Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Robust Regression over Averaged Uncertainty

Dimitris Bertsimas
Sloan School of Management and Operations Research Center, MIT, MA, 02139 dbertsim@mit.edu

Yu Ma
Sloan School of Management and Operations Research Center, MIT, MA, 02139, midsumer@mit.edu

We propose a new formulation of robust regression by integrating all realizations of the uncertainty set and taking an averaged approach to obtain the optimal solution for the ordinary least-squared regression problem. We show that this formulation surprisingly recovers ridge regression and establishes the missing link between robust optimization and the mean squared error approaches for existing regression problems. We first prove the equivalence for four uncertainty sets: ellipsoidal, box, diamond, and budget, and provide closed-form formulations of the penalty term as a function of the sample size, feature size, as well as perturbation protection strength. We then show in synthetic datasets with different levels of perturbations, a consistent improvement of the averaged formulation over the existing worst-case formulation in out-of-sample performance. Importantly, as the perturbation level increases, the improvement increases, confirming our method's advantage in high-noise environments. We report similar improvements in the out-of-sample datasets in real-world regression problems obtained from UCI datasets.

Key words: robust optimization, uncertainty, linear regression, machine learning, ridge regression

1. Introduction

Protecting against data uncertainty is at the center of modern machine learning modeling in both the predictive and generative paradigms Bertsimas and Sim (2004), Bertsimas et al. (2019), Hastie et al. (2009), Hariri et al. (2019). Uncertainties in both the input and outcome data could be attributed to implementation, recording, and manual errors. Examples such as incorrect vital readings during hospital patient stay, as well as manual mistakes on temperature recordings for climate change, are ubiquitous and inherent problems in most real-world applications that can impact the solution quality of the original problem if solved directly. Furthermore, issues such as

over-fitting may lead to worse performances in out-of-sample validations Bühlmann and van de Geer (2011), Goodfellow et al. (2016).

The most prominent approach to address this problem is the use of regularization by incorporating an additional penalty term that either penalizes or encourages certain structures of the solution Wang et al. (2006), Kratsios and Hyndman (2020). Two of the most common regularization techniques for linear regression are ridge regression Hoerl and Kennard (1970) and the least absolute shrinkage and selection operator (LASSO) regression Tibshirani (1996). Specifically, ridge regression is useful to mitigate the problem of multicollinearity, or highly correlated independent variables, in problems with a large number of parameters. It also provides a smaller variance and mean square estimator Kennedy (2003). On the other hand, LASSO encourages sparse solutions, (i.e., only a small subset of features coefficients are nonzero) Natarajan (1995), Tibshirani et al. (2004) and is also computationally efficient with abundant literature on efficient algorithms for its implementation Bento et al. (2018).

Another classical approach to account for adversarial noise in the input data is by following the robust optimization paradigm and formulating the original least square problem as a robust optimization problem Bertsimas et al. (2017, 2011), Ghaoui and Lebret (1997), Lewis (2002), Lewis and Pang (2010), Xu et al. (2008), Ben-Tal et al. (2009). The robust regression solution protects against the worst-case noise perturbation in the input data by solving a min-max problem using efficient algorithms such as the cutting-plane approach. Robust regression offers several advantages. By concretely defining the adversarial perturbations the regression model is protecting against, this framework provides additional insights into the behaviors of solutions and beliefs of the original data. This worst-case approach also leads to a more straightforward analysis of the estimators Xu et al. (2008) as well as algorithms for finding the estimators Ben-Tal et al. (2015).

Not surprisingly, there exists a wealth of work that demonstrated a deeper connection between the robust optimization framework and the traditional regularization techniques, where a main result from Bertsimas and Copenhaver (2018) characterizes the equivalence of the robust regression with regularization uncertainty certain norm-induced uncertainty sets. A key observation of this equivalence is that, although in practice we aim to solve ordinary least-squares regression, which offers advantages such as computational simplicity, we only proved robust regression's equivalence to a regularized root-mean-squared regression. This curiosity begs the natural question if there exists a missing link between the traditional robust regression with the regularization techniques. An additional disadvantage of the existing robust optimization formulation is that the solutions it recovers protect against the worst-case uncertainty of the defined uncertainty set.

An intuitive remedy is that instead of protecting against worst-case perturbations, we recover the solution by protecting against an averaged perturbation over all realization on the uncertainty set. This approach avoids over-protecting extreme perturbations and takes a more holistic perspective. Surprisingly, we show this approach establishes a missing connection between traditional robust regression and the practically solved least square regression. By reformulating the known min-max approach to a min-average approach, we find that among all the uncertainty sets we considered, this averaged approach recovers ridge regression in the ordinary least-squared formulation. Our contributions are as follows:

(a) We provide a principled, natural, and theoretical justification for why we should solve the least square problem under a robust optimization lens in addition to its known computational advantages.

(b) We justify the squared formulation as an appropriate model to solve by providing evidence of some of its empirical advantages using both synthetic and real-world datasets.

2. Robust Regression over Averaged Uncertainty

In this section, we outline the robust regression over averaged uncertainty formulation as well as the main theorem demonstrating its equivalence with the least square regression formulation.

2.1. Norms and Uncertainty Sets

To capture our belief of the structure of the noise we aim to protect against, we construct uncertainty sets that obey certain boundedness conditions. We first introduce the necessary background of matrix norms, which are used to define uncertainty sets. Given $\mathbf{\Delta} \in \mathbb{R}^{n\times k}$, the p-Frobenius norm is the entrywise ℓ_p norm on the entries of $\mathbf{\Delta}$ defined as:

$$\|\mathbf{\Delta}\|_{F_p} = \left(\sum_{i=1}^{n}\sum_{j=1}^{k}|\Delta_{ij}|^p\right)^{1/p},$$

where when $p \to \infty$, we have,

$$\|\mathbf{\Delta}\|_{\infty} = \max_{1\le i\le n, 1\le j\le k}|\Delta_{ij}|.$$

Given the p-Frobenius norm, we consider the following four uncertainty sets defined as follows:

- Ellipsoidal Uncertainty refers to $\mathcal{U}_1 = \{\mathbf{\Delta} : \|\mathbf{\Delta}\|_{F_2} \le \rho\}$.
- Box Uncertainty refers to $\mathcal{U}_2 = \{\mathbf{\Delta} : \|\mathbf{\Delta}\|_{\infty} \le \rho\}$.
- Diamond Uncertainty refers to $\mathcal{U}_3 = \{\mathbf{\Delta} : \|\mathbf{\Delta}\|_{F_1} \le \rho\}$.
- Budget Uncertainty refers to $\mathcal{U}_4 = \{\mathbf{\Delta} : \|\mathbf{\Delta}\|_{F_1} \le \rho, \|\mathbf{\Delta}\|_{\infty} \le \Gamma\}$.

We also remark that by protecting perturbations against every individual entry of $\mathbf{\Delta}$, we are conducting global noise protection. Other types of noise protection, such as feature-wise protection, can be done by using induced norms.

2.2. Known Root-Mean-Square Equivalence

We first outline the known result of the equivalence between robust regression under the Forbenius norm and root-mean-square, ridge regularization regression. Specifically, under certain norm conditions, this equivalence no longer holds. We show in the later section that this equivalence is maintained even under these circumstances in the new averaged uncertainty formulation.

THEOREM 1. *Bertsimas and Copenhaver (2018), Bertsimas et al. (2011) The following relations hold:*

(a) *For any* $p, q \in [1, \infty]$*:*

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{F_p}} \|\boldsymbol{y} - (\boldsymbol{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_p = \min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_p + \lambda \|\boldsymbol{\beta}\|_{p*}.$$

(b) *For* $p \neq q$ *and* $p \in (1, \infty)$*, and with* $\delta_m(a, b) := \max\{\|\boldsymbol{u}\|_a : \boldsymbol{u} \in \mathbb{R}^m, \|\boldsymbol{u}\|_b = 1\}$*,*

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_p + \frac{\lambda}{\delta_m(q,p)} \|\boldsymbol{\beta}\|_{q*} \leq \min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{F_q}} \|\boldsymbol{y} - (\boldsymbol{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_p,$$

and the inequality can be strict, i.e., in this case, robust regression and ridge regularized regression are not equivalent.

2.3. Characterization of Averaged Uncertainty

We provide the theorem that characterizes the equivalence between robust regression over averaged uncertainty and ridge regression over four norm-induced, commonly considered uncertainty sets below. For simplicity, we use the notation that $\int_{\boldsymbol{\Delta} \in \mathcal{U}} \mathrm{d}\boldsymbol{\Delta}$ refers to $\int_{\{\boldsymbol{\Delta}: \boldsymbol{\Delta} \in \mathcal{U}\}} \mathrm{d}\boldsymbol{\Delta}$. Similarly, $\int_{\|\boldsymbol{\Delta}\|_q \leq \rho} \mathrm{d}\boldsymbol{\Delta}$ refers to $\int_{\{\boldsymbol{\Delta}: \|\boldsymbol{\Delta}\|_{F_q} \leq \rho\}} \mathrm{d}\boldsymbol{\Delta}$.

THEOREM 2. *Given a data matrix* $\boldsymbol{X} \in \mathbb{R}^{n \times k}$*, where* n *is the number of samples and* k *is the number of features and an outcome data vector* $\boldsymbol{y} \in \mathbb{R}^n$*, the problem*

$$\min_{\boldsymbol{\beta}} \left(\int_{\boldsymbol{\Delta} \in \mathcal{U}} \|\boldsymbol{y} - (\boldsymbol{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2^2 \, \mathrm{d}\boldsymbol{\Delta} \right)$$

is equivalent to

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2.$$

(a) *For the ellipsoidal uncertainty set where* $\mathcal{U} = \mathcal{U}_1 : \lambda = \dfrac{\rho^2}{k}$*.*

(b) *For the box uncertainty set where* $\mathcal{U} = \mathcal{U}_2 : \lambda = \dfrac{n\rho^2}{3}$*.*

(c) *For the diamond uncertainty set where* $\mathcal{U} = \mathcal{U}_3 : \lambda = \dfrac{2n\rho^2}{(nk+2)(nk+1)}$*.*

(d) *For the budget uncertainty set where* $\mathcal{U} = \mathcal{U}_4$*:*

$$\lambda = \frac{2n\rho^2}{(nk+1)(nk+2)} - \frac{n(\rho - \Gamma)^{nk}((n^2k^2 + 3nk - 2)\Gamma^2 + (4 - 2nk)\rho\Gamma)}{(nk+1)(nk+2)((\rho^{nk} - (\rho - \Gamma)^{nk})}.$$

We show the proof for each uncertainty set in their respective section. This result establishes a connection between averaged-uncertainty robust regression and least square ridge regression. Across all uncertainty sets we consider, the final characterizations all arrive at ridge regression but with different leading constant penalty terms. This implies that ridge regression is a general regularization method that protects against global perturbations of different noise structures, and by understanding the selected penalty strengths, we may gain insights into the distributions of the original perturbations.

3. Ellipsoidal Uncertainty Set

LEMMA 1. *Folland (2001) If some α_j is odd, then* $\int_{x_1^2+\cdots+x_n^2\leq 1} x_1^{\alpha_1}\cdots x_n^{\alpha_n}\,\mathrm{d}x_1\cdots\mathrm{d}x_n = 0$.

LEMMA 2. *If $\mathbf{\Delta}\in\mathbb{R}^{n\times k}$ and $\mathbf{\Delta}=\left[\boldsymbol{a_1}\ \boldsymbol{a_2}\ \cdots \boldsymbol{a_k}\right]$, where $\boldsymbol{a_i}$ are column vectors and $V(nk,\rho)$ is the volume of $\mathcal{U}_1$ in dimension nk with radius ρ.*

$$\int_{\|\mathbf{\Delta}\|_2\leq\rho}\mathbf{\Delta}^T\mathbf{\Delta}\,\mathrm{d}\mathbf{\Delta}=\begin{bmatrix}\frac{V(nk,\rho)}{k} & 0 & \cdots & 0\\ 0 & \frac{V(nk,\rho)}{k} & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & \cdots & \frac{V(nk,\rho)}{k}\end{bmatrix}.$$

Proof.

$$\int_{\|\mathbf{\Delta}\|_2\leq\rho}\mathbf{\Delta}^T\mathbf{\Delta}\,\mathrm{d}\mathbf{\Delta}=\int_{\|\mathbf{\Delta}\|_2\leq\rho}\begin{bmatrix}\boldsymbol{a_1}^T\boldsymbol{a_1} & \boldsymbol{a_1}^T\boldsymbol{a_2} & \cdots & \boldsymbol{a_1}^T\boldsymbol{a_k}\\ \boldsymbol{a_2}^T\boldsymbol{a_1} & \boldsymbol{a_2}^T\boldsymbol{a_2} & \cdots & \boldsymbol{a_2}^T\boldsymbol{a_k}\\ \cdots & \cdots & \cdots & \cdots\\ \boldsymbol{a_k}^T\boldsymbol{a_1} & \boldsymbol{a_k}^T\boldsymbol{a_2} & \cdots & \boldsymbol{a_k}^T\boldsymbol{a_k}\end{bmatrix}\mathrm{d}\mathbf{\Delta}.$$

All entries of this matrix except those on the diagonal are polynomials of elements of $\boldsymbol{\Delta}$ with exponent 1. Thus, using Lemma 1, this expression can be simplified to:

$$\int_{\|\boldsymbol{\Delta}\|_2\le\rho} \boldsymbol{\Delta}^T\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{\Delta} = \int_{\|\boldsymbol{\Delta}\|_2\le\rho} \begin{bmatrix} \boldsymbol{a_1}^T\boldsymbol{a_1} & 0 & \cdots & 0 \\ 0 & \boldsymbol{a_2}^T\boldsymbol{a_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \boldsymbol{a_k}^T\boldsymbol{a_k} \end{bmatrix} \mathrm{d}\boldsymbol{\Delta}.$$

By symmetry, we also have that $\int_{\|\boldsymbol{\Delta}\|_2\le\rho} \boldsymbol{a_1}^T\boldsymbol{a_1}\,\mathrm{d}\boldsymbol{\Delta} = \int_{\|\boldsymbol{\Delta}\|_2\le\rho} \boldsymbol{a_2}^T\boldsymbol{a_2}\,\mathrm{d}\boldsymbol{\Delta} = \cdots = \int_{\|\boldsymbol{\Delta}\|_2\le\rho} \boldsymbol{a_k}^T\boldsymbol{a_k}\,\mathrm{d}\boldsymbol{\Delta} = \frac{V(nk,\rho)\rho^2}{k}$, and thus

$$\int_{\|\boldsymbol{\Delta}\|_2\le\rho} \boldsymbol{\Delta}^T\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{\Delta} = \begin{bmatrix} \frac{V(nk,\rho)\rho^2}{k} & 0 & \cdots & 0 \\ 0 & \frac{V(nk,\rho)\rho^2}{k} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{V(nk,\rho)\rho^2}{k} \end{bmatrix}. \qquad \square$$

Now, putting everything together, for the ellipsoidal uncertainty set, we have that, let $\boldsymbol{\Delta}\in\mathbb{R}^{n\times k}$ $= \begin{bmatrix} \boldsymbol{a_1} & \boldsymbol{a_2} & \cdots & \boldsymbol{a_n} \end{bmatrix}^T$, where $\boldsymbol{a_i}\in\mathbb{R}^{1\times k}$ and $\boldsymbol{\beta}\in\mathbb{R}^{k\times 1}$. We have

$$\begin{aligned}
&\min_{\boldsymbol{\beta}} \int_{\|\boldsymbol{\Delta}\|_2\le\rho} \|\boldsymbol{y}-(\boldsymbol{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_2^2\,\mathrm{d}\boldsymbol{\Delta} \\
&= \min_{\boldsymbol{\beta}} \int_{\|\boldsymbol{\Delta}\|_2\le\rho} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\Delta}\boldsymbol{\beta}\|_2^2\,\mathrm{d}\boldsymbol{\Delta} \\
&= \min_{\boldsymbol{\beta}} \left(\int_{\|\boldsymbol{\Delta}\|_2\le\rho} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2^2 - 2(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Delta}\boldsymbol{\beta}\,\mathrm{d}\boldsymbol{\Delta} \right) \\
&= \min_{\boldsymbol{\beta}} \left(\int_{\|\boldsymbol{\Delta}\|_2\le\rho} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2\,\mathrm{d}\boldsymbol{\Delta} + \int_{\|\boldsymbol{\Delta}\|_2\le\rho} \|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2^2\,\mathrm{d}\boldsymbol{\Delta} - 2(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T \left(\int_{\|\boldsymbol{\Delta}\|_2\le\rho} \boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{\Delta} \right)\boldsymbol{\beta} \right).
\end{aligned}$$

By Lemma 2, for the second term, we have

$$\begin{aligned}
&= \int_{\|\boldsymbol{\Delta}\|_2\le\rho} \|\boldsymbol{\Delta}^T\boldsymbol{\beta}\|_2^2\,\mathrm{d}\boldsymbol{\Delta} \\
&= \int_{\|\boldsymbol{\Delta}\|_2\le\rho} \boldsymbol{\beta}^T\boldsymbol{\Delta}^T\boldsymbol{\Delta}\boldsymbol{\beta}\,\mathrm{d}\boldsymbol{\Delta} \\
&= \boldsymbol{\beta}^T \left(\int_{\|\boldsymbol{\Delta}\|_2\le\rho} \boldsymbol{\Delta}^T\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{\Delta} \right)\boldsymbol{\beta}
\end{aligned}$$

$$= \frac{V(nk,\rho)\rho^2}{k}\|\boldsymbol{\beta}\|_2^2.$$

For the third term,

$$\int_{\|\boldsymbol{\Delta}\|_2\le\rho} (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{\Delta}\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\Delta} = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\left(\int_{\|\boldsymbol{\Delta}\|_2\le\rho}\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{\Delta}\right)\boldsymbol{\beta}.$$

We are integrating over a polynomial of elements of $\boldsymbol{\Delta}$ where all exponents are 1, and thus this term is 0. We obtain

$$\begin{aligned}
&\min_{\boldsymbol{\beta}} \int_{\|\boldsymbol{\Delta}\|_2\le\rho} \|\boldsymbol{y}-(\boldsymbol{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_2^2\,\mathrm{d}\boldsymbol{\Delta}\\
&= \min_{\boldsymbol{\beta}} \left(\int_{\|\boldsymbol{\Delta}\|_2\le\rho} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2\,\mathrm{d}\boldsymbol{\Delta} + \int_{\|\boldsymbol{\Delta}\|_2\le\rho} \|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2^2\,\mathrm{d}\boldsymbol{\Delta} - 2(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\left(\int_{\|\boldsymbol{\Delta}\|_2\le\rho}\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{\Delta}\right)\boldsymbol{\beta}\right)\\
&= \min_{\boldsymbol{\beta}} \left(V(nk,\rho)\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{V(nk,\rho)\rho^2}{k}\|\boldsymbol{\beta}\|_2^2 - 0\right)\\
&= V(nk,\rho)\min_{\boldsymbol{\beta}} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{\rho^2}{k}\|\boldsymbol{\beta}\|_2^2,
\end{aligned}$$

and Theorem 2(a) follows.

4. Box Uncertainty Set

LEMMA 3. *Given $\boldsymbol{x}\in\mathbb{R}^n$, and x_i being a component of the vector $\boldsymbol{x}$, we have $\int_{\|\boldsymbol{x}\|_\infty\le\rho} x_i^2\,\mathrm{d}\boldsymbol{\Delta} = \frac{(2\rho)^n\rho}{3}$.*

Proof. Without loss of generality, we consider $x_i = x_n$.

$$\begin{aligned}
&\int_{\|\boldsymbol{x}\|_\infty\le\rho} x_n^2\,\mathrm{d}\boldsymbol{\Delta}\\
&= \underbrace{\int_{-\rho}^{\rho}\cdots\int_{-\rho}^{\rho}}_{n-1}\int_{-\rho}^{\rho} x_n^2\,\mathrm{d}x_n\,\underbrace{\mathrm{d}x_1\cdots\mathrm{d}x_{n-1}}_{n-1}\\
&= \underbrace{\int_{-\rho}^{\rho}\cdots\int_{-\rho}^{\rho}}_{n-1}\frac{2\rho^3}{3}\,\underbrace{\mathrm{d}x_1\cdots\mathrm{d}x_{n-1}}_{n-1}\\
&= (2\rho)^{n-1}\frac{2\rho^3}{3}\\
&= \frac{(2\rho)^n\rho^2}{3}.
\end{aligned}$$

□

LEMMA 4. *Let* $\mathcal{U}_2 = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_\infty \leq \rho\}$. *Given* $\boldsymbol{x} \in \mathbb{R}^n$, *and* x_i, x_j *being different components of the vector* $\boldsymbol{x}$, *we have* $\int_{\|\boldsymbol{x}\|_\infty \leq \rho} x_i x_j \, \mathrm{d}\boldsymbol{\Delta} = 0$.

Proof.

$$\begin{aligned}
&\int_{\|\boldsymbol{x}\|_\infty \leq \rho} x_i x_j \, \mathrm{d}\boldsymbol{\Delta} \\
&= \underbrace{\int_{-\rho}^{\rho} \cdots \int_{-\rho}^{\rho}}_{n-2} \int_{-\rho}^{\rho} \int_{-\rho}^{\rho} x_i x_j \, \mathrm{d}x_i \, \mathrm{d}x_j \underbrace{\mathrm{d}x_1 \cdots \mathrm{d}x_n}_{n-2} \\
&= \underbrace{\int_{-\rho}^{\rho} \cdots \int_{-\rho}^{\rho}}_{n-2} 0 \underbrace{\mathrm{d}x_1 \cdots \mathrm{d}x_n}_{n-2} = 0.
\end{aligned}$$

□

LEMMA 5. *If* $\boldsymbol{\Delta} \in \mathbb{R}^{n \times k}$ *and* $\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{a_1} & \boldsymbol{a_2} & \cdots & \boldsymbol{a_k} \end{bmatrix}$, *where* $\boldsymbol{a_i}$ *are column vectors, then*

$$\int_{\|\boldsymbol{\Delta}\|_\infty \leq \rho} \boldsymbol{\Delta}^T \boldsymbol{\Delta} \, \mathrm{d}\boldsymbol{\Delta} = \begin{bmatrix} \frac{(2\rho)^{nk}\rho^2 n}{3} & 0 & \cdots & 0 \\ 0 & \frac{(2\rho)^{nk}\rho^2 n}{3} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{(2\rho)^{nk}\rho^2 n}{3} \end{bmatrix}.$$

Proof. We have

$$\int_{\|\boldsymbol{\Delta}\|_\infty \leq \rho} \boldsymbol{\Delta}^T \boldsymbol{\Delta} \, \mathrm{d}\boldsymbol{\Delta} = \int_{\|\boldsymbol{\Delta}\|_\infty \leq \rho} \begin{bmatrix} \boldsymbol{a_1}^T \boldsymbol{a_1} & \boldsymbol{a_1}^T \boldsymbol{a_2} & \cdots & \boldsymbol{a_1}^T \boldsymbol{a_k} \\ \boldsymbol{a_2}^T \boldsymbol{a_1} & \boldsymbol{a_2}^T \boldsymbol{a_2} & \cdots & \boldsymbol{a_2}^T \boldsymbol{a_k} \\ \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{a_k}^T \boldsymbol{a_1} & \boldsymbol{a_k}^T \boldsymbol{a_2} & \cdots & \boldsymbol{a_k}^T \boldsymbol{a_k} \end{bmatrix} \mathrm{d}\boldsymbol{\Delta}.$$

We observe that by Lemma 4, we have for the off-diagonal entries,

$$\begin{aligned}
&\int_{\|\boldsymbol{\Delta}\|_\infty \leq \rho} \boldsymbol{a_i}^T \boldsymbol{a_j} \, \mathrm{d}\boldsymbol{\Delta} \\
&= \int_{\|\boldsymbol{\Delta}\|_\infty \leq \rho} \sum_{\ell=1}^{n} a_{i\ell} a_{j\ell} \, \mathrm{d}\boldsymbol{\Delta} \\
&= \sum_{\ell=1}^{n} \int_{\|\boldsymbol{\Delta}\|_\infty \leq \rho} a_{i\ell} a_{j\ell} \, \mathrm{d}\boldsymbol{\Delta} = 0.
\end{aligned}$$

By Lemma 3, for the diagonal entries,

$$\begin{aligned}
&\int_{\|\boldsymbol{\Delta}\|_\infty \le \rho} \boldsymbol{a_i}^T \boldsymbol{a_i} \,\mathrm{d}\boldsymbol{\Delta} \\
&= \int_{\|\boldsymbol{\Delta}\|_\infty \le \rho} \sum_{\ell=1}^{n} a_{i\ell}^2 \,\mathrm{d}\boldsymbol{\Delta} \\
&= \sum_{\ell=1}^{n} \int_{\|\boldsymbol{\Delta}\|_\infty \le \rho} a_{i\ell}^2 \,\mathrm{d}\boldsymbol{\Delta} \\
&= \sum_{\ell=1}^{n} \frac{(2\rho)^{nk}\rho^2}{3} \\
&= \frac{(2\rho)^{nk}\rho^2 n}{3}.
\end{aligned}$$

□

Putting everything together, for the box uncertainty set we have

$$\begin{aligned}
&\min_{\boldsymbol{\beta}} \int_{\|\boldsymbol{\Delta}\|_\infty \le \rho} \|\boldsymbol{y} - (\boldsymbol{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_2^2 \,\mathrm{d}\boldsymbol{\Delta} \\
&= \min_{\boldsymbol{\beta}} \left(\int_{\|\boldsymbol{\Delta}\|_\infty \le \rho} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \,\mathrm{d}\boldsymbol{\Delta} + \int_{\|\boldsymbol{\Delta}\|_\infty \le \rho} \|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2^2 \,\mathrm{d}\boldsymbol{\Delta} - 2\int_{\|\boldsymbol{\Delta}\|_\infty \le \rho} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Delta}\boldsymbol{\beta} \,\mathrm{d}\boldsymbol{\Delta} \right) \\
&= \min_{\boldsymbol{\beta}} \left((2\rho)^{nk} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{(2\rho)^{nk}\rho^2 n}{3} \|\boldsymbol{\beta}\|_2^2 - 0 \right) \\
&= (2\rho)^{nk} \min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{\rho^2 n}{3} \|\boldsymbol{\beta}\|_2^2,
\end{aligned}$$

and Theorem 2(b) follows.

5. Diamond Uncertainty Set

LEMMA 6. *Coxeter (1973) Let V_n denote the hypervolume of $\mathcal{U}_3$, then $V_n = \frac{(2\rho)^n}{n!}$.*

LEMMA 7. *We have $\int_{\|\boldsymbol{x}\|_1 \le \rho} x_i^2 \,\mathrm{d}\boldsymbol{\Delta} = \frac{(2\rho)^{n+1}\rho}{(n+2)!}$.*

Proof. Let V_{n-1} be the volume of the diamond uncertainty set $\{\boldsymbol{x} \in \mathbb{R}^{n-1} : \|\boldsymbol{x}\|_1 \le \rho\}$ in the $(n-1)$-th dimension. Let $y_i = \frac{x_i}{\rho - x_n}$, and $z_i = \rho y_i$ and without loss of generality, consider $x_i = x_n$.

$$\begin{aligned}
&\int_{\|\boldsymbol{x}\|_1 \le \rho} x_n^2 \,\mathrm{d}\boldsymbol{\Delta} \\
&= 2^n \int_{x_1 + \cdots + x_n \le \rho, x_i \ge 0 \ \forall i} x_n^2 \,\mathrm{d}\boldsymbol{\Delta} \\
&= 2^n \int_0^\rho \left(\int_{x_1 + \cdots + x_{n-1} \le \rho - x_n, x_i \ge 0} 1 \,\mathrm{d}x_1 \cdots \mathrm{d}x_{n-1} \right) x_n^2 \,\mathrm{d}x_n)
\end{aligned}$$

$$
\begin{aligned}
&= 2^n \int_0^\rho \left(\int_{y_1+\cdots+y_{n-1}\le 1, y_i\ge 0} (\rho - x_n)^{n-1} \,\mathrm{d}y_1 \cdots \mathrm{d}y_{n-1} \right) x_n^2 \,\mathrm{d}x_n) \\
&= 2^n \int_0^\rho \left(\int_{z_1+\cdots+z_{n-1}\le \rho, z_i\ge 0} \frac{(\rho - x_n)^{n-1}}{\rho^{n-1}} \mathrm{d}z_1 \cdots \mathrm{d}z_{n-1} \right) x_n^2 \,\mathrm{d}x_n) \\
&= 2^n \int_0^\rho \left(\int_{z_1+\cdots+z_{n-1}\le \rho, z_i\ge 0} \left(1 - \frac{x_n}{\rho}\right)^{n-1} \mathrm{d}z_1 \cdots \mathrm{d}z_{n-1} \right) x_n^2 \,\mathrm{d}x_n) \\
&= 2^n \int_0^\rho \left(1 - \frac{x_n}{\rho}\right)^{n-1} \frac{V_{n-1}}{2^{n-1}} x_n^2 \,\mathrm{d}x_n \\
&= 2V_{n-1} \int_0^\rho \left(1 - \frac{x_n}{\rho}\right)^{n-1} x_n^2 \,\mathrm{d}x_n \\
&= 2\frac{(2\rho)^{n-1}}{(n-1)!} \frac{2\rho^3}{n(n+1)(n+2)} \\
&= \frac{(2\rho)^{n+1}\rho}{(n+2)!}.
\end{aligned}
$$

□

LEMMA 8. *Let* $\mathcal{U}_3 = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_1 \le \rho\}$. *Then, for* $i \ne j$ $\int_{\|\boldsymbol{x}\|_1\le\rho} x_i x_j \,\mathrm{d}\boldsymbol{\Delta} = 0$.

Proof. We will prove by induction. We first prove the base case of $n = 2$ where $\int_{|x_1|+|x_2|\le\rho} x_1 x_2 \,\mathrm{d}x_1 \,\mathrm{d}x_2 = 0$. We decompose the integral

$$\int_{|x_1|+|x_2|\le\rho} x_1 x_2 \,\mathrm{d}x_1 \,\mathrm{d}x_2 = \int_A x_1 x_2 \,\mathrm{d}A + \int_B x_1 x_2 \,\mathrm{d}B + \int_C x_1 x_2 \,\mathrm{d}C + \int_D x_1 x_2 \,\mathrm{d}x_1 \,\mathrm{d}D,$$

where the regions $A,\ B,\ C,\ D$ are depicted in Figure 1. From symmetry we have

$$\int_A x_1 x_2 \,\mathrm{d}A = \int_C x_1 x_2 \,\mathrm{d}C = -\left(\int_B x_1 x_2 \,\mathrm{d}B\right) = -\left(\int_D x_1 x_2 \,\mathrm{d}D\right).$$

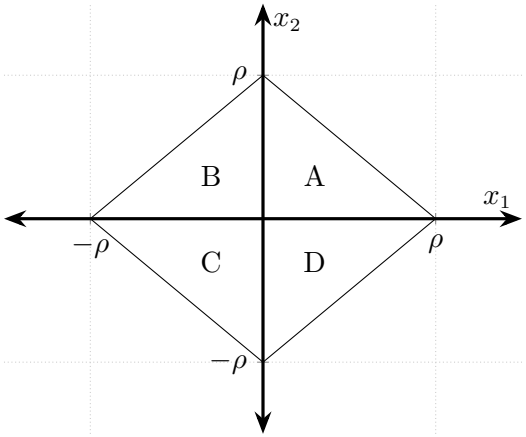

Figure 1 The decomposition of the set $\{(x_1, x_2) : |x_1| + |x_2| \le \rho\}$ into the sets $A,\ B,\ C,\ D$.

thus $\int_{|x_1|+|x_2|\le\rho} x_1x_2\,\mathrm{d}x_1\,\mathrm{d}x_2=0$.

Suppose the lemma holds for all values of k where $2\le k\le n$. Now let $k=n+1$, we consider $\mathcal{U}_3=\{\boldsymbol{x}\in\mathbb{R}^{n+1}:\|\boldsymbol{x}\|_1\le\rho\}$ be separated into two regions defined by $x_{n+1}\ge 0$ and $x_{n+1}<0$. Then we are solving for

$$\begin{aligned}
&\int_{\substack{|x_1|+\cdots+|x_{n+1}|\le\rho,\\ x_{n+1}<0}} x_1x_2\,\mathrm{d}x_1\cdots\mathrm{d}x_{n+1}+\int_{\substack{|x_1|+\cdots+|x_{n+1}|\le\rho,\\ x_{n+1}\ge 0}} x_1x_2\,\mathrm{d}x_1\cdots\mathrm{d}x_{n+1}\\
&=\int_{-\rho}^{0}\left(\int_{|x_1|+\cdots+|x_n|\le\rho+x_{n+1}} x_1x_2\,\mathrm{d}x_1\cdots\mathrm{d}x_n\right)\mathrm{d}x_{n+1}+\\
&\int_{0}^{\rho}\left(\int_{|x_1|+\cdots+|x_n|\le\rho-x_{n+1}} x_1x_2\,\mathrm{d}x_1\cdots\mathrm{d}x_n\right)\mathrm{d}x_{n+1}=0,
\end{aligned}$$

since both integrals are both 0. This shows that the lemma holds for $k=n+1$ and concludes the proof. □

LEMMA 9. *If* $\boldsymbol{\Delta}\in\mathbb{R}^{n\times k}$ *and* $\boldsymbol{\Delta}=\begin{bmatrix}\boldsymbol{a_1}\ \boldsymbol{a_2}\ \cdots\ \boldsymbol{a_k}\end{bmatrix}$, *where* $\boldsymbol{a_i}$ *are column vectors, then*

$$\int_{\|\boldsymbol{\Delta}\|_1\le\rho}\boldsymbol{\Delta}^T\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{\Delta}=\begin{bmatrix}\frac{(2\rho)^{nk+1}\rho n}{(nk+2)!} & 0 & \cdots & 0\\ 0 & \frac{(2\rho)^{nk+1}\rho n}{(nk+2)!} & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & \cdots & \frac{(2\rho)^{nk+1}\rho n}{(nk+2)!}\end{bmatrix}.$$

Proof.

$$\int_{\|\boldsymbol{\Delta}\|_1\le\rho}\boldsymbol{\Delta}^T\boldsymbol{\Delta}\,\mathrm{d}\boldsymbol{\Delta}=\int_{\|\boldsymbol{\Delta}\|_1\le\rho}\begin{bmatrix}\boldsymbol{a_1}^T\boldsymbol{a_1} & \boldsymbol{a_1}^T\boldsymbol{a_2} & \cdots & \boldsymbol{a_1}^T\boldsymbol{a_k}\\ \boldsymbol{a_2}^T\boldsymbol{a_1} & \boldsymbol{a_2}^T\boldsymbol{a_2} & \cdots & \boldsymbol{a_2}^T\boldsymbol{a_k}\\ \cdots & \cdots & \cdots & \cdots\\ \boldsymbol{a_k}^T\boldsymbol{a_1} & \boldsymbol{a_k}^T\boldsymbol{a_2} & \cdots & \boldsymbol{a_k}^T\boldsymbol{a_k}\end{bmatrix}\mathrm{d}\boldsymbol{\Delta}.$$

We observe that the elements off-diagonal can all be expressed as

$$\begin{aligned}
&\int_{\|\boldsymbol{\Delta}\|_1\le\rho}\boldsymbol{a_i}^T\boldsymbol{a_j}\,\mathrm{d}\boldsymbol{\Delta}\\
&=\int_{\|\boldsymbol{\Delta}\|_1\le\rho}\sum_{\ell=1}^{n}a_{i\ell}a_{j\ell}\,\mathrm{d}\boldsymbol{\Delta}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\ell=1}^{n} \int_{\|\boldsymbol{\Delta}\|_1 \le \rho} a_{i\ell} a_{j\ell} \, \mathrm{d}\boldsymbol{\Delta} \\
&= 0,
\end{aligned}
$$

where the last equality is obtained by Lemma 8. For the terms in the diagonal, by Lemma 7 we have

$$
\begin{aligned}
&\int_{\|\boldsymbol{\Delta}\|_1 \le \rho} \boldsymbol{a_i}^T \boldsymbol{a_i} \, \mathrm{d}\boldsymbol{\Delta} \\
&= \int_{\|\boldsymbol{\Delta}\|_1 \le \rho} \sum_{\ell=1}^{n} a_{i\ell}^2 \, \mathrm{d}\boldsymbol{\Delta} \\
&= \sum_{\ell=1}^{n} \int_{\|\boldsymbol{\Delta}\|_1 \le \rho} a_{i\ell}^2 \, \mathrm{d}\boldsymbol{\Delta} \\
&= \frac{(2\rho)^{nk+1} \rho n}{(nk+2)!}.
\end{aligned}
$$

□

Putting everything together, for the polyhedral uncertainty set we have

$$
\begin{aligned}
&\min_{\boldsymbol{\beta}} \int_{\|\boldsymbol{\Delta}\|_1 \le \rho} \|\boldsymbol{y} - (\boldsymbol{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2^2 \, \mathrm{d}\boldsymbol{\Delta} \\
&= \min_{\boldsymbol{\beta}} \left(\int_{\|\boldsymbol{\Delta}\|_1 \le \rho} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \, \mathrm{d}\boldsymbol{\Delta} + \int_{\|\boldsymbol{\Delta}\|_1 \le \rho} \|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2^2 \, \mathrm{d}\boldsymbol{\Delta} - 2 \int_{\|\boldsymbol{\Delta}\|_1 \le \rho} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{\Delta}\boldsymbol{\beta}) \, \mathrm{d}\boldsymbol{\Delta} \right) \\
&= \min_{\boldsymbol{\beta}} \left(\frac{(2\rho)^{nk}}{nk!} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \int_{\|\boldsymbol{\Delta}\|_1 \le \rho} \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{\Delta}^{\boldsymbol{T}} \boldsymbol{\Delta} \boldsymbol{\beta} \, \mathrm{d}\boldsymbol{\Delta} - 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \left(\int_{\|\boldsymbol{\Delta}\|_1 \le \rho} \boldsymbol{\Delta} \, \mathrm{d}\boldsymbol{\Delta} \right) \boldsymbol{\beta} \right) \\
&= \min_{\boldsymbol{\beta}} \left(\frac{(2\rho)^{nk}}{nk!} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{(2\rho)^{nk+1} \rho n}{(nk+2)!} \|\boldsymbol{\beta}\|_2^2 - 0 \right) \\
&= \frac{(2\rho)^{nk}}{nk!} \min_{\boldsymbol{\beta}} \left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{2n\rho^2}{(nk+2)(nk+1)} \|\boldsymbol{\beta}\|_2^2 \right).
\end{aligned}
$$

The 2nd equality follows from the fact that the volume of a hypercube with a dimension of nk and a side length of 2ρ is $(2\rho)^{nk}$. Thus, Theorem 2(c) follows.

6. Budget Uncertainty Set

LEMMA 10. *The volume V_n of the region $\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_1 \le \rho, \|\boldsymbol{x}\|_\infty \le \Gamma\}$, is $V_n = \frac{(2\rho)^n - n(2(\rho-\Gamma))^n}{n!}$.*

Proof. V_n is the volume of a polytope defined by the region $A_n = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_1 \le \rho\}$ truncated out by $2n$ corners of the regions $\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_1 \le \rho, \|\boldsymbol{x}\|_\infty > \Gamma\}$. From our analysis of $\mathcal{U}_3$, the volume of A_n is $\frac{(2\rho)^n}{n!}$ and two of these $2n$ corners from opposing sides can be combined into a polytope of volume $\frac{(2(\rho-\Gamma)^n)}{n!}$, and thus we have $V_n = \frac{(2\rho)^n}{n!} - n\frac{(2(\rho-\Gamma))^n}{n!}$. □

LEMMA 11. *We have*

$$\int_{\|\boldsymbol{x}\|_1\le\rho,\ \|\boldsymbol{x}\|_\infty\le\Gamma} x_i^2\,\mathrm{d}\boldsymbol{\Delta} = \frac{2\rho^2}{(n+1)(n+2)}\frac{(2\rho)^n - n(2(\rho-\Gamma))^n}{n!} - \frac{2(\rho-\Gamma)^n}{n!}\frac{(n^2+3n-2)\Gamma^2+(4-2n)\rho\Gamma}{(n+1)(n+2)}.$$

Proof. Without loss of generality, we compute the case of $x_i = x_n$. If we assume that all $x_i \ge 0$, then depending on the value of x_n, there can be two cases where

$$x_1+\cdots+x_n\le\rho, x_i\le\Gamma \quad \forall i\in[1:n]\begin{cases} 0\le x_n\le\rho-\Gamma, & \text{Case 1, denote as region } A_n \\ \rho-\Gamma\le x_n\le\Gamma, & \text{Case 2, denote as region } B_n \end{cases}$$

In Case 1, where $y_i = \frac{x_i}{\rho-x_n}$, $z_i = \rho y_i$, and V_{n-1} is the volume defined by Lemma 10,

$$\begin{aligned}
&\int_{\|\boldsymbol{x}\|_1\le\rho,\|\boldsymbol{x}\|_\infty\le\Gamma,0\le x_n\le\rho-\Gamma} x_n^2\,\mathrm{d}A_n \\
&= 2^n\int_0^{\rho-\Gamma}\int_{x_1+\cdots+x_n\le\rho,0\le x_i\le\Gamma\ \forall i\in[1:n-1]} x_n^2\,\mathrm{d}A_{n-1} \\
&= 2^n\int_0^{\rho-\Gamma}\int_{x_1+\cdots+x_{n-1}\le\rho-x_n,0\le x_i\le\Gamma\ \forall i\in[1:n-1]} x_n^2\,\mathrm{d}x_1\cdots\mathrm{d}x_{n-1} \\
&= 2^n\int_0^{\rho-\Gamma}\int_{y_1+\cdots+y_{n-1}\le1,0\le y_i\le\frac{\Gamma}{\rho-x_n}\ \forall i\in[1:n-1]} (\rho-x_n)^{n-1}x_n^2\mathrm{d}y_1\cdots\mathrm{d}y_{n-1} \\
&= 2^n\int_0^{\rho-\Gamma}\int_{z_1+\cdots+z_{n-1}\le1,0\le z_i\le\frac{\rho\Gamma}{\rho-x_n}\ \forall i\in[1:n-1]} \frac{(\rho-x_n)^{n-1}x_n^2}{\rho^{n-1}}\mathrm{d}z_1\cdots\mathrm{d}z_{n-1} \\
&= 2^n\int_0^{\rho-\Gamma}\frac{(\rho-x_n)^{n-1}x_n^2}{\rho^{n-1}}\frac{V_{n-1}}{2^{n-1}}\,\mathrm{d}x_n \\
&= 2\int_0^{\rho-\Gamma}\frac{(\rho-x_n)^{n-1}x_n^2}{\rho^{n-1}}\frac{(2\rho)^n-(n-1)(2(\rho-\frac{\rho\Gamma}{\rho-x_n}))^{n-1}}{(n-1)!}\,\mathrm{d}x_n \\
&= \frac{2^n}{(n-1)!}\int_0^{\rho-\Gamma}\frac{(\rho-x_n)^{n-1}x_n^2}{\rho^{n-1}}\rho^{n-1}\left(1-(n-1)(\frac{\rho-\Gamma-x_n}{\rho-x_n})^{n-1}\right)\mathrm{d}x_n \\
&= \frac{2^n}{(n-1)!}\int_0^{\rho-\Gamma}(\rho-x_n)^{n-1}x_n^2\frac{(\rho-x_n)^{n-1}-(n-1)(\rho-\Gamma-x_n)^{n-1}}{(\rho-x_n)^{n-1}}\,\mathrm{d}x_n \\
&= \frac{2^n}{(n-1)!}\int_0^{\rho-\Gamma}(\rho-x_n)^{n-1}x_n^2-(n-1)(\rho-\Gamma-x_n)^{n-1}x_n^2\,\mathrm{d}x_n \\
&= \frac{2^n}{(n+2)!}(\Gamma^n((4n^2+2n)\rho\Gamma-(n^2+n)\Gamma^2-(n^2+3n+2)\rho^2)+2\rho^{n+2}+ \\
&(\rho-\Gamma)^n((2-2n)\Gamma^2+(4n-4)\rho\Gamma+(2-2n)\rho^2)).
\end{aligned}$$

In Case 2, since we have $\rho-\Gamma\le x_n\le\Gamma$, we can rewrite $x_1+\cdots+x_{n-1}\le\rho-x_n$ as $x_1+\cdots+x_{n-1}\le\Gamma$. This implies that in case 2, $x_i\le\Gamma\ \ \forall i\in[1:n-1]$ will be automatically satisfied. We thus instead

are dealing with the problem of $x_1 + \cdots + x_{n-1} \le \rho - x_n, x_i \ge 0$. We recognize that this is exactly the diamond uncertainty set case. Thus we have the following:

$$\begin{aligned}
&\int_{\|\boldsymbol{x}\|_1 \le \rho, \|\boldsymbol{x}\|_\infty \le \Gamma, \rho - \Gamma \le x_n \le \Gamma} x_n^2 \,\mathrm{d}B_n \\
&= 2^n \int_{\rho-\Gamma}^{\Gamma} \int_{x_1 + \cdots + x_n \le \rho, x_i \ge 0,\ \forall i \in [1:n-1]} x_n^2 \,\mathrm{d}B_{n-1} \\
&= 2^n \int_{\rho-\Gamma}^{\Gamma} \left(1 - \frac{x_n}{\rho}\right)^{n-1} \frac{V_{n-1}}{2^{n-1}} x_n^2 \,\mathrm{d}x_n \\
&= 2 \int_{\rho-\Gamma}^{\Gamma} \left(1 - \frac{x_n}{\rho}\right)^{n-1} \frac{(2\rho)^{n-1}}{(n-1)!} x_n^2 \,\mathrm{d}x_n \\
&= \frac{2^n}{(n-1)!} \int_{\rho-\Gamma}^{\Gamma} (\rho - x_n)^{n-1} x_n^2 \,\mathrm{d}x_n \\
&= \frac{2^n}{(n+2)!} (\Gamma^n((-2n^2 - 4n)\rho\Gamma + (n^2+n)\Gamma^2 + (n^2+3n+2)\rho^2) + \\
&(\rho - \Gamma)^n((-n^2 - n)\Gamma^2 - 2n\rho\Gamma - 2\rho^2)).
\end{aligned}$$

Combing both cases, we have

$$\begin{aligned}
&\int_{\|\boldsymbol{x}\|_1 \le \rho, \|\boldsymbol{x}\|_\infty \le \Gamma} x_n^2 \,\mathrm{d}\boldsymbol{\Delta} \\
&= \int_{\|\boldsymbol{x}\|_1 \le \rho, \|\boldsymbol{x}\|_\infty \le \Gamma, 0 \le x_n \le \rho - \Gamma} x_n^2 \,\mathrm{d}A_n + \int_{\|\boldsymbol{x}\|_1 \le \rho, \|\boldsymbol{x}\|_\infty \le \Gamma, \rho - \Gamma \le x_n \le \Gamma} x_n^2 \,\mathrm{d}B_n \\
&= \frac{2^n}{(n+2)!} (2\rho^{n+2} - (\rho - \Gamma)^n((n^2 + 3n - 2)\Gamma^2 + (4 - 2n)\rho\Gamma + 2n\rho^2) \\
&= \frac{2\rho^2}{(n+1)(n+2)} \frac{(2\rho)^n - n(2(\rho - \Gamma))^n}{n!} - \frac{(2(\rho - \Gamma))^n}{n!} \frac{(n^2 + 3n - 2)\Gamma^2 + (4 - 2n)\rho\Gamma}{(n+1)(n+2)}.
\end{aligned}$$

□

LEMMA 12. *Let* $\mathcal{U}_4 = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_1 \le \rho, \|\boldsymbol{x}\|_\infty \le \Gamma\}$. *We have* $\int_{\|\boldsymbol{x}\|_1 \le \rho, \|\boldsymbol{x}\|_\infty \le \Gamma} x_i x_j \,\mathrm{d}\boldsymbol{\Delta} = 0$.

Proof. We prove this by induction. First for the base case of $n = 2$, we show that $\int_{\|\boldsymbol{x}\|_1 \le \rho, \|\boldsymbol{x}\|_\infty \le \Gamma} x_1 x_2 \,\mathrm{d}x_1 \mathrm{d}x_2 = 0$. We decompose the integral

$$\int_{\|\boldsymbol{x}\|_1 \le \rho, \|\boldsymbol{x}\|_\infty \le \Gamma} x_1 x_2 \,\mathrm{d}x_1 \mathrm{d}x_2 = \int_A x_1 x_2 \,\mathrm{d}A + \int_B x_1 x_2 \,\mathrm{d}B + \int_C x_1 x_2 \,\mathrm{d}C + \int_D x_1 x_2 \,\mathrm{d}D,$$

where the sets $A,\ B,\ C,\ D$ are depicted n Figure 2. Due to symmetry, we have

$$\int_A x_1 x_2 \,\mathrm{d}A = \int_C x_1 x_2 \,\mathrm{d}C = -\left(\int_B x_1 x_2 \,\mathrm{d}B\right) = -\left(\int_D x_1 x_2 \,\mathrm{d}D\right).$$

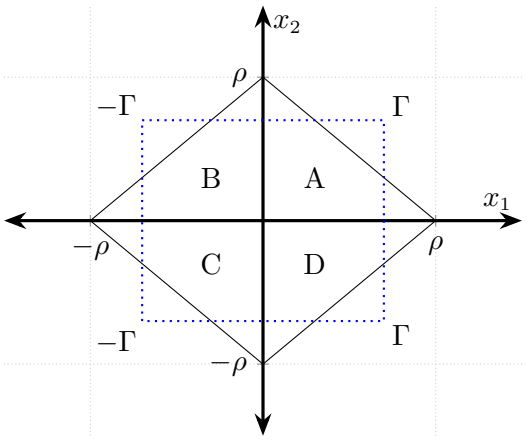

Figure 2 Demonstration of $\mathcal{U}_4$ in $\mathbb{R}^2$ defined by the $\ell 1$-norm set of $\{(x_1,x_2):\|\boldsymbol{x}\|_1\le\rho,\|\boldsymbol{x}\|_\infty\le\Gamma\}$, as observed, due to symmetry, we see that the value of x_1x_2 is the same in A and C, and in B and D.

Suppose the lemma holds for all values of k where $2\le k\le n$, let us prove the case of $k=n+1$. The space defined by $\mathcal{U}_4=\{\boldsymbol{x}\in\mathbb{R}^{n+1}:\|\boldsymbol{x}\|_1\le\rho,\|\boldsymbol{x}\|_\infty\le\Gamma\}$ can be separated to two spaces defined by $-\Gamma\le x_{n+1}<0$ and $0\le x_{n+1}\le\Gamma$. Then, similar to Lemma 11 where we consider two cases depending on the value of x_{n+1}, we are solving for

$$\begin{aligned}
&\int_{\substack{|x_1|+\cdots+|x_{n+1}|\le\rho,\\ |x_1|\le\Gamma,\cdots,|x_n|\le\Gamma,\\ -\Gamma\le x_{n+1}\le 0}} x_1x_2\,\mathrm{d}x_1\cdots\mathrm{d}x_{n+1}+\int_{\substack{|x_1|+\cdots+|x_{n+1}|\le\rho,\\ |x_1|\le\Gamma,\cdots,|x_n|\le\Gamma,\\ 0\le x_{n+1}\le\Gamma}} x_1x_2\,\mathrm{d}x_1\cdots\mathrm{d}x_{n+1}\\
&\qquad=\int_{-\Gamma}^{-(\rho-\Gamma)}\left(\int_{|x_1|+\cdots+|x_n|\le\rho+x_{n+1}} x_1x_2\,\mathrm{d}x_1\cdots\mathrm{d}x_n\right)\mathrm{d}x_{n+1}\\
&\qquad+\int_{-(\rho-\Gamma)}^{0}\left(\int_{\substack{|x_1|+\cdots+|x_n|\le\rho+x_{n+1},\\ |x_1|\le\Gamma,\cdots,|x_n|\le\Gamma}} x_1x_2\,\mathrm{d}x_1\cdots\mathrm{d}x_n\right)\mathrm{d}x_{n+1}\\
&\qquad+\int_{0}^{\rho-\Gamma}\left(\int_{\substack{|x_1|+\cdots+|x_n|\le\rho-x_{n+1},\\ |x_1|\le\Gamma,\cdots,|x_n|\le\Gamma}} x_1x_2\,\mathrm{d}x_1\cdots\mathrm{d}x_n\right)\mathrm{d}x_{n+1}+\\
&\qquad\int_{\rho-\Gamma}^{\Gamma}\left(\int_{|x_1|+\cdots+|x_n|\le\rho-x_{n+1}} x_1x_2\,\mathrm{d}x_1\cdots\mathrm{d}x_n\right)\mathrm{d}x_{n+1}=0,
\end{aligned}$$

where the second and third inner integrals, as shown above, are 0, and the first and fourth inner integrals denote diamond uncertainty sets, which by Lemma 8, are also 0. We thus proved this lemma also holds for $n+1$, and conclude the proof. □

Putting everything together, for the budget uncertainty set, we have the following:

$$\begin{aligned}
&\min_{\boldsymbol{\beta}} \int_{\|\boldsymbol{\Delta}\|_1 \le \rho, \|\boldsymbol{\Delta}\|_\infty \le \Gamma} \|\boldsymbol{y} - (\boldsymbol{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2^2 \, \mathrm{d}\boldsymbol{\Delta} \\
&= \min_{\boldsymbol{\beta}} \int_{\|\boldsymbol{\Delta}\|_1 \le \rho, \|\boldsymbol{\Delta}\|_\infty \le \Gamma} \left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{\Delta}\boldsymbol{\beta}\|_2^2 - 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{\Delta}\boldsymbol{\beta}) \right) \mathrm{d}\boldsymbol{\Delta} \\
&= V_n \min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + f(n, k, \rho, \Gamma)\|\boldsymbol{\beta}\|_2^2,
\end{aligned}$$

where

$$f(n, k, \Gamma, \rho) = \frac{2n\rho^2}{(nk+1)(nk+2)} - \frac{n(\rho - \Gamma)^{nk}((n^2k^2 + 3nk - 2)\Gamma^2 + (4 - 2nk)\rho\Gamma)}{(nk+1)(nk+2)(\rho^{nk} - (\rho - \Gamma)^{nk})},$$

and Theorem 2(d) follows.

7. Computational Results

In this section, we study the performance of averaged uncertainty robust regression (AUR) against worst-case uncertainty robust regression (WUR) using both synthetic and real-world data and found that AUR outperforms WUR across all datasets. All experiments are run using Gurobi 0.11.5, Julia 1.9.3, and Python 3.10.6 using a Mac Intel i7 core. The Homogenous Barrier algorithm was used for the optimization formulation to avoid numerical issues. Our codebase is publicly available for those interested in reproducing results presented in this paper Ma and Bertsimas (2023).

7.1. Computational Experiment Set-up

The main goal of the experiment is to compare AUR against WUR

- Worst-case uncertainty robust regression (WUR): $\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_2$.
- Averaged uncertainty robust regression (AUR): $\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$.

We use mean squared error (MSE) as the evaluation metric for all experiments.

To select λ, we adopt either the computed constant from Theorem 2 or cross-validation (CV), which retrieves λ that achieves the best validation loss and retraining it on the original training set. The CV grids are defined by choice of λ that ranges from 0 to 1 with 0.05 increments for all experiments to ensure a fine-grain grid for comparison.

7.2. Real-World Data

We selected ten publicly available UCI regression datasets Dua and Graff (2017) to analyze the performance of AUR. When missing data is present in the original data, we drop the entire sample. If a feature contains more than 20% missing values, we drop this feature. We also pre-process the datasets by removing features that do not contain useful information. The final dataset is then standardized using min-max scaling. The information of the datasets is summarized in Table 1. To simulate different real-world noises, we added perturbations generated using the hit-and-run Zabinsky (2009) method from the ellipsoidal, box, diamond, and budget uncertainty sets with values of $\rho \in [0.001, 0.01, 0.05, 0.1, 0.2, 0.3]$. For each dataset and each perturbation strength ρ, 10 perturbations are generated using different random seeds to ensure our results account for a diverse perturbation of noise under the same condition. We then split each dataset into 80/20 training and testing sets and applied AUR and WUR respectively to study their out-of-sample MSE performances. Overall, we conducted 2400 experiments that vary across different uncertainty sets (4), different datasets (10), different perturbation strengths (6), and different perturbation randomness (10).

7.3. Performance on Real World Data

Below we report the MSE of AUR over WUR on the real-world datasets over different uncertainty sets in Figure 3. We observe that across all different perturbation levels, AUR outperforms WUR by 0.4% - 0.9% on average, with improvements increasing as perturbation increases. This result confirms our belief that AUR is able to protect against noise more holistically than traditional WUR, and can be especially useful for real-world datasets when there is strong noise perturbation. We note an exception of the box uncertainty set, which decreases as perturbation increases. We argue that this is because box uncertainty set by nature protects against worst-case global perturbation of every data entry, and is inherently an over-protection. We obtain high λ values as the size of the sample size grows, which over-regularizes the training and attributes to this behavior.

Dataset Name	Number of Samples	Number of Features
Abalone	4177	9
Auto-MPG	398	8
Automobile	193	25
Breast Cancer Wisconsin	194	34
Computer Hardware	209	9
Concrete	1030	9
Wine Quality (red)	1599	12
Wine Quality (white)	4898	12
Energy Efficiency	768	9
Synchronous Machine	557	5

Table 1 UCI datasets used in real-world experiments, where sample sizes and feature sizes range different scales.

Another important observation is the advantage of using regularization terms obtained by Theorem 2 in comparison to those obtained by cross-validation (CV). As seen in Figure 3, we achieve a 0.6-0.8% MSE improvement. Their improvements are relatively equivalent across different perturbation levels across budget, diamond, and ellipsoidal uncertainty sets, confirming a consistent advantage. Besides the performance improvement offered by using the regularization terms computed according to Theorem 2, we also observe that CV is susceptible to the randomness of the training procedure when choosing the optimal regularization term. Given the same UCI dataset as well as the same uncertainty set, we expect to see the same regularization terms selected as they protect against the same set of noises. However, we show in Table 2 that the number of different regularization terms selected for the same dataset can be as large as 6 using CV, whereas we only need to consider 1 regularization term using Theorem 2. This implies that CV is not the most reliable methodology for computing regularization terms, as it provides unstable selections as we are exposed to randomness. We note that budget and diamond uncertainty sets give consistent CV-selected regularization terms, and this is because in practice, when the dimension of the problem becomes large, in order

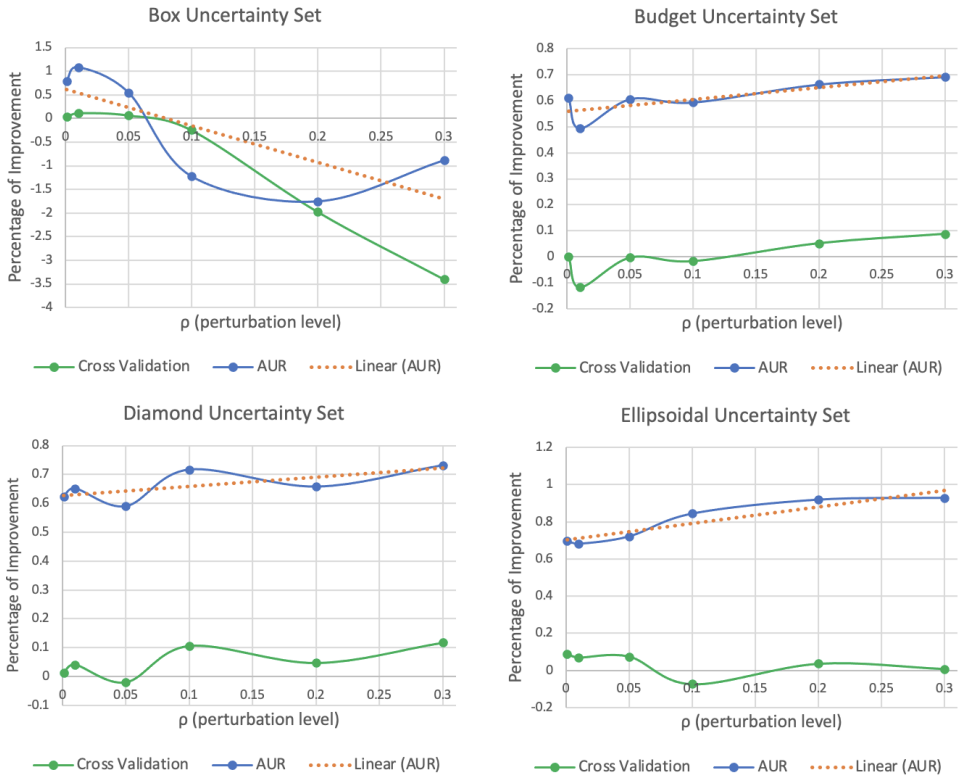

Figure 3 Percentage of AUR over WUR across 10 UCI datasets, where the orange line is the trend line for AUC improvements from Theorem 2 computed regularization term.

# of Different λ	Box	Budget	Diamond	Ellipsoidal
1	23.3%	100%	100%	80.0%
2	23.3%	0%	0%	15.0%
3	23.3%	0%	0%	3.33%
4	15.0%	0%	0%	1.67%
5	10.0%	0%	0%	0%
6	5.0%	0%	0%	0%

Table 2 Frequency of experiments that have different regularization terms (λ) obtained with 10 different perturbation noise of the same UCI dataset with the same perturbation strength. It demonstrates the instability of CV regularization term selection

for the perturbation to be contained within the diamond and budget uncertainty sets, the scale of noise becomes smaller than those contained in ellipsoidal or box uncertainty sets, and randomness has a diminished effect on the regularization term selection.

7.4. Synthetic Data

We study more closely the behavior of AUR in comparison to WUR as the number of informative features and the number of samples vary using synthetic datasets. We generated synthetic regression datasets where the regression target is a random linear combination of random features that are well-defined, centered, and unbiased. We vary the number of data samples (300, 400, 500, 600, 700, 900), as well as the number of informative features (3, 4, 5, 6, 7, 8, 10) to study the effects of these factors on the performance. Additive perturbations are then generated using the hit-and-run method to simulate noise from the ellipsoidal, box, diamond, and budget uncertainty sets.

Specifically, we test the noise level of the uncertainty set of $\rho \in [0.001, 0.01, 0.05, 0.1, 0.2, 0.3]$, where for the budget uncertainty set, we choose $\Gamma = 0.8\rho$. For our samples to truly reflect the monotonically increasing perturbation level, we enforce that samples generated from a higher perturbation level must not reside in the space from the previously smaller perturbation level (i.e., generated perturbation matrix from $\rho = 0.3$ cannot reside in the uncertainty set defined by $\rho = 0.2$). To achieve stability of our results, we repeated each experiment 20 times with a different random seed.

7.5. Performance on Synthetic Data

We observe that AUR improves over WUR across all uncertainty sets, sample sizes, as well as number of informative features. This is in accordance with what we have seen in the real-world datasets. Importantly, the improvements across all uncertainty sets decrease as the number of samples increases, and as the number of informative features increases as shown in Figure 4. This observation implies that AUR's advantage diminishes as the scale and complexity of the regression problem increases.

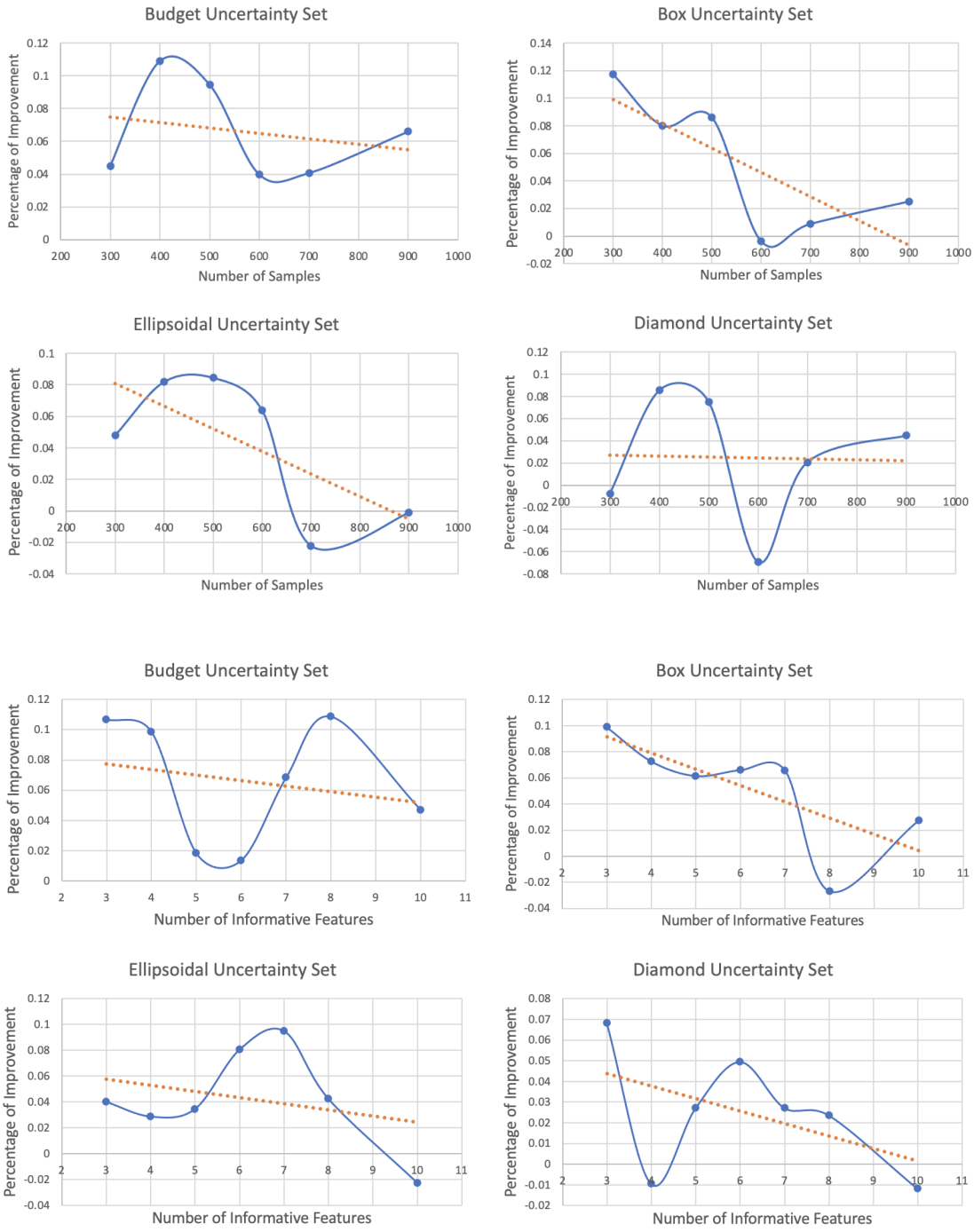

Figure 4 Percentage of improvement of AUR from WUR across different synthetic datasets with different sample sizes and different informative feature sizes. The orange line indicates the trend of improvement, where it monotonically decreases as the sample size increases, and as the number of informative features increases.

8. Conclusions

In this work, we introduced robust regression over averaged uncertainty sets and found that it is equivalent to the mean squared regression with ridge regularization for ellipsoidal, box, diamond, and budget uncertainty sets. We thus established a natural explanation via averaged uncertainty sets of ridge regression. On both synthetic and real-world datasets we found that empirically, the averaged uncertainty set approach outperforms the worst-case uncertainty case out-of-sample in all experiments, both real-world and synthetic. We also observe that the improvements of AUR decrease as the number of samples and the number of informative features increase, meaning that as the scale and complexity of the problem become harder, the advantage of robust regression under averaged uncertainty set over the worst-case uncertainty set diminishes.

References

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, December 2009. doi: 10.1515/9781400831050. URL https://doi.org/10.1515/9781400831050.

Aharon Ben-Tal, Elad Hazan, Tomer Koren, and Shie Mannor. Oracle-based robust optimization via online learning. *Operations Research*, 63(3):628–638, June 2015. doi: 10.1287/opre.2015.1374. URL https://doi.org/10.1287/opre.2015.1374.

Jose Bento, Ralph Furmaniak, and Surjyendu Ray. On the complexity of the weighted fused lasso. *IEEE Signal Processing Letters*, 25(10):1595–1599, October 2018. doi: 10.1109/lsp.2018.2867800. URL https://doi.org/10.1109/lsp.2018.2867800.

Dimitris Bertsimas and Martin S. Copenhaver. Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 270(3):931–942, November 2018. doi: 10.1016/j.ejor.2017.03.051. URL https://doi.org/10.1016/j.ejor.2017.03.051.

Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations Research*, 52(1):35–53, feb 2004. doi: 10.1287/opre.1030.0065. URL https://doi.org/10.1287%2Fopre.1030.0065.

Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, January 2011. doi: 10.1137/080734510. URL https://doi.org/10.1137/080734510.

Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, February 2017. doi: 10.1007/s10107-017-1125-8. URL https://doi.org/10.1007/s10107-017-1125-8.

Dimitris Bertsimas, Jack Dunn, Colin Pawlowski, and Ying Daisy Zhuo. Robust classification. *INFORMS Journal on Optimization*, 1(1):2–34, January 2019. doi: 10.1287/ijoo.2018.0001. URL https://doi.org/10.1287/ijoo.2018.0001.

Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data*. Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-20192-9. URL https://doi.org/10.1007%2F978-3-642-20192-9.

H. S. M. Coxeter. *Regular Polytopes*. Dover Publications, 1973.

Dheeru Dua and Casey Graff. Uci machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Gerald B. Folland. How to integrate a polynomial over a sphere. *The American Mathematical Monthly*, 108 (5):446–448, 2001. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2695802.

Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1035–1064, oct 1997. doi: 10.1137/s0895479896298130. URL https://doi.org/10.1137%2Fs0895479896298130.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Reihaneh H. Hariri, Erik M. Fredericks, and Kate M. Bowers. Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, 6(1), jun 2019. doi: 10.1186/s40537-019-0206-3. URL https://doi.org/10.1186%2Fs40537-019-0206-3.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2009. doi: 10.1007/978-0-387-84858-7. URL https://doi.org/10.1007%2F978-0-387-84858-7.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, feb 1970. doi: 10.1080/00401706.1970.10488634. URL https://doi.org/10.1080%2F00401706.1970.10488634.

Peter E. Kennedy. *A Guide to Econometrics, Fifth Edition*. The MIT Press, 2003.

Anastasis Kratsios and Cody Hyndman. Deep arbitrage-free learning in a generalized HJM framework via arbitrage-regularization. *Risks*, 8(2):40, apr 2020. doi: 10.3390/risks8020040. URL https://doi.org/10.3390%2Frisks8020040.

Adrian Lewis. Robust regularization. Technical report, Simon Fraser University, 2002.

Adrian S. Lewis and C. H. Jeffrey Pang. Lipschitz behavior of the robust regularization. *SIAM Journal on Control and Optimization*, 48(5):3080–3104, jan 2010. doi: 10.1137/08073682x. URL https://doi.org/10.1137%2F08073682x.

Yu Ma and Dimitris Bertsimas. Averaged Robust Regression, 11 2023. URL https://github.com/yuma-sudo/RO-average.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, apr 1995. doi: 10.1137/s0097539792240406. URL https://doi.org/10.1137%2Fs0097539792240406.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, jan 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://doi.org/10.1111%2Fj.2517-6161.1996.tb02080.x.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(1):91–108, dec 2004. doi: 10.1111/j.1467-9868.2005.00490.x. URL https://doi.org/10.1111%2Fj.1467-9868.2005.00490.x.

Li Wang, Michael D. Gordon, and Ji Zhu. Regularized least absolute deviations regression and an efficient algorithm for parameter tuning. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 690–700, 2006. doi: 10.1109/ICDM.2006.134.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'08, page 1801–1808, Red Hook, NY, USA, 2008. Curran Associates Inc. ISBN 9781605609492.

Zelda B. Zabinsky. Global optimization: Hit and run methods. In Christodoulos A. Floudas and Panos M. Pardalos, editors, *Encyclopedia of Optimization, Second Edition*, pages 1342–1346. Springer, 2009. doi: 10.1007/978-0-387-74759-0_236. URL https://doi.org/10.1007/978-0-387-74759-0_236.