(\xi_j\right)^\top x - v(\xi_i)^\top x}{\left\|\xi_i - \xi_j\right\|}$ giving indices $i = k$, $j = l$. The optimal solution to WP is then to set

$$w_{ij} = \begin{cases} \delta / \left\|\xi_i - \xi_j\right\|, & i = k, j = l, \\ \frac{1}{N} - \delta / \left\|\xi_i - \xi_j\right\|, & i = k, j = k, \\ \frac{1}{N}, & k \neq i = j, \\ 0, & \text{otherwise} \end{cases}.$$

Observe that the choice of indices k and l is dependent on x, but for small δ we expect under Assumption 4 that for almost every sample S, $\frac{v\left(\xi_j\right)^\top x_\delta(S) - v(\xi_i)^\top x_\delta(S)}{\left\|\xi_i - \xi_j\right\|}$ will give the same indices as $\frac{v\left(\xi_j\right)^\top x_0(S) - v(\xi_i)^\top x_0(S)}{\left\|\xi_i - \xi_j\right\|}$. The net effect of robustification applied to almost every sample is to increase q_l to $\frac{1}{N} + \delta / \left\|\xi_l - \xi_k\right\|$ and decrease q_k to $\frac{1}{N} - \delta / \left\|\xi_l - \xi_k\right\|$. This means that

$$\begin{aligned} x_\delta(S) - x_0(S) &= -H^{-1} \sum_i q_i v\left(\xi_i\right) + H^{-1} \sum_i \frac{1}{N} v\left(\xi_i\right) \\ &= \delta H^{-1} \frac{v\left(\xi_k\right) - v\left(\xi_l\right)}{\left\|\xi_l - \xi_k\right\|}, \end{aligned}$$

so $\bar{y}(S) = H^{-1} \frac{v(\xi_k) - v(\xi_l)}{\left\|\xi_l - \xi_k\right\|}$, where k and l are indices giving the highest value of $\frac{v(\xi_l)^\top x_0(S) - v(\xi_k)^\top x_0(S)}{\left\|\xi_i - \xi_j\right\|}$, and Lemma 2 yields

$$\text{MVRS} = \mathbb{E}_S[\left(v_0(S) - \bar{v}\right)^\top H^{-1} \frac{v\left(\xi_k\right) - v\left(\xi_l\right)}{\left\|\xi_l - \xi_k\right\|}].$$

In the previous two sections the structure of the cost function with respect to the random variable ξ has not been critical to incremental improvement; everything has been determined by the set of cost functions $c(x, \xi_i)$ evaluated at the sample points. For Wasserstein robustification we need to pay attention to the behavior of $c(x, \xi)$ with respect to changes in ξ. For example if $c(x, \xi) = \frac{1}{2}x^2 - g(\xi)x$ where g is a strictly convex and non-negative function of ξ then for positive values of x Wasserstein robustification moves weight from ξ_k with a high value of g to some ξ_l with a lower value that is not too far away. In this case

$$\text{MVRS} = \mathbb{E}_S[\left(g_0(S) - \bar{g}\right) \frac{\left|g\left(\xi_l\right) - g\left(\xi_k\right)\right|}{\left|\xi_l - \xi_k\right|}].$$

We now move to the general case where the support of $\mathbb{Q}$ is not constrained, so robustification will allow movements in the sample points ξ_i. As observed above, the behavior of $c(x,\xi)$ with respect to changes in ξ affects the out-of-sample performance of Wasserstein robustification. Often we will take x fixed and it is convenient to write $c_x(z)$ for $c(x,z)$, where we use z to denote an element of the set M that contains ξ_i. We assume throughout that $c_x(z)$ is differentiable at all $z \in \mathbb{R}^m$.

Using (15), the inner maximization problem P is equivalent to solving

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

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 P as

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

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

We make use of a result of Gao and Kleywegt (2016, Corollary 2, part iii).

PROPOSITION 10. *(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 $\lambda^* \ge 0$ 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 11. *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) $\|\nabla c_x(z_i)\| \ge \|\nabla c_x(\xi_k)\|$ 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 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 and each $x_j \geq 0$. 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 12. *With the Wasserstein distance metric if every component of $v(z)$ is strictly concave then for $x \in R^n_+$, 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 12 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 13. *(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

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

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

6. 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.

We have shown how to quantify incremental improvement from robustification in univariate examples using the "ground-truth" probability distribution**.** We may ask whether these results give some guidance to a risk-neutral decision maker 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. When analytical techniques are not applicable (and assuming the DRO problem has linear variation), it may be possible to use statistical estimation of MVRS from the data available. An alternative may be to carry out tests on synthetic data. The key observation from our work is that different robustification techniques have different behaviors that can depend on characteristics of the distribution. So it will be important to check whether the underlying distribution is symmetric or skewed, and whether it has heavy tails. Then any tests should be carried out using different robustifications on synthetic data that share the appropriate characteristics.

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.

Acknowledgments

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.

References

Bayraksan G, Love D (2015) Data-driven stochastic programming using phi-divergences. *The Operations Research Revolution*, 1–19 (INFORMS).

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

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

Chopra VK, Ziemba WT (2013) The effect of errors in means, variances, and covariances on optimal portfolio choice. *Handbook of the Fundamentals of Financial Decision Making: Part I*, 365–373 (World Scientific).

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

Drela M (1998) Pros & cons of airfoil optimization. *Frontiers of Computational Fluid Dynamics 1998*, 363–381 (World Scientific).

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

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

Gao R, Kleywegt A (2016) Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199* .

Gotoh JY, Kim M, Lim A (2017) Calibration of distributionally robust empirical optimization models. *arXiv preprint arXiv:1711.06565* .

Gotoh JY, Kim M, Lim A (2018) Robust empirical optimization is almost the same as mean–variance optimization. *Operations Research Letters* 46(4):448–452.

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

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

Lawrence S, Giles C, Tsoi A (1997) Lessons in neural network training: Overfitting may be harder than expected. *AAAI/IAAI*, 540–545 (Citeseer).

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

Luo F, Mehrotra S (2017) Decomposition algorithm for distributionally robust optimization using Wasserstein metric. *arXiv preprint arXiv:1704.03920* .

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

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

Scarf H (1958) A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production* .

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

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

Shapiro A, Ruszczynski A, Dentcheva D (2014) *Lectures on Stochastic Programming: Modeling and Theory* (SIAM).

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

Van Parys B, Esfahani P, Kuhn D (2017) From data to decisions: Distributionally robust optimization is optimal. *arXiv preprint arXiv:1704.04118* .

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

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

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

Xu H, Caramanis C, Mannor S (2009) Robust regression and Lasso. *Advances in Neural Information Processing Systems*, 1801–1808.

Appendix 1: Proofs of propositions

Proof of Lemma 1

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 Lemma 2

Linear variation with direction $\bar{y}(S)$ and Lemma 1 gives

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

for almost every sample S. 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

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

We have

$$C_\delta(S) = \frac{1}{2} x_\delta(S)^\top H x_\delta(S) + \bar{v}^\top x_\delta(S) + \bar{u}.$$

so

$$\begin{aligned} C_\delta(S) - C_0(S) &= \frac{1}{2} x_\delta(S)^\top H x_\delta(S) - \frac{1}{2} x_0(S)^\top H x_0(S) + \bar{v}^\top (x_\delta(S) - x_0(S)) \\ &= \frac{1}{2} x_\delta(S)^\top H x_\delta(S) - \frac{1}{2} x_0(S)^\top H x_0(S) - x^{*\top} H (x_\delta(S) - x_0(S)) \\ &= (x^* - x_0(S))^\top H (x_0(S) - x_\delta(S)) + \frac{1}{2} (x_\delta(S) - x_0(S))^\top H (x_\delta(S) - x_0(S)). \end{aligned}$$

Substituting $\bar{v}$ for $-Hx^*$ and $\bar{v}_0(S)$ for $-Hx_0(S)$ yields

$$\begin{aligned}\mathbb{E}_S[C_\delta(S) - C_0(S)] = \mathbb{E}_S[&(\bar{v}_0(S) - \bar{v})^\top (x_0(S) - x_\delta(S)) \\ &+ \frac{1}{2}(x_\delta(S) - x_0(S))^\top H(x_\delta(S) - x_0(S))].\end{aligned} \tag{16}$$

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 (16) 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 Proposition 1

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

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

A sample S where there is equality $v(\xi_i)^\top x_0(S) = v(\xi_{i+1})^\top x_0(S)$ for some i, will have $\{v(\xi_1), v(\xi_2), \ldots, v(\xi_N)\}$ satisfying

$$(v(\xi_{i+1}) - v(\xi_i))^\top H^{-1} \sum_{j=1}^{N} v(\xi_j) = 0. \tag{17}$$

The equation (17) defines a manifold in $\mathbb{R}^{nN}$ of dimension strictly less than nN. Since $v(\xi)$ has an n-dimensional density from which we sample independently, the joint density on $(v(\xi_1), v(\xi_2), \ldots, v(\xi_N))$ has dimension nN. Thus the probability measure of the set of samples satisfying (17) is zero. 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 Lemma 3

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

$$0 \in Hx_\delta(S) + \bar{v}_0(S) + (\delta/2)G(x_\delta(S)). \tag{18}$$

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 under Assumption 4, Proposition 1 implies that all samples S apart from a set with probability 0 are strictly ordered by SAA, so

$$v(\xi_1)^\top x_0(S) < v(\xi_2)^\top x_0(S) < ... < v(\xi_N)^\top x_0(S), \tag{19}$$

and so for the samples that are strictly ordered by SAA, and for δ small enough we have (19) with $x_\delta(S)$ replacing $x_0(S)$. It follows that for all samples S strictly ordered by SAA, the relationship (18) 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

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

So

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

with the remark on incremental improvement being immediate. □

Proof of Proposition 2

Since $H = 1$, Lemma 3 gives for sufficiently small $\delta > 0$,

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

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

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

so

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

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

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 $\mathbb{E}_S[C_\delta(S)] = \mathbb{E}_S[C_0(S)] + (\delta^2/8)\,\mathbb{E}_S[R(S)^2]$, giving $\text{VRS}(\delta) < 0$ for all $\delta > 0$. □

Proof of Proposition 3

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 12 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 4

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

$$0 = \int_{-\infty}^{\infty} z f(z) dz = \int_{-\infty}^{\infty} \frac{z}{\tau'(\tau^{-1}(z))} \widetilde{f}(\tau^{-1}(z)) dz = \int_{-\infty}^{\infty} \tau(w) \widetilde{f}(w) dw \tag{20}$$

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

$$\int_{w_0}^{\infty} (\tau(z) + 2w_0 - z)\, \widetilde{f}(z) dz = 0. \tag{21}$$

We will use Proposition 3 and we begin by rewriting the required expression in terms of $\widetilde{F}$. From (9) 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

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

Let $T(z) = \int_z^\infty \tau(u)\widetilde{f}(u)du \geq 0$ for $z \geq w_0$. From (21)

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

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

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

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 (21) 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 $z \leq w_0$ 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 (21), $V(\infty)=0$. Now $\tau(u)+2z_0-u$ has derivative $\tau'(u)-1\ge 0$ and because the integral in (21) 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 $z\le w_0$ when $\widetilde{F}(z)^{N-1}-(1-\widetilde{F}(z))^{N-1}<0$, then

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

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

Proof of Proposition 5

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

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

The problem IP has Lagrangian

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

which has first derivative

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

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

$$-v(\xi_i)^\top x+\lambda(\delta,x)\phi'\left(1+r_i(\delta,x)\right)+\mu=0,\quad i=1,2,\ldots,N. \tag{22}$$

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 (22) 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 $i = 1,2,\ldots,N$,

$$\begin{array}{l} -v(\xi_i)^\top x + \lambda(\delta,x)r_i(\delta,x)\left(\phi''(1) + g_0(r_i(\delta,x))\right) \\ \quad +\bar{v}_0(S)^\top x - \frac{1}{N}\sum_j \lambda(\delta,x)r_j(\delta,x)g_0(r_j(\delta,x)) = 0. \end{array} \tag{23}$$

Moreover from the constraint

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

we get

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{r_i(\delta,x)^2}{2}(\phi''(1) + g(r_i(\delta,x)))\right) = \delta^2. \tag{24}$$

We will establish the proposition via several lemmas.

LEMMA 5. *The solution $\lambda(\delta,x), r_i(\delta,x)$, $i = 1,2,\ldots,N$ to (23) and (24) 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 (23) and (24) 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

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

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

$$\sum_{i=1}^{N} w_i = 0. \tag{25}$$

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 (23)

$$\begin{aligned}&r_i(\delta, x)\left(\phi''(1) + g_0(r_i(\delta, x))\right) - \frac{1}{N}\sum_j r_j(\delta, x) g_0(r_j(\delta, x)) \\ &= \frac{1}{\lambda(\delta, x)}(v(\xi_i) - \bar{v}_0(S))^\top x.\end{aligned} \tag{26}$$

From (25) and (26) we deduce

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

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 $w_i\le 0$ for $v(\xi_i)^\top x\le\widehat{v}$ 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 (25).

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 $(x_0(S),0)$. □

LEMMA 6. *Suppose* $k=\sqrt{\frac{2}{\phi''(1)}}$. *Then*

$$r_i(\delta,x)=\delta k\frac{(v(\xi_i)-\bar{v}_0(S))^\top x}{\left(x^\top V(S)x\right)^{1/2}}+\delta^2 h_i(\delta,x) \tag{27}$$

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

Proof Rearranging (23) gives

$$r_i(\delta,x)=\frac{1}{\lambda(\delta,x)}\frac{v(\xi_i)^\top x-\bar{v}_0(S)^\top x}{\phi''(1)+g_0(r_i(\delta,x))}+\frac{\sum_j r_j(\delta,x)g_0(r_j(\delta,x))}{N(\phi''(1)+g_0(r_i(\delta,x)))}. \tag{28}$$

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

$$\lim_{\delta\to 0} r_i(\delta, x)\lambda(\delta, x) = \lim_{\delta\to 0} \eta_i(\delta, x) = \frac{(v(\xi_i)^\top x - \bar{v}_0(S)^\top x)}{\phi''(1)}. \tag{29}$$

Multiplying (24) 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

$$\lim_{\delta\to 0} \delta\lambda(\delta, x) = \left(\frac{x^\top V(S)x}{2\phi''(1)}\right)^{1/2}. \tag{30}$$

Since for almost all S we will have $\frac{x^\top V(S)x}{2\phi''(1)} > 0$, (29) and (30) give

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

whereby applying Lemma 5 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 5 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 6 gives

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

and by Lemma 5 $h_i(\delta, x)$ and $\frac{\partial}{\partial x_j} h_i(\delta, x)$ are bounded as $\delta \to 0$, so it follows that

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

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

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

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

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

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

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

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

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

as required. □

Proof of Proposition 6

Immediately from the improvement lemma and Proposition 5 we know that robustification with smooth ϕ divergence incrementally improves SAA if

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

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

Proof of Lemma 4

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

$$0\in\partial\rho(c(x,S))=Hx+(1-\delta)\bar{v}_0+\delta G_{\text{CVaR}} \tag{31}$$

which gives (11) and (12) when the subgradient at the optimal solution is unique. □

Proof of Proposition 7

Consider all samples S satisfying (13). Suppose that $\alpha\in(\frac{m}{N},\frac{m+1}{N}]$ 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 $k\leq m+1\leq\ell$. When $k\neq\ell$ we have non-differentiability of CVaR at x and the subdifferential $\partial\text{CVaR}_{1-\alpha}[\{v(\xi_i)^\top x\}]$ 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 4

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

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

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

Lemma 4 then gives

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

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

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

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 8

We apply Proposition 7 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 7 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

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

as required. □

Proof of Proposition 9

We will use Proposition 8 and show that

$$-\mathbb{E}_S[(\bar{g}_0(S) - \bar{g})\mathrm{CVaR}_{1-\alpha}[\{-g(\xi_i)\}]] = \int_{-\infty}^{\infty} Q(z)(1 - F(z))^{N-1}\Lambda_\alpha(z)dz. \tag{32}$$

First observe that

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

so

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

where w_i is the sample from W. We have that $-\mathrm{CVaR}_{1-\alpha}[\{-w_i\}]$ 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

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

where z_i are the order statistics. So

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

Since Lemma 12 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

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

where $m_\alpha = \lceil \alpha N \rceil - 1$, the identity (32) now follows. □

Proof of Proposition 11

We suppose that $c(x,\xi)$ is strictly concave in ξ. We first observe by Proposition 10 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

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

The Lagrangian of P1 is

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

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 7. *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 7

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

Now

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

So

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

Hence

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

which yields the result. □

Proof of Proposition 12

Recall

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

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 11 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 11 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 11.

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 7 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

$$H(x - x_0(S)) = -\delta \frac{J_v(\xi^*)J_v(\xi^*)^\top x_0(S)}{\|J_v(\xi^*)^\top x_0(S)\|} + O(\delta^2). \tag{33}$$

From (33) 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 13

(a) In the univariate example $J_v(\xi^*) = -\nabla g(\xi^*(S))$ so from Proposition 12 we have

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

and $v(\xi) = -g(\xi)$, so Lemma 2 gives

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

(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 9 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

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

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

We have

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

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

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

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 $z_1\le z_2\le\ldots\le z_N$. The sample mean is $\bar{z}=\frac{1}{N}\sum_{i=1}^N z_i$.

LEMMA 8. $\mathbb{E}[z_i^2] = \frac{N!}{(N-i)!(i-1)!}\int_{-\infty}^\infty z^2F(z)^{i-1}f(z)(1-F(z))^{N-i}dz.$

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

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

as required. □

LEMMA 9. *If $i<j$ then*

$$\mathbb{E}[z_i z_j] = \frac{N(N-1)\dots(N-j+1)}{(i-1)!(j-i-1)!}\int_{-\infty}^{\infty}\int_{x_a}^{\infty} x_a x_b B(x_a,x_b)dx_b dx_a \tag{34}$$

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 $A_{ij} = \{z_i \in (x_a, x_a+\varepsilon) \text{ and } z_j \in (x_b, x_b+\varepsilon')\}$.

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

But

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

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

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

as required. □

LEMMA 10.

$$\sum_{j=i+1}^{N}\mathbb{E}[z_i z_j] = \frac{N!}{(i-1)!(N-i-1)!}\int_{-\infty}^{\infty} zF(z)^{i-1}(1-F(z))^{N-i-1}f(z)Q(z)dz. \tag{35}$$

Proof

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

Substituting in (34) and substituting for $Q(z)$ yields (35). □

Lemma 11.

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

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 10 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 12.

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

Proof Applying Lemmas 8, 10, and 11, 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

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

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 (36) as required. □

LEMMA 13. *Suppose z has density f and cumulative distribution function F. For all $\alpha\in(0,1]$,*

$$\int_{-\infty}^{\infty}f(z)(1-F(z))^{N-1}\Lambda_\alpha(z)dz=\frac{1}{N}. \tag{37}$$

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

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

Now

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

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

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

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. □