Distributionally Robust Optimization for Engineering Design under Uncertainty

Michael G. Kapteyn*
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
Karen E. Willcox†
The University of Texas at Austin, Austin, TX 78712, USA
Andy B. Philpott‡
The University of Auckland, Auckland 1010, New Zealand

This paper addresses the challenge of design optimization under uncertainty when the designer only has limited data to characterize uncertain variables. We demonstrate that the error incurred when estimating a probability distribution from limited data affects the out-of-sample performance (i.e., performance under the true distribution) of optimized designs. We demonstrate how this can be mitigated by reformulating the engineering design problem as a distributionally robust optimization (DRO) problem. We present computationally efficient algorithms for solving the resulting DRO problem. The performance of the DRO approach is explored in a practical setting by applying it to an acoustic horn design problem. The DRO approach is compared against traditional approaches to optimization under uncertainty, namely, sample-average approximation and a multi-objective optimization incorporating a risk reduction objective. In contrast with the multi-objective approach, the proposed DRO approach does not use an explicit risk reduction objective but rather specifies a so-called ambiguity set of possible distributions, and optimizes against the worst-case distribution in this set. Our results show that the DRO designs in some cases significantly outperform those designs found using the sample-average or multi-objective approaches.

I. Introduction

The importance of considering the nature of the uncertain parameters in engineering design has long been established,[1–3] the primary reason being that simply considering all parameters to be deterministic tends to over-fit designs to the chosen parameter values. This results in a need to introduce ad-hoc safety factors and/or degraded performance of the system when it is exposed to the true uncertain operating environment.[4,5]

In design under uncertainty, the designer must decide how to characterize and represent the uncertainty. The most prevalent approach is to treat the uncertain parameters as random variables, endowed with some probability distribution.[1,4] Other treatments and representations of uncertainty have also been established,[6] such as interval or set-based uncertainty,[7–9] and possibility theory.[10] In this work we focus on the case of a probabilistic representation of uncertainty, where the designer must specify a probability distribution for the uncertain parameters. In practice the designer will never have perfect knowledge of how the underlying system parameters vary, and thus will lack access to the true probability distribution governing an uncertain parameter. Even if the true probability distribution were known, evaluating expected performance using this distribution would likely involve many evaluations of the computational models representing the system of interest, which for complex engineering systems becomes computationally infeasible.

*PhD Candidate, Department of Aeronautics and Astronautics.
†Director, Oden Institute for Computational Engineering and Sciences.
‡Professor, Department of Engineering Science.

A common approach is to extract a sample of the uncertain parameters and model the expected performance using the sample average. Optimizing the sample average as a function of the design parameters yields the sample-average approximation (SAA) method of stochastic programming. In this paper we consider problems for which a black-box computational model can be employed in the solution of an SAA problem, but this model is tractable for only small values of the sample size. In other words, the designer's computational budget admits the solution of a SAA model with only a modest sample size.

Distributionally robust optimization (DRO) has emerged from within the optimization community as an approach that explicitly accounts for the fact that one is never able to exactly specify a probability distribution in practice.[11–17] DRO weakens the requirement to specify a single probability distribution for the uncertain parameters. Instead, it requires the designer to select a *set* of possible probability distributions—we then optimize for the worst-case distribution within this set. Evidence from the literature shows that for some problems, when compared with the SAA approach, a DRO approach can improve the performance of a design when evaluated "out-of-sample", i.e., when using the true probability distribution.[18,19]

The first contribution of this paper is to show how engineering design under uncertainty problems can be formulated and solved using DRO. A subtlety in our application is that typically DRO is used when only a finite number of data samples are available from some real world process. In the engineering design context, we wish to characterize the distribution in performance of designs over some range of uncertain parameters. In this context, it is often the case that samples of the uncertain parameter are readily available, but analyzing a sample (i.e., computing the performance of a design at a data sample) is computationally expensive. Thus, in our context *limited data* refers to the fact that our finite computational budget only allows the analysis of a modest number of samples. The DRO formulation mitigates the effect of a limited sample size by explicitly seeking designs that are robust to deviations in the sample.

The second contribution is to provide computationally efficient algorithms for solving the engineering design DRO problem, with computational cost comparable to solving an SAA problem, when the set of distributions is constructed using either the L_2-norm, the Kullback-Leibler (KL) divergence, or a Wasserstein distance.

Much of the supporting theory presented in the DRO literature involves asymptotic results, or results that rely on strict assumptions.[12,14,15,19,20] The final contribution of this paper is to showcase the effectiveness of DRO as a practical approach to engineering design optimization. To this end we apply the formulation to a black-box model design problem using small samples of the uncertain parameters. We compare the DRO approach to the traditional SAA approach, and show that DRO is able to outperform SAA in this setting.

A commonly used competing approach to introducing robustness in engineering design is to augment the expected performance objective with a variance reduction objective and perform multi-objective optimization (MOO). Recent work has shown that under certain assumptions the DRO is almost the same as a mean-variance optimization.[19] In this paper we show that, for our design problem, the DRO designs outperform those found using mean-variance multi-objective optimization (MOO).

The outline of this paper is as follows. Section II formulates the problem of optimizing a design based on limited data using SAA. Section III presents an alternative formulation of the design problem as a DRO problem, and presents algorithms for solving the resulting optimization problems. Section IV presents the results of the DRO approach applied to an illustrative design problem. These results are used to analyze the performance of the approach, and compare it to the baseline SAA approach. Section V compares the DRO approach with an MOO approach commonly used in engineering design. This mean-variance optimization approach introduces robustness through variance reduction, rather than through distributional ambiguity. Section VI presents a discussion of the results and suggests possible areas to explore in future work. Finally, Section VII concludes the paper.

II. Design Under Uncertainty Using Limited Data

This section formulates the problem of design optimization under uncertainty adapted to the setting in which the designer does not know the true distribution of the uncertain parameters, but instead has access to a sample consisting of a limited number of realizations from this unknown distribution. A methodology for optimizing the mean performance of designs using SAA is also introduced, and applied to an illustrative design problem in order to highlight the potential consequences of designing using limited data.

II.A. Problem Formulation

The vector of design variables is denoted by $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ is the feasible design space encoding constraints on the design. To represent the uncertain parameters, a probability space $(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{u}})$ is defined, with sample space Ω, σ-algebra $\mathcal{F}$, and probability measure $\mathbb{P}_{\mathbf{u}}$. The uncertain parameters are defined as a random variable $\mathbf{u} : \Omega \to \mathcal{U}$ which has dimension equal to the number of uncertain parameters in the design problem. The uncertainty space, $\mathcal{U}$, represents the space of all possible realizations of the random variable, $\mathbf{u}$. Subscripts denote particular realizations of the uncertain parameters, i.e., $\mathbf{u}_i = \mathbf{u}(\omega_i)$ for some $\omega_i \in \Omega$. The quantity of interest (QoI) of our system is represented as a function, $Q(\mathbf{x}, \mathbf{u})$, of the design variables $\mathbf{x}$, and the uncertain variables, given by the random variable $\mathbf{u}$. As the QoI, $Q(\mathbf{x}, \mathbf{u})$, is a function of the random variable $\mathbf{u}$, it too is a random variable, with a probability measure denoted $\mathbb{P}_{\mathbf{Q}}$. Note that the randomness in the QoI is induced only by the randomness in $\mathbf{u}$. In a slight overload of notation, $Q(\mathbf{x}, \mathbf{u}_i)$ will be used to denote the deterministic value of the QoI evaluated at a realization of the uncertain parameters, $\mathbf{u}_i$. For complex engineering systems, evaluating the function $Q(\mathbf{x}, \mathbf{u}_i)$ typically requires solving a computationally expensive model of the system that could take hours or even weeks to solve on a supercomputer.

The problem of design optimization under uncertainty is to find a vector of optimal design variables that produces a distribution of the QoI that is favorable in some sense. In this work we focus on optimizing the mean of this distribution. Hence, the performance of a design, $\mathbf{x}$ is defined to be

$$Z(\mathbf{x}) = \mathbb{E}_{\mathbb{P}_{\mathbf{Q}}}\left[Q(\mathbf{x}, \mathbf{u})\right], \tag{1}$$

where $\mathbb{E}_{\mathbb{P}_{\mathbf{Q}}}$ denotes the expectation under the distribution $\mathbb{P}_{\mathbf{Q}}$. Assuming that a lower QoI is favorable, the design optimization problem can be written

$$\mathcal{P}: \qquad \min_{\mathbf{x} \in \mathcal{X}} Z(\mathbf{x}). \tag{2}$$

The optimal objective value of $\mathcal{P}$ is denoted by $Z_\infty \equiv Z(\mathbf{x}_\infty)$, where $\mathbf{x}_\infty$ is a corresponding optimal design.

A key limitation in solving $\mathcal{P}$ is that it requires many evaluations of $Z(\mathbf{x})$, for different designs, $\mathbf{x}$, evaluated during the design optimization loop. Computing $Z(\mathbf{x})$ exactly is generally impossible using only a black-box function $Q(\mathbf{x}, \mathbf{u}_i)$. This is because the designer is unable to exactly propagate the continuous distribution over $\mathbf{u}$ through the black-box function, which only takes discrete values as input. A common approach is to instead use a sample-based approximation of $Z(\mathbf{x})$. To do so accurately would require many evaluations of the expensive black-box function $Q(\mathbf{x}, \mathbf{u}_i)$ for samples $\mathbf{u}_i$. Repeating this precise approximation process for every iteration of the design optimization is generally computationally intractable. In practice, the designer can usually only afford a small number of black-box function evaluations to analyze each iteration of the design. Consequently, $Z(\mathbf{x})$ must be approximated using only a small sample of m realizations of the random variable Q, namely $Q(\mathbf{x}, \mathbf{u}_i)$ for $i = 1, \ldots, m$, where $\mathbf{u}_1, \ldots \mathbf{u}_m$ are randomly sampled from $\mathbb{P}_{\mathbf{u}}$.

Due to this limitation, the designer will be forced to approximate the optimization problem $\mathcal{P}$, and solve a different optimization problem that *can* be solved using the limited data available. Solving this approximate problem produces a design, denoted by $\mathbf{x}_m$, with corresponding performance $Z_m \equiv Z(\mathbf{x}_m)$.

In this paper we will discuss existing methodologies for using a limited number of black-box function evaluations to produce the design $\mathbf{x}_m$, and propose a new methodology based on DRO. In any case, there will inevitably be error in approximating $\mathcal{P}$ using only a small number of function evaluations. Consequently, the solution $\mathbf{x}_m$ obtained by the designer will, on average, perform worse than the true solution $\mathbf{x}_\infty$, i.e., $Z_m \geq Z_\infty$. With this in mind, we seek design methodologies that are robust to this error in the sense that over all possible sample draws the methodology produces designs that exhibit good performance when realized under the true distribution of uncertainty.

For the analysis throughout this paper, since it is impossible to enumerate all possible sample draws, we instead simulate $T = 500$ random sample draws, each consisting of m random realizations from $\mathbb{P}_{\mathbf{u}}$. We evaluate the design methodology by analyzing the average performance of designs over all sample draws, defined by

$$\mu(Z_m) = \frac{1}{T}\sum_{t=1}^{T} Z_m^t, \tag{3}$$

where Z_m^t denotes the value of the performance, Z_m, for sample draw t. We will also evaluate the risk level of the methodology, as indicated by the performance in the worst-case sample draws. For this we use the

95th percentile of the performance over all sample draws. Assuming we have ordered the sample indices so that $Z_m^1, ..., Z_m^T$ is in ascending order, we denote the 95th percentile of Z_m by $\rho(Z_m)$, defined by

$$\rho(Z_m) = \min\left\{ Z_m^t \mid t > 0.95T \right\}. \tag{4}$$

II.B. Motivation: Performance of a sample-average approximation approach

One existing approach for approximating the full optimization problem $\mathcal{P}$ (Eqn. 2) utilizes the sample-average approximation (SAA).[21] For a given random sample from $\mathbb{P}_{\mathbf{u}}$, with realizations denoted $\mathbf{u}_1, ..., \mathbf{u}_m$, the SAA estimate of the performance of a design, $\mathbf{x}$, is given by

$$\hat{Z}_m(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} Q(\mathbf{x}, \mathbf{u}_i). \tag{5}$$

The SAA approach is justified by the fact that if the sample is drawn i.i.d from $\mathbb{P}_{\mathbf{u}}$, then $\hat{Z}_m(\mathbf{x})$ is an unbiased estimate of $Z(\mathbf{x})$. Furthermore, by the strong law of large numbers $\hat{Z}_m(\mathbf{x})$ will converge almost surely to $Z(\mathbf{x})$, and by the Central Limit Theorem, the variance of the estimate $\hat{Z}_m(\mathbf{x})$ will decrease with a rate $O(1/m)$. The SAA approach aims to optimize the true performance $Z(\mathbf{x})$, by optimizing the SAA estimate of performance. This gives the SAA problem

$$\mathcal{S}: \qquad \min_{\mathbf{x} \in \mathcal{X}} \hat{Z}_m(\mathbf{x}). \tag{6}$$

A design that solves $\mathcal{S}$ is denoted by $\mathbf{x}_m$. The optimal objective value of $\mathcal{S}$ is denoted $\hat{Z}_m \equiv \hat{Z}(\mathbf{x}_m)$. Note that $\hat{Z}_m$ is the *in-sample* performance of the design, i.e., the average performance over the sample of uncertain variables, $\mathbf{u}_1, ..., \mathbf{u}_m$. In practice, what is more important is Z_m, the performance of the design $\mathbf{x}_m$ *out-of-sample*, i.e., under the *true* distribution $\mathbb{P}_{\mathbf{u}}$. Note that Z_m is not known by the designer, since it requires the calculation of the right-hand side of Eqn. 1 which is usually computationally intractable in this setting, as explained in the previous section.

To demonstrate potential pitfalls of the SAA methodology, we apply it to an illustrative design problem. We consider the problem of designing an acoustic horn to achieve maximum efficiency. In this problem the two design variables, x_1 and x_2, determine the shape of the acoustic horn flare. The uncertain parameter $\mathbf{u}$ is the wave number (an operating condition of the horn). The wave number is assumed to follow a fixed truth distribution, $\mathbb{P}_u = \text{Uniform}(1.3, 1.5)$. We stress that the design methodologies described in this paper assume that the designer does not know this truth distribution, but they are able to sample from it. The QoI, $Q(\mathbf{x}, \mathbf{u})$ is the amount of internal reflection generated by the horn design, operating at a given wave number. A smaller QoI is more favorable. A complete description of this design problem is provided in Appendix A.

We simulate $T = 500$ realizations of a practical design process. In each case we draw a sample of size m from $\mathbb{P}_{\mathbf{u}}$, and use this to solve $\mathcal{S}$ using an interior point method[a]. We run the optimization algorithm for 100 iterations to simulate a scenario where the designer has a fixed computational budget allowing $100m$ function evaluations. We denote the resulting design $\mathbf{x}_m^t$. We evaluate the out-of-sample performance of this design, Z_m^t, by computing Eqn. 1 using a 1000-point Monte-Carlo estimate for the expectation. For the sake of comparison, we also compute an estimate for the true optimal design $\mathbf{x}_\infty$, and its out-of-sample performance, Z_∞. This is done by solving $\mathcal{S}$ using a very large sample size of $m = 1000$. Note that Z_m^t, $\mathbf{x}_\infty$, and Z_∞ are all considered to be too computationally expensive to be computed during a practical design process, but are computed here to enable comparison. Figure 1 gives histograms showing the distributions of Z_m values across the $T = 500$ sample draws, for sample sizes $m = 5, 10, 20$. Also indicated on each histogram are the estimated true optimal performance, Z_∞, the mean performance across all sample draws, $\mu(Z_m)$, and the 95th percentile of performance across all sample draws, $\rho(Z_m)$.

We see that for some sample draws, Z_m is close to our estimate of the true optimal objective $Z_\infty = 0.029$. However, the long right tails of the distributions indicate that there are also many sample draws for which Z_m is far from Z_∞. This is especially the case for small sample sizes. In the $m = 5$ case, the 95th percentile is $\rho = 0.044$. This indicates that, in the worst 5% of sample draws, the SAA method results in designs that produce a QoI over 50% higher (worse) than the true optimum.

[a] As implemented in the function `fmincon`, included in the `MATLAB` Optimization Toolbox[22]

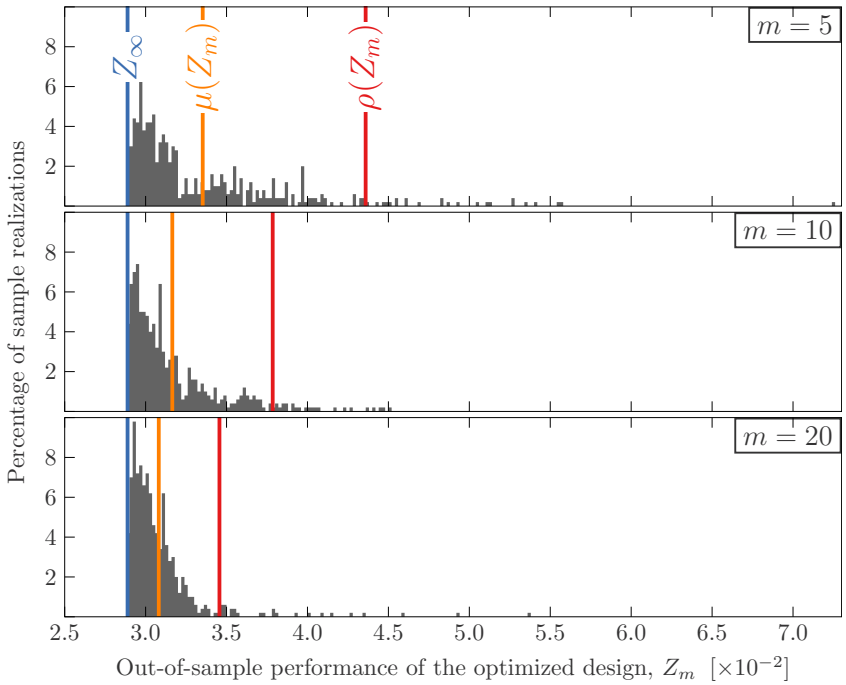

Figure 1: Histograms showing the performance of SAA designs computed using each of the $T = 500$ sample draws of size m. The optimal, mean, and 95th percentile of performance over all sample draws are also indicated.

To see how this poor performance arises, Figure 2 displays the QoI over the range of the uncertainty, for the best and worst performing horn designs obtained over all the sample draws of size $m = 5$. Recall that we define the performance, Z_m to be the mean value of the QoI over this range. We see that in the best sample draw, good in-sample performance translates into good out-of-sample performance (i.e., performance under the true uniform distribution). Using SAA with this sample thus produces a design with near optimal performance ($Z_m = 2.9 \times 10^{-2}$). In contrast, the realizations in the worst sample are clustered around one end of the range of the uncertainty. As a result, the design computed using this sample is over-fitted to the sample, ultimately leading to poor out-of-sample performance ($Z_m = 7.2 \times 10^{-2}$). These results show that with small sample size the SAA approach is susceptible to over-fitting, motivating the question of whether an alternative methodology can produce designs that exhibit better out-of-sample performance when provided with the same data.

III. Distributionally Robust Design Optimization

In this section we present a principled method for introducing the notion of robustness against variability in sample distributions into the design problem, using the mathematical framework of DRO. Section III.A formulates the problem of finding distributionally robust designs from a given sample of the uncertain parameters. These are designs that achieve good mean performance under the sample distribution (in-sample performance), while requiring that this performance be robust to deviations in the sample distribution used. Section III.B discusses how to select a set of probability distributions against which to robustify the design. Section III.C presents efficient algorithms for solving the resulting distributionally robust design problem using three different notions of distance. Finally, Section III.D discusses solving the outer optimization problem to find an optimal design in the DRO problem.

III.A. Overview of the Distributionally Robust Approach

The central idea behind a distributionally robust approach is to optimize the design while considering a set of possible distributions, rather than a single distribution.[12,23] The set of distributions used in the

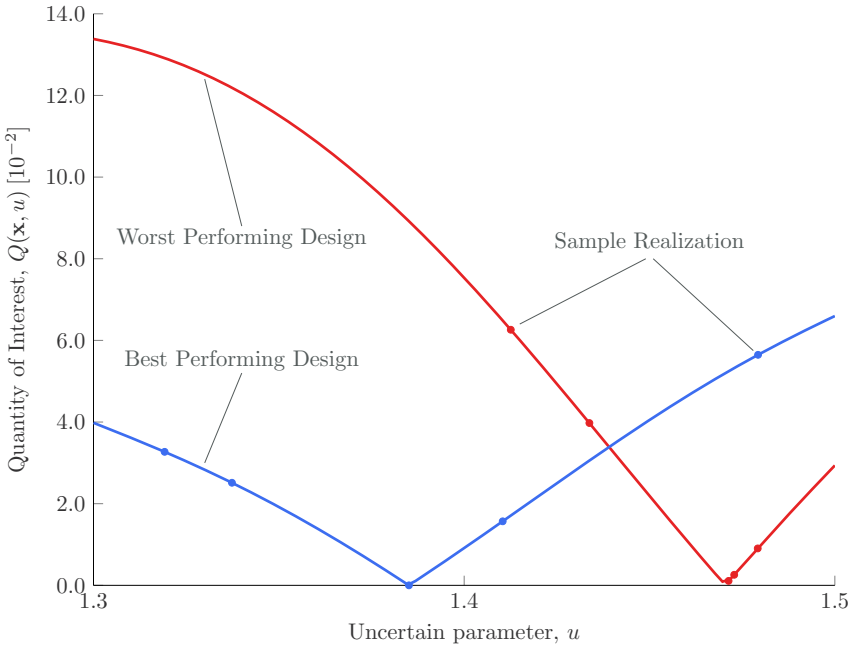

Figure 2: Performance of two SAA designs over the uncertainty space $\mathcal{U}$. These are the designs that exhibit the best and worst mean performance over all designs computed using different sample draws of $m = 5$ realizations of the random variables.

optimization is termed the *ambiguity set*,[13] which we denote by $\mathcal{P}$. In the context of engineering design, the distribution we are interested in is the distribution of the QoI, namely, $\mathbb{P}_{\mathbf{Q}}$. As the QoI is computed using an expensive black-box function, we generally do not know the true distribution $\mathbb{P}_{\mathbf{Q}}$. Instead we have a finite computational budget which we use to generate a small sample from this distribution, $Q(\mathbf{x}, \mathbf{u}_i), i = 1, ..., m$. This sample is used to construct an ambiguity set of distributions that are considered in the DRO approach. Note that specifying the ambiguity set in this way does not require knowledge of the true distributions $\mathbb{P}_{\mathbf{Q}}$ or $\mathbb{P}_{\mathbf{u}}$ or knowledge about properties like their support or moments.

The DRO method seeks a design that performs well for all distributions within the ambiguity set. This is achieved by solving a minimax problem to optimize the worst-case expected performance under any distribution within the ambiguity set. Using the notation introduced in Section II.A, the distributionally robust design optimization problem can be written

$$\mathcal{D}(\mathcal{P}): \qquad \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{p}\in\mathcal{P}} \quad \sum_{i=1}^{m} p_i Q(\mathbf{x}, \mathbf{u}_i). \tag{7}$$

where $\mathbf{p} \in \mathcal{P}$ denotes probability distributions within the ambiguity set $\mathcal{P}$. We denote the optimal design found in $\mathcal{D}(\mathcal{P})$ by $\mathbf{x}_m$, where the ambiguity set, $\mathcal{P}$, used in the optimization will be clear from the context, or explicitly denoted using the notation $\mathbf{x}_m|_{\mathcal{P}}$.

The inner maximization problem in $\mathcal{D}$ involves finding the worst-case expectation over all distributions in the ambiguity set. This inner maximization problem gives the worst-case distribution

$$\mathbf{p}^*(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{p}\in\mathcal{P}} \sum_{i=1}^{m} p_i Q(\mathbf{x}, \mathbf{u}_i). \tag{8}$$

The outer minimization problem finds the design that achieves the best possible mean performance under the worst-case distribution $\mathbf{p}^*(\mathbf{x})$.

The following subsections describe the key challenges in employing the distributionally robust approach described above. In particular, we outline an approach to constructing the ambiguity set, solving the inner optimization problem to find the worst-case distribution in the ambiguity set, and solving the outer optimization problem to find the design that optimizes the worst-case performance within the ambiguity set.

III.B. Constructing the ambiguity set

The tractability and success of the distributionally robust approach relies on a careful selection of the ambiguity set $\mathcal{P}$. In the context of engineering design, the function $Q(\mathbf{x}, \mathbf{u}_i)$ is typically a computationally expensive black-box function. This makes solving the inner optimization problem (Eqn. 8) intractable for a general ambiguity set $\mathcal{P}$. To see this, note that if we allowed the support of $\mathbf{p}$ to change, we would have to draw a new sample $\mathbf{u_i}, i = 1, ..., m$ at each iteration of the inner optimization problem. This would consequently require evaluating $Q(\mathbf{x}, \mathbf{u}_i)$ m times at *every* iteration of the inner optimization problem, rendering the DRO approach significantly more computationally expensive than the traditional SAA approach, which only requires m function evaluations for the *entire* inner optimization problem.
In order for the DRO approach to be an attractive alternative to the SAA approach, we wish to formulate the DRO inner optimization problem in a way that allows us to solve the problem with only m function evaluations. The way we achieve this is by fixing the sample $\mathbf{u}_i, i = 1, ..., m$, and restricting the ambiguity set to contain only distributions with support on this sample. This allows us to evaluate $Q(\mathbf{x}, \mathbf{u}_i)$ for $\mathbf{u}_i, i = 1, ..., m$ only once for the entire inner optimization problem.

Under the aforementioned restriction to distributions with fixed, finite support, the ambiguity set can be defined as a set of discrete probability distributions represented by vectors $\mathbf{p}$ that assign probabilities to each of the m realizations in an associated sample, i.e., p_i gives the probability that the random variable $\mathbf{u}$ takes the value $\mathbf{u}_i$, for $i = 1, ..., m$. If we consider only the empirical distribution associated with the sample, then the ambiguity set becomes a singleton, $\mathcal{P} = \{\hat{\mathbf{p}}\}$. This reduces the problem to the SAA problem $\mathcal{S}$ (Eqn. 6). In the DRO approach, we acknowledge the fact that the empirical distribution $\hat{\mathbf{p}}$ does not perfectly reflect the true probability distribution. This is done by including additional distributions in the ambiguity set.

To assist with notation, we define $\Omega^m \subseteq \mathbb{R}^m$ to be the set of all possible discrete probability distributions that could be associated with a sample of size m, i.e.,

$$\Omega^m = \{\mathbf{p} \in \mathbb{R}^m \ : \ \sum_{i=1}^{m} p_i = 1, p_i \geq 0 \ i = 1, ..., m\}. \tag{9}$$

In general, any ambiguity set we consider will be a subset of Ω^m. We aim to include in the ambiguity set all distributions that are a plausible reflection of the true distribution. To this end, we define a function $D(\cdot, \cdot) : \Omega^m \times \Omega^m \to \mathbb{R}_{\geq 0}$, that measures the distance (i.e., a non-negative scalar that measures the degree of similarity) between two discrete probability distributions in the set Ω^m. We then construct the ambiguity set $\mathcal{P} \subseteq \Omega^m$ to contain all distributions that are sufficiently similar to the empirical distribution given by the sample data. The allowable distance from the empirical distribution, r, is termed the *radius of ambiguity*. Using these definitions, the ambiguity set constructed using the distance function D, is given by

$$\mathcal{P}_D(\hat{\mathbf{p}}, r) = \{\mathbf{p} \in \Omega^m : D(\mathbf{p}, \hat{\mathbf{p}}) \leq r\}. \tag{10}$$

One decision to make when constructing the ambiguity set is the choice of r, the radius of ambiguity, which determines the size of the set. Increasing the size of the ambiguity set allows the designer to be almost certain that it will include the true distribution. However, this leads to over-conservatism and, consequently, poor mean performance of the resulting design. To see this, consider expanding the ambiguity set to include every possible distribution. This gives rise to the usual worst-case optimization problem, since the worst-case distribution will always be one in which all the probability density is assigned to the worst possible realization of the uncertainty. In many real situations this is an unrealistic candidate for the true distribution, and so accounting for this distribution in an optimization for mean performance is generally an overly conservative approach. The effect of the radius of ambiguity on the performance of the resulting DRO designs is explored in Section IV.

The other important decision to make when constructing the ambiguity set is the choice of the distance function, $D(\cdot, \cdot)$. A good function is one that accurately measures how plausible a given distribution is, given our known data on the uncertainty, so that as r increases we add distributions that are increasingly implausible. The following subsection introduces three of the distance functions studied in this work.

III.C. Finding the worst-case distribution

In this section we present three distance functions that are amenable to the construction of an ambiguity set using a sample of realizations of the uncertain variables. In particular, these distance functions compare

discrete probability distributions supported on the sampled values. As described in the previous section, we leverage this restriction of the support in order to present computationally efficient algorithms for solving the inner optimization over the resulting ambiguity sets.

III.C.1. L_2-norm ambiguity

One possible distance function utilizes the L_2 or Euclidean norm between two vectors in Ω^m, and is defined by

$$D_{\mathrm{L2}}(\hat{\mathbf{p}}, \mathbf{p}) = \|\hat{\mathbf{p}} - \mathbf{p}\|_2 = \sqrt{\sum_{i=1}^{m} (\hat{p}_i - p_i)^2}. \tag{11}$$

The L_2-norm can be interpreted as a modified χ^2 distance. Consequently, the ambiguity set generated using this distance can be interpreted as the set of probabilities satisfying a goodness of fit test in relation to the empirical distribution (see Ref. 24 for details on this interpretation).

Using the L_2-norm, we define the ambiguity set as the set of all distributions within a given L_2 distance of the empirical distribution. We continue to denote the radius of ambiguity by r, where it will be clear from the context that this is measured in the L_2-norm. To this end, we define the L_2-norm ambiguity set as

$$\mathcal{P}_{\mathrm{L2}}(\ \hat{\mathbf{p}},\ r\ ) = \{\ \mathbf{p} \in \Omega^m\ :\ D_{\mathrm{L2}}(\hat{\mathbf{p}}, \mathbf{p}) \leq r\ \}. \tag{12}$$

Note that after a particular value of $r = r_{\max}$, the ambiguity set will contain all possible distributions, i.e. $\mathcal{P}_{\mathrm{L2}}(\hat{\mathbf{p}}, r_{\max}) = \Omega^m$. Increasing r beyond $r_{\max}$ will thus have no effect on $\mathcal{P}_{\mathrm{L2}}$. Note also that $r_{\max}$ depends on m. We introduce a normalized form of the radius of ambiguity to assist in the selection of r, and facilitate size comparisons between ambiguity sets generated using different distance functions. The normalized radius of ambiguity is defined as

$$\bar{r} = r/r_{\max} \in [0, 1]. \tag{13}$$

The allowable range of the radius of ambiguity can be expressed as $\bar{r} \in [0, 1]$, regardless of m.

Recall that when $\hat{\mathbf{p}}$ corresponds to the empirical distribution for an i.i.d. sample, we have the special case that $\hat{\mathbf{p}} = \frac{\mathbb{1}}{m}$. In this case we can derive $r_{\max}$ analytically:

$$r_{\max}(m) = D_{\mathrm{L2}}(\hat{\mathbf{p}}, \mathbf{e_1}) = \sqrt{\left(\frac{1}{m} - 1\right)^2 + \sum_{i=2}^{m} \left(\frac{1}{m}\right)^2} = \sqrt{1 - \frac{1}{m}}, \tag{14}$$

where the degenerate distribution $\mathbf{e}_i \in \Omega^m$ has $e_i = 1$, and all other entries set to 0.

Beyond a certain value of $\bar{r}$, $\mathcal{P}_{\mathrm{L2}}$ will not be contained within the interior of the probability simplex Ω^m, and will thus include distributions that assign a probability of zero to one or more outcomes in the sample. A criticism of the L_2-norm distance is that as the ambiguity set grows, it will include such distributions before other distributions that are more plausible, i.e. those that assign all outcomes in the sample a non-zero probability.

Solving the inner maximization problem given by Eq. (8) using an L_2-norm ambiguity set involves finding the distribution $\mathbf{p} \in \mathcal{P}_{\mathrm{L2}}$ that generates the worst possible expected performance of a given design. Given a design, $\mathbf{x}$, a sample of the uncertain parameters, $\mathbf{u}_i$, for $i = 1, ..., m$, and a radius of ambiguity r, this can be formulated as a convex optimization problem, involving a linear objective function subject to linear and second-order cone constraints:

$$\begin{aligned} \max_{\mathbf{p}} \quad & \sum_{i=1}^{m} p_i Q(\mathbf{x}, \mathbf{u}_i) \\ \text{s.t.} \quad & \|\hat{\mathbf{p}} - \mathbf{p}\|_2 \leq r \\ & \mathbf{p} \in \Omega^m \end{aligned} \tag{15}$$

Much of the appeal of using an ambiguity set defined by the L_2-norm is that this problem can be solved in closed-form using Algorithm 1. For a full discussion, derivation, and proof of this algorithm see Ref. 24.

Algorithm 1 Compute the worst-case probability distribution in a given L2 ball.

Input:

Discrete reference distribution $\hat{\mathbf{p}}$,
Function values at distribution support $Q(\mathbf{x}, \mathbf{u}_i) \equiv Q_i$,
Radius of L2 ambiguity set r.

Output:

Worst-case discrete distribution $\mathbf{p}^*$, on the same support as $\hat{\mathbf{p}}$.

1: $m = \text{length}(\hat{\mathbf{p}})$
2: $K = \{1, 2, ..., m\}$
3: **while** $|K| > 1$ **do**
4: $\quad k = m - |K|$
5: $\quad \bar{Q} = \dfrac{1}{(m-k)} \sum\limits_{i \in K} Q_i$
6: $\quad s = \sqrt{\dfrac{1}{(m-k)} \sum\limits_{i \in K} (Q_i^2 - \bar{Q}^2)}$
7: $\quad$ **if** $k = 0$ **then**
8: $\quad\quad p_i = \hat{p}_i + \dfrac{Q_i - \bar{Q}}{\sqrt{m}\, s} r,\ i \in K$
9: $\quad$ **else**
10: $\quad\quad$ **for** $i \in K$ **do**
11: $\quad\quad\quad p_i = \hat{p}_i + \dfrac{1}{(m-k)} \left(\sum\limits_{i \notin K} \hat{p}_i + \sqrt{(m-k)(r^2 - \sum\limits_{i \notin K} \hat{p}_i^2) - (\sum\limits_{i \notin K} \hat{p})^2} \dfrac{Q_i - \bar{Q}}{s} \right)$
12: $\quad\quad$ **end for**
13: $\quad\quad p_i = 0,\ i \notin K$
14: $\quad$ **end if**
15: $\quad$ **if** $p_i \geq 0 \quad \forall i \in K$ **then**
16: $\quad\quad \mathbf{p}^* = \mathbf{p}$
17: $\quad\quad$ **return** $\mathbf{p}^*$
18: $\quad$ **else**
19: $\quad\quad$ Find critical $j \in K$. This is the last index of $p_i < 0$ to become positive as we decrease r. This can be done by setting $p_i = 0$ in line 10 and solving for r.
20: $\quad\quad$ Set $K = K \setminus \{j\}$.
21: $\quad$ **end if**
22: **end while**
23: $p_i = 0, \quad i \notin K$
24: $p_i = 1, \quad i \in K$
25: $\mathbf{p}^* = \mathbf{p}$
26: **return** $\mathbf{p}^*$

III.C.2. KL divergence ambiguity

Another function that can be used to compare the degree of similarity between two discrete probability distributions is the KL divergence (or relative entropy). The KL divergence has been used extensively in the literature to construct ambiguity sets for DRO.[14–17] The KL divergence between the reference distribution, $\hat{\mathbf{p}}$, and a distribution (of equal dimension), $\mathbf{p}$, is defined by

$$D_{\mathrm{KL}}(\hat{\mathbf{p}}, \mathbf{p}) = \sum_{i=1}^{m} \hat{p}_i \log\left(\frac{\hat{p}_i}{p_i}\right). \tag{16}$$

The KL divergence has two cases that require special consideration. In this work, we adopt the conventions that for any non-zero a,

$$0 \log \frac{0}{a} = 0, \tag{17}$$

$$a \log \frac{a}{0} = \infty. \tag{18}$$

The first convention (17) is adopted because $\lim_{x \to 0} x \log(x) = 0$. The second convention (18) ensures that a distribution that assigns a zero probability to a realization that was observed in the sample will result in an infinite KL divergence. This ensures that such distributions will not be included in the KL divergence ambiguity set. This is a favorable contrast with the L_2-norm ambiguity set, which can include distributions that assign realizations observed in the sample a likelihood of zero.

Since the KL divergence function is not symmetric, the ordering of the arguments matters. In this work we choose the reference distribution $\hat{\mathbf{p}}$ as the first argument when constructing the KL ambiguity set. This is motivated by a recent paper by Van Parys et al., which showed that this choice is favorable (see Ref. 15 for more details).

We define the KL ambiguity set in a way that is analogous to the L_2-norm ambiguity set. We denote the radius of ambiguity by r, where it will be clear from the context when this is measured in terms of the KL divergence, and define

$$\mathcal{P}_{\mathrm{KL}}(\ \hat{\mathbf{p}},\ r\ ) = \{\ \mathbf{p} \in \Omega^m\ :\ D_{\mathrm{KL}}(\hat{\mathbf{p}}, \mathbf{p}) \leq r\ \}. \tag{19}$$

As in the previous section we define a normalized radius of ambiguity $\bar{r}$. However in this case there is no well-defined maximum radius $r_{\max}$. In fact, we could increase r indefinitely, with the boundaries of $\mathcal{P}_{\mathrm{KL}}$ getting asymptotically close to the probability simplex Ω^m. Instead of normalizing by the maximum radius as we did in the L_2-norm case, we instead define the normalized KL divergence radius of ambiguity to be

$$\bar{r} = \frac{\max_{\mathbf{p} \in \mathcal{P}_{\mathrm{KL}}} \|\mathbf{p} - \hat{\mathbf{p}}\|_\infty}{1 - \frac{1}{m}} \in [0, 1]. \tag{20}$$

This is essentially a normalized L_∞-norm radius. The maximum L_∞ norm gives the maximum difference that any single probability is allowed to vary, over all distributions within the ambiguity set. Note that there are other valid ways to parameterize the size of the ambiguity set, e.g., by measuring the volume proportion of Ω^m contained in the ambiguity set. However, care must be taken that the chosen parameterization is consistent across different dimensions, m, and distance function D.

Given a design $\mathbf{x}$, a sample of the uncertain parameters $\mathbf{u}_i$, for $i = 1, ..., m$, and a KL divergence radius of ambiguity r, the inner maximization problem given by Eq. (8) can be formulated as a convex optimization problem. This is done by introducing new variables $z_i, i = 1, ..., m$ and performing simple algebraic manipulation in order to reformulate the KL divergence constraint to yield linear and exponential cone type constraints:

$$\sum_{i=1}^{m} \hat{p}_i \log\left(\frac{\hat{p}_i}{p_i}\right) \leq r \iff \begin{cases} p_i \geq \hat{p}_i \exp\left(\dfrac{-z_i}{\hat{p}_i}\right), & i = 1, \ldots, m \\ \displaystyle\sum_{i=1}^{m} z_i \leq r \end{cases} \tag{21}$$

Under this reformulation, the maximization can be written as

$$\begin{aligned} \max_{\mathbf{p},\mathbf{z}} \quad & \sum_{i=1}^{m} p_i Q(\mathbf{x}, \mathbf{u}_i) \\ \text{s.t.} \quad & p_i \geq \hat{p}_i \exp\left(\frac{-z_i}{\hat{p}_i}\right), \quad i = 1, \ldots, m \\ & \sum_{i=1}^{m} z_i \leq r \\ & \mathbf{p} \in \Omega^m \end{aligned} \tag{22}$$

The problem can be solved in this form by many commercial convex optimization packages, for example the CVX package,[25,26] which uses a successive approximation method. However, since this problem needs to be re-solved at every iteration of the design optimization, the computational cost of a brute force optimization such as this would quickly accumulate. To achieve further speedup, the optimization problem in Eqn. 22 has been analyzed in detail in the literature, and has been shown to be reducible to a scalar root-finding problem. For details on this analysis, see Ref. 27. Algorithm 2 outlines how to solve the optimization problem using this approach, which is far more computationally efficient than solving Eqn. 22 directly.

Algorithm 2 Compute worst-case distribution within a KL ball

Input:

Discrete reference distribution $\hat{\mathbf{p}}$,
Function values at distribution support $Q(\mathbf{x}, \mathbf{u}_i) \equiv Q_i$,
Radius of KL ambiguity set r.

Output:

Worst-case discrete distribution $\mathbf{p}^*$, on the same support as $\hat{\mathbf{p}}$.

$m = \text{length}(\hat{\mathbf{p}})$
$K = \{i : \hat{p}_i = 0\}$
$\tilde{K} = \{i : \hat{p}_i > 0\}$
Define $f(\nu) = \sum_{i \in \tilde{K}} \hat{p}_i \log(\nu - Q_i) + \log\left(\sum_{i \in \tilde{K}} \frac{\hat{p}_i}{\nu - Q_i}\right)$, for $\nu > \max_{i \in \tilde{K}} Q_i$
for $i \in K$ **do**
    $p_i = 0$
end for
$I^* = K \cap \text{argmax}_i \, Q_i$
if $\exists \, k \in I^*$ **such that** $f(Q_k) < r$ **then** ▷ If multiple k exist, choose any one
    $\nu = Q_k$
    $r = 1 - \exp(f(\nu) - r)$
    $p_k = r$
else
    Find ν such that $f(\nu) = r$ using a line search
    $r = 0$
end if
for $i \in \tilde{K}$ **do**
    $q_i = \frac{\hat{p}_i}{\nu - Q_i}$
end for
for $i \in \tilde{K}$ **do**
    $p_i = \frac{(1 - r) q_i}{\sum_{i=1}^{m} q_i}$
end for
$\mathbf{p}^* = \mathbf{p}$
return $\mathbf{p}^*$

III.C.3. Wasserstein ambiguity

The third function we consider for the construction of the ambiguity set is the Wasserstein distance. Ambiguity sets based on Wasserstein distance are common in the DRO literature, and have been shown to produce tractable formulations with good performance in a variety of settings.[20,28–30] A notable point of difference in our setting is that we restrict our attention to considering distributions with a fixed discrete support, namely the sample $\mathbf{u}_i, i = 1, ..., m$.

Informally, the Wasserstein distance measures the degree of dissimilarity between two probability distributions by computing the cost of transforming one distribution into the other by transporting probability mass. The Wasserstein distance of order k is defined as the k^{th} root of the cost incurred when performing such a transformation in an optimal way. In this case, we define the cost of transporting a unit of probability mass from $\mathbf{u}_i$ to $\mathbf{u}_j$ to be the k^{th} power of the Euclidean distance, i.e., $d_{ij} = \|\mathbf{u}_i - \mathbf{u}_j\|_2^k$. In this work we will consider the Wasserstein distance with $k = 1$, which is often referred to as the earth-movers distance.

Mathematically, we can compute the Wasserstein distance between two discrete distributions, $\hat{\mathbf{p}}$ and $\mathbf{p}$, by solving a transportation problem in the form of a linear program:

$$\begin{aligned} D_{\text{W}}(\hat{\mathbf{p}}, \mathbf{p}, \mathbf{u}_1, ..., \mathbf{u}_m) = \min_{f_{ij}} & \sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij} d_{ij} \\ \text{s.t.} & \sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij} = 1 \\ & \sum_{i=1}^{m} f_{ij} \le \hat{p}_i \\ & \sum_{j=1}^{m} f_{ij} \le p_i \\ & f_{ij} \ge 0 \qquad \forall\, i, j \end{aligned} \tag{23}$$

We define the Wasserstein ambiguity set in a way that is analogous to the L_2-norm and KL divergence ambiguity sets. We again denote the radius of ambiguity by r, where it will be clear from the context when this is measured in terms of the Wasserstein distance, and define

$$\mathcal{P}_{\text{W}}(\, \hat{\mathbf{p}}, \mathbf{u}_1, ..., \mathbf{u}_m,\ r\,) = \{\, \mathbf{p} \in \Omega^m \ :\ D_{\text{W}}(\hat{\mathbf{p}}, \mathbf{p}, \mathbf{u}_1, ..., \mathbf{u}_m) \le r \,\}\,. \tag{24}$$

As in the previous sections we wish to define, for convenience, a normalized radius of ambiguity

$$\bar{r} = r / r_{\max} \in [0, 1]\,. \tag{25}$$

In this case, $r_{\max}$ represents the maximum possible Wasserstein distance between the empirical distribution $\hat{\mathbf{p}}$, and another distribution $\mathbf{p}$, both supported on a random sample $\mathbf{u}_1, ..., \mathbf{u}_m$ drawn from $\mathbb{P}_{\mathbf{u}}$. It follows that

$$r_{\max} = \max \|\mathbf{u}_i - \mathbf{u}_j\|_2^k \left(1 - \frac{1}{m}\right) \tag{26}$$

where the multiplier on the right follows from the fact that $\hat{\mathbf{p}} = \frac{\mathbb{1}}{m}$ is fixed, and the maximizing $\mathbf{p}$ will be always be a degenerate distribution. If the true distribution $\mathbb{P}_{\mathbf{u}}$ is known, then the maximization is taken over all possible realizations of $\mathbf{u}_i$ and $\mathbf{u}_j$ from $\mathbb{P}_{\mathbf{u}}$. If $\mathbb{P}_{\mathbf{u}}$ has finite support, then $\max \|\mathbf{u}_i - \mathbf{u}_j\|_2^k$ can often be found analytically by taking $\mathbf{u}_i$ and $\mathbf{u}_j$ to lie on the boundary of the support. If $\mathbb{P}_{\mathbf{u}}$ has infinite support, $r_{\max}$ will be unbounded in general. In this case, we choose to define $r_{\max}$ by truncating the support of $\mathbb{P}_{\mathbf{u}}$ appropriately (e.g., using a 99% confidence interval or 3-σ bounds). In the general case that $\mathbb{P}_{\mathbf{u}}$ is not known, we define r_{max} by taking the maximimum over realizations $\mathbf{u}_i$ and $\mathbf{u_j}$ observed in our sample.

Given a design $\mathbf{x}$, a sample of the uncertain parameters $\mathbf{u}_i$, for $i = 1, ..., m$, and a Wasserstein distance radius of ambiguity r, the inner maximization problem given by Eq. (8) can be formulated as a convex optimization problem:

$$\begin{aligned} \max_{\mathbf{p}} \quad & \sum_{i=1}^{m} p_i Q(\mathbf{x}, \mathbf{u}_i) \\ \text{s.t.} \quad & D_{\text{W}}(\hat{\mathbf{p}}, \mathbf{p}, \mathbf{u}_1, ..., \mathbf{u}_m) \le r \end{aligned} \tag{27}$$

Note that the implicit minimization that arises in the constraint via the definition of D_W (in (23)) can be ignored. This is because finding a feasible (not necessarily optimal) objective value in (23) that satisfies the constraint in (27), guarantees that the optimal objective value in (23) will also satisfy the constraint in (27).

III.D. Solving the outer optimization problem

Recall that the outer optimization in the distributionally robust design optimization problem (7) involves finding the design that optimizes the worst-case performance over all distributions within the ambiguity set.

Using the methods outlined in the previous sections, we can solve the inner problem to compute a worst-case distribution $\mathbf{p}^*(\mathbf{x})$, for a given design, $\mathbf{x}$. We can thus write the outer optimization problem in the following form:

$$\min_{\mathbf{x}\in\mathcal{X}} \sum_{i=1}^{m} p_i^*(\mathbf{x}) Q(\mathbf{x}, \mathbf{u}_i). \tag{28}$$

Note that this is similar to the form of the SAA problem, $\mathcal{S}$ (Eqn. 6), with the worst-case distribution $\mathbf{p}^*(\mathbf{x})$ taking the place of the empirical distribution $\hat{\mathbf{p}}$.

We adopt an iterative approach to solving this optimization problem. The initial design, $\mathbf{x} = \mathbf{x}_0$, is typically the current or nominal design that we wish to further optimize. For example, in the acoustic horn model problem (detailed in appendix A) we set the initial design to be a simple linearly expanding horn flare, and then optimize this further by considering more complex geometries. At each iteration of the outer design loop, we compute $Q(\mathbf{x}, \mathbf{u}_i)$ for $i = 1, ..., m$. We then solve the inner optimization problem to compute $\mathbf{p}^*(\mathbf{x})$, using the approaches outlined in previous sections. In the engineering design optimization setting, the computational cost is dominated by the cost of evaluating the expensive black-box function $Q(\mathbf{x}, \mathbf{u}_i)$. Note that by our construction, solving the inner problem requires no additional evaluations of the expensive function $Q(\mathbf{x}, \mathbf{u}_i)$, so that we are able to compute the objective function of the outer problem (28) using only m evaluations of the expensive black-box function. This is the same number of function evaluations per iteration as would be required by an SAA approach. By this argument we make the case that the computational cost of the DRO approach is roughly equivalent to the SAA approach.

Given the ability to compute the objective function as described above, there exist many different strategies for iteratively improving the design. The overall computational cost of the design process will depend on how many iterations the strategy requires to reach an optimal design. Comparison of different algorithms for solving this outer problem is outside the scope of the current work. Details of the specific algorithm we use to solve the acoustic horn illustrative design problem are presented in the following section.

IV. Performance of the Distributionally Robust Approach

This section explores the performance of the DRO design methodology outlined in the previous section. In Section IV.A we investigate how the method trades off mean performance with the risk level in performance, and also compare and contrast the performance of distance functions used to construct the ambiguity set. Section IV.B then analyzes how the performance of the DRO approach depends on the distribution of the underlying uncertainty.

IV.A. Mean-risk tradeoff

Section III.B discussed how enlarging the ambiguity set trades off mean performance over all sample draws, versus robustness in performance between draws. In this section, we study how effectively the DRO methodology is able to make this trade-off, for the L_2-norm, KL divergence, and Wasserstein distance ambiguity sets. We do this using computational experiments on the acoustic horn design problem (described in Appendix A). In an effort to minimize the effect of random sampling, we use the same $T = 500$ random samples from $\mathbb{P}_{\mathbf{u}}$ when comparing the different ambiguity sets. For each sample, we compute distributionally robust designs $\mathbf{x}_m$ by solving problem $\mathcal{D}$ using each of the three formulations of the ambiguity set. In the problem we study, the black-box function that returns $Q(\mathbf{x}, \mathbf{u}_i)$ also returns derivative information, namely, $\frac{\partial Q((\mathbf{x}, \mathbf{u}_i))}{\partial \mathbf{x}}$. This allows us to adopt a gradient-based iterative procedure to solve the outer optimization problem described in Section III.D. In particular, we use an interior point method[b]. To compute each design we run the algorithm

[b] As implemented in the function `fmincon`, included in the `MATLAB` Optimization Toolbox[22]

for 100 iterations. This corresponds to a scenario where the designer has a fixed computational budget of $100m$ black-box function evaluations. This approach works well for this particular application, since the QoI is generally smooth and its derivatives are readily available. In cases where derivative information is not available, or the black-box function $Q(\mathbf{x}, \mathbf{u}_i)$ is highly non-linear and/or non-convex, a more careful treatment of the outer problem would be required. However, we note that such careful treatment would also be required to use the SAA approach to optimize such a function.

For each design we evaluate the out-of-sample performance using Eqn. 1. These designs will in general differ depending on the radius of ambiguity used. Hence we repeat the experiment (using the same samples) for 21 different values of $\bar{r}$ encompassing the full range $\bar{r} \in [0, 1]$. For each $\bar{r}$ value we compute the mean performance, $\mu(Z_m)$, and the 95th percentile or risk level in performance, $\rho(Z_m)$, over the T sample draws (see Eqn. 3 and Eqn. 4). Additionally, in order to provide further information about the spread in performance of the resulting designs over the T sample draws, Table 1 in Appendix B presents additional statistics of Z_m^t over the T sample draws, for each methodology studied in this section.

Using this information the performance of the DRO methodology for each formulation of the ambiguity set can be described by a curve, parameterized by $\bar{r}$, of mean performance versus risk in performance over all sample draws. Figure 3 below plots these mean-risk trade-off curves for sample sizes of $m = 5, 10$, and 20. Also shown for comparison is our estimate of the true optimum performance, Z_∞, that would be obtainable if the designer had an infinite computational budget.

To analyze these results, it is worth recalling that when the radius of ambiguity is zero, the DRO problem, $\mathcal{D}$, is equivalent to the SAA problem, $\mathcal{S}$. As we increase the radius of ambiguity, $\bar{r}$, we increase the amount of distributional robustness enforced in the designs. At first, adding robustness leads to an improvement in both the mean and risk objectives, for all sample sizes studied. This means that using a small value of $\bar{r} > 0$ produces better average out-of-sample performance than using SAA and optimizing directly for average in-sample performance. However, after a critical radius of ambiguity is reached, increasing the radius further results in worse mean performance, and a greater risk level in performance. The potential improvement gained by using the DRO methodology over the SAA methodology increases as the sample size, m, is reduced. This suggests that the DRO method is most effective when available data are most scarce. The mean-risk curve corresponding to the KL divergence ambiguity set reaches the critical point at a better mean and risk level than the mean-risk curve for the L_2-norm or Wasserstein ambiguity sets. This means that if the optimal value of $\bar{r}$ is chosen for each sample size and ambiguity set combination, the DRO approach using a KL divergence ambiguity set will, on average, produce designs with the best mean performance and lowest risk level in performance.

We define the optimality gap as the difference between the best performance attainable by a method and the true optimum performance. For $m = 5$, the DRO approach with KL divergence ambiguity set is able to reduce the opimality gap in mean performance by 66% compared to SAA. For $m = 10$ and $m = 20$, the reductions in the mean performance optimality gap are 65% and 46%, respectively. The corresponding reductions in the optimality gap in terms of the risk level in performance are 56%, 69%, and 51% for $m = 5, 10, 20$ respectively.

IV.B. Effect of the underlying distribution

In this section we investigate how the shape of the probability distribution governing the uncertain parameters affects the performance of the DRO approach for design using small sample sizes. When applying the methodology to the acoustic horn model problem in the previous sections, we used a uniform truth distribution for the uncertain parameter. In this section, we repeat the previous experiments, this time with a new truth distribution, given by:

$$\mathbf{u} \sim \mathbb{P}_u^{\mathrm{N}} = \mathrm{Normal}(1.4, 0.0577). \tag{29}$$

The mean and standard deviation of the normal distribution were chosen to be equal to those of the uniform distribution used in the previous sections. The computational experiments outlined in Section IV.A were repeated using the normal distribution. Figure 4 shows the mean performance vs. the risk level in performance over all T samples. The analogous mean-risk curves for the uniform distribution (as seen in Figure 3) are shown for comparison. Also shown is the estimated true optimum performance, Z_∞, for the uniform and normal distributions respectively. Table 2 in Appendix B presents additional statistics of Z_m^t over the T sample draws, for the normally distributed uncertainty.

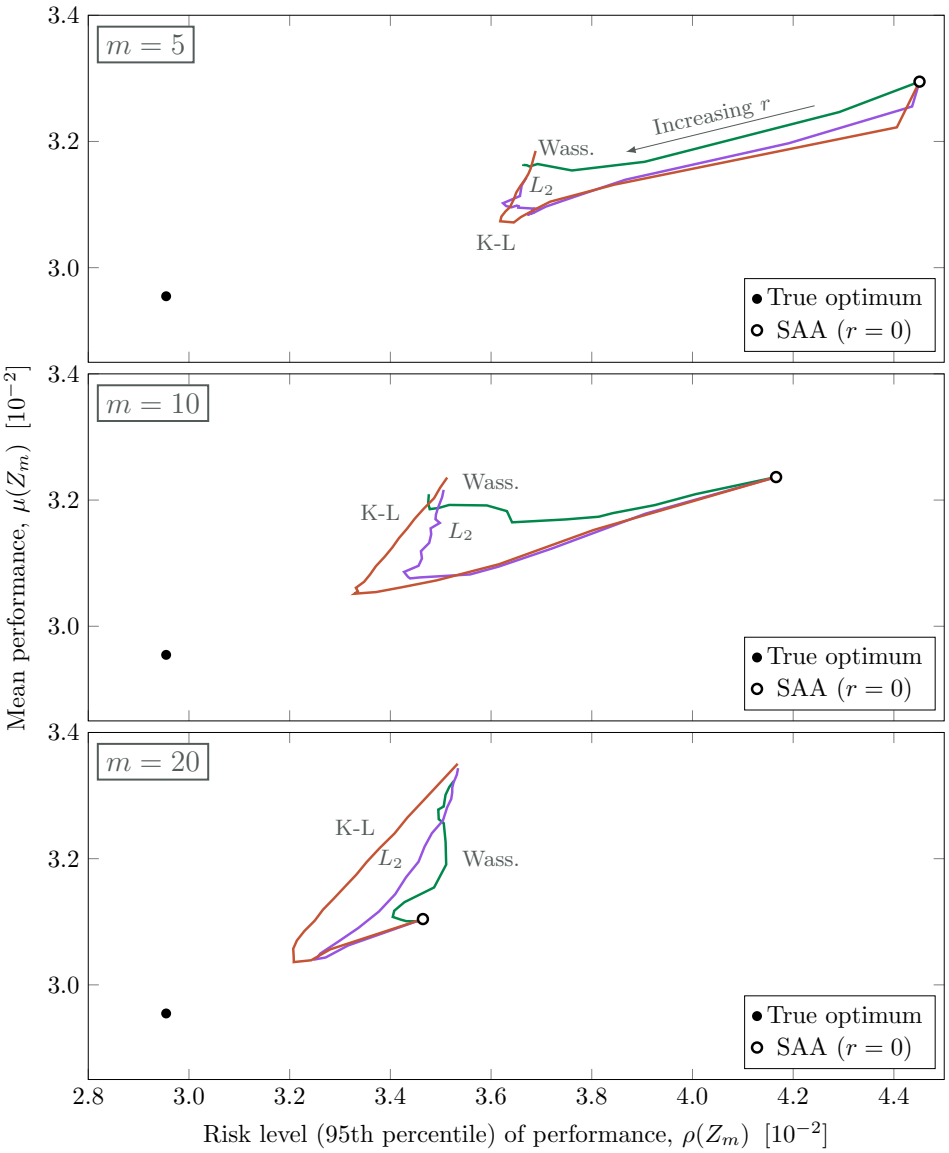

Figure 3: Mean-risk tradeoff curves for designs computed using sample draws of size $m = 5, 10, 20$ and the DRO approach with L_2-norm, KL divergence, and Wasserstein distance ambiguity sets. Also shown is the true optimal performance.

The mean-risk curves for the two distributions are qualitatively similar. Under the normally distributed uncertainty, increasing the radius of ambiguity from zero still almost always results in an improvement in both mean performance across the samples, and the risk level in performance across the samples. However, in comparison to the uniformly distributed uncertainty, the critical radius–after which a further increase is detrimental–occurs much sooner. As a result, the potential performance gain by using the DRO approach compared to the SAA approach (which is equivalent to $r = 0$) is smaller for the normally distributed uncertainty. Furthermore, there appears to be almost no difference in performance between the L_2-norm and K-L divergence ambiguity sets under normally distributed uncertainty. In this case the Wasserstein distance ambiguity set again performs worse than the L_2-norm and K-L divergence ambiguity sets, giving only a slight improvement in mean and risk level for the $m = 5$ and $m = 10$ cases. In the $m = 20$ case, the Wasserstein ambiguity set provides negligible benefit over the SAA approach.

These results show that for the acoustic horn design problem we study, the DRO methodology is still able to outperform the SAA approach under normally distributed uncertainty, but to a lesser extent than under uniformly distributed uncertainty.

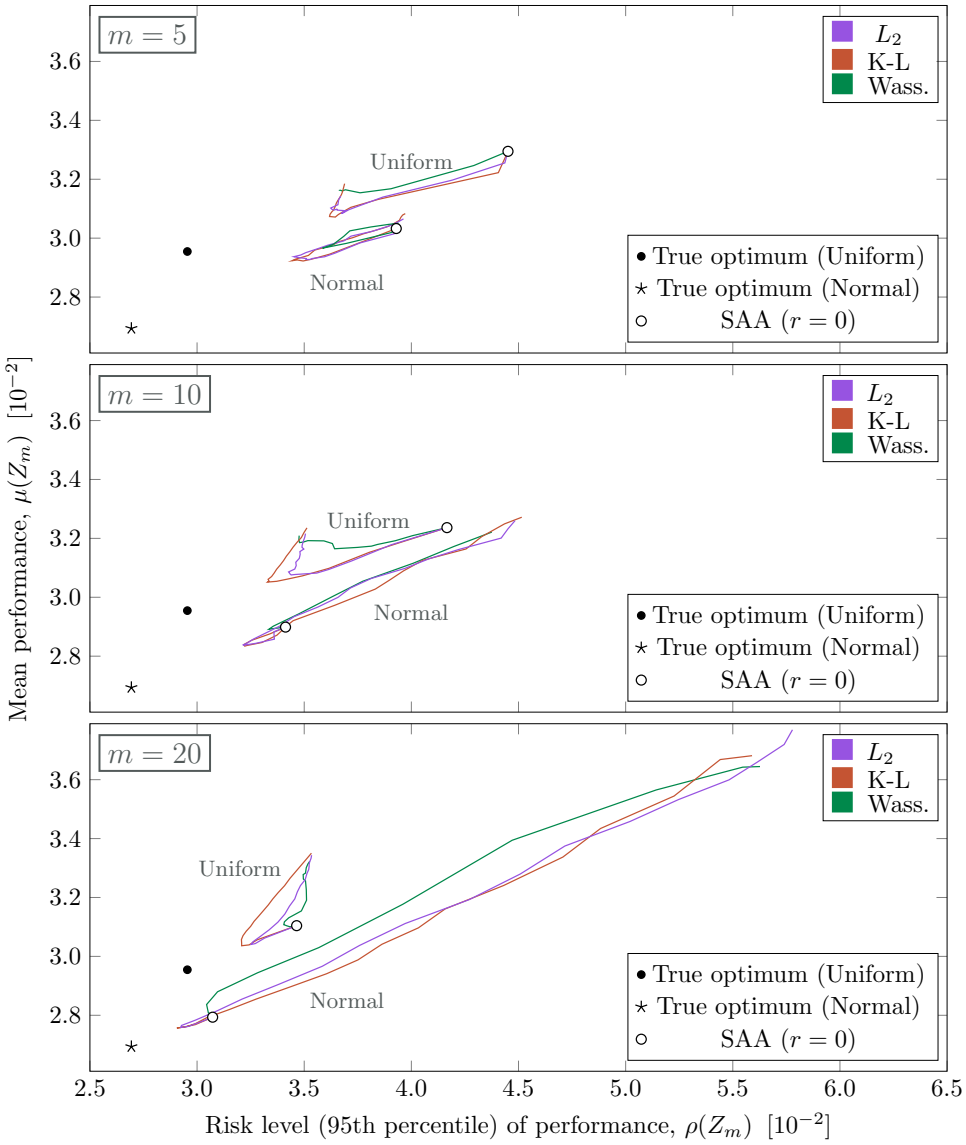

Figure 4: Comparison between mean-risk trade-off curves for designs computed using normally distributed and uniformly distributed uncertainty. Results are for sample sizes $m = 5, 10, 20$ and the DRO approach with both L_2-norm and K-L divergence ambiguity sets.

V. Comparison Between Distributional Robustness and Robustness Through Variance Reduction

In this section we compare the DRO methodology described in Section III, with a commonly used formulation for design under uncertainty that introduces robustness through variance reduction. To this end, we first analyze whether the addition of a variance reduction objective results in a reduction in the risk level of out-of-sample performance, and how this compares to the results obtained using a DRO approach. We then show that designs found using the DRO methodology naturally exhibit reduced out-of-sample variance, despite the absence of an explicit variance reduction objective in the formulation.

V.A. Formulation of variance reduction through multi-objective optimization (MOO)

A common approach to introducing robustness against uncertainty into a design is to augment the SAA objective in $\mathcal{S}$ (Eqn. 6) by adding a weighted penalty on the variability in the QoI. This amounts to casting the problem as a MOO problem, in which the designer optimizes for mean performance, while simultaneously

minimizing the variability in performance, under the distribution of the uncertainty. Note that in our context, we are ultimately not interested in reducing the out-of-sample variance in the performance of a given design. Instead, we are interested in how reducing in-sample variance during the design optimization can ultimately result in an improvement in the out-of-sample mean performance of designs, which is our primary interest. This section formulates a MOO problem adapted to the setting in which the designer only has access to a random sample of values of the uncertain variables.

In order to optimize for both the mean and variability of in-sample performance, we introduce a trade-off parameter λ, which governs the relative importance placed on the mean objective versus the standard deviation objective. Note that in the case $\lambda = 0$, the designer optimizes for the mean performance under the empirical distribution, and thus the MOO problem reduces to the SAA problem, $\mathcal{S}$ (Eqn. 6). Increasing λ increases the degree of importance placed on achieving robustness through variance reduction. In this way, selecting the λ parameter is analogous to selecting the $\bar{r}$ parameter for the ambiguity set in the DRO problem, $\mathcal{D}$. The MOO problem for jointly optimizing the in-sample mean and standard deviation in performance over a sample of the uncertain parameters, $\mathbf{u}_1, ..., \mathbf{u}_m$, with associated empirical distribution $\hat{\mathbf{p}}$, can be written as

$$\mathcal{M}(\lambda): \quad \min_{\mathbf{x}\in\mathcal{X}} \quad (1-\lambda)\underbrace{\mathbb{E}_{\hat{\mathbf{p}}}\left[Q(\mathbf{x},\mathbf{u})\right]}_{\text{Mean}} + \lambda\underbrace{\sigma_{\hat{\mathbf{p}}}\left[Q(\mathbf{x},\mathbf{u})\right]}_{\text{Std. Dev.}}, \tag{30}$$

where

$$\begin{aligned}
\mathbb{E}_{\hat{\mathbf{p}}}\left[Q(\mathbf{x},\mathbf{u})\right] &= \sum_{i=1}^{m} \hat{p}_i Q(\mathbf{x},\mathbf{u}_i), \\
\sigma_{\hat{\mathbf{p}}}\left[Q(\mathbf{x},\mathbf{u})\right] &= \sqrt{\sum_{i=1}^{m} \hat{p}_i (Q(\mathbf{x},\mathbf{u}_i) - \mathbb{E}_{\hat{\mathbf{p}}}\left[Q(\mathbf{x},\mathbf{u})\right])^2}, \\
\lambda &\in [0,1].
\end{aligned}$$

For a given value of λ, we solve $\mathcal{M}(\lambda)$ (using the same gradient-based procedure as was used to solve the DRO and SAA problems), and denote the resulting optimal design by $\mathbf{x}_m(\lambda)$. As in Section II.A, we are interested in the out-of-sample performance of designs, i.e., how $\mathbf{x}_m(\lambda)$, performs under the true distribution $\mathbb{P}_{\mathbf{u}}$, rather than the sample distribution $\hat{\mathbf{p}}$. In addition to the out-of-sample mean performance of a design, $Z(\mathbf{x})$ (Eqn. 1), we also define the out-of-sample standard deviation in the performance of a design, and denote this by

$$S(\mathbf{x}) = \sigma_{\mathbb{P}_{\mathbf{u}}}\left[Q\left(\mathbf{x},\mathbf{u}\right)\right]. \tag{31}$$

Thus the out-of-sample standard deviation of a MOO design, $\mathbf{x}_m(\lambda)$, is denoted $S_m(\lambda) \equiv S(\mathbf{x}_m(\lambda))$. For comparison, we also define the true optimal MOO designs, $\mathbf{x}_\infty(\lambda)$, that the designer would only be able to obtain if they had an infinite computational budget, and could thus fully characterize the true distribution $\mathbb{P}_{\mathbf{Q}}$. The mean and standard deviation in performance of these designs are denoted by $Z_\infty \equiv Z_m(\mathbf{x}_\infty(\lambda))$ and $S_\infty \equiv S_m(\mathbf{x}_\infty(\lambda))$ respectively. Throughout this section, we suppose that the fixed truth distribution of the uncertainty is $\mathbb{P}_{\mathbf{u}} \sim \text{Uniform}\left[\,1.3,\,1.5\,\right]$, and we use the same $T = 500$ random sample draws as those used in Sections II and III.

V.B. Mean-Risk Tradeoff

In Section III we showed that when supplied with limited data the DRO methodology is able to produce designs with better out-of-sample performance than the SAA approach. It could be argued that the variance reduction objective in the MOO approach prevents the optimization from over-fitting to the given sample, thus leading to better out-of-sample performance. To investigate whether this is the case, we repeat the analysis of Section IV.A, this time using the MOO approach, and compare the results to those obtained using the DRO approach. These mean-risk trade-off curves for the MOO and DRO approaches are shown in Figure 5. In this figure, we show results for values of $\lambda \in [0, 0.5]$. Values of $\lambda \in (0.5, 1]$ are omitted, as they were found to result in increasingly poor out-of-sample mean performance and are therefore not of interest in our context. We see that the mean-risk curves for the MOO approach are qualitatively similar to those of the DRO approach, in the sense that increasing λ from zero (i.e., adding some importance to the variability objective) improves both the mean and the risk level in out-of-sample performance over repeated trials. As

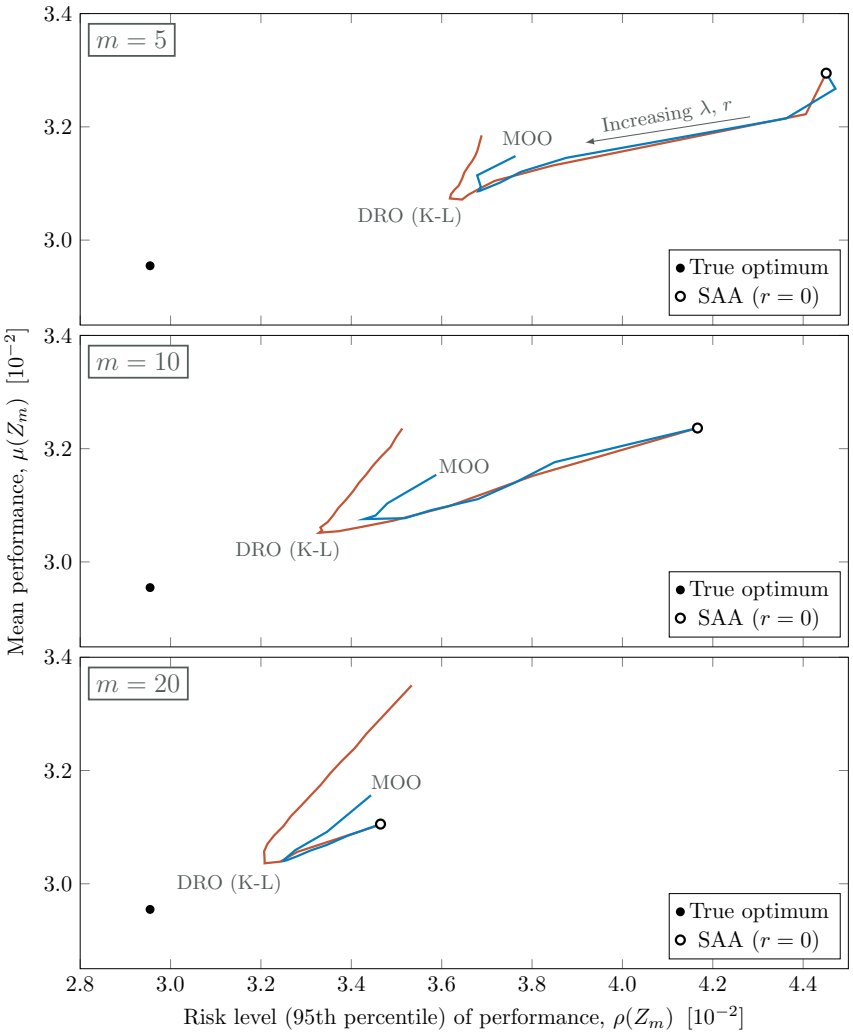

Figure 5: Mean versus 95th percentile curves generated using MOO, with varying sample size, for the acoustic horn design problem. Each curve corresponds to a sample size $m = 5, 10, 20$, and is computed by averaging over $T = 500$ samples, respectively. Each point on a curve corresponds to a particular value of $\lambda \in [\,0,\, 0.5\,]$.

in the DRO approach, increasing λ past a critical value leads to over-conservatism and consequently poor performance. We see that when the optimal values of λ and r are used, the DRO approach is able to achieve better mean performance, and a lower risk level in performance than the MOO approach.

From these results, we can conclude that adding a small variance reduction objective does make the design methodology robust to poor realizations of the sample, but to a lesser degree than optimizing for distributional robustness directly.

V.C. Mean-Variance Tradeoff

Traditionally when MOO is used in engineering design, the canonical result sought is the trade-off curve between the two objectives, in this case mean performance and variability in performance, parameterized by the trade-off parameter λ. When the true distribution $\mathbb{P}_{\mathbf{Q}}$ is available this curve is termed the *Pareto frontier*, and describes the performance of the set of designs that achieve an optimal convex combination of $Z_\infty(\lambda)$ and $S_\infty(\lambda)$. When $\mathbb{P}_{\mathbf{Q}}$ can not be perfectly characterized, e.g., due to a constrained computational budget, the true Pareto frontier must be approximated. In this section we investigate how the DRO and MOO methods compare in their ability to approximate the Pareto frontier using small samples.

It is important to note that the objective function in $\mathcal{M}(\lambda)$ involves only the *in-sample* mean and variability in performance. Thus, being an optimizer of $\mathcal{M}(\lambda)$ does not necessarily guarantee good *out-of-sample*

mean and variance in performance, and thus the MOO method may not result in the best approximation of the Pareto frontier.

In the case of the MOO approach, for each sample we compute a set of designs that solve $\mathcal{M}(\lambda)$ for values of $\lambda \in [0,\, 0.5]$. For each of these designs we compute the out-of-sample mean performance, $Z_m(\mathbf{x}_m(\lambda))$, and variability in performance, $S_m(\mathbf{x}_m(\lambda))$. These quantities are then averaged over the $T = 500$ samples. Similarly, using the DRO approach, for each sample we compute a set of designs that optimize $\mathcal{D}$, using a KL ambiguity set and 20 uniformly spaced values of $\bar{r} \in [0,\, 1]$. Again we compute $Z_m(\mathbf{x}_m)$ and $S_m(\mathbf{x}_m)$ for each design, and average these over all $T = 500$ samples.

Figure 6 shows the mean versus standard deviation trade-off curves, for the MOO and DRO approaches, for sample sizes $m = 5, 10, 20$. The Pareto frontier is included for comparison. Recall that the $\mu(\cdot)$ operator indicates the mean of these values over all of the T realizations (see Eqn. 3). In this figure, we once again show results for values of $\lambda \in [0, 0.5]$. Values of $\lambda \in (0.5, 1]$ are omitted, as they were found to result in increasingly poor out-of-sample mean performance, which is our primary interest.

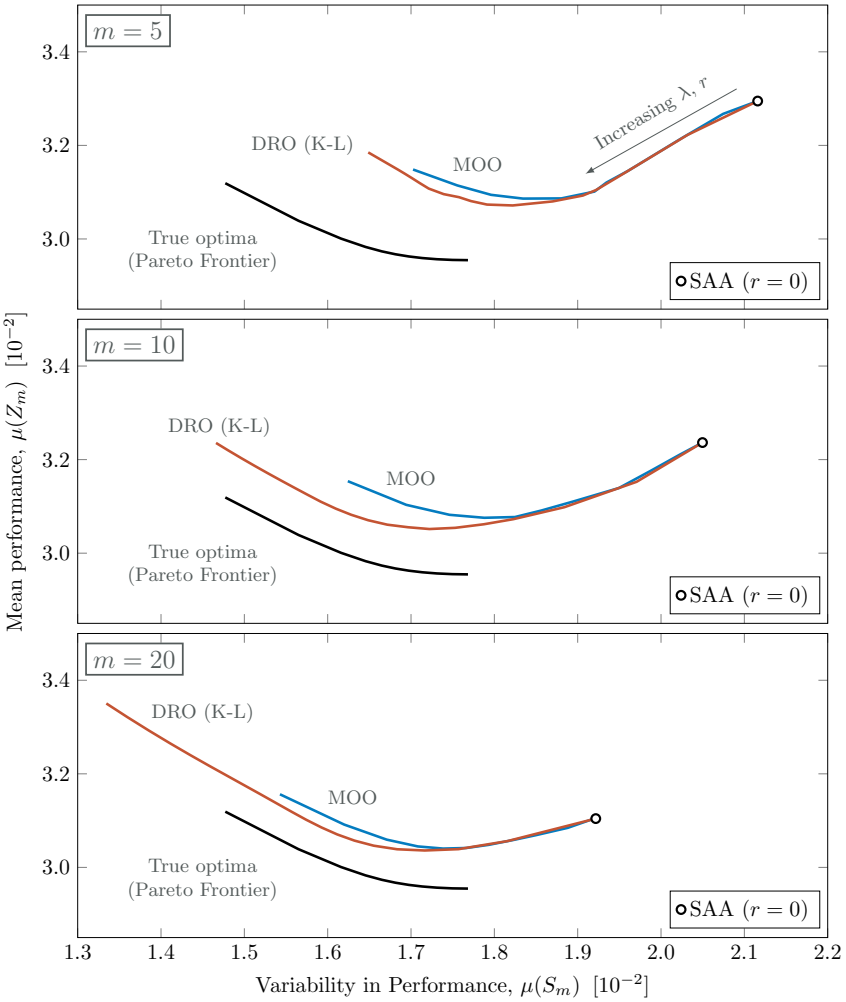

Figure 6: Mean versus standard deviation curves generated using MOO, with varying sample size, for the acoustic horn design problem. Each curve corresponds to a sample size $m = 5, 10, 20$, and is computed by averaging over $T = 500$ samples, respectively. Each point on a curve corresponds to a particular value of $\lambda \in [0,\, 0.5]$.

Focusing first on the MOO approach, we see that the mean-variability curves for a finite sample size, m, do not coincide with the Pareto frontier (which effectively corresponds to an infinite sample size). The shape of the curve also changes as the sample size is decreased. For small m, we see that optimizing for the in-sample mean performance does not give the best out-of-sample mean performance. Instead, shifting some importance onto the standard deviation objective (i.e., increasing λ from zero) actually results in an

improvement in the out-of-sample mean performance, while *also* achieving lower out-of-sample variability in performance. In this way it appears that more consistently performing designs are less susceptible to over-fitting, and are thus better able to generalize from in-sample to out-of-sample performance.

The mean-variability curves for the DRO approach are in-fact closer to the Pareto frontier than those obtained using the MOO approach. This means that for the acoustic horn problem we studied, designs found using DRO are able to achieve a closer to optimal combination of out-of-sample mean performance and variability in performance.

VI. Discussion

In Section IV we compared the DRO approach with the SAA approach, for different values of the radius of ambiguity r. In all the cases studied, we see that the marginal value of the DRO approach, i.e., the improvement in design performance using the DRO approach over the SAA approach in the limit as the radius of ambiguity tends to zero, is positive. This means that adding some robustness to the design problem was always beneficial. Furthermore, this improvement in performance occurs in both the average performance and the risk level in performance. Although robust optimization is often viewed as a risk averse approach, these results suggest that the DRO approach is also beneficial to a risk neutral designer.
Given that increasing the radius of ambiguity away from zero appears to be generally beneficial, a natural question to ask is whether the optimal radius of ambiguity can be selected a-priori. Although it is difficult to answer this concretely, our results reveal some trends that could guide this decision. In particular, we see that optimal value for the radius of ambiguity increases as the sample size is decreased. This agrees with the intuition that a greater degree of robustness is beneficial when the designer has less information with which to characterize the uncertainty. One difficulty that became apparent in this work is that the optimal value of the radius of ambiguity heavily depends on how it is parameterized. In this work we suggest mitigating this by defining a normalized radius of ambiguity that is consistent across different formulations of the ambiguity set, and different sample sizes.

In Section V we analyzed the performance of the DRO approach against a mean-variance MOO approach. These results showed that the DRO designs are able to achieve a better mean and variance in performance when compared with the MOO designs. This is in spite of the fact that DRO never explicitly optimizes for a reduction of variability. This suggests that when the designers objective is to minimize out-of-sample variability in performance, a DRO formulation could perform better than minimizing in-sample variance.

Another important question is how the benefits of the DRO approach manifest in the resulting designs. To provide some insight and a possible answer to this question, consider that a single sample draw can be used to compute a single design $\mathbf{x}_m$, which is an estimate of the true optimal design $\mathbf{x}_\infty$. Since each design $\mathbf{x}_m$ obtained by SAA depends on the random sample drawn, the estimator $\mathbf{x}_m$ is a random variable. In other words, every time we draw a sample, we will compute a different realization of $\mathbf{x}_m$. Depending on the optimization problem, and underlying distribution of uncertainty, the estimator $\mathbf{x}_m$ can be biased or unbiased with a variance that decreases linearly with sample size. If we compute $\mathbf{x}_m$ using DRO, then the variance of the estimator for a given sample size will in general be smaller than the variance of the SAA estimator due to estimator shrinkage.[31] The DRO estimator will have a bias even if the SAA estimator is unbiased, but the detrimental effect of this bias on out-of-sample performance will be dominated by the improvements due to shrinkage. In summary, the DRO approach outperforms the SAA approach on average by way of reducing the variance in the resulting designs, across possible sample draws.

Although the distributionally robust approach in this paper has been demonstrated on a single problem, the formulation and algorithms make no specific assumptions about the structure of the problem. The approach is therefore applicable to a wide range of design problems, and is particularly suitable for problems where only a small number of samples can be evaluated to solve the optimization under uncertainty problem. The effects of uncertainty on the quantity of interest vary depending on the system in question. Testing the DRO approach on additional design problems could reveal whether certain characteristics of the system of interest influence the performance of the method.

In this paper we have investigated uniformly and normally distributed uncertainty. Our results suggest that normally distributed uncertainty is less amenable to the DRO approach than uniformly distributed uncertainty for the design problem we studied. This suggests that characteristics of the underlying distribution (e.g., skewness, heavy tails) of the uncertainty could influence the performance of the method. Further study into different distributions would be beneficial in order to gain reliable insights as to which distributions are

most amenable to the DRO methodology.

We considered the setting in which the designer has access to a sample of realized values of the uncertain parameters. Another avenue for future work is to extend the distributionally robust design formulation to alternative settings. For example, the formulation could be applied to the situation where the designer has limited knowledge about the distribution of uncertain parameters. For example, if the designer has knowledge about moments of the distribution, they could construct an ambiguity set that includes moment constraints, This type of ambiguity set has been explored previously in the literature.[12]

VII. Conclusions

This paper has demonstrated how engineering design under uncertainty problems can be formulated and solved using DRO. The DRO approach addresses the particular challenge of design under uncertainty using limited data and the overfitting that can occur. We have demonstrated the potential benefits of a distributionally robust approach using a practical design optimization under uncertainty problem. Our approach outperforms the widely used SAA and MOO-based approaches. In particular, the DRO designs have better out-of-sample performance than SAA or MOO designs, with greater gains when the number of samples to solve the optimization under uncertainty problem is small. The DRO formulation is computationally tractable, as the algorithms presented in this paper require the same computational budget per design iteration as SAA and MOO.

VIII. Acknowledgements

This work was supported by AFOSR grant FA9550-16-1-0108 under the Dynamic Data Driven Application Systems Program, by the Defense Advanced Research Projects Agency [EQUiPS program, award W911NF-15-2-0121, Program Manager F. Fahroo], by a New Zealand Marsden Fund grant under contract UOA1520, and by the SUTD-MIT International Design Center.

References

[1]Kennedy, M. C. and O'Hagan, A., "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 63, No. 3, 2001, pp. 425–464.

[2]Beyer, H.-G. and Sendhoff, B., "Robust Optimization–A Comprehensive Survey," *Computer Methods in Applied Mechanics and Engineering*, Vol. 196, No. 33, 2007, pp. 3190–3218.

[3]Du, X. and Chen, W., "Efficient Uncertainty Analysis Methods for Multidisciplinary Robust Design," *AIAA journal*, Vol. 40, No. 3, 2002, pp. 545–581.

[4]Du, X. and Chen, W., "Methodology for Managing the Effect of Uncertainty in Simulation-Based Design," *AIAA Journal*, Vol. 38, No. 8, 2000, pp. 1471–1478.

[5]Keane, A. and Nair, P., *Computational Approaches for Aerospace Design: The Pursuit of Excellence*, John Wiley & Sons, 2005.

[6]Helton, J. C., Johnson, J. D., and Oberkampf, W. L., "An Exploration of Alternative Approaches to the Representation of Uncertainty in Model Predictions," *Reliability Engineering & System Safety*, Vol. 85, No. 1, 2004, pp. 39–71.

[7]Rao, S. and Cao, L., "Optimum Design of Mechanical Systems Involving Interval Parameters," *Transactions-American Society of Mechanical Engineers Journal of Mechanical Design*, Vol. 124, No. 3, 2002, pp. 465–472.

[8]Ben-Tal, A., El Ghaoui, L., and Nemirovski, A., *Robust Optimization*, Princeton University Press, 2009.

[9]Zaman, K., Rangavajhala, S., McDonald, M. P., and Mahadevan, S., "A Probabilistic Approach for Representation of Interval Uncertainty," *Reliability Engineering & System Safety*, Vol. 96, No. 1, 2011, pp. 117–130.

[10]Smets, P., "Probability, Possibility, Belief: Which and Where?" *Quantified Representation of Uncertainty and Imprecision*, Springer, 1998, pp. 1–24.

[11]Goh, J. and Sim, M., "Distributionally Robust Optimization and its Tractable Approximations," *Operations Research*, Vol. 58, No. 4-part-1, 2010, pp. 902–917.

[12]Delage, E. and Ye, Y., "Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems," *Operations research*, Vol. 58, No. 3, 2010, pp. 595–612.

[13]Wiesemann, W., Kuhn, D., and Sim, M., "Distributionally Robust Convex Optimization," *Operations Research*, Vol. 62, No. 6, 2014, pp. 1358–1376.

[14]Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G., "Robust Solutions of Optimization Problems Affected by Uncertain Probabilities," *Management Science*, Vol. 59, No. 2, 2013, pp. 341–357.

[15]Van Parys, B. P., Esfahani, P. M., and Kuhn, D., "From Data to Decisions: Distributionally Robust Optimization is Optimal," *arXiv preprint arXiv:1704.04118*, 2017.

[16]Hu, Z. and Hong, L. J., "Kullback-Leibler Divergence Constrained Distributionally Robust Optimization," *Optimization Online*, 2013.

[17]Calafiore, G. C., "Ambiguous Risk Measures and Optimal Robust Portfolios," *SIAM Journal on Optimization*, Vol. 18, No. 3, 2007, pp. 853–877.

[18]Bertsimas, D., Gupta, V., and Kallus, N., "Robust sample average approximation," *Mathematical Programming*, Vol. 171, No. 1-2, 2018, pp. 217–282.

[19]Gotoh, J.-y., Kim, M. J., and Lim, A. E., "Robust Empirical Optimization is Almost the Same as Mean–variance Optimization," *Operations Research Letters*, 2018.

[20]Esfahani, P. M. and Kuhn, D., "Data-Driven Distributionally Robust Optimization using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations," *arXiv preprint arXiv:1505.05116*, 2015.

[21]Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T., "The Sample Average Approximation Method for Stochastic Discrete Optimization," *SIAM Journal on Optimization*, Vol. 12, No. 2, 2002, pp. 479–502.

[22]"MATLAB Optimization Toolbox," MathWorks, Natick, MA, USA, Version 7.6, MATLAB 2017a.

[23]Xu, H., Caramanis, C., and Mannor, S., "A Distributional Interpretation of Robust Optimization," *Mathematics of Operations Research*, Vol. 37, No. 1, 2012, pp. 95–110.

[24]Philpott, A., de Matos, V., and Kapelevich, L., "Distributionally robust SDDP," *Computational Management Science*, 2017, pp. 1–24.

[25]Grant, M. and Boyd, S., "CVX: Matlab Software for Disciplined Convex Programming, Version 2.1," `http://cvxr.com/cvx`, March 2014.

[26]Grant, M. and Boyd, S., "Graph Implementations for Nonsmooth Convex Programs," *Recent Advances in Learning and Control*, edited by V. Blondel, S. Boyd, and H. Kimura, Lecture Notes in Control and Information Sciences, Springer-Verlag Limited, 2008, pp. 95–110, `http://stanford.edu/~boyd/graph_dcp.html`.

[27]Filippi, S., Cappé, O., and Garivier, A., "Optimism in Reinforcement Learning and Kullback-Leibler divergence," *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, IEEE, 2010, pp. 115–122.

[28]Pflug, G. and Wozabal, D., "Ambiguity in Portfolio Selection," *Quantitative Finance*, Vol. 7, No. 4, 2007, pp. 435–442.

[29]Shafieezadeh-Abadeh, S., Esfahani, P. M., and Kuhn, D., "Distributionally Robust Logistic Regression," *Advances in Neural Information Processing Systems*, NIPS, 2015, pp. 1576–1584.

[30]Gao, R. and Kleywegt, A. J., "Distributionally Robust Stochastic Optimization with Wasserstein distance," *arXiv preprint arXiv:1604.02199*, 2016.

[31]Copas, J. B., "Regression, Prediction and Shrinkage," *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 45, No. 3, 1983, pp. 311–335.

[32]Ng, L. W. and Willcox, K. E., "Multifidelity Approaches for Optimization Under Uncertainty," *International Journal for Numerical Methods in Engineering*, Vol. 100, No. 10, 2014, pp. 746–772.

[33]Eftang, J. L., Huynh, D., Knezevic, D. J., and Patera, A. T., "A Two-step Certified Reduced Basis Method," *Journal of Scientific Computing*, Vol. 51, No. 1, 2012, pp. 28–58.

[34]Udawalpola, R. and Berggren, M., "Optimization of an Acoustic Horn with Respect to Efficiency and Directivity," *International Journal for Numerical Methods in Engineering*, Vol. 73, No. 11, 2008, pp. 1571–1606.

[35]Ng, L. W., Huynh, D. P., and Willcox, K., "Multifidelity Uncertainty Propagation for Optimization Under Uncertainty," *12th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference and 14th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, AIAA, 2012, p. 5602.

A. Model Problem Formulation

In this appendix we introduce a motivating example design problem: designing an acoustic horn for maximum efficiency subject to an uncertain operating condition. This design problem is used throughout the paper to explore different design methodologies. Namely, it is used to generate the computational results in Sections II and IV.

We suppose that we are tasked with optimizing the design of an acoustic horn in order to minimize the amount of internal reflection present, and thus maximize the efficiency of the horn. In order to analyze how the design of the horn affects the amount of internal reflection, we utilize a computational model. A brief summary of the model is presented in Figure 7 below and the description that follows. For more details on the theory behind the acoustic horn and the computational model used, we refer the reader to References 32,33 and 34.

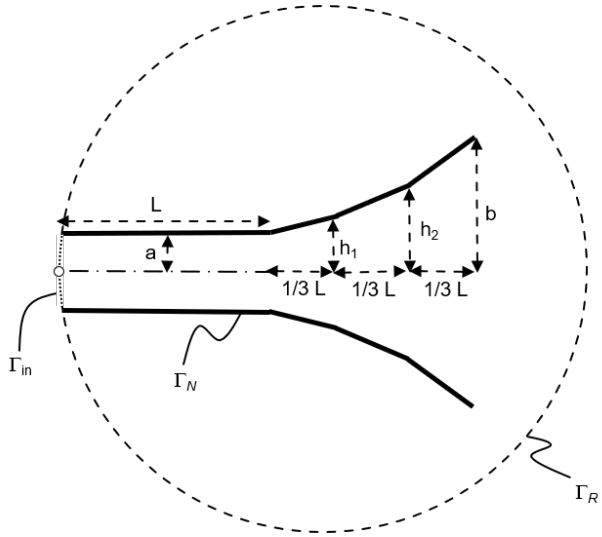

Figure 7: The geometry of the acoustic horn model (Adapted from Ng et al.[35]). The design variables used in this thesis are h_1 and h_2, while the remaining parameters shown are considered fixed.

The exterior domain is truncated by a circular absorbing boundary of radius $R = 25$. The non-dimensional horn geometry is axisymmetric and is parameterized by five variables. We consider three of these to be fixed parameters: the horn inlet length $L = 5$, and half-width $a = 0.5$, and the outlet half-width $b = 3$. The remaining two variables are design variables h_1 and h_2, corresponding to the half-widths at two uniformly spaced points in the horn flare (see Figure 7). Both design variables h_1 and h_2 are constrained to lie within the interval $[\,a,\, b\,] = [\,0.5,\, 3\,]$. The governing equation is the non-dimensional Helmholtz equation,

$$\nabla^2 v + k^2 v = 0, \tag{32}$$

where v is the non-dimensionalized pressure, and k is the wave-number, which we treat as the uncertain operating condition of the horn. The boundary conditions on the horn inlet Γ_{in}, horn surface Γ_N, and far field boundary Γ_R are given by

$$\Gamma_{in}: \qquad \mathrm{i}kv + \tfrac{\partial v}{\partial n} = 2\mathrm{i}k, \tag{33}$$

$$\Gamma_N: \qquad \tfrac{\partial v}{\partial n} = 0, \tag{34}$$

$$\Gamma_R: \quad \tfrac{\partial v}{\partial n} - \left(\mathrm{i}k - \tfrac{1}{2R} + \tfrac{1}{8R(1-\mathrm{i}kR)}\right) v = 0, \tag{35}$$

where n is a unit vector normal to the corresponding boundary, and $\mathrm{i} = \sqrt{-1}$. The governing equation is solved to compute v using a reduced basis finite element model with $n = 116$ basis vectors. The QoI for the acoustic horn model is the reflection coefficient:

$$s = \left| \int_{\Gamma_{\mathrm{in}}} v \, d\Gamma - 1 \right|, \tag{36}$$

which is a fractional measure describing how much of an incoming wave is internally reflected in the horn, as opposed to being transmitted out into the environment. It is thus considered a measure of the horn efficiency, with a lower reflection coefficient giving more favorable performance. Thus, framing the acoustic horn design problem using the notation introduced in Section II.A, we have:

$$\text{Design variables:} \quad \mathbf{x} = [\, h_1,\ h_2 \,]^\top, \tag{37}$$

$$\text{Design space:} \quad \mathcal{X} = [\, 0.5\ ,\ 3 \,] \times [\, 0.5\ ,\ 3 \,], \tag{38}$$

$$\text{Uncertain parameters:} \quad \mathbf{u} = k, \tag{39}$$

$$\text{Uncertainty space:} \quad \mathcal{U} = [1.3, 1.5], \tag{40}$$

$$\text{Output QoI} \quad Q\,(\mathbf{x}, \mathbf{u}) = s. \tag{41}$$

We perform an exploratory analysis of the acoustic horn design space to show how the horn design affects the reflection coefficient for different operating wave-numbers. For this study, we suppose that the wave number follows a fixed truth probability distribution

$$\mathbf{u} \sim \mathbb{P}_u = \text{Uniform}(1.3, 1.5). \tag{42}$$

We compute three point-wise optimal designs. These are the designs that achieve the lowest QoI for $\mathbf{u} = k = 1.3$, 1.4, and 1.5, respectively. These are denoted $\mathbf{x}_{1.3}$, $\mathbf{x}_{1.4}$, and $\mathbf{x}_{1.5}$ respectively. We also compute the design that achieves minimum mean performance,$\mathbf{x}_\mu$, and the design that achieves minimum variance in performance, $\mathbf{x}_\sigma$, over the uncertainty space. Figure 8 shows the horn flare geometry of these horn designs, and the corresponding QoI over the uncertainty space.

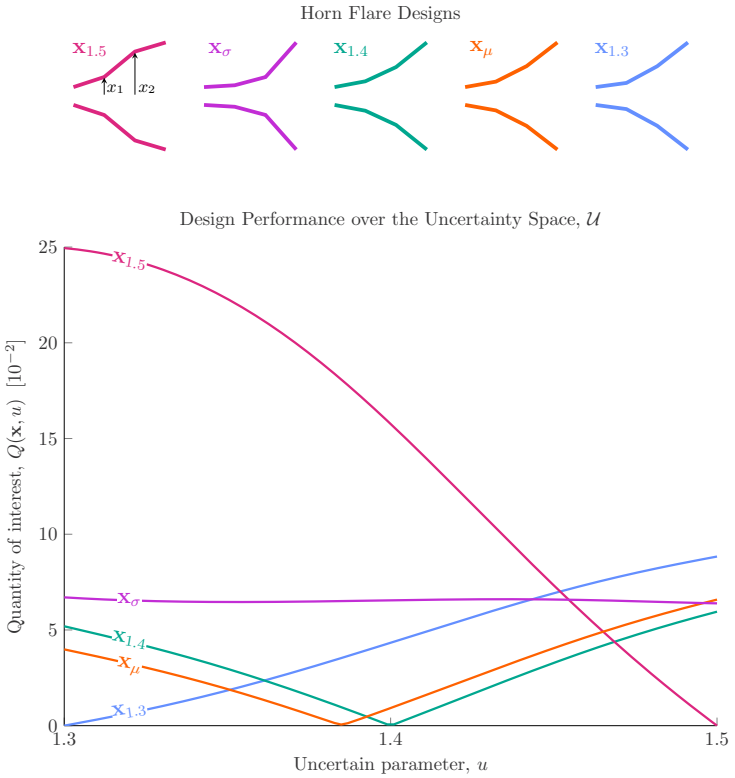

Figure 8: Performance of designs that exhibit optimal mean performance ($\mathbf{x}_\mu$), minimal variation in performance ($\mathbf{x}_\sigma$), or optimal performance at a particular value of the uncertain variable ($\mathbf{x}_{1.3}$, $\mathbf{x}_{1.4}$, $\mathbf{x}_{1.5}$).

We see that designs $\mathbf{x}_{1.3}$, $\mathbf{x}_{1.4}$ and $\mathbf{x}_{1.5}$ all exhibit low reflection at the corresponding design wave numbers, but also exhibit reduced off-design performance. We also see from the performance of $\mathbf{x}_\sigma$ that it is possible to obtain a horn design with consistent performance over the range of wave numbers, however, this comes with the downside that QoI is relatively high everywhere. In fact, the performance of $\mathbf{x}_\mu$ ends up being superior to $\mathbf{x}_\sigma$ at almost all wave numbers.

B. Tabulated Results

This appendix presents tabulated statistics of the data used in Figures 3 and 4 respectively. These tables provide further insight (beyond the mean and 95th percentile information presented in Figures 3 and 4) into the spread in design performances over the $T = 500$ different sample draws we studied.

		L_2					KL					Wasserstein				
	$\bar{r}$	Min.	25th	50th	75th	Max.	Min.	25th	50th	75th	Max.	Min.	25th	50th	75th	Max.
$m = 5$	0.0	2.836	2.887	3.068	3.458	9.754	2.836	2.887	3.068	3.458	9.754	2.836	2.887	3.068	3.458	9.754
	0.2	2.836	2.868	2.970	3.263	7.834	2.835	2.872	2.965	3.221	7.589	2.836	2.900	3.065	3.359	7.400
	0.4	2.835	2.862	2.949	3.197	7.681	2.835	2.869	2.968	3.178	7.449	2.836	2.907	3.084	3.340	7.400
	0.6	2.836	2.870	2.970	3.215	7.608	2.836	2.891	3.020	3.250	7.416	2.836	2.903	3.086	3.335	7.400
	0.8	2.836	2.879	2.979	3.239	7.334	2.836	2.902	3.076	3.325	7.407	2.836	2.903	3.086	3.335	7.400
	1.0	2.835	2.888	3.052	3.320	7.225	2.836	2.909	3.109	3.402	7.405	2.836	2.903	3.086	3.335	7.400
$m = 10$	0.0	2.928	2.955	3.052	3.311	6.261	2.928	2.955	3.052	3.311	6.261	2.928	2.955	3.052	3.311	6.261
	0.2	2.928	2.945	2.993	3.116	4.985	2.928	2.942	2.990	3.104	4.900	2.928	3.014	3.128	3.340	4.040
	0.4	2.928	2.953	3.016	3.171	4.154	2.928	2.957	3.016	3.119	4.055	2.928	3.044	3.211	3.360	4.029
	0.6	2.928	2.977	3.080	3.242	4.037	2.928	2.992	3.080	3.196	4.032	2.928	3.040	3.214	3.364	4.029
	0.8	2.928	3.004	3.137	3.305	4.025	2.928	3.023	3.164	3.295	4.029	2.928	3.040	3.216	3.363	4.029
	1.0	2.928	3.043	3.219	3.380	4.027	2.928	3.061	3.242	3.400	4.029	2.928	3.040	3.216	3.363	4.029
$m = 20$	0.0	2.955	2.972	3.021	3.158	5.367	2.955	2.972	3.021	3.158	5.367	2.955	2.972	3.021	3.158	5.367
	0.2	2.955	2.972	3.006	3.093	3.925	2.955	2.972	3.011	3.082	3.656	2.956	3.083	3.140	3.263	4.041
	0.4	2.955	3.030	3.127	3.237	3.649	2.955	3.028	3.077	3.132	3.428	2.966	3.195	3.335	3.447	3.584
	0.6	2.955	3.131	3.240	3.349	3.597	2.956	3.100	3.155	3.214	3.449	2.966	3.230	3.344	3.440	3.593
	0.8	2.956	3.201	3.321	3.420	3.595	2.957	3.172	3.249	3.325	3.536	2.966	3.230	3.344	3.440	3.593
	1.0	2.958	3.261	3.367	3.467	3.617	2.961	3.269	3.372	3.472	3.619	2.966	3.230	3.344	3.440	3.593

Table 1: Tabulated results for the data shown in Figure 3. Namely, this table shows percentiles of Z_m^t (see Eqn. 1) over $T = 500$ sample draws, with sample sizes $m = 5, 10, 20$, and uniform underlying distribution $\mathbb{P}_{\mathbf{u}}$. The DRO methodology is used with varying radius of ambiguity $\bar{r}$, and ambiguity sets constructed using L_2-norm, KL divergence, or Wasserstein distance. All values have units of 10^{-2}, consistent with the figures throughout.

		L_2					KL					Wasserstein				
	$\bar{r}$	Min.	25th	50th	75th	Max.	Min.	25th	50th	75th	Max.	Min.	25th	50th	75th	Max.
$m = 5$	0.0	2.694	2.748	2.868	3.134	5.771	2.694	2.748	2.868	3.134	5.771	2.694	2.748	2.868	3.134	5.771
	0.2	2.694	2.730	2.811	3.039	4.521	2.694	2.728	2.809	3.019	4.719	2.694	2.743	2.872	3.117	4.407
	0.4	2.693	2.735	2.831	3.019	5.499	2.693	2.728	2.813	3.038	6.045	2.694	2.737	2.891	3.166	7.099
	0.6	2.694	2.744	2.843	3.076	6.409	2.694	2.727	2.844	3.138	6.632	2.694	2.737	2.890	3.172	7.099
	0.8	2.694	2.747	2.852	3.134	6.853	2.694	2.731	2.878	3.191	7.049	2.694	2.738	2.891	3.172	7.099
	1.0	2.694	2.741	2.879	3.196	7.099	2.694	2.740	2.913	3.220	7.099	2.694	2.739	2.891	3.172	7.099
$m = 10$	0.0	2.694	2.716	2.795	2.985	4.638	2.694	2.716	2.795	2.985	4.638	2.694	2.716	2.795	2.985	4.638
	0.2	2.693	2.714	2.766	2.885	3.896	2.693	2.714	2.765	2.878	3.894	2.695	2.777	2.870	3.029	4.096
	0.4	2.694	2.729	2.807	2.948	3.837	2.694	2.719	2.783	2.997	5.743	2.694	2.776	3.005	3.463	5.527
	0.6	2.694	2.753	2.872	3.142	7.036	2.694	2.733	2.836	3.123	6.898	2.694	2.781	3.043	3.446	7.179
	0.8	2.694	2.773	2.954	3.318	7.138	2.694	2.755	2.940	3.331	7.141	2.694	2.781	3.044	3.430	7.179
	1.0	2.695	2.788	3.063	3.515	7.179	2.694	2.788	3.072	3.538	7.179	2.694	2.781	3.044	3.431	7.179
$m = 20$	0.0	2.693	2.703	2.741	2.832	3.820	2.693	2.703	2.741	2.832	3.820	2.693	2.703	2.741	2.832	3.820
	0.2	2.694	2.707	2.735	2.786	3.398	2.694	2.704	2.731	2.789	3.207	2.696	2.804	2.861	2.924	3.711
	0.4	2.694	2.747	2.820	2.962	4.716	2.694	2.718	2.775	2.898	4.708	2.694	2.916	3.301	3.749	7.153
	0.6	2.694	2.816	3.059	3.383	6.355	2.694	2.749	2.896	3.161	7.070	2.694	2.963	3.332	3.901	7.396
	0.8	2.694	2.910	3.304	3.871	7.396	2.693	2.818	3.132	3.593	7.390	2.694	2.974	3.308	3.857	7.396
	1.0	2.695	2.981	3.518	4.245	7.296	2.694	2.956	3.415	4.083	7.396	2.694	2.978	3.335	3.867	7.297

Table 2: Tabulated results for the data shown in Figure 4. Namely, this table shows percentiles of Z_m^t (see Eqn. 1) over $T = 500$ sample draws, with sample sizes $m = 5, 10, 20$, and normal underlying distribution $\mathbb{P}_{\mathbf{u}}$. The DRO methodology is used with varying radius of ambiguity $\bar{r}$, and ambiguity sets constructed using L_2-norm, KL divergence, or Wasserstein distance. All values have units of 10^-2, consistent with the figures throughout.