Robust sample average approximation with small sample sizes*

E.J. Anderson† A.B. Philpott‡

February 27, 2019

Abstract

We consider solving stochastic optimization problems in which we seek to minimize the expected value of an objective function with respect to an unknown distribution of random parameters. Our focus is on models that use sample average approximation (SAA) with small sample sizes. We analyse the out-of-sample performance of solutions obtained by solving a robust version of the SAA problem, and derive conditions under which these solutions are improved in comparison with SAA. We analyse three different mechanisms for constructing a robust solution: a CVaR-based risk measure, phi-divergence using total variation, and a Wasserstein metric.

1 Introduction

An important class of stochastic programming problems involves the optimization of the expectation of some objective function such as cost or profit, where the outcome depends both on a decision variable and some random variable, Y. Formally this can be written as

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

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

†University of Sydney

‡University of Auckland

where the decision variable x is constrained to lie in $X \subseteq \mathbb{R}^n$, and expectations are taken over the random variable Y, with instance $y \in \mathbb{R}^m$. We denote an optimal solution of P by x^* and its optimal value by C^*. A key assumption for the problem class of interest to us is that the probability distribution $\mathbb{P}$ of Y is independent of the choice of x. In the simplest case Y has a known distribution $\mathbb{P}$ and the expectation can be evaluated through integration to give an explicit expression for the expected value in terms of the decision variable, x, and then the stochastic programming problem can be solved analytically.

In many cases the distribution of the random variable is unknown and the decision maker must make her choice on the basis of a sample of data points drawn from Y, but without access to an explicit description of the distribution $\mathbb{P}$. We will suppose that we have a sample, S, of N points from Y where $S = \{y_1, y_2, ..., y_N\}$ but we do not know the distribution of Y. The problem P can then be approximated by the *sample average approximation problem*

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

Given a sample S we will write $\mathbb{E}[c(x, S)]$ as a shorthand for the objective function of SAA, where the expectation assigns equal probability to sample points. We denote an optimal solution to SAA by $x_{SAA}(S)$.

A common environment leading to problems of the form SAA has a historical record of outcomes that can be assumed to be drawn from a fixed distribution, and where the performance of a decision made now depends on the uncertain outcome that occurs in the future. For example we may be looking at financial data and assume that the underlying distribution of prices for the last three months is a good guide to the distribution of the uncertain prices that will occur next week; or we may have a historical record of annual demand for a set of products and need to make some capacity decisions; or we may be considering optimizing the performance of a hydroelectric system where the record of annual rainfall over a period of fifty years can be taken as indicative of the distribution of rainfall for next year.

A related problem occurs when the distribution of Y is known but is not available in a form that allows the calculation of the value of $\mathbb{E}[c(x, Y)]$, most frequently because the dimension, m, of the random variable is too high to allow the calculation of the required integral for the expectation. In these circumstances we may wish to approximate the expectation by averaging over a sample of points drawn from Y. This arrangement matches our set up and our results will apply. However, typically such problems are solved by taking samples with a size measured in thousands or tens of thousands and it is the asymptotic results that are critical. Our focus is on problems with

relatively small samples (tens or hundreds), and we wish to look at behaviors that occur well before the asymptotic limit.

A third type of problem sometimes occurs when the cost function c is very expensive to evaluate requiring some kind of simulation. For such problems we are forced to use the sample average approximation to estimate the expectation. However for these problems the decision maker has an opportunity to choose the set of points $y_1, y_2, ..., y_N$ at which the cost is calculated. In our model the sample is drawn randomly from the distribution $\mathbb{P}$.

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. [19], [14], [13]). Here the coefficients of a model are estimated using a training set of data, and a model with many coefficients can choose these to match the training set very well. When applied to out-of-sample test data the model often performs worse than a simpler model with fewer coefficients. The solutions from sample average approximations with small sample sizes can also perform poorly out of sample (see e.g. [4],[7],[25]).

We will explore the circumstances in which a robust approach may end up doing better when applied to out-of-sample test data. To motivate this we consider a simple example, where

$$c(x, y) = x^2 + 40x - xy + 80,$$

and y has a lognormal density with mean 60 and standard deviation 27. The expected cost of any candidate solution x is

$$x^2 + 40x - \mathbb{E}[y]x + 80 = (x - 10)^2 - 20$$

and the optimal solution to P with these data is $x^* = 10$.

A sample $S = \{y_1, y_2, \ldots, y_N\}$ of y yields the problem (1), which in this example has optimal solution

$$x_{SAA}(S) = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i}{2} - 20.$$

Suppose a decision maker instead solves a robust version of (1)

$$\min_x \{(1 - \delta)\mathbb{E}[c(x, S)] + \delta \mathrm{CVaR}_{1-\alpha}[c(x, S)]\} \tag{2}$$

minimizing a convex combination of expectation and conditional value at risk. Here we write $\mathrm{CVaR}_{1-\alpha}[c(x, S)]$ for the conditional value at risk of the

discrete distribution $\{c(x, y_i) : y_i \in S\}$, where each sample point has equal probability (see [17]). If for example $N = 10$ and $\alpha = 0.1$ then

$$\text{CVaR}_{0.9}[c(x, S)] = x^2 + 40x - xy_{\min} + 80,$$

where $y_{\min}$ is the minimum element of the sample S. The solution to (2) is then

$$x_R(S) = \frac{1}{2}\left(\frac{(1-\delta)}{N}\sum_{i=1}^{N} y_i + \delta y_{\min} - 40\right).$$

Suppose we construct 20000 samples each of 10 observations of y. Each sample S yields an SAA solution $x_{SAA}(S)$ and a robust solution $x_R(S)$, each of which has an expected cost with respect to the underlying lognormal distribution. These expected costs depend on the sample S. Averaging over possible samples, the expected cost of $x_{SAA}(S)$ (taken with respect to the sampling distribution for S) is -1.90697, and the expected cost of $x_R(S)$ is shown in Table 1 for increasing values of δ. The expected cost of the candidate solution $x_R(S)$ improves as δ increases up to an optimal value (around 0.05) after which it starts getting worse.

δ	$\mathbb{E}_S[c(x_R(S), Y)]$
0.00	-1.90697
0.01	-2.11933
0.02	-2.28048
0.03	-2.39041
0.04	-2.44914
0.05	-2.45665
0.06	-2.41296

Table 1. Comparison of expected cost of $x_R(S)$ as δ increases.

Observe that $x_{SAA}(S)$ is an unbiased estimator of x^*, but $x_R(S)$ is biased below x^*. At first sight we might think that since $x_{SAA}(S)$ is an unbiased estimator of the correct minimum x^*, and $x_R(S)$ is not, then $x_R(S)$ will on average be a worse solution. But the opportunity for improvement occurs through shrinkage [5]. Different samples S give rise to different values $x_R(S)$, which in this example has a variance 16.573 for $\delta = 0.06$, that is lower than the variance of $x_{SAA}(S)$ (17.938). The samples that lead to poor choices of $x_{SAA}(S)$ and higher costs are adjusted through robustification in a way that gives values $x_R(S)$ that are moved towards the centre and closer to the correct value. The improvements from reducing the spread of the distribution of $x_R(S)$ can over-ride the additional cost from shifting the average value of $x_R(S)$ away from the correct value.

As a contrast, let us consider the problem in which

$$c(x, y) = x^2 - 80x + xy + 80$$

where y has the same lognormal density with mean 60 and standard deviation 27. As before the optimal solution to P with these data is $x^* = 10$ with objective value -20.

Given a sample S of 10 points of y, we have

$$x_{SAA}(S) = -\frac{\frac{1}{N}\sum_{i=1}^{N} y_i}{2} + 40.$$

We continue to assume $\alpha = 0.1$ so it is just the highest cost sample element that contributes to $\text{CVaR}_{1-\alpha}[c(x, S)]$. However here the highest cost occurs for the largest y_i value rather than the smallest and so

$$x_R(S) = -\frac{1}{2}\left(\frac{(1-\delta)}{N}\sum_{i=1}^{N} y_i + \delta y_{\max} - 80\right)$$

where $x_R(S)$ solves

$$\min_x (1-\delta)\mathbb{E}[c(x, S)] + \delta(x^2 - 80x + xy_{\max} + 80).$$

The expected cost of the solution $x_R(S)$ depends on δ. Using the same 20000 samples of $N = 10$ observations, we obtain the figures in Table 2.

δ	$\mathbb{E}_S[c(x_R(S), Y)]$
0.00	-1.90697
0.01	-1.33165
0.02	-0.60401
0.03	0.27593
0.04	1.30818
0.05	2.49274
0.06	3.82961

Table 2. The expected cost of the candidate solution x_R gets worse as δ increases.

The variance of $x_R(S)$ (21.425 for $\delta = 0.06$) is larger than that of $x_{SAA}(S)$ (17.938) so this example shows that robustification and bias combine to yield a solution that is worse when tested out of sample. It is easy to confirm by repeating the above experiment that a negative δ, if chosen small enough, will improve out-of-sample performance. This means that a decision maker who pursues a mildly risk-seeking objective in this example will do better on average than one who acts as if neutral to risk.

The example is instructive for decision makers who are always risk averse in the hope of reducing variation in costs. In both examples the risk-averse action $x_R(S)$ has a lower variance than $x_{SAA}(S)$ when evaluated using the sample S, but might have a higher variance when evaluated using the true distribution, and this translates into higher out-of-sample expected costs. So applying a risk-averse model to sample data might give an action that increases risk out of sample. Whether this happens depends on the form of $c(x, y)$ and the characteristics of the underlying true distribution; one goal of this paper is to seek to understand this relationship.

The risk measure in the example is a *coherent* risk measure as defined by [1]. Optimizing with a coherent risk measure is a special case of *distributionally robust optimization* [24], in which the decision maker chooses x to solve

$$\text{DRO:} \quad \min_{x \in X} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} \left[c(x, Z) \right], \tag{3}$$

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 Z with distribution $\mathbb{Q}$.

In the setting of sample average approximation, we write ν_S for the sample distribution which has probability $1/N$ of being at each of the sample points in S, i.e. the uniform distribution on the finite set $S = \{y_1, y_2, ..., y_N\}$. Now $\mathcal{P}$ is defined to be a region $\mathcal{P}_\delta$ containing the sample distribution and parametrized by δ, so that it increases in size for larger δ and when $\delta = 0$ we have $\mathcal{P}_\delta = \{\nu_S\}$ and the distributionally robust optimization collapses back to the SAA problem of minimizing the expectation under ν_S of $c(x, y)$. When $\delta > 0$ the distributionally robust approach assumes that Nature plays against the decision maker selecting a distribution close to the sample distribution but designed so that the decision x gives a bad result in 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 ([18], [8]) 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, [6] constructs a confidence set for the first and second moments of $\mathbb{P}$ based on a sample, whereas [23] constructs $\mathcal{P}$ in terms of a likelihood function, and [3] chooses $\mathcal{P}$ to be the confidence region of a goodness-of-fit test.

A number of authors consider a DRO model where the set $\mathcal{P}_\delta$ is obtained from looking at distributions within a distance δ of the sample distribution under some metric on the space of distributions. One choice is to use ϕ-

divergence (such as the Kullback-Leibler divergence) to define the distance, and we give more details in the following section. 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 [2] give a tutorial discussion of the use of ϕ-divergence in this setting, and Shapiro [20] also discusses the different types of ϕ-divergence and their links with coherent risk measures. Gotoh, Kim and Lim 2018 [12] 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. [22] 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 [16] 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 [25]. The paper by Gao and Kleywegt [10] gives a comparison of the Wasserstein and ϕ-divergence approaches arguing for the better performance of the former and including some detailed comparisons on a newsvendor problem. An important consideration in the choice of approach is the computational burden involved in carrying out the inner maximization of DRO. The work by Esfahani and Kuhn [9] demonstrates how this can be done in the Wasserstein case for a wide variety of objective function forms.

Much of this work has been motivated by the search for policies that do well when applied to out-of-sample data. For example, as shown in [4], solutions to financial optimization problems are very sensitive to sampling errors in estimated returns. Out-of-sample performance of solutions to such problems is often much better when a distributionally robust approach is used [6],[25]. 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 a penalty will be paid for the reduction in variance making the average cost higher. 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 [9] 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. [11] when using Kullback-Leibler divergence in an inventory problem and a logistic regression problem. Luo and Mehrotra [15] 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. [11] show that in their setup a portfolio optimization problem never sees an improvement in mean.

The paper is laid out as follows. The next section establishes our notation and terminology, formally defines the marginal value of robust solution, and defines the three classes of robustification that we will study. Section 3 gives some results for sample average approximation. Section 4 studies robustification using a CVaR-based coherent risk measure (as in the example discussed above). In section 5 we revisit the results of the previous section when total variation is used to robustify P. Section 6 repeats this analysis for robustification using the Wasserstein distance. In section 7 we conclude the paper with some general observations. The proofs of all the propositions in the paper are deferred to an appendix.

2 Preliminaries

Our interest is in the solution of the stochastic optimization problem P using sample average approximation (1) and its distributionally robust version. Note $x_{SAA}(S)$ (the SAA solution) depends on the sample S. For N large it can be shown (see [21]) that $x_{SAA}(S)$ will approach the solution set of P. We use $C_{SAA}(S) = \mathbb{E}_{\mathbb{P}}[c(x_{SAA}(S), Y)]$, to denote the expected cost of $x_{SAA}(S)$ given the sample S. Taking expectations over $\mathbb{P}$ amounts to looking at the out-of-sample performance of the solution $x_{SAA}(S)$ under the real distribution.

A robust version of this problem generates a solution $x_R(S)$, 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 $x_R(S) = x_{SAA}(S)$. Fundamentally we are interested in the quality of the solution as measured by $C_R(S) = \mathbb{E}_{\mathbb{P}}[c(x_R(S), Y)]$ in comparison with the SAA alternative $C_{SAA}(S)$ as δ varies. Since the solution quality depends on what sample is chosen, we are interested in the expectations of $C_{SAA}(S)$ and $C_R(S)$ 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

$$VRS(\delta) = \mathbb{E}_S[C_{SAA}(S) - C_R(S)],$$

and the *marginal value of the robust solution* (MVRS) is

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

where this limit exists.

The value of $VRS(0)$ is zero, and we are interested in circumstances in which VRS(δ) is positive for small positive δ, which means that $x_R(S)$ performs better out of sample than $x_{SAA}(S)$. 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 Y. Thus, given a fixed distribution for the random variable Y, we begin by taking a sample of size N, construct our choice of decision variable $x_R(S)$ and then evaluate $c(x_R(S), Y)$ out of sample. We are interested in the distribution under $\mathbb{P}$ of the resulting costs $c(x_R(S), Y)$. In some circumstances this distribution will have reduced variance in comparison with the equivalent distribution that arises from the SAA solution $x_{SAA}(S)$.

It is important to be specific here about how we construct a robust version of the SAA problem. Suppose $c(x, y)$ can be separated into $\widetilde{c}(x, y) + d(y)$ where $\widetilde{c}(x, y)$ contains terms that depend on x and $d(y)$ does not. Then x^* solves P if and only if x^* solves

$$\tilde{\text{P}}: \quad \min_{x \in X} \mathbb{E}[\widetilde{c}(x, Y)]$$

and the optimal values differ by $\mathbb{E}[d(y)]$. Without loss of generality then we can assume that $c(x, y)$ contains no terms of the form $d(y)$. This will mean that SAA and the robust versions of SAA that we formulate can also be assumed to contain no such terms. Observe that this makes a difference when robustifying as robust versions of P and $\tilde{\text{P}}$ will in general have different solutions**.**

As we have observed, the robust problem is determined in terms of a set $\mathcal{P}_\delta$ of possible probability distributions for Y that are close to the nominal distribution ν_S having probability $1/N$ on each of the points in $S = \{y_1, y_2, ..., y_N\}$. There are various approaches to robustification depending on the form taken by $\mathcal{P}_\delta$. We consider three of these:

1. Robustification based on a coherent risk measure ρ solves

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

which can be reformulated as

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

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

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

which corresponds to $\mathcal{P}$ being a polyhedral set of probability measures $\mathcal{P}_\delta$ that depends on δ. For example when $\alpha = \frac{1}{N}$, $\mathcal{P}_\delta$ 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.

2. 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

$$d_\phi(\nu_q, \nu_p) = \sum_{i=1}^{N} p_i \phi \left(\frac{q_i}{p_i} \right) \tag{4}$$

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

$$\mathcal{P}_\delta = \{\nu : d_\phi(\nu, \nu_S) \leq \delta\}$$

or

$$\mathcal{P}_\delta = \{\nu : d_\phi(\nu_S, \nu) \leq \delta\}.$$

Note that in general this is not symmetric, so different sets $\mathcal{P}_\delta$ are obtained depending on whether ν_S is chosen to be ν_p or ν_q in (4). In this paper we consider a symmetric instance $\phi(t) = |t - 1|$ which gives $d_\phi(\nu_q, \nu_p) = \sum_{i=1}^{N} |q_i - p_i|$ (called total variation).

3. Distributionally robust optimization using a Wasserstein metric chooses

$$\mathcal{P}_\delta = \{\nu : d_W(\nu, \nu_S) \leq \delta\},$$

where $d_W(\nu, \nu_S)$ 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

$$d_W(\nu_1, \nu_2) = \min_{\gamma \in \Gamma(\nu_1, \nu_2)} \int_{M \times M} \|z_1 - z_2\| \; d\gamma(z_1, z_2) \tag{5}$$

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 can use this definition with different metrics corresponding to different transportation costs, but in our discussion we will use a Euclidean metric.

3 Sample average approximation

Suppose we seek to improve SAA using robustification. As we have seen in the introduction, the success of this will depend on the form of the cost function $c(x, y)$. We will assume throughout this paper that $c(x, y)$ is smooth and strictly convex in x for every realization y. Thus $\mathbb{E}_{\mathbb{P}}[c(x, Y)]$ is also strictly convex in x. For any y we can approximate $c(x, y)$ using a Taylor series expansion in x around $x_0 = \arg\min_x \mathbb{E}_{\mathbb{P}}[c(x, Y)]$, so

$$c(x, y) = c_0(y) + h(y)^\top (x - x_0) + \frac{1}{2}(x - x_0)^\top H(y)(x - x_0) + o(\|x - x_0\|^2),$$

where constants $c_0(y) \in \mathbb{R}$, $h(y) \in \mathbb{R}^n$ and $H(y)$ is a positive definite $n \times n$ matrix. The objective function we wish to minimize is then

$$\mathbb{E}_{\mathbb{P}}[c_0(Y) + h(Y)^\top (x - x_0) + \frac{1}{2}(x - x_0)^\top H(Y)(x - x_0) + o(\|x - x_0\|^2)]$$

which (neglecting higher order terms) is of the form

$$\mathbb{E}_{\mathbb{P}}\left[\frac{1}{2}x^\top H(Y)x + v(Y)^\top x + u(Y)\right] \tag{6}$$

where we have written $v(Y) = h(Y) - H(Y)x_0$ and $u(Y) = c_0(Y) + \frac{1}{2}x_0^\top H(Y)x_0$.

The Taylor series expansion is around x_0, which will not be available to us. But this discussion shows that close to the optimal x there is a quadratic approximation of the cost function in the form appearing in (6). This motivates us to consider cost functions of the form $c(x, y) = \frac{1}{2}x^\top H(y)x + v(y)^\top x + u(y)$. As mentioned above we will ignore the term $u(y)$ as it makes no difference to the optimization. In what follows, for different forms of robustification we study conditions on $\mathbb{P}$, H and v that guarantee positive MVRS for objective functions of this form.

For simplicity, we will take $X = \mathbb{R}^n$ in problem P. Then an optimal x^* that minimizes (6), will solve the first order conditions

$$\partial \mathbb{E}_{\mathbb{P}}[c(x, Y)]/\partial x_i = \mathbb{E}_{\mathbb{P}}[H(Y)x + v(Y)] = 0,$$

so

$$x^* = -\overline{H}^{-1}\overline{v}$$

where $\overline{H} = \mathbb{E}_{\mathbb{P}}[H(Y)]$ and $\overline{v} = \mathbb{E}_{\mathbb{P}}[v(Y)]$.

Given a sample $S = \{y_1, y_2, \ldots, y_N\}$ drawn from Y we have sample means $\overline{H}_S = (1/N)\sum_{i=1}^N H(y_i)$, $\overline{v}_S = (1/N)\sum_{i=1}^N v(y_i)$ and $\overline{u}_S = (1/N)\sum_{i=1}^N u(y_i)$. Thus we can calculate the SAA solution $x_{SAA}(S)$ which minimizes $(1/N)\sum_{i=1}^N c(x, y_i)$. Taking derivatives we get

$$\overline{H}_S x_{SAA}(S) + \overline{v}_S = 0,$$

so that $x_{SAA}(S) = -\overline{H}_S^{-1}\overline{v}_S$. The fact that both $\overline{H}_S$ and $\overline{v}_S$ depend on the sample will imply that this is typically a biased estimate.

Example 1 (SAA with bias)**:**

Suppose that $c(x, y) = (1 + y)(x^2 - 20xy)$, and Y is uniform on $[0, 1]$. We have

$$\begin{aligned} \mathbb{E}[c(x, Y)] &= \int_0^1 (1 + y)(x^2 - 20xy)dy \\ &= \frac{3}{2}x^2 - \frac{50}{3}x \end{aligned}$$

which is minimized at $x^* = \frac{50}{9} = 5.556$. To consider an extreme case suppose that there is a sample size of $N = 1$. So the optimal solution varies according to a single sample point y_1 and is given by $x_{SAA}(\{y_1\})$. By taking derivatives we have

$$x_{SAA}(\{y_1\}) = 10y_1$$

and the expected value of this is given by

$$\int_0^1 10y_1 dy_1 = 5.$$

Hence there is a negative bias. This is reduced as the sample size increases (for example the expected value of $x_{SAA}(S)$ when $N = 10$ is 5.5038). We have seen in the introduction how solving a CVaR-based robust version of the problem will lead to solutions with a lower value, i.e. $x_R(S) < x_{SAA}(S)$. Though we will not discuss the details here, it turns out that the same is true in this case and a robust solution will tend to increase the negative bias of $x_{SAA}(S)$ and make the solution worse. Numerically, we can check the case with sample size $N = 10$, and $\alpha = 0.1$ to show that the expected value of $x_R(S)$ becomes smaller as δ increases from 0, and $\text{VRS}(\delta) < 0$ for all $\delta \in (0, 1]$. □

Henceforth in the paper we will restrict attention to the case where H does not depend on y, (and $u(y)$ does not appear) so

$$c(x,y) = \frac{1}{2}x^\top Hx + v(y)^\top x \tag{7}$$

with $v(y) \in \mathbb{R}^n$ and H a symmetric positive definite matrix. The analysis above gives $x^* = -H^{-1}\overline{v}$ and $x_{SAA}(S) = -H^{-1}\overline{v}_S$, which is unbiased since $\mathbb{E}_S[\overline{v}_S] = \overline{v}$. We can easily quantify the expected additional cost introduced by using the sample average approximation.

Lemma 1 *If* $c(x,y) = \frac{1}{2}x^\top Hx + v(y)^\top x$ *then* $C^* = -\frac{1}{2}\overline{v}^\top H^{-1}\overline{v}$, *and* $C_{SAA}(S) = C^* + \frac{1}{2}(\overline{v} - \overline{v}_S)^\top H^{-1}(\overline{v} - \overline{v}_S)$.

Lemma 1 simplifies in the scalar case, where we assume $c(x,y) = x^2 - g(y)x$, and denote $\mathbb{E}_{\mathbb{P}}(g(Y))$ by $\bar{g}$. Then we have $H = 2$, $\bar{u} = 0$, and $\overline{v} = -\bar{g}$, so $x^* = \bar{g}/2$, and by Lemma 1, $C^* = -\bar{g}^2/4$. Given a sample $S = \{y_1, y_2, \ldots, y_N\}$ let $\overline{g}_S = (1/N)\sum_{i=1}^N g(y_i)$, and $\overline{v}_S = -\overline{g}_S$. Lemma 1 then gives

$$\begin{aligned} C_{SAA}(S) &= C^* + \frac{1}{2}(\overline{v} - \overline{v}_S)^\top H^{-1}(\overline{v} - \overline{v}_S) \\ &= C^* + \frac{1}{4}(\bar{g} - \overline{g}_S)^2. \end{aligned}$$

Hence the expected additional cost from using the sample average approximation is $(1/4)\mathbb{E}_S[(\overline{g}_S - \overline{g}))^2]$ where the expectation is taken over samples S.

In the following sections we will explore what happens when we compute robust solutions to P using the three different approaches outlined in the introduction.

4 CVaR based robustness

Given the sample S, we consider a risk-averse version of SAA

$$\text{RSAA}(\delta) : \min_x \rho[c(x,S)]$$

in which $\mathbb{E}[c(x,S)]$ is replaced by evaluation with the coherent risk measure

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

where $\delta \in [0,1]$ and for $\alpha \in (0,1]$, $\text{CVaR}_{1-\alpha}[c(x,S)]$ is the conditional value at risk at level $1-\alpha$ of the discrete distribution $\{c(x,y_i) : y_i \in S\}$.

We will make use of the first order condition for the robust solution $x_R(S)$ to RSAA(δ), which is unique in the case that $\delta < 1$ from our assumptions on strict convexity of c. The following result captures these conditions for the quadratic case.

Lemma 2 *Suppose* $c(x,y) = \frac{1}{2}x^\top Hx + v(y)^\top x$. *The solution to RSAA($\delta$) satisfies*

$$x_R(S) \in -H^{-1}\left((1-\delta)\overline{v}_S + \delta G_{CVaR}(x_R(S))\right) \tag{8}$$

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

$$x_R(S) = -H^{-1}\left((1-\delta)\overline{v}_S + \delta\overline{v}_{CVaR}\right). \tag{9}$$

In the lemma below we are concerned with the gradient of CVaR at the point $x_{SAA}(S) = -H^{-1}\overline{v}_S$. In the case that $\text{CVaR}_{1-\alpha}\left[\{v(y_i)^\top x\}\right]$ is differentiable at $x_{SAA}(S)$ then we write $\overline{v}_{\text{CVaR}}$ for this vector. In the case that $\text{CVaR}_{1-\alpha}\left[\{v(y_i)^\top x\}\right]$ is not differentiable at $x_{SAA}(S)$ then we can define $\overline{v}_{\text{CVaR}}$ as an element of the subdifferential at $x_{SAA}(S)$ (we give details in the proof). In our examples we will consider continuous distributions for $\mathbb{P}$ and hence when taking expectations over samples S, the outcomes with CVaR not differentiable at $x_{SAA}(S)$ will have zero measure and the choice of $\overline{v}_{\text{CVaR}}$ at these points will not have an impact on MVRS.

Lemma 3 *Suppose* $c(x,y) = \frac{1}{2}x^\top Hx + v(y)^\top x$. *Then MVRS is well defined and*

$$MVRS \in \mathbb{E}_S[(\overline{v} - \overline{v}_S)^\top H^{-1}(\overline{v}_{CVaR} - \overline{v}_S)].$$

Corollary 4 *Suppose* $c(x,y) = x^2 - g(y)x$, *where* $g(Y)$ *has a continuous distribution with mean* $\bar{g}$ *and variance* σ^2, *and let* $\overline{g}_S = \frac{1}{N}\sum_{i=1}^N g(y_i)$ *and* $\alpha \in (0,1]$. *Then*

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

The expression for MVRS in the Corollary includes a CVaR term which takes the value zero in the event that $\overline{g}_S = 0$ and so $\text{sgn}(\overline{g}_S)g(y_i) = 0$, for each sample point. Since $g(Y)$ has a continuous distribution we can deduce that there is zero probability of the sample mean $\overline{g}_S$ being exactly zero. Thus this event will not affect the expectation $\mathbb{E}_S$.

In the scalar case where $c(x,y) = x^2 - g(y)x$, it is possible to derive a more explicit form of MVRS if we know the distribution of $g(y)$, and we can assume that $\overline{g}_S$ is always positive. Suppose $W = g(Y) - \bar{g}$ has a density $f(w)$ and cumulative distribution function $F(w)$. We define $Q(z) = \int_z^\infty w f(w) dw$, and for any $\alpha \in (0,1]$ we define the function

$$\begin{aligned}\Lambda_\alpha(z) &= \frac{1}{\alpha N} + \frac{1}{\alpha N}(N-1)\frac{F(z)}{(1-F(z))} \\ &\quad + \frac{1}{\alpha N}\frac{(N-1)(N-2)}{2}\frac{F(z)^2}{(1-F(z))^2} \\ &\quad + \ldots + (1 - \frac{m-1}{\alpha N})\left(\begin{array}{c} N-1 \\ m-1 \end{array}\right)\frac{F(z)^{m-1}}{(1-F(z))^{m-1}},\end{aligned}$$

where m is the unique integer for which $\alpha \in (\frac{m-1}{N}, \frac{m}{N}]$. This gives the following result.

Proposition 5 *Suppose $c(x,y) = x^2 - g(y)x$ where $\overline{g}_S > 0$, and we solve RSAA(δ) where $\alpha \in (0,1]$. Then*

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

There are some observations we can make in relation to the condition $\overline{g}_S > 0$. This is included in order to ensure that $x_{SAA}(S) > 0$ and hence that it is the left rather than right tail of the $g(Y)$ distribution that appears in the CVaR term. We can usually assume that $\overline{g}_S$ is close to the mean of the $g(Y)$ distribution for reasonable sample sizes. This is often enough to make the probability of $\overline{g}_S < 0$ 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 $\overline{g}_S < 0$ with probability close to 1. When this happens there are alternative formulae (which we will not give here) obtained through defining $W = \bar{g} - g(Y)$.

We now study some examples of MVRS. The formula in Proposition 5 shows that MVRS will be positive if the second term is small. There is a connection here to the skew in the distribution of $g(Y)$. We consider an example with a large right-hand skew and show that MVRS is positive.

Example 2 (exponential distribution):

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

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

When $N = 10$, this gives a value of MVRS equal to $\frac{9}{200}$. □

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.

Example 3 (uniform distribution)**:**

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

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

which is zero when $N = 2$, and positive when $N > 2$. When $N = 10$, this gives a value of $\frac{a^2}{60} - \frac{a^2}{132} = \frac{a^2}{110}$. □

Example 4 (normal distribution)**:**

Consider univariate and multivariate examples with a normal distribution. In the univariate case suppose that $g(Y)$ 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 $f(z) = \frac{1}{\sigma\sqrt{2\pi}}\exp(\frac{-z^2}{2\sigma^2})$. 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

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

In the Appendix, Lemma 18, we show that

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

which gives MVRS = 0.

The same result can be obtained in the multivariate case when there is rotational symmetry for the distribution, and H is the identity matrix. Thus we consider the case where $v(y)$ is a vector where each component is drawn

from an independent $N(0, \sigma^2)$ distribution all with the same variance σ^2. Thus $\overline{v} = 0$ and

$$\begin{aligned} \text{MVRS} &= \mathbb{E}_S[(\overline{v} - \overline{v}_S)^\top (\overline{v}_{\text{CVaR}} - \overline{v}_S)] \\ &= \mathbb{E}_S[(-\overline{v}_S)^\top (\overline{v}_{\max} - \overline{v}_S)] \end{aligned}$$

where $\overline{v}_{\max}$ is the v value from the sample with the highest value for $v^\top x$. Because of rotational symmetry this expectation is independent of the choice of (non-zero) vector x. Hence we consider $x = e_1$ and we can set $\overline{v}_{\max}$ to be the v value from the sample with the highest value for v_1. Writing $\overline{v}_{S,i}$ and $\overline{v}_{\max,i}$ for the i'th components, we have

$$\text{MVRS} = \mathbb{E}_S[-\overline{v}_{S,1}(\max_{y \in S}\{v_1(y)\} - \overline{v}_{S,1})] - \sum_{j=2}^{n} \mathbb{E}_S[\overline{v}_{S,j}(\overline{v}_{\max,j} - \overline{v}_{S,j})].$$

The first term here matches the univariate analysis and is zero. We may reorder the sample elements so that $\overline{v}_{\max} = v(y_1)$. Then

$$\mathbb{E}_S[\overline{v}_{S,j}(\overline{v}_{\max,j} - \overline{v}_{S,j})] = \mathbb{E}_S[\sum_{i=1}^{N}(1/N)v_j(y_i)(v_j(y_1) - \sum_{i=1}^{N}(1/N)v_j(y_i))].$$

Using independence between $v_j(y_i)$ and $v_j(y_k)$ for $i \neq k$ we have

$$\mathbb{E}_S[\overline{v}_{S,j}(\overline{v}_{\max,j} - \overline{v}_{S,j})] = \mathbb{E}_S[(1/N)v_j(y_1)^2] - \mathbb{E}_S[\sum_{i=1}^{N}(1/N^2)v_j(y_i)^2] = 0.$$

Hence MVRS $= 0$ in this case too. □

The analysis here is specific to the multivariate normal where rotational symmetry is achieved at the same time as independence between different components.

Example 5 (mixture of univariate normal distributions):

We consider a case where Y is univariate and $g(Y)$ is formed as a mixture of two normal distributions having the same mean (large enough to ensure that $\bar{g}_S > 0$). 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 $\sigma^2 = \int_{-\infty}^{\infty} w^2 f(w) dw = \frac{(\sigma_1^2 + \sigma_2^2)}{2}$,

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

and

$$\begin{aligned} Q(z) &= (1/2)\int_z^\infty w\,(f_1(w)+f_2(w))\,dw \\ &= (1/2)(Q_1(z)+Q_2(z)) \\ &= (1/2)(\sigma_1^2 f_1(z)+\sigma_2^2 f_2(z)). \end{aligned}$$

Taking $\alpha = 1/N$, we obtain

$$\begin{aligned} \text{MVRS} &= \frac{(\sigma_1^2+\sigma_2^2)}{4N} - \frac{1}{4}\int_{-\infty}^{\infty}(\sigma_1^2 f_1(z)+\sigma_2^2 f_2(z))(1-(F_1(z)+F_2(z))/2)^{N-1}dz \\ &= \frac{(\sigma_1^2+\sigma_2^2)}{4N} - \frac{1}{4}\int_{-\infty}^{\infty}\left(\frac{\sigma_1}{\sqrt{2\pi}}\exp(\frac{-z^2}{2\sigma_1^2})+\frac{\sigma_2}{\sqrt{2\pi}}\exp(\frac{-z^2}{2\sigma_2^2})\right) \\ &\quad \times\left(1-\frac{1}{2\sqrt{2\pi}}\int_{-\infty}^{z}\left(\frac{1}{\sigma_1}\exp(\frac{-u^2}{2\sigma_1^2})+\frac{1}{\sigma_2}\exp(\frac{-u^2}{2\sigma_2^2})\right)du\right)^{N-1}dz. \end{aligned}$$

We can evaluate MVRS numerically. For example if $\sigma_1 = 1$, $\sigma_2 = 2$ and $N = 3$ we obtain MVRS $= -2.166\,3 \times 10^{-2}$, and if $N = 5$, MVRS $= -3.748\,0 \times 10^{-2}$.

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

5 Total variation

As before we consider the problem P with

$$c(x,y) = \frac{1}{2}x^\top Hx + v(y)^\top x.$$

In this section we consider evaluating MVRS when robustifying the SAA solution using a ϕ-divergence approach with L_1 norm, often called total variation. Thus the inner optimization solves

$$\max_{p\in\mathcal{P}_\delta}\mathbb{E}[\sum_{i=1}^N p_i v(y_i)^\top x]$$

where

$$\mathcal{P}_\delta = \{p : \sum_{i=1}^N \left|p_i - \frac{1}{N}\right| \le \delta\}.$$

In the case where each of the N observations $y_1, y_2, ..., y_N$ has a different value for $v(y_i)^\top x$, and $\delta < 2/N$ it is easy to see that the inner maximization will involve taking probability $\delta/2$ from the y_i with the lowest $c(x, y_i)$ value in the sample, $y_{\min} = \arg\min_{y_i} v(y_i)^\top x$, and moving it to the y_i with the highest $c(x, y_i)$ value, $y_{\max} = \arg\max_{y_i} v(y_i)^\top x$.

This robustification is similar to the CVaR approach of the previous section in the special case $\alpha = 1/N$, but instead of moving weight from all the points in the sample to the worst point, we move weight from the best point in the sample to the worst point. The solution we obtain, $x_R(S)$, satisfies first order conditions

$$Hx + (1/N)\sum_{i=1}^{N}(v(y_i)) + (\delta/2)(v(y_{\max}) - v(y_{\min})) = 0,$$

so

$$x_R(S) = -H^{-1}(\overline{v}_S + \delta R_S/2),$$

where $R_S = v(y_{\max}) - v(y_{\min})$.

In the same way that we allowed for points where CVaR was non-differentiable, we can drop our assumption on different values for each $v(y_i)^\top x$ and consider the possibility that $\arg\min_{y_i} v(y_i)^\top x$ or $\arg\max_{y_i} v(y_i)^\top x$ might have more than one element. Then the (unique) optimum $x_R(S)$ is a solution of

$$Hx + (1/N)\sum_{i=1}^{N}(v(y_i)) + (\delta/2)(\overline{v}_{\max}(x) - \overline{v}_{\min}(x)) = 0,$$

where $\overline{v}_{\max}(x) \in \text{conv}(\{v(y_i) : i \in \arg\max_i\{v(y_i)^\top x\}\})$, $\overline{v}_{\min}(x) \in \text{conv}(\{v(y_i) : i \in \arg\min\{v(y_i)^\top x\}\})$. Now $x_R(S)$ depends on δ and we set

$$R_S^{(\delta)} = -(2/\delta)(Hx_R(S) + \overline{v}_S)$$

and observe that $R_S^{(\delta)} = \overline{v}_{\max}(x_R(S)) - \overline{v}_{\min}(x_R(S))$. We let $R_S = \lim_{\delta\to 0} R_S^{(\delta)}$ (with a similar argument to that in the proof of Lemma 3 to show that this limit exists). Note that $x_R(S) \to x_{SAA}(S)$ and $R_S = \overline{v}_{\max} - \overline{v}_{\min}$ for some $\overline{v}_{\max} \in \text{conv}(\{v(y_i) : i \in \arg\max_i\{v(y_i)^\top x_{SAA}(S)\}\})$, $\overline{v}_{\min} \in \text{conv}(\{v(y_i) : i \in \arg\min\{v(y_i)^\top x_{SAA}(S)\}\})$.

Lemma 6 *Suppose $c(x, y) = \frac{1}{2}x^\top Hx + v(y)^\top x$. Then MVRS is well defined and*

$$MVRS = \mathbb{E}_S[(1/2)(\overline{v} - \overline{v}_S)^\top H^{-1} R_S]. \tag{10}$$

From the Lemma

$$\text{MVRS} = \mathbb{E}_S[(\overline{v} - \overline{v}_S)^\top H^{-1} (\overline{v}_{\max} - \overline{v}_{\min})]$$

for some $\overline{v}_{\max} \in \text{conv}(\{v(y_i) : i \in \arg\max_i \{v(y_i)^\top x_{SAA}(S)\}\})$, and $\overline{v}_{\min} \in \text{conv}(\{v(y_i) : i \in \arg\min\{v(y_i)^\top x_{SAA}(S)\}\})$. With continuous distributions except on a set of measure zero these sets are singletons and we have $\overline{v}_{\max} = \arg\max_{v(y_i)}\{v(y_i)^\top x_{SAA}(S)\})$, $\overline{v}_{\min} = \arg\min_{v(y_i)}\{v(y_i)^\top x_{SAA}(S)\})$. Thus the more complex definitions will not have an impact in evaluating MVRS as an expectation over samples.

In the scalar case where $c(x, y) = x^2 - g(y)x$, we have $H = 2$ and $v(y) = -g(y)$, and we may take $S = \{y_1, y_2, \ldots, y_N\}$ ordered so that

$$g(y_1) \leq g(y_2) \leq \ldots \leq g(y_N).$$

From (10) with $H = 2$ we get

$$\text{MVRS} = \frac{1}{4}\mathbb{E}_S[(\overline{g}_S - \overline{g})R_S].$$

Hence, writing $\overline{R} = \mathbb{E}_S[R_S]$

$$\begin{aligned}
\mathbb{E}_S[C_R(S)] &= \mathbb{E}_S[C_{SAA}(S)] + (\delta/4)\mathbb{E}[(\overline{g} - \overline{g}_S)R_S] + \left(\delta^2/16\right)\mathbb{E}_S[R_S^2] \\
&= \mathbb{E}_S[C_{SAA}(S)] + (\delta/4)\mathbb{E}_S[(\overline{g} - \overline{g}_S)(R_S - \overline{R})] \\
&\quad + \left(\delta^2/16\right)\mathbb{E}[R_S^2] + (\delta/4)\mathbb{E}_S[(\overline{g} - \overline{g}_S)]\overline{R} \\
&= \mathbb{E}_S[C_{SAA}(S)] - (\delta/4)\text{cov}(\overline{g}_S, R_S) + \left(\delta^2/16\right)\mathbb{E}_S[R_S^2].
\end{aligned}$$

In the case that the distribution of prices $g(y)$ is symmetric about its mean then we can condition on R_S and observe that for any sample with prices $\{g(y_1), g(y_2), ..., g(y_N)\}$ there is another sample with prices $\{2\overline{g} - g(y_1), 2\overline{g} - g(y_2), ..., 2\overline{g} - g(y_N)\}$ which is equally likely, in which each price is replaced by a price at the same distance but on the opposite side of $\overline{g}$. This mirror sample has the same range but $(\overline{g} - \overline{g}_S)$ is reversed in sign since $\overline{g} - g(y_i)$ is replaced by $\overline{g} - (2\overline{g} - g(y_i)) = g(y_i) - \overline{g}$. From this we deduce that MVRS is zero, and thus $\mathbb{E}_S[C_R(S)] = \mathbb{E}_S[C_{SAA}(S)] + \left(\delta^2/16\right)\mathbb{E}_S[R_S^2]$. So, in this case, robustification always makes things worse. However when there is a skew in the distribution of prices we can expect to see $\text{cov}(\overline{g}_S, R_S) \neq 0$. For small δ this is the dominant term and will determine whether MVRS is positive or negative.

A key observation is that when there is a right skew in the distribution of $g(y)$ 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 implies that the range is

positively correlated with the mean, and hence $\mathbb{E}_S[(\overline{g}-\overline{g}_S)R_S] < 0$. A robust solution takes weight from a high outlier and moves it to the lowest value. On average these moves improve the solution.

As in the previous section, we will work with the random variable $W = g(Y) - \overline{g}$ which has mean 0. Let W have density $f(w)$ and cumulative distribution function $F(w)$, and recall $Q(z) = \int_z^\infty wf(w)dw$.

Lemma 7 *Suppose $c(x, y) = x^2 - g(y)x$ where $\overline{g}_S > 0$, then total variation robustification gives*

$$MVRS = \frac{1}{4}\int_{-\infty}^{\infty} \left(F(z)^{N-1} - (1 - F(z))^{N-1}\right) Q(z)dz. \tag{11}$$

We already gave an argument to show that $\mathbb{E}_S[(\overline{g}_S - \overline{g})R_S] = 0$ when W has a symmetric distribution. Now consider a distribution F for W with a right skew. It is possible to precisely identify a set of distributions where the right skew will guarantee a positive value for MVRS. Note that the condition that we derive for MVRS > 0 in this total variation case has no equivalent for the CVaR form of robustification.

We take the point z_0 where $F(z_0) = 1/2$ and construct a symmetric distribution $\widetilde{F}$ where $\widetilde{F}(z_0 + y) = 1 - \widetilde{F}(z_0 - y)$. Let $\widetilde{f}$ be the density for $\widetilde{F}$. Then we write $\tau(z)$ for the translation in z that moves $\widetilde{F}$ onto F, so that $\widetilde{F}(z) = F(\tau(z))$. Thus our construction ensures that $\tau(z) \geq z$ with equality when $z \leq z_0$. As usual we will assume that $\overline{g}_S > 0$.

Proposition 8 *If τ is differentiable with $\tau'(z) \geq 1$ for z in the support of F then MVRS ≥ 0 with strict inequality if F is not identical to $\widetilde{F}$.*

6 Wasserstein metric

We turn now to the third robustness approach that is commonly used, where the uncertainty sets are δ balls in a Wasserstein metric as defined by (5). We are interested in the case 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 ν_S (which has equal probabilities at each of the sample points $y_1, y_2, ..., y_N$). This gives the following inner problem:

$$\begin{array}{ll} \mathrm{P}_{\mathrm{inner}} : \max_\nu & \mathbb{E}_\nu[c(x, Z)] \\ \text{subject to} & d_W(\nu, \nu_S) \leq \delta \end{array}$$

in which the expectation is taken over the random variable Z with distribution ν.

In the previous two sections the structure of the cost function with respect to the random variable Y has not been critical; everything has been determined by the set of cost functions $c(x, y_i)$ evaluated at the sample points. With Wasserstein we will consider moves in the sample points y_i and we need to pay much more attention to the behavior of $c(x, y)$ with respect to changes in y. Often we will take x fixed and it is convenient to write $c_x(y)$ for $c(x, y)$. Using (5), the inner maximization problem is equivalent to solving

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

Since $\gamma \in \Gamma(\nu, \nu_S)$ 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 y_i is moved to under γ. More precisely we can rewrite $\gamma \in \Gamma(\nu, \nu_S)$ in terms of components γ_i that are measures on M with $\gamma_i = \gamma(\cdot, y_i)$. Since ν_S has mass $1/N$ at y_i we have $\gamma_i(M) = 1/N$, and the measure ν is simply formed by adding together the components from each sample point, so that $\nu = \sum_{i=1}^N \gamma_i$.

By writing $\nu_i = N\gamma_i$ for the probability distribution on M (with the scaling of N applied so that total mass of ν_i is 1) we can write this problem as

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

where Z_i is a random variable with distribution ν_i.

We make use of a result of Gao and Kleywegt [10, Corollary 2, part iii].

Proposition 9 *(Gao and Kleywegt) If there is an optimal solution to $\bar{P}$ then there is an optimal solution where some particular index i_0 has the property that for every $i \neq i_0$, ν_i has weight 1 on a point $z_i^* \in \arg\max_{z\in M}\{c_x(z) - \lambda^* \|z - y_i\|\}$ where $\lambda^* \geq 0$ is the Lagrange multiplier for the constraint in $\bar{P}$, and ν_{i_0} has weight on at most two points in $\arg\max_{z\in M}\{c_x(z) - \lambda^* \|z - y_{i_0}\|\}$.*

There are two cases (c_x is concave in y and c_x is convex in y) where we can be more explicit about the solution of $\bar{\text{P}}$. In the concave case we can think about contour surfaces in M which have the same gradient modulus. The solution to $\bar{\text{P}}$ has such a surface where all the points outside the surface are moved inwards to lie on that surface and points inside the surface are not moved.

Proposition 10 *When $c_x(y)$ is concave then $\bar{P}$ has a solution in which each ν_i has support at a single point $z_i \in M$. If we write $\overline{J} = \{i : z_i \neq y_i\}$ for the points that move, then (a) $\nabla c_x(z_i) = \alpha_i(z_i - y_i)$ for some scalar α_i for $i \in \overline{J}$; (b) $\|\nabla c_x(z_i)\| = \|\nabla c_x(z_j)\|$ for $i \in \overline{J}$ and $j \in \overline{J}$; and (c) $\|\nabla c_x(z_i)\| \geq \|\nabla c_x(y_k)\|$ for $i \in \overline{J}$ and $k \notin \overline{J}$.*

When we do not have a concave cost function c_x then the types of move that occur are in general more complex. The resulting behavior will be less appropriate for the underlying problem and robustification using a Wasserstein metric may behave poorly. In order to give insight into the type of behavior that may occur we analyze the extreme case when c_x is convex.

It is helpful to consider the slope (in terms of costs) of the line from y_i to $z \in M$, i.e. the slope of the line from $(y_i, c_x(y_i))$ to $(z, c_x(z))$. We write $\varphi_i(z)$ for this so $\varphi_i(z) = (c_x(z) - c_x(y_i))/\|z - y_i\|$. The result below shows that when $c_x(y)$ is convex then the inner maximization sends weight at y_i to a point z_i on the boundary of the region M starting with the point y_i with the highest value of $\max_{z \in M}(\varphi_i(z))$ (over i), and moving on at each step making changes to the points where this maximum slope is greatest and continuing until the distance limit is reached. This is achieved by showing that the points y_i can be ordered by $\max_{z \in M}(\varphi_i(z))$ with only the points with higher values being moved. Note that we may have one ν_i having support on two different points, and these may both be at the boundary of M.

Proposition 11 *When $c_x(y)$ is convex then the optimal solution to $\bar{P}$ has each ν_i with mass either at y_i or at the boundary of M (or both). If ν_i has mass at y_i and for some other j we have $\nu_j \neq y_j$ then $\max_{z \in M} \varphi_i(z) \leq \varphi_j(\nu_j)$.*

We observe that using a robust approach with the Wasserstein metric is unlikely to do well when $c_x(y)$ is convex. The result of adding weight to points at the boundary of M will often depend on some arbitrary choices on how the set M of all possible y values is defined. Moreover even when the bound on the Wasserstein distance is small the points introduced into the sample may be far from the existing points. Finally we note that for small bounds on the Wasserstein distance the first point to be changed is defined by the slope maximum, which depends on the way that M is chosen, rather than being closely related to the characteristics of the point itself. So we will look at the case where $c(x, y)$ is concave in y.

Now we wish to calculate the value of MVRS for the Wasserstein metric. As before we assume that $c(x, y) = \frac{1}{2}x^\top Hx + v(y)^\top x$. Note that $\nabla c_x(y) = \sum_j x_j \nabla v_j(y)$ and $c_x(y)$ is concave if each component of v is concave. We will assume that the sample $S = \{y_1, y_2, ...y_N)$ has each

point with a different value for $\left\|\sum_j x_j \nabla v_j(y_i)\right\|$ and $\|\nabla v_j\|$ is continuous. Then for small δ we will move just one point, we deduce this from Proposition 10 part (b), since for small δ it is impossible for two different points to end up with the same value for $\|\nabla c_x(y_i)\|$ without moving a combined distance more than δ. Moreover part (c) of the Proposition shows that the point that is moved is y^*, the sample point with the highest gradient norm, $\left\|\sum_j x_j \nabla v_j(y^*)\right\| = \max_i \left\|\sum_j x_j \nabla v_j(y_i)\right\|$. The solution we obtain after robustification, given a distance limit δ, moves y^* to a point $y^* + N\delta \frac{\sum_j x_j \nabla v_j(y^*)}{\left\|\sum_j x_j \nabla v_j(y^*)\right\|}$, where the term $N\delta$ arises from the way that we define the Wasserstein distance. So the optimal solution, x, for DRO satisfies

$$Hx + (1/N)\sum_{i=1}^{N}(v(y_i)) - (1/N)v(y^*) + (1/N)v\left(y^* + N\delta \frac{\sum_j x_j \nabla v_j(y^*)}{\left\|\sum_j x_j \nabla v_j(y^*)\right\|}\right) = 0.$$

It is simpler to deal with the scalar case where we have $c(x,y) = x^2 - g(y)x$ and this becomes

$$2x - (1/N)\sum_{i=1}^{N} g(y_i) + (1/N)g(y^*) - (1/N)g\left(y^* - N\delta \frac{\nabla g(y^*)}{\|\nabla g(y^*)\|}\right) = 0.$$

But for small δ

$$g\left(y^* - N\delta \frac{\nabla g(y^*)}{\|\nabla g(y^*)\|}\right) - g(y^*) = -N\delta \|\nabla g(y^*)\| + o(\delta),$$

so to first order we have

$$x_R(S) = \frac{1}{2N}\sum_{i=1}^{N} g(y_i) - \frac{\delta}{2}\|\nabla g(y^*)\|.$$

Proposition 12 *(a) When $c(x,y) = x^2 - g(y)x$ and g is a convex function of y then*

$$MVRS = \mathbb{E}_S[(\overline{g}_S - \overline{g})\|\nabla g(y_S^*)\|]/2$$

where $y_S^* = \arg\max_{y_i \in S}\{\|\nabla g(y_i)\|\}$ and $\overline{g}_S = (1/N)\sum_{i=1}^{N} g(y_i)$.
(b) In the case that $c(x,y) = x^2 - y^2 x$ and the y values are realizations of a random variable Y which is non-negative and has density and cdf given by f and F, then

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

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 size of the largest g absolute derivative. There is very similar behavior here to that we have seen in other cases. If there is a skew in the underlying distribution of y towards values with high values for $g(y)$, then we can expect to see samples where there is an outlier producing both a high value for $\overline{g}_S - \overline{g}$ and also a high value for $\|\nabla g(y_S^*)\|$. 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 $c(x, y) = x^2 - y^2 x$ and the underlying distribution of the random variable Y is exponential with mean 1, so $f(u) = e^{-u}$, $F(u) = 1 - e^{-u}$.

Thus

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

since $\int_0^\infty u^2 e^{-u}du = 2$. When $N = 5$ we can numerically evaluate the integrals and obtain MVRS = 1.2488. □

7 Conclusions and discussion

The application of robustification to stochastic optimization problems to improve out-of-sample performance has been widely reported in the literature. This paper contributes to our understanding of why this is the case. It also identifies circumstances in which robustification will make average out-of-sample performance deteriorate. We have defined the MVRS parameter as a first order measure of this improvement, and calculated this for a number of univariate examples. MVRS depends on the form of the objective function, the version of robustification applied, and the underlying "ground-truth" probability distribution.

The comparisons we make are with sample average approximation which makes no assumptions on the underlying probability distribution. In the context of improving out-of-sample performance SAA already does well in comparison with parametric methods which are more likely to overfit. Though

robustification may be valuable from a risk reduction point of view, our work demonstrates that it may also have value for a risk neutral decision maker.

There are often circumstances when a decision maker has some knowledge of the underlying distribution that can be helpful in predicting how robustification will perform. Two characteristics of the distribution are particularly relevant: Is the distribution symmetric or skewed? And does the distribution have heavy tails?

To understand the impact of small amounts of robustification of different forms we can summarize the changes made on the SAA problem as follows:

1. For the 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.

2. 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.

3. For Wasserstein robustification, provided $c(x, y)$ is concave in y, the sample point with the largest value for the norm of the gradient with respect to y is moved incrementally to a higher cost position (the exact move depends on the function c).

It is simplest to interpret our results in a univariate framework, when we have $c(x, y) = x^2 - g(y)x$. The value of $x_{SAA}(S)$ is $\overline{g}_S/2$. Since each of the different robustification approaches move weight to lower values of $g(y_i)$ (corresponding to higher costs) we have $x_R(S) < x_{SAA}(S)$. Though this introduces a bias in the value of $\mathbb{E}_S[x_R(S)]$ we can obtain improvement through shrinkage when there are larger moves to the left for samples with high values of $\overline{g}_S$ (and hence high values for $x_{SAA}(S)$) than there are for samples with low values of $\overline{g}_S$ (and hence low values for $x_{SAA}(S)$). 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. This observation provides some intuition explaining why we end up with MVRS having the form $\mathbb{E}_S[(\overline{g}_S - \overline{g})(x_{SAA}(S) - x_R(S))/\delta]$.

For CVaR robustification the change in optimal solution, $x_{SAA}(S) - x_R(S)$, 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 $g(y_i)$. This produces the term $\sigma^2/(2N)$ 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 $g(Y)$. 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 $\overline{g}_S - \overline{g}$ and $x_{SAA}(S) - x_R(S)$. This effect works in the opposite direction to the $\sigma^2/(2N)$ term.

Examples for CVaR robustification show that when the distribution is uniform over an interval, the $\sigma^2/(2N)$ term dominates and MVRS is positive; when the distribution is normal the effect from the left hand tail exactly cancels the positive term and MVRS is 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.

When we consider the total variation form of the robustification it is only the tails that influence the change that is made, and $(x_{SAA}(S) - x_R(S))/\delta$ is simply half the range of values in the sample. Here any skew to the right in the distribution of $g(Y)$ will induce a correlation that yields a positive value for MVRS. We note that MVRS is zero for symmetric distributions under the total variation robustification, which does not hold for the other two types of robustification.

For the Wasserstein robustification and convex $g(y)$ the point where g has the highest gradient is moved. This will be a point towards the extremities of the y_i values (that in general occur in a multivariate space) - and hence is likely to be where $g(y_i)$ is large and so costs are low. In the special case of y scalar and $g(y) = y^2$ 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 y^2 has a positive skew.

The sign of MVRS has been our focus in this discussion. We need to be cautious in directly comparing values of MVRS between different robustifications. The value will clearly be determined by the way that changes are parameterized by δ, and there is an arbitrariness in this.

There are a number of aspects in which it would be desirable to extend our discussion. First our primary application is to historical data, it would be natural to consider auto-correlation between sample points, and its implications for the methods we consider, but this lies beyond the scope of the present paper. Second our analysis has been restricted to the optimization of smooth strictly convex functions, and specifically to the case where the SAA approach is unbiased. A more complete treatment would consider cases where cost functions are less well-behaved. Finally our concrete classifications are essentially restricted to the univariate case, and it would be valuable to extend this work to give a more comprehensive treatment of multivariate

problems.

References

[1] P. Artzner, F. Delbaen, J. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

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

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

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

[5] J. B. Copas. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3):311–354, 1983.

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

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

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

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

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

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

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

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

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

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

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

[17] R.T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

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

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

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

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

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

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

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

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

Appendix : Proofs of propositions

Proof of Lemma 1

Using $x^* = -H^{-1}\overline{v}$ and $x_{SAA}(S) = -H^{-1}\overline{v}_S$ we can deduce

$$C^* = \mathbb{E}_{\mathbb{P}}[c(x^*, y)] = \frac{1}{2}(x^*)^\top H x^* + \overline{v}^\top x^* = -\frac{1}{2}\overline{v}^\top H^{-1}\overline{v},$$

and

$$\begin{aligned} C_{SAA}(S) &= \mathbb{E}_{\mathbb{P}}[c(x_{SAA}(S), y)] = \frac{1}{2}x_{SAA}(S)^\top H x_{SAA}(S) + \overline{v}^\top x_{SAA}(S) \\ &= \frac{1}{2}\overline{v}_S^\top H^{-1}\overline{v}_S - \overline{v}^\top H^{-1}\overline{v}_S \\ &= C^* + \frac{1}{2}(\overline{v} - \overline{v}_S)^\top H^{-1}(\overline{v} - \overline{v}_S) \end{aligned}$$

as required. □

Proof of Lemma 2

The translation equivariance of ρ yields

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

The first order conditions for RSAA(δ) become

$$0 \in \partial\rho(c(x,S)) = Hx + (1-\delta)\overline{v}_S + \delta G_{\mathrm{CVaR}} \tag{12}$$

which gives (8) and (9) when the subgradient at the optimal solution is unique. □

Proof of Lemma 3

We begin by establishing the value that we will use for $\overline{v}_{\mathrm{CVaR}}$ in the case that $\mathrm{CVaR}_{1-\alpha}\left[\{v(y_i)^\top x\}\right]$ is not differentiable at $x_{SAA}(S)$. We suppose that $\alpha \in (\frac{m-1}{N}, \frac{m}{N}]$ for some integer m. For a given x, we suppose that

$$v(y_1)^\top x \geq v(y_2)^\top x \geq ..v(y_k)^\top x = v(y_{k+1})^\top x = ... = v(y_\ell)^\top x$$

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

$$G_{\mathrm{CVaR}}(x) = \frac{1}{\alpha N}\sum_{i=1}^{k-1} v(y_i) + (1 - \frac{k-1}{\alpha N})\mathrm{conv}\{v(y_k), v(y_{k+1}), ..., v(y_\ell)\}.$$

We write $x_R^{(\delta)}$ for $x_R(S)$ at a particular value δ. Suppose that the ordering (including equalities) of the elements $v(y_i)^\top x$, $i = 1, 2, ..., N$ is the same at $x_R^{(\delta_1)}$ and $x_R^{(\delta_2)}$. Hence $G_{\text{CVaR}}(x_R^{(\delta_1)}) = G_{\text{CVaR}}(x_R^{(\delta_2)}) = G$. So we can find $g_i \in G$ with $x_R^{(\delta_i)} = -H^{-1}\left((1-\delta_i)\overline{v}_S + \delta_i g_i\right)$. Thus

$$\begin{aligned}\theta x_R^{(\delta_1)} + (1-\theta)x_R^{(\delta 2)} &= -\theta(H^{-1}\left((1-\delta_1)\overline{v}_S + \delta_1 g_1\right)) \\ &\quad -(1-\theta)H^{-1}\left((1-\delta_2)\overline{v}_S + \delta_2 g_2\right) \\ &= -H^{-1}\left((1-\beta)\overline{v}_S + \beta z\right)\end{aligned}$$

where $\beta = \theta\delta_1 + (1-\theta)\delta_2$ and $z = (\theta\delta_1/\beta)g_1 + ((1-\theta)\delta_2/\beta)g_2$. Now z is a convex combination of g_1 and g_2 so is in G. Also $G_{\text{CVaR}}(\theta x_R^{(\delta_1)} + (1-\theta)x_R^{(\delta_2)}) = G$. (This follows since the ordering relation between any $v(y_i)^\top x$ and $v(y_j)^\top x$ is preserved by taking convex combinations of the x.) Thus $x_R^{(\beta)} = \theta x_R^{(\delta_1)} + (1-\theta)x_R^{(\delta_2)}$ since the optimality condition is satisfied. Hence we have established that for any δ between δ_1 and δ_2 the optimal solution lies on the line between $x_R^{(\delta_1)}$ and $x_R^{(\delta_2)}$ and the ordering between the elements $v(y_i)^\top x$, $i = 1, 2, \cdots, N$ is preserved.

Consider $\delta \to 0$. There are only a finite number of possible orderings and by our argument above orderings cannot be repeated. Hence the ordering is constant for δ close enough to 0. Now we can use the analysis above (setting $\delta_1 = \Delta$ fixed, $\delta_2 = 0$, $\mu = \delta$) to show $x_R^{(\varepsilon\Delta)} = \varepsilon x_R^{(\Delta)} + (1-\varepsilon)x_{SAA}(S)$ provided Δ is chosen so that the ordering of the elements $v(y_i)^\top x_R^{(\delta)}$ remains constant for $\delta < \Delta$. Now, from optimality of $x_R^{(\varepsilon\Delta)}$, there is some $g^{(\varepsilon\Delta)} \in G_{\text{CVaR}}(x_R^{(\Delta)})$ with

$$x_R^{(\varepsilon\Delta)} = -H^{-1}\left((1-\varepsilon\Delta)\overline{v}_S + (\varepsilon\Delta)g^{(\varepsilon\Delta)}\right).$$

We define

$$\overline{v}_{\text{CVaR}} = \lim_{\varepsilon \to 0} g^{(\varepsilon\Delta)}.$$

Since $G_{\text{CVaR}}(x_R^{(\Delta)})$ is compact this limit point exists and because $x_R^{(\varepsilon\Delta)} \to x_{SAA}(S)$ as $\varepsilon \to 0$, and the graph of the subdifferential for a convex function is closed, we can deduce that $\overline{v}_{\text{CVaR}} \in G_{\text{CVaR}}(x_{SAA}(S))$.

More generally we will define $\overline{v}_{\text{CVaR}}(x_R(S))$ for $\delta > 0$ as

$$\overline{v}_{\text{CVaR}}(x_R(S)) = -(1/\delta)Hx_R(S) - ((1/\delta) - 1)\overline{v}_S.$$

The optimality conditions ensure that this is in $\partial\text{CVaR}_{1-\alpha}\left[\{v(y_i)^\top x\}\right]$, and so matches the gradient when CVaR is differentiable. Recall $x_{SAA}(S) = -H^{-1}\overline{v}_S$, so we get

$$x_R(S) = x_{SAA}(S) - \delta H^{-1}(\overline{v}_{\text{CVaR}}(x_R(S)) - \overline{v}_S)$$

and

$$C_{SAA}(S) = \mathbb{E}_{\mathbb{P}}[c(x_{SAA}(S), y)] = \frac{1}{2} x_{SAA}(S)^\top H x_{SAA}(S) + \overline{v}^\top x_{SAA}(S).$$

If we write R for $(\overline{v}_{\mathrm{CVaR}}(x_R(S)) - \overline{v}_S)$, then the expected cost of the risk-averse solution given the sample S is

$$\begin{aligned} C_R(S) &= \mathbb{E}_{\mathbb{P}}[c(x_R(S), y)] \\ &= \frac{1}{2}\left(x_{SAA}(S) - \delta H^{-1} R\right)^\top H \left(x_{SAA}(S) - \delta H^{-1} R\right) + \overline{v}^\top \left(x_{SAA}(S) - \delta H^{-1} R\right) \\ &= C_{SAA}(S) - \delta x_{SAA}(S)^\top R + \frac{\delta^2}{2} R^\top H^{-1} R - \delta \overline{v}^\top H^{-1} R \\ &= C_{SAA}(S) - \delta(\overline{v} - \overline{v}_S)^\top H^{-1} R + \frac{\delta^2}{2} R^\top H^{-1} R. \end{aligned}$$

Thus we obtain

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

In this expression R depends on $x_R(S)$ and hence δ. From our discussion above we know that $\overline{v}_{\mathrm{CVaR}}(x_R(S))$ approaches $\overline{v}_{\mathrm{CVaR}}$ as δ approaches 0. Hence

$$\begin{aligned} \mathrm{MVRS} &= \mathbb{E}_S[(\overline{v} - \overline{v}_S)^\top H^{-1} R] \\ &= \mathbb{E}_S[(\overline{v} - \overline{v}_S)^\top H^{-1} (\overline{v}_{\mathrm{CVaR}} - \overline{v}_S)], \end{aligned}$$

as required. □

Proof of Corollary 4

Applying Lemma 3 with $H = 2$ and $v(y) = -g(y)$, so $\overline{v}_S = -\overline{g}_S$. Now $\overline{v}_{\mathrm{CVaR}}$ is the derivative of $\mathrm{CVaR}_{1-\alpha}[\{v(y_i)^\top x\}]$ evaluated at $x_{SAA}(S) = \overline{g}_S/2$. Thus

$$\overline{v}_{\mathrm{CVaR}} = \mathrm{CVaR}_{1-\alpha}[\{-\mathrm{sgn}(\overline{g}_S) g(y_i)\}].$$

As in Lemma 3 there is a need for care when $x_{SAA}(S) = 0$ since at that point we have $\mathrm{CVaR}_{1-\alpha}[\{v(y_i)^\top x\}]$ non differentiable. The formulation here makes $\overline{v}_{\mathrm{CVaR}} = 0$ in this case. But since the Corollary statement involves an expectation over a continuous distribution we can see that $x_{SAA}(S) = 0$ with probability zero and our definition at this point will have no impact.

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

$$\begin{aligned} x_R(S) &= x_{SAA}(S) - \delta H^{-1}(\overline{v}_{\mathrm{CVaR}} - \overline{v}_S) \\ &= x_{SAA}(S) - (\delta/2)(\mathrm{CVaR}_{1-\alpha}[\{-\mathrm{sgn}(\overline{g}_S) g(y_i)\}] + \overline{g}_S) \\ &= x_{SAA}(S) - \frac{\delta}{2}(\overline{g}_S + \mathrm{CVaR}_{1-\alpha}[\{-\mathrm{sgn}(\overline{g}_S) g(y_i)\}]) \end{aligned}$$

and

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

as required. □

We now prove some results on order statistics for 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 13 $\mathbb{E}[z_i^2]=\frac{N!}{(N-i)!(i-1)!}\int_{-\infty}^{\infty} z^2F(z)^{i-1}f(z)(1-F(z))^{N-i}dz.$

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

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

■

Lemma 14 *If $i < j$ then*

$$\mathbb{E}[z_i z_j] = \frac{N(N-1)...(N-j+1)}{(i-1)!(j-i-1)!} \int_{-\infty}^{\infty} \int_{x_a}^{\infty} x_a x_b B(x_a, x_b) dx_b dx_a \qquad (13)$$

where

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

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

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

But

$$\begin{aligned}
&\left(\begin{array}{c} N \\ i-1 \end{array} \right) (N - i + 1) \left(\begin{array}{c} N - i \\ j - i - 1 \end{array} \right) (N - j + 1) \\
= \ & \frac{N(N-1)...(N-j+1)}{(i-1)!(j-i-1)!},
\end{aligned}$$

and

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

Thus

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

as required. ■

Lemma 15

$$\sum_{j=i+1}^{N} \mathbb{E}[z_i z_j] = \frac{N!}{(i-1)!(N-i-1)!} \int_{-\infty}^{\infty} zF(z)^{i-1}(1-F(z))^{N-i-1} f(z)Q(z)dz. \tag{14}$$

Proof.

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

Substituting in (13) and substituting for $Q(z)$ yields (14). ■

Lemma 16

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

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 15 we have

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

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 $Q_W(z) = \int_z^{\infty} uf(-u)du$ we can change variables $v = -z$ and obtain

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

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

Proposition 17

$$\mathbb{E}[z_j \bar{z}] = \frac{(N-1)!}{(N-j)!(j-1)!} \int_{-\infty}^{\infty} Q(z)(1-F(z))^{N-j} F(z)^{j-1} dz. \qquad (15)$$

Proof. Applying Lemmas 13, 15, and 16, 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

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

Integrating the third term of A by parts gives

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

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 (15) as required. ■

Proof of Proposition 5

We will use Corollary 4 and show that

$$-\mathbb{E}_S[(\overline{g}_S - \overline{g})\mathrm{CVaR}_{1-\alpha}[\{-g(y_i)\}]] = \int_{-\infty}^{\infty} Q(z)(1-F(z))^{N-1}\Lambda_\alpha(z)dz. \quad (16)$$

First observe that

$$\mathbb{E}_S[(\overline{g}_S - \overline{g})\overline{g}] = 0$$

so

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

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 $\alpha \in (\frac{m-1}{N}, \frac{m}{N}]$ then

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

so

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

Since Proposition 17 gives

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

and

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

where m is the unique integer for which $\alpha \in (\frac{m-1}{N}, \frac{m}{N}]$, the identity (16) now follows. □

Lemma 18 *Suppose z has density f and cumulative distribution function F. For all $\alpha \in (0,1]$,*

$$\int_{-\infty}^{\infty} f(z)(1-F(z))^{N-1}\Lambda_\alpha(z)dz = \frac{1}{N}. \tag{17}$$

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

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

Now

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

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

So

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

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

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

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

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

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

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

Proof of Lemma 6

Recall that

$$C_{SAA}(S) = C^* + \frac{1}{2}(\overline{v} - \overline{v}_S)^\top H^{-1}(\overline{v} - \overline{v}_S).$$

The expected cost of the robust solution has the same form as $C_{SAA}(S)$ but with $\overline{v}_S$ replaced by $\overline{v}_S + \delta R_S^{(\delta)}/2$ (given by the adjusted weights on the sample points). Thus

$$\begin{aligned} C_R(S) &= C^* + \frac{1}{2}(\overline{v} - \overline{v}_S - \delta R_S^{(\delta)}/2)^\top H^{-1}(\overline{v} - \overline{v}_S - \delta R_S^{(\delta)}/2) \\ &= C_{SAA}(S) + (\delta/2)(\overline{v} - \overline{v}_S)^\top H^{-1} R_S^{(\delta)} + (\delta^2/8) R_S^{(\delta)\top} H^{-1} R_S^{(\delta)}. \end{aligned}$$

Thus

$$\text{VRS}(\delta) = \mathbb{E}_S[(\delta/2)(\overline{v} - \overline{v}_S)^\top H^{-1} R_S^{(\delta)} + (\delta^2/8) R_S^{(\delta)\top} H^{-1} R_S^{(\delta)}].$$

We know that $R_S^{(\delta)}$ approaches R_S as δ approaches 0. Hence

$$\text{MVRS} = \mathbb{E}_S[(1/2)(\overline{v} - \overline{v}_S)^\top H^{-1} R_S]$$

as required. □

Proof of Lemma 7

If z_i is the i'th order statistic of $\{w_i : i = 1, ...N\}$ then

$$\begin{aligned} \text{MVRS} &= \frac{1}{4}\mathbb{E}[(\overline{g}_S - \overline{g}) R_S] \\ &= \frac{1}{4}\mathbb{E}\left[(z_N - z_1)\bar{z}\right]. \end{aligned}$$

Let

$$Q(z) = \left(\int_z^\infty w f(w) dw\right).$$

By Lemma 17,

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

and

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

which yields the result. □

Proof of Proposition 8

From the definition of τ we have $F(z) = \widetilde{F}(\tau^{-1}(z))$, and

$$f(z) = \widetilde{f}(\tau^{-1}(z))/\tau'(\tau^{-1}(z)).$$

As z_0 is the mean for $\widetilde{F}$, we have $z_0 = \int_{-\infty}^{\infty} w\widetilde{f}(w)dw$. We know that F has mean 0 and hence

$$0 = \int_{-\infty}^{\infty} zf(z)dz = \int_{-\infty}^{\infty} \frac{z}{\tau'(\tau^{-1}(z))}\widetilde{f}(\tau^{-1}(z))dz = \int_{-\infty}^{\infty} \tau(w)\widetilde{f}(w)dw \quad (18)$$

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

$$\begin{aligned}\int_{-\infty}^{\infty} \tau(w)\widetilde{f}(w)dw &= \int_{-\infty}^{z_0} w\widetilde{f}(w)dw + \int_{z_0}^{\infty} \tau(w)\widetilde{f}(w)dw \\ &= \int_{z_0}^{\infty} (2z_0 - z)\widetilde{f}(z)dw + \int_{z_0}^{\infty} \tau(w)\widetilde{f}(w)dw\end{aligned}$$

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

$$\int_{z_0}^{\infty} (\tau(z) + 2z_0 - z)\,\widetilde{f}(z)dy = 0. \quad (19)$$

We rewrite our required expression in terms of $\widetilde{F}$:

$$\begin{aligned}\mathbb{E}[(\overline{g}_S - \overline{g})R_S] &= \int_{-\infty}^{\infty} \left(F(z)^{N-1} - (1-F(z))^{N-1}\right)\left(\int_z^{\infty} uf(u)du\right)dz \\ &= \int_{-\infty}^{\infty} \left(\widetilde{F}(\tau^{-1}(z))^{N-1} - (1-\widetilde{F}(\tau^{-1}(z)))^{N-1}\right) \\ &\quad \times \left(\int_z^{\infty} u\widetilde{f}(\tau^{-1}(z))du\right)dz \\ &= \int_{-\infty}^{\infty} \left(\widetilde{F}(w)^{N-1} - (1-\widetilde{F}(w))^{N-1}\right) \\ &\quad \times \left(\int_w^{\infty} \tau(z)\widetilde{f}(z)dz\right)\tau'(w)dw\end{aligned}$$

using a change of variable $w = \tau^{-1}(z)$ and $z = \tau^{-1}(u)$.

Hence

$$\begin{aligned}\mathbb{E}[(\overline{g}_S - \overline{g})R_S] &= \int_{-\infty}^{z_0} \left(\widetilde{F}(z)^{N-1} - (1-\widetilde{F}(z))^{N-1}\right)\left(\int_z^{z_0} u\widetilde{f}(u)dz + \int_{z_0}^{\infty} \tau(u)\widetilde{f}(u)du\right)dz \\ &\quad + \int_{z_0}^{\infty} \left(\widetilde{F}(z)^{N-1} - (1-\widetilde{F}(z))^{N-1}\right)\left(\int_z^{\infty} \tau(u)\widetilde{f}(u)du\right)\tau'(z)dz.\end{aligned}$$

We want to replace $\tau'(z)$ with 1 in the second term and to do this we need to establish that $T(z) = \int_z^\infty \tau(u)\widetilde{f}(u)du \geq 0$ for $z > z_0$. We have $T'(z) = -\tau(z)\widetilde{f}(z)$ and as $z_0 < 0$ the function $T(z)$ increases as z increases from z_0 and then decreases to zero. It is enough to show that $T(z_0) > 0$ to show $T(z) \geq 0$ for $z > z_0$. Now from (18)

$$\begin{aligned} T(z_0) &= -\int_{-\infty}^{z_0} \tau(z)\widetilde{f}(z)dz = -\int_{-\infty}^{z_0} z\widetilde{f}(z)dz \\ &= -z_0 + \int_{z_0}^{\infty} u\widetilde{f}(u)du \\ &= -\frac{z_0}{2} + \int_{z_0}^{\infty} (u - z_0)\widetilde{f}(u)du > 0. \end{aligned}$$

Hence $\left(\widetilde{F}(z)^{N-1} - (1-\widetilde{F}(z))^{N-1}\right)\left(\int_z^\infty \tau(u)\widetilde{f}(u)du\right) \geq 0$ for $z > z_0$. Thus using $\tau'(z) \geq 1$ we have shown that

$$\begin{aligned} \mathbb{E}[(\overline{g}_S - \overline{g})R_S] \geq & \int_{-\infty}^{z_0} \left(\widetilde{F}(z)^{N-1} - (1-\widetilde{F}(z))^{N-1}\right) \\ & \times \left(\int_z^{z_0} z\widetilde{f}(z)dz + \int_{z_0}^{\infty} \tau(u)\widetilde{f}(u)du\right) dz \\ & + \int_{z_0}^{\infty} \left(\widetilde{F}(z)^{N-1} - (1-\widetilde{F}(z))^{N-1}\right)\left(\int_z^\infty \tau(u)\widetilde{f}(u)du\right) dz. \end{aligned}$$

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:

$$\int_{-\infty}^{\infty} \left(\widetilde{F}(z)^{N-1} - (1-\widetilde{F}(z))^{N-1}\right)\left(\int_z^\infty (u - z_0)\widetilde{f}(u)du\right) dw = 0.$$

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

$$\begin{aligned} \mathbb{E}[(\overline{g}_S - \overline{g})R_S] \geq & \int_{-\infty}^{z_0} \left(\widetilde{F}(z)^{N-1} - (1-\widetilde{F}(z))^{N-1}\right)\left(\int_z^{z_0} z_0\widetilde{f}(u)du + A\right) dz \\ & + \int_{z_0}^{\infty} \left(\widetilde{F}(w)^{N-1} - (1-\widetilde{F}(w))^{N-1}\right) \\ & \times \left(\int_z^\infty (\tau(u) - u + z_0)\widetilde{f}(u)du\right) dz \end{aligned}$$

where $A = \int_{z_0}^{\infty}(\tau(u) - z + z_0)\widetilde{f}(u)du$. Hence

$$\mathbb{E}[(\overline{g}_S - \overline{g})R_S] \geq \int_{-\infty}^{\infty} \left(\widetilde{F}(z)^{N-1} - (1 - \widetilde{F}(z))^{N-1}\right)(U(z) + V(z))dz$$

where $U(z) = \int_z^{z_0} z_0\widetilde{f}(u)du + A$ for $z \leq z_0$ and $U(z) = \int_{z_0}^{z} z_0\widetilde{f}(u)du + A$ for $z > z_0$. Note that U is symmetric around z_0 and is maximized at z_0 since $z_0 < 0$. Hence

$$\begin{aligned} & \left(\widetilde{F}(z_0 - k)^{N-1} - (1 - \widetilde{F}(z_0 - k))^{N-1}\right) U(z_0 - k) \\ = & -\left(\widetilde{F}(z_0 + k)^{N-1} - (1 - \widetilde{F}(z_0 + k))^{N-1}\right) U(z_0 + k), \end{aligned}$$

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 $z \leq z_0$ and for $z > z_0$ we have

$$\begin{aligned} V(z) &= \int_z^{\infty}(\tau(z) - u + z_0)\widetilde{f}(u)du - \left(\int_{z_0}^{z} z_0\widetilde{f}(u)du + A\right) \\ &= \int_{z_0}^{z}(-2z_0 - \tau(u) + u)\widetilde{f}(u)du. \end{aligned}$$

So, from (19), $V(\infty) = 0$. Now $\tau(u) + 2z_0 - u$ has derivative $\tau'(u) - 1 \geq 0$ and because the integral in (19) is zero, we can deduce that $\tau(u) + 2z_0 - u$ starts negative and becomes positive. Since

$$\frac{d}{dz}V(z) = (-2z_0 - \tau(z) + z)\widetilde{f}(z),$$

we know that V starts by increasing and then decreases to zero. Moreover $V(z_0) = 0$. Hence it is always non-negative. Since $V(z)$ is zero for $z \leq z_0$ when $\widetilde{F}(z)^{N-1} - (1 - \widetilde{F}(z))^{N-1} < 0$, then

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

and so $\mathbb{E}[(\overline{g}_S - \overline{g})R_S] \geq 0$. In the case that F and $\widetilde{F}$ are not identical then $\tau'(z) > 1$ on some region and the inequality becomes strict. □

Proof of Proposition 10

We suppose that $c(x, y)$ is concave in y. We first observe by Proposition 9 that this is enough to show that the solution to $\bar{\mathrm{P}}$ has each v_i supported on a single point (if v_i has weight p on z_{i1} and $(1 - p)$ on z_{i2} then setting

v_i to have weight 1 on $pz_{i1} + (1-p)z_{i2}$ increases the objective of $\bar{\mathrm{P}}$ and still satisfies the constraint). Thus $\bar{\mathrm{P}}$ becomes

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

The Lagrangian of P1 is

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

which is maximized at z_i. So

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

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

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

$$\varepsilon \|\nabla c_x(y_i)\|^2 - \lambda\varepsilon \|\nabla c_x(y_i)\| \leq 0,$$

and $\|\nabla c_x(y_i)\| \leq \lambda$. And hence for any choice of z_j with $z_j \neq y_j$, $\|\nabla c_x(y_i)\| \leq \|\nabla c_x(z_i)\|$, as required. □

Proof of Proposition 11

Suppose that the solution to $\bar{\mathrm{P}}$ has γ_i with support in J_i discrete points, and probability γ_{ij} on point z_{ij}, $j = 1, 2, ...J_i$. ($J_i = 1$ or 2). For some i consider choosing a new point z_0 instead of z_{ij}. More precisely we replace the term $\gamma_{ij} c_x(z_{ij})$ with $\gamma_0 c_x(z_0) + (\gamma_{ij} - \gamma_0) c_x(y_i)$ where we choose γ_0 to make the Wasserstein distance the same, or less:

$$\gamma_0 = \min(\gamma_{ij}, \gamma_{ij} \|z_{ij} - y_i\| / \|z_0 - y_i\|).$$

There will be an increase in expected cost if $\gamma_0 c_x(z_0) + (\gamma_{ij} - \gamma_0) c_x(y_i) - \gamma_{ij} c_x(z_{ij}) > 0$. In the case that z_{ij} is not at the boundary of M then we move z_0 out to the boundary by choosing $z_0 = y_i + \alpha(z_{ij} - y_i)$ for some $\alpha > 1$, so that $\|z_{ij} - y_i\| / \|z_0 - y_i\| = 1/\alpha$. By convexity $c_x(z_0) > c_x(y_i) + \alpha(c(z_{ij}) - c(y_i))$ and thus

$$\begin{array}{rl} & \gamma_0 c_x(z_0) + (\gamma_{ij} - \gamma_0) c_x(y_i) - \gamma_{ij} c_x(z_{ij}) \\ = & c_x(z_0)\gamma_{ij}/\alpha + (\gamma_{ij} - \gamma_{ij}/\alpha) c_x(y_i) - \gamma_{ij} c_x(z_{ij}) \\ = & \gamma_{ij} \left(c_x(z_0) - c_x(y_i) - \alpha(c(z_{ij}) - c(y_i))\right)/\alpha > 0. \end{array}$$

Hence we deduce that one of the points on which weight is placed must be at the boundary (to avoid a contradiction).

Suppose now that ν_i has mass at y_i and for some other j with $\nu_j \neq y_j$ we have $\max_{z\in M}(\varphi_i(z)) > \max_{z\in M}(\varphi_j(z))$. Suppose that ν_j has weight at $z_{jh} \neq y_j$. Choose a point z^*_{ik} with $\varphi_i(z^*_{ik}) > \varphi_j(z_{jh})$. Then for small $\varepsilon > 0$ we set $\nu'_j(z_{jh}) = \nu_j(z_{jh}) - \varepsilon$, $\nu'_j(y_j) = \nu_j(y_j) + \varepsilon$, $\nu'_i(z^*_{ik}) = k\varepsilon$, $\nu'_i(y_i) = \nu_i(y_i) - k\varepsilon$ where $k = \|z_{jh} - y_j\| \,/\, \|z^*_{ik} - y_i\|$. The change in the objective function is given by

$$\begin{aligned} \varepsilon\left(c_x(y_j) - c_x(z_{jh})\right) + \varepsilon k\left(c_x(z^*_{ik}) - c_x(y_i)\right) &= \varepsilon\|z_{jh} - y_j\|\left(\varphi_i(z^*_{ik}) - \varphi_j(z_{jh})\right) \\ &> 0, \end{aligned}$$

so that there is an improvement and it is easy to check that the overall value of $\sum_{i=1}^{N}\sum_{j=1}^{J_i}\|z_{ij} - y_i\|\ \gamma_{ij}$ is unchanged. This gives a contradiction and establishes what we require. □

Proof of Proposition 12

(a) The first order conditions are

$$2x - \overline{g}_S + \delta\,\|\nabla g(y^*_S)\| = 0.$$

Hence

$$x^*_R(S) = \frac{\overline{g}_S}{2} - \delta\,\|\nabla g(y^*_S)\|/2 = x^*_{SAA}(S) - \delta\,\|\nabla g(y^*_S)\|/2.$$

The expected cost for the robust solution can be calculated in the same way as we have seen in the other two cases.

$$\begin{aligned} C_R(S) &= \mathbb{E}_{\mathbb{P}}[c(x^*_R(S), y)] \\ &= bx^*_R(S) + x^*_R(S)^2 - x^*_R(S)\overline{g} \\ &= C_{SAA}(S) + (\overline{g} - \overline{g}_S)\delta\,\|\nabla g(y^*_S)\|/2 + \delta^2\,\|\nabla g(y^*_S)\|^2/4. \end{aligned}$$

So

$$\mathrm{VRS}(\delta) = \mathbb{E}_S[(\overline{g}_S - \overline{g})\delta\,\|\nabla g(y^*_S)\|/2 + \delta^2\,\|\nabla g(y^*_S)\|^2/4],$$

and

$$\mathrm{MVRS} = \mathbb{E}_S\left[(\overline{g}_S - \overline{g})\,\|\nabla g(y^*_S)\|\right]/2.$$

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

$$\begin{aligned} \mathrm{MVRS} &= \mathbb{E}_S\left[((1/N)\sum_{i=1}^{N} y_i^2 - \mathbb{E}[y^2])y_N\right] \\ &= \mathbb{E}_S[\frac{y_N}{N}\sum_{i=1}^{N} y_i^2] - \mathbb{E}[y^2]\mathbb{E}[y_N]. \end{aligned}$$

Writing y_i for the order statistics we have, for $i < N$, (essentially this is the result of Lemma 14 with $j = N$)

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

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

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

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

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

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

We have

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

as required. □