Robust Optimization using Machine Learning for Uncertainty Sets

Theja Tulabandhula and Cynthia Rudin

CSAIL, MIT, Cambridge, MA 02139, USA

Abstract. Our goal is to build robust optimization problems for making decisions based on complex data from the past. In robust optimization (RO) generally, the goal is to create a policy for decision-making that is robust to our uncertainty about the future. In particular, we want our policy to best handle the the worst possible situation that could arise, out of an *uncertainty set* of possible situations. Classically, the uncertainty set is simply chosen by the user, or it might be estimated in overly simplistic ways with strong assumptions; whereas in this work, we learn the uncertainty set from data collected in the past. The past data are drawn randomly from an (unknown) possibly complicated high-dimensional distribution. We propose a new uncertainty set design and show how tools from statistical learning theory can be employed to provide probabilistic guarantees on the robustness of the policy.
Keywords: machine learning, uncertainty sets, robust optimization, data-driven decision making, decision making under uncertainty.

1 Introduction

In this work, we consider a situation often faced by decision makers: a policy needs to be created for the future that would be a best possible reaction to the worst possible uncertain situation; this is a question of *robust optimization*. In our case, the decision maker does not know what the worst situation might be, and uses complex data to estimate the *uncertainty set*, which is the set of uncertain future situations. Here we are interested in answering questions such as: How might we construct a principled uncertainty set from these complex data? Can we ensure that with high probability our policy will be robust to whatever the future brings?

The uncertainty set $\mathcal{U}$ can be defined in many ways, and the central goal of this work is how to model $\mathcal{U}$ from complex data from the past. The data $\{(\mathbf{x}^i, y^i)\}_{i=1}^n$ take the form of features and labels, with $\mathbf{x}^i \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y^i \in \mathcal{Y}$. Some of the different ways uncertainty sets can be constructed are:

• Using a priori assumptions: We may have *a priori* knowledge about the range of possible future situations. This knowledge can guide us in constructing the uncertainty set $\mathcal{U}$ using, for instance, interval constraints.

• Using empirical statistics: We could create an uncertainty set using empirical statistics of the labels ignoring the feature vectors altogether.

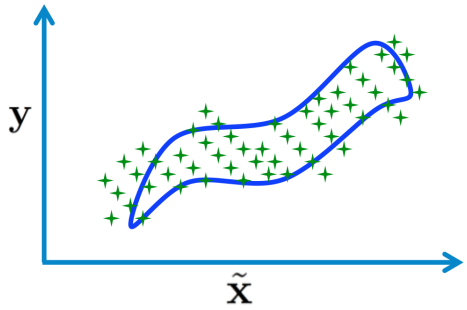
• Using linear regression to model complex data: Here, we use the complex past data $\{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^n$, but we make strong (potentially incorrect) assumptions on the probability distribution these data are drawn from.
• Using machine learning to model complex data, which is the topic of this work: This setting is more general than linear regression and with much weaker assumptions. We provide two principled ways to construct set $\mathcal{U}$ using historical data. In both, we optimize prediction models over the data $\{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^n$, and use them to construct uncertainty set $\mathcal{U}$. $\mathcal{U}$ is used within the robust optimization problem to construct $\boldsymbol{\pi}^*$, and Theorem 1 provides a guarantee on its robustness; this guarantee is derived using statistical learning theory. Theorem 1 describes the guarantee for a generic class of prediction models and Theorem 2 specializes the guarantee for a specific set of prediction models, namely, the conditional quantile models. The only assumption made in this approach is that the data are drawn i.i.d from an unknown source distribution. In particular, there is no normality assumption. Let us give examples of how the two methods we propose for this approach would work when $\mathcal{U}$ is constructed from a regression problem:

– For the first method, for every $\tilde{\mathbf{x}}$ the uncertainty set $\mathcal{U}$ corresponds to the domain of a indicator function on part of the set $\mathcal{Y}$. It is 1 on most of the training examples and is 0 farther away from them. Figure 1(a) shows an illustration of this.
– For the second method, we estimate the 95^{th} and 5^{th} percentiles of $\mathbf{y}$ given $\tilde{\mathbf{x}}$ and set $\mathcal{U}$ to be all values of $\mathbf{y} \in \mathcal{Y}$ between the two estimates. Figure 1(b) illustrates this.

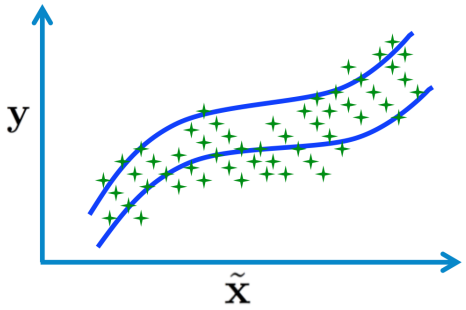
Being able to define uncertainty sets from predictive models is important: the uncertainty sets can now be specialized to a given new situation $\tilde{\mathbf{x}} \in \mathcal{X}$, and this is true even if we have never seen $\tilde{\mathbf{x}}$ before. For instance, when ordering daily supplies $\mathbf{y}^i$ for an ice cream parlor in Boston, an uncertainty set that depends on the weather might be much smaller than one that does not; planning for too much uncertainty in the weather can be too conservative and very costly: it would not be wise to budget for the largest possible summer sales in the middle of the winter.

Our approaches for constructing uncertainty sets are flexible, intuitive, easy to understand from a practitioner's point of view, and at the same time can bring all the rich theoretical results of learning theory to justify the data-driven methodology. Our uncertainty set designs can handle prediction models for classification, regression, ranking and other supervised learning problems. A main theme of this work is that RO is a new context in which many learning theory results naturally apply and can be directly used.

The closest work to ours is possibly that of [1], where the authors provide a linear-regression-based robust decision making paradigm for portfolio allocation problems, where they assume a multivariate linear regression model for the learning step. A big departure from this approach is that in our work, we are able to design uncertainty sets for a general class of decision making problems while making weak assumptions about the distributional aspects of the historical data.



(a) Using optimized set function



(b) Using optimized conditional quantile functions

Fig. 1. The empirical data $\{\mathbf{x}^i, \mathbf{y}^i\}_{i=1}^n$ is shown along with the boundaries created by the proposed methods in each of the above figures. Evaluation of these boundaries at a given $\tilde{\mathbf{x}}$ produces an uncertainty set. In (a), a set function is optimized over the sample and its evaluation at every $\tilde{\mathbf{x}}$ is plotted. In (b), we use optimized conditional quantile models to get the boundaries.

We base our uncertainty set design on regularized empirical risk minimization, which is quite a bit more general than regression.

2 Formulation

Let all the uncertain parameters of the decision problem be denoted by a vector $\mathbf{u} \in \mathbb{R}^m$. Given a realization of $\mathbf{u}$, let the (basic non-robust) decision making problem be written as:

$$\min_{\boldsymbol{\pi}} \rho(\boldsymbol{\pi}, \mathbf{u}) \quad \text{s.t. } F(\boldsymbol{\pi}, \mathbf{u}) \in \mathcal{K}. \tag{1}$$

Here $\boldsymbol{\pi} \in \mathit{\Pi} \subseteq \mathbb{R}^{d_1}$ is the decision vector and $f : \mathit{\Pi} \times \mathbb{R}^m \to \mathbb{R}$ is the objective function. Function $F : \mathit{\Pi} \times \mathcal{U} \to \mathcal{K}$ and convex cone $\mathcal{K} \subseteq \mathbb{R}^{d_2}$ describe the constraints of the problem.

The robust version of the decision problem in Equation (1) is thus:

$$\min_{\boldsymbol{\pi}} \max_{\mathbf{u} \in \mathcal{U}} f(\boldsymbol{\pi}, \mathbf{u}) \quad \text{s.t. } F(\boldsymbol{\pi}, \mathbf{u}) \in \mathcal{K} \text{ for all } \mathbf{u} \in \mathcal{U}, \tag{2}$$

where $\mathcal{U} \subset \mathbb{R}^m$ represents the uncertainty set.

To solve Equation (2), we prescribe the following steps:

Step 1: Construct $\mathcal{U}$ using any of the four methods listed in this section.

Step 2: Obtain a robust solution, using either of the two options below:

Option 1: If $\mathcal{U}$ is a "nice" set, then there are natural ways [2] to transform it into a relaxed set $\mathcal{U}'$ so that the robust optimization problem can be solved to obtain a robust solution π^*. For instance, if U can be bounded using a box or an ellipsoid, that box or ellipsoid can be $\mathcal{U}'$.

Option 2: If $\mathcal{U}$ is not a "nice" set, then sample L elements from $\mathcal{U}$ uniformly. Then solve the sampled version of Equation (2) to obtain a robust solution $\boldsymbol{\pi}^*$.

We focus on **Step 1**. The goal is to ensure that the true realization of parameter $\mathbf{u} \in \mathbb{R}^m$ belongs to set $\mathcal{U}$ with a high likelihood. Let $\mathbf{u}$ be equal to an m-dimensional vector of unknown labels $[\tilde{y}^1 \ \ldots \ \tilde{y}^m]^T$, where each label $\tilde{y}^j \in \mathcal{Y}$ can be predicted given a corresponding feature vector $\tilde{\mathbf{x}}^j \in \mathcal{X}$. Thus m labels $\{\tilde{y}^j\}_{j=1}^m$, which can be forecasted from $\{\tilde{\mathbf{x}}^j\}_{j=1}^m$, feed into the decision problem of Equation (2).

General prediction models:

Let $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ represent a feature vector and $y \in \mathcal{Y} \subseteq \mathbb{R}$ represent a label. Consider a class of set functions $I \in \mathcal{I}$, where $I : \mathcal{X} \to \mathfrak{M}_{\mathbb{R}}$, where $\mathfrak{M}_{\mathbb{R}}$ is the set of all measurable sets of $\mathbb{R}$. Let us say that we have a procedure that picks a function I^{Alg} so that most of the labels of the training examples obey $y^i \in I^{\text{Alg}}(\mathbf{x}^i)$, $i = 1, ..., n$. As long as I^{Alg} belongs to a set of "simple" functions, we have a guarantee on how well I^{Alg} will generalize to new observations. Specifically, consider the following empirical risk minimization procedure:

$$\min_{I \in \mathcal{I}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[y^i \notin I(\mathbf{x}^i)], \tag{3}$$

where $\mathbf{1}[\cdot]$ is the indicator function. Let an optimal solution to the above problem be I^{Alg}. Then, define the uncertainty set $\mathcal{U}$ as:

$$\mathcal{U} = \Pi_{j=1}^{m} I^{\text{Alg}}(\tilde{\mathbf{x}}^j), \tag{4}$$

where $\mathcal{U}$ is a product of m measurable sets.

The above setting is quite general. In particular, since the range of function I^{Alg} is $\mathfrak{M}_{\mathbb{R}}$, we can capture sets that are arbitrarily more complicated than simple intervals. For instance, if $\mathbb{P}_{y^j|\tilde{\mathbf{x}}^j}$ is bimodal, then for certain values of $\tilde{\mathbf{x}}^j$, $I^{\text{Alg}}(\tilde{\mathbf{x}}^j)$ can be the union of two disjoint intervals.

Conditional quantile models:

In this method, we specialize the generic function class $\mathcal{I}$ to the class of set functions defined using conditional quantile models. We will estimate an upper quantile of $\tilde{y}$ for each $\tilde{\mathbf{x}}$, and a lower quantile of $\tilde{y}$ for each $\tilde{\mathbf{x}}$. The uncertainty set will be the interval between the two quantile estimates. This method is applicable when our prediction task is a regression problem.

When $y \sim \mathbb{P}_y$, the τ^{th} quantile of y, denoted by μ^τ, is defined as $\mu^\tau := \inf\{\mu : \mathbb{P}_y(y \leq \mu) = \tau\}$. Here τ can vary between 0 and 1. In the special case when τ is set to 0.5, this defines the median. Similarly, when $(\mathbf{x}, y) \sim \mathbb{P}_{\mathbf{x},y}$, the conditional quantile μ^τ can be defined as a function from $\mathcal{X}$ to $\mathcal{Y}$, $\mu^\tau(\mathbf{x}) := \inf\{\mu : \mathbb{P}_{y|\mathbf{x}}(y \leq \mu) = \tau\}$.

In our setting, $\tilde{y}^j$ conditioned on $\tilde{\mathbf{x}}^j$ is distributed according to $\mathbb{P}_{\tilde{y}^j|\tilde{\mathbf{x}}^j}$. Thus, given a value of $\tau \in [0, 1]$, $\mathbb{P}_{\tilde{y}^j|\mathbf{x}=\tilde{\mathbf{x}}^j}(\tilde{y}^j \leq \mu^\tau(\tilde{\mathbf{x}}^j)) = \tau$ where $\mu^\tau(\mathbf{x})$ is the conditional quantile defined earlier. Our method picks two values of τ, $\delta_p \leq \delta_q$ such that:

$$\mathbb{P}_{\tilde{y}^j|\tilde{\mathbf{x}}^j}(\tilde{y}^j \leq \mu^{\delta_p}(\tilde{\mathbf{x}}^j)) = \delta_p, \text{ and } \mathbb{P}_{\tilde{y}^j|\tilde{\mathbf{x}}^j}(\tilde{y}^j \leq \mu^{\delta_q}(\tilde{\mathbf{x}}^j)) = \delta_q.$$

For example, a typical value for the pair (δ_p, δ_q) can be $(0.05, 0.95)$ which makes $\mu^{\delta_p}(\tilde{\mathbf{x}}^j)$ correspond to the 5% conditional quantile and $\mu^{\delta_q}(\tilde{\mathbf{x}}^j)$ correspond to the 95% conditional quantile. Given these two conditional quantiles, we have:

$$\mathbb{P}_{\tilde{y}^j|\tilde{\mathbf{x}}^j}(\mu^{\delta_p}(\tilde{\mathbf{x}}^j) < \tilde{y}^j \leq \mu^{\delta_q}(\tilde{\mathbf{x}}^j)) = \delta_q - \delta_p.$$

Thus, the unknown future realization of $\tilde{y}^j$ belongs to the interval $[\mu^{\delta_p}(\tilde{\mathbf{x}}^j), \mu^{\delta_q}(\tilde{\mathbf{x}}^j)]$ with high probability if δ_p and δ_q are chosen appropriately.

Quantile regression can be seen as an empirical risk minimization algorithm where the loss function is defined appropriately to obtain a conditional quantile function. That is, we aim to obtain an estimator function $\beta(\mathbf{x})$ of the true conditional quantile function $\mu^\tau(\mathbf{x})$ given a predefined quantile parameter τ. In particular, the *pinball* loss (or newsvendor loss) function defined below is used.

$$l^\tau(\beta(\mathbf{x}), y) = \begin{cases} \tau \cdot (y - \beta(\mathbf{x})) & \text{if } y - \beta(x) \geq 0, \\ (\tau - 1) \cdot (y - \beta(\mathbf{x})) & \text{otherwise.} \end{cases}$$

Let $l^\tau_{\mathbb{P}}(\beta) = \mathbb{E}_{\mathbf{x},y}[l^\tau(\beta(\mathbf{x}), y)]$. In our setting, we will let $\mathcal{B}_0$ be our hypothesis class that we want to pick conditional quantile functions from.

Let the empirical risk minimization procedure using the pinball loss output a conditional quantile model $\beta^{Alg,\tau}$ when given the historical sample $S = \{(\mathbf{x}^i, y^i)\}_{i=1}^n$ of size n and a parameter τ. That is, let $l_S^\tau(\beta) = \frac{1}{n}\sum_{i=1}^n l^\tau(\beta(\mathbf{x}^i), y^i)$ and $\beta^{Alg,\tau} \in \arg\min_{\beta\in\mathcal{B}_0} l_S^\tau(\beta)$. The following definition of $\mathcal{U}$ uses two empirical conditional quantile functions with $\tau = \delta_p$ and $\tau = \delta_q$ respectively:

$$\mathcal{U} = \Pi_{j=1}^m \left[\min\left(\beta^{\mathrm{Alg},\delta_p}(\tilde{\mathbf{x}}^j), \beta^{\mathrm{Alg},\delta_q}(\tilde{\mathbf{x}}^j)\right), \max\left(\beta^{\mathrm{Alg},\delta_p}(\tilde{\mathbf{x}}^j), \beta^{\mathrm{Alg},\delta_q}(\tilde{\mathbf{x}}^j)\right)\right]. \quad (5)$$

Here $\mathcal{U}$ is again a product of m intervals, each one constructed so that it contains the unknown $\tilde{y}^j$ with high probability.

3 Robustness guarantee using general prediction functions

Consider the setting described in Section 2, where we have a class of general set functions $\mathcal{I}$. Let $S := \{(\mathbf{x}^i, y^i)\}_{i=1}^n$ be the training data which are independent and identically distributed. Let algorithm A represent a generic learning procedure. That is, it takes S as an input and outputs I^{Alg}. Since I^{Alg} is a function of sample S, we will show that the unknown $\tilde{y}^j$ belong to the interval $I^{\mathrm{Alg}}(\tilde{\mathbf{x}}^j)$ with high probability over S as long as the set of functions $\mathcal{I}$ from which I^{Alg} is picked is "simple". Note that we do not assume anything about the source distribution.

In order to state our result, we will define the following quantity known as the empirical Rademacher average [3]. For a set $\mathcal{F}$ of functions, the *empirical Rademacher average* is defined with respect to a given random sample $S' = \{z^i\}_{i=1}^n$ as

$$\mathcal{R}_{S'}(\mathcal{F}) = \mathbb{E}_{\sigma^1,\ldots,\sigma^n}\left[\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \sigma^i f(z^i)\right|\right],$$

where for each $i = 1, .., n$, $\sigma^i = \pm 1$ with equal probability. The *Rademacher average* is defined to be the expectation of the empirical Rademacher average over the random sample S: $\mathcal{R}(\mathcal{H}) = \mathbb{E}_{z^1,\ldots,z^n}[\mathcal{R}_S(\mathcal{H})]$. The interpretation of the Rademacher average is that it measures the ability of function class $\mathcal{F}$ to fit noise, coming from the random $\sigma_i's$. If the function class can fit noise well, it is a highly complex class. The Rademacher average is one of many ways to measure the richness of a function class, including covering numbers, fat-shattering dimensions [4] and the Vapnik-Chervonenkis dimension [5].

Theorem 1. *If $\mathcal{U}$ is defined as in Equation (4), then with probability at least $1-\delta$ over training sample S, we have robustness guarantee*

$$\mathbb{P}_{\{\tilde{\mathbf{x}}^j,\tilde{y}^j\}_{j=1}^m}\left(F(\pi^*, [\tilde{y}^1 \ldots \tilde{y}^m]^T) \in \mathcal{K}\right) \geq \left(\left[1 - \frac{1}{n}\sum_{i=1}^n \mathbf{1}[y^i \notin I^{Alg}(\mathbf{x}^i)] - 2\mathcal{R}(l\circ\mathcal{I}) - \sqrt{\frac{\log\frac{1}{\delta}}{2n}}\right]_+\right)^m,$$

where $\epsilon > 0$ *is a pre-determined constant, and* $\left[\cdot\right]_+$ *is shorthand for* $\max(0, \cdot)$.

The result is a lower bound on the probability of infeasibility. This bound depends on the performance of the data dependent set function I^{Alg}. If I^{Alg} is such that its performance, measured in terms of $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[y^i \notin I^{\text{Alg}}(\mathbf{x}^i)]$ is good (i.e., lower in value), then the right hand side of the inequality increases, resulting in a higher chance of feasibility. This probability of feasibility also depends on the number of estimates m that enter the decision problem of Equation (2). When $n \to \infty$, the Rademacher term and the square root terms become zero and the probability of feasibility depends on the asymptotic performance of I^{Alg} (which converges to I^*, the 'best-in-class' set function), as desired.

4 Robustness guarantee using conditional quantile functions

Theorem 2. *If* $\mathcal{U}$ *is defined as in Equation (5), then with probability at least* $1-\delta$ *over training sample* S*, we have*

$$\mathbb{P}_{\{\tilde{\mathbf{x}}^j, \tilde{y}^j\}_{j=1}^m}\Big(F(\pi^*, [\tilde{y}^1 ... \tilde{y}^m]^T) \in \mathcal{K}\Big) \geq$$

$$\left(\left[\frac{1}{n}\sum_{i=1}^{n}\Big(r_\epsilon^-(y^i - \beta^{Alg,\delta_q}(\mathbf{x}^i)) - r_\epsilon^+(y^i - \beta^{Alg,\delta_p}(\mathbf{x}^i))\Big) - \frac{8}{\epsilon}\mathcal{R}(\mathcal{B}_0) - 2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\right]_+\right)^m, \tag{6}$$

where $\epsilon > 0$ *is a pre-determined constant,* $\left[\cdot\right]_+$ *is shorthand for* $\max(0, \cdot)$*,* $r_\epsilon^-(z) := \min\Big(1, \max\big(0, -\frac{z}{\epsilon}\big)\Big)$ *and* $r_\epsilon^+(z) := \min\Big(1, \max\big(0, 1-\frac{z}{\epsilon}\big)\Big)$.

The lower bound is a function of the empirical performance of the two conditional quantile estimators and the Rademacher average of the hypothesis set. As $n \to \infty$, the Rademacher average and the square-root term tend to zero at a rate $O(\frac{1}{\sqrt{n}})$. The term $\frac{1}{n}\sum_{i=1}^{n}\Big(r_\epsilon^-(y^i - \beta^{\text{Alg},\delta_q}(\mathbf{x}^i)) - r_\epsilon^+(y^i - \beta^{\text{Alg},\delta_p}(\mathbf{x}^i))\Big)$ converges to $\mathbb{P}_{\mathbf{x},y}(\beta^{\text{Alg},\delta_p}(\mathbf{x}) \leq y \leq \beta^{\text{Alg},\delta_q}(\mathbf{x}))$.

5 Conclusion

In this paper, we considered a class of single-stage decision making problems where the uncertainty is derived from statistical modeling. We present two principled approaches to design uncertainty sets in the robust optimization framework for these problems using statistical learning theory. In the first approach, we use a general class of set functions and define the uncertainty set using them. The second approach develops this idea further using the notion of quantiles to define the uncertainty set. For both approaches, we give probabilistic guarantees on the feasibility of the robust solutions thus obtained.

Bibliography

[1] Donald Goldfarb and Garud Iyengar. Robust portfolio selection problems. *Mathematics of Operations Research*, 28(1):1–38, 2003.

[2] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton University Press, 2009.

[3] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Localized rademacher complexities. In *Computational Learning Theory*, pages 44–58. Springer, 2002.

[4] Peter L Bartlett, Philip M Long, and Robert C Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996.

[5] Vladimir N Vapnik. *Statistical learning theory*. Wiley, 1998.