Beyond the Norms: Detecting Prediction Errors in Regression Models

Andres Altieri [1] **Marco Romanelli** [2] **Georg Pichler** [3] **Florence Alberge** [4] **Pablo Piantanida** [5]

Abstract

This paper tackles the challenge of detecting unreliable behavior in regression algorithms, which may arise from intrinsic variability (e.g., aleatoric uncertainty) or modeling errors (e.g., model uncertainty). First, we formally introduce the notion of *unreliability in regression*, i.e., when the output of the regressor exceeds a specified discrepancy (or error). Then, using powerful tools for probabilistic modeling, we estimate the discrepancy density, and we measure its statistical diversity using our proposed metric for *statistical dissimilarity*. In turn, this allows us to derive a data-driven score that expresses the uncertainty of the regression outcome. We show empirical improvements in error detection for multiple regression tasks, consistently outperforming popular baseline approaches, and contributing to the broader field of uncertainty quantification and safe machine learning systems. Our code is available at `https://zenodo.org/records/11281964`.

1. Introduction

In recent years, machine learning has gained ground in fields such as automatic processing, and autonomous decision-making, generating a crucial need for operational safety and reliability, aiming to prevent catastrophic errors (Amodei et al., 2016; Antun et al., 2020; Li et al., 2023). The need for trustworthy models has generated a lot of research in areas concerned with the identification of anomalous patterns that may trigger undesired, and potentially dangerous, predictions such as out-of-distribution detection (Dadalto et al., 2022; Ming et al., 2023), selective classification (Corbière et al., 2019; Huang et al., 2020), misclassification detection (Granese et al., 2021), and adversarial attacks (Goodfellow et al., 2015; Picot et al., 2023), among others.

Anomalies may stem from intrinsic variability in the data, i.e., *aleatoric uncertainty* (Kiureghian & Ditlevsen, 2009). For instance, certain inputs may exhibit a larger dispersion of the dependent variable than expected, i.e., *heteroscedastic uncertainty* (Kendall & Gal, 2017). Additional sources of unreliable behavior may be associated with approximation errors or model uncertainty (Kiureghian & Ditlevsen, 2009), occurring when the regressor is not statistically close to potential observations. This discrepancy may arise from various factors such as sub-optimal model selection, inadequate model fitting, and insufficient training, among others.

The challenge of identifying anomalous predictions in machine learning models is closely connected to the task of uncertainty quantification, which aims to assess the uncertainty inherent in model predictions (Kotelevskii et al., 2022; Liu et al., 2020). In classification tasks, the process typically involves estimating a probability distribution across labels, introducing a clear distinction between correct and incorrect decisions and a straightforward understanding of uncertainty. In stark contrast, regression problems often present a more convoluted landscape: a single regressor output is not sufficient to evaluate the model uncertainty. Thus, an additional uncertainty estimation in the form of a conditional distribution or prediction interval is usually incorporated.

This work aims to develop a simple yet effective framework to evaluate the reliability of the predictions of a given regressor and detect potentially anomalous situations that may arise. Our approach involves data-driven techniques to address the inherent challenges in the detection problem, particularly in compensating for estimation inaccuracies associated with baseline methods relying on the estimation of the conditional distribution of the dependent target variable.

1.1. Contributions

Our main contributions can be summarized as follows:

[1]Laboratoire des signaux et systèmes (L2S), Université Paris-Saclay CNRS CentraleSupélec, Gif-sur-Yvette, France [2]New York University, New York, NY, USA [3]Institute of Telecommunications, TU Wien, Vienna, Austria [4]Systèmes et applications des technologies de l'information et de l'énergie (SATIE), CNRS Université Paris-Saclay, Gif-sur-Yvette, France [5]International Laboratory on Learning Systems (ILLS) and Quebec AI Institute (Mila), McGill ETS CNRS Université Paris-Saclay CentraleSupélec, Montreal (QC), Canada. Correspondence to: Andres Altieri <andres.altieri@centralesupelec.fr>, Marco Romanelli <mr6852@nyu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. We study a novel detection challenge focused on identifying inaccuracies in the predictions made by a regressor. Our definition of "incorrect" predictions references instances in which the regressor output may deviate beyond a predefined threshold from typical observations.

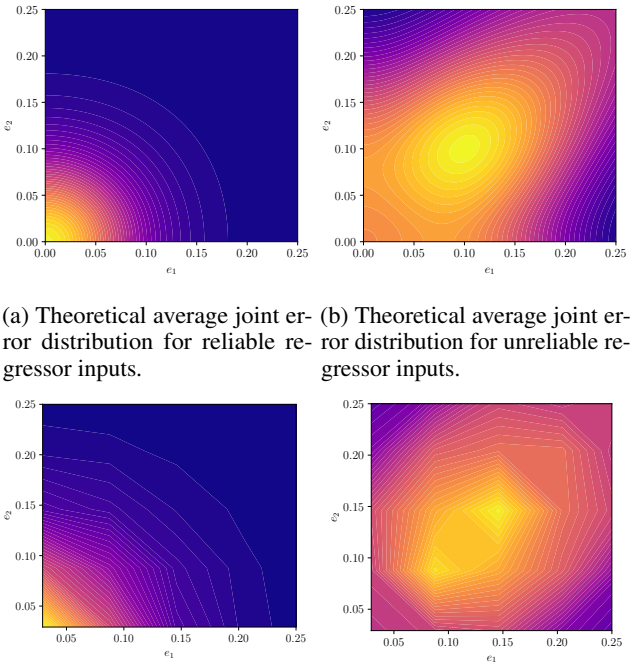
2. We formalize two fundamental base predictors that serve as benchmark models for the broader detection problem. These are constructed through approximation of the optimal detector and rely on estimates of the conditional distributions of the target and error variables.

3. We utilize an empirical measure of diversity inspired by (Rao, 1982) to develop two novel data-driven algorithms. Crucially, the advantage rooted in their inherent capability to compensate for inaccuracies in the underlying probability estimates, allows for them to achieve superior performance in detecting incorrect predictions of regressor models across various datasets and different error metrics.

1.2. Related Work

Over recent years, a multitude of methods have been introduced aiming to quantify the uncertainty in model predictions, each grounded in distinct underlying hypotheses, and generally producing an output conditional distribution or prediction interval. Noteworthy examples encompass techniques based on the assumptions of both parametric models, where pertinent parameters are learned from the data (Lakshminarayanan et al., 2017; Pichler et al., 2022; Qiu et al., 2020), Bayesian approaches (Hernandez-Lobato & Adams, 2015; Ritter et al., 2018; Sharma et al., 2021), non-parametric quantile function estimates (Chung et al., 2021; Fasiolo et al., 2021; Tagasovska & Lopez-Paz, 2019), regression with reject option (Zaoui et al., 2020), estimation of conditional variance (Fan & Yao, 1998), confidence intervals (Pearce et al., 2018), and conformal intervals (Papadopoulos et al., 2002; Romano et al., 2019). Additionally, a related challenge lies in gauging the accuracy of these predictions, involving concepts such as calibration and sharpness (Kuleshov et al., 2018; Zhao et al., 2020), among others.

The challenges of detecting anomalous behavior in regression algorithms and computing a conditional estimate for dependent variables are related, yet not equivalent. Constructing a detector does not necessarily demand a comprehensive and accurate statistical characterization of dependent variables. Instead, it requires an approach that excels in the specified detection task, capable of extracting relevant information even from suboptimal estimates. To draw an analogy, in classification problems, obtaining well-calibrated estimates of class probabilities (Guo et al., 2017) aids in detecting classification errors or out-of-distribution samples. However, the performance of such detectors can be enhanced through pre-processing techniques and constructing statistics that capture the pertinent information needed to detect specific conditions (Gomes et al., 2023; Granese et al., 2021; Liang et al., 2018).



(a) Theoretical average joint error distribution for reliable regressor inputs.

(b) Theoretical average joint error distribution for unreliable regressor inputs.

(c) Regions from (a) but estimated from data using SQR.

(d) Regions from (b) but estimated from data using SQR.

Figure 1. An optimal detector for regression errors effectively partitions the regressor input space into two sets: one where the regressor is considered reliable, and another where it is considered unreliable. Surprisingly, the joint error distributions in each set exhibit distinct (diversity) behaviors, enabling the identification of points that lead to effective detectors in real applications. The visual representation above depicts the theoretical distributions in both sets, illustrating clear differences between the two classes. The corresponding approximate distributions learned from the data, shown below, exhibit the same distinctive behavior. These plots correspond to the example in Section 2.

2. Main Definitions and Example

Let $\mathbf{X}$ be a random vector on $\mathcal{X} \subseteq \mathbb{R}^K$ with distribution $p_{\mathbf{X}}$, and let $\mathbf{Y}$ be a dependent random vector on $\mathcal{Y} \subseteq \mathbb{R}^M$ with conditional distribution $p_{\mathbf{Y}|\mathbf{X}}$. Consider a regressor $f_{\mathcal{D}_n}(\mathbf{x})$ that has been trained on a dataset $\mathcal{D}_n = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ drawn independently from the joint distribution $p_{\mathbf{XY}}$. Let us consider a scenario where an alert needs to be triggered whenever, for a given input $\mathbf{x}$, there is a significant risk that the output of the regressor exceeds an absolute error threshold ϵ. We may define a prediction error random variable $E \triangleq \|\mathbf{Y} - f_{\mathcal{D}_n}(\mathbf{X})\|$ and state the condition to be detected as $E > \epsilon$. The most powerful detector (cf. Proposition 4.2)

would report the regressor as unreliable at $\mathbf{x}$, when

$$\mathbb{P}\left(E > \epsilon | \mathbf{X} = \mathbf{x}\right) > \gamma, \tag{1}$$

where $\gamma \in [0,\ 1]$ is a constant design parameter which controls the sensitivity of the detector. The optimal detector splits the input space $\mathcal{X}$ into two sets: a set $\mathcal{X}_B$ that contains *bad* points, i.e.

$$\mathcal{X}_B \triangleq \{\mathbf{x} : \mathbb{P}\left(E > \epsilon | \mathbf{X} = \mathbf{x}\right) > \gamma\}, \tag{2}$$

where the decision is that the regressor is unreliable, and a complementary set of good points $\mathcal{X}_G = \mathcal{X}_B^{\mathsf{c}}$ where the decision is that the regressor is reliable.

Remark. Notice that this problem is not always well-posed: if for a given regressor the probability on the l.h.s. of (1) is independent of $\mathbf{x}$, then the optimal criterion becomes trivial. The performance that we can achieve will thus be defined by the dependence of this probability on $\mathbf{x}$.

In practice, the optimal detector cannot be implemented since the underlying conditional distribution is not known, so we propose a different approach. For a given input $\mathbf{x}$, we consider two independent realizations of the error variables E_1, E_2 drawn independently according to the conditional distribution $p_{E|\mathbf{X}=\mathbf{x}}$. The joint distribution of these two variables $p_{E_1,E_2|\mathbf{X}=\mathbf{x}}$ is the product distribution:

$$p_{E_1 E_2|\mathbf{X}=\mathbf{x}}(e_1, e_2) = p_{E|\mathbf{X}=\mathbf{x}}(e_1) p_{E|\mathbf{X}=\mathbf{x}}(e_2). \tag{3}$$

The distribution in (3) is a measure of the relation of the error event over two independent observations. Our detection approach is based on the hypothesis that the joint error distributions in (3) are different enough for the values of $\mathbf{x}$ in the set of bad points $\mathcal{X}_B$ and those in the set of good points. Therefore, by considering these distributions, we can effectively construct a reliable detector.

Example. To showcase the validity of this assumption, consider $X \sim \mathcal{N}(0, 1)$ and the simple additive noise model

$$Y = \phi(x) + W, \tag{4}$$

where $\phi(x)$ is a deterministic function and $p_{W|X=x}(w|x) = \mathcal{N}(w; 0, \sigma^2(x))$. The underlying distribution of the difference $Y - f_{\mathcal{D}_n}(x)$ given $X = x$ in this example is: $p_{Y-f_{\mathcal{D}_n}(x)|X=x}(e|x) = \mathcal{N}(e; b(x), \sigma^2(x))$, where $b(x) \triangleq \phi(x) - f_{\mathcal{D}_n}(x)$ is the bias error of the regressor. From these expressions, we can calculate the probability in (1) and obtain the optimal detector. Utilizing this information, we can determine the set of good and bad points, enabling us to construct the product distribution (3) for both classes.

In a numerical experiment, we draw 50 realizations of X, categorize them into good and bad points, and compute the respective empirical error distributions (3). We then average these distributions to represent the average behavior of each class. The results are depicted in Figs. 1(a) and (b) for the good and bad points, respectively. These plots illustrate that the product distributions exhibit distinct behaviors for values of x where the regressor would be deemed reliable or unreliable. Thus, these product functions can be employed to construct a discriminator. In practice, however, these distributions are unknown and should be estimated from data. In Figs. 1 (c) and (d), we repeat the experiment for the same samples of X, but we estimate the distributions using *Simultaneous Quantile Regression* (Tagasovska & Lopez-Paz, 2019). As observed, the estimated distributions align well with the theoretical ones, indicating that a discriminator can be effectively constructed from real data[1].

3. Formal Statement of the Problem

We now establish the formal framework, slightly generalizing the problem introduced in Section 2. We use the concept of a discrepancy or error function, which quantifies the "deviation" of a regressor output $f_{\mathcal{D}_n}(\mathbf{x})$ from the target $\mathbf{Y}$. Then, we shall formulate our detection problem.

3.1. Regression Error Framework

Definition 3.1 (Discrepancy function)**.** Let us consider a function $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ such that, for a pair $(\mathbf{x}, \mathbf{y})$, the value $d(\mathbf{y}, f_{\mathcal{D}_n}(\mathbf{x}))$ serves as an error measure of the difference between the target variable and the prediction. We call any such function a *discrepancy function*.

The choice of the discrepancy function depends on the regression task at hand, and could be different according to the application and its requirements. Given a suitable discrepancy function we may now introduce a notion of good or normal behavior for a regressor with respect to this metric.

Definition 3.2 (ϵ-goodness)**.** Given a discrepancy function d, an error threshold $\epsilon > 0$ and a data pair $(\mathbf{x}, \mathbf{y})$ we say that the trained regressor $f_{\mathcal{D}_n}$ is ϵ-good with respect to d for the data pair when $d(\mathbf{y}, f_{\mathcal{D}_n}(\mathbf{x})) \leq \epsilon$.

Based on this notion we define a binary random variable which indicates that the regressor is operating beyond the expected limits for a certain data pair:

$$A_\epsilon \triangleq \mathbb{1}\{d(\mathbf{Y}, f_{\mathcal{D}_n}(\mathbf{X})) \geq \epsilon\}. \tag{5}$$

This variable takes the value 0 for those data pairs for which the regressor is ϵ-good and 1 for those pairs that are ϵ-bad. We denote by $P_B(\mathbf{x}; \epsilon)$ the probability that the regressor is ϵ-bad given that $\mathbf{X} = \mathbf{x}$. This probability can be written as:

$$P_B(\mathbf{x}; \epsilon) = \mathbb{E}\left[A_\epsilon | \mathbf{X} = \mathbf{x}\right]. \tag{6}$$

[1]Additional details on the experiment and results based on real datasets can be found in Appendix A.

In a similar fashion we can compute the probability that the regressor is ϵ-good at $\mathbf{x}$, which we denote as $P_G(\mathbf{x};\epsilon)$ and can be computed as $P_G(\mathbf{x};\epsilon) = 1 - P_B(\mathbf{x};\epsilon)$.

Our goal is to detect inputs for which the regressor is operating beyond specification, that is, inputs for which there is a high risk that the regressor output be ϵ-bad. The goal is then to design a discriminator to detect this scenario.

4. Vanilla Predictor: Baseline

From a theoretical standpoint, if we are able to compute the exact probability of being ϵ-bad given by (6), we could construct the most powerful discriminator as follows.

4.1. Optimal (oracle) Discriminator

Definition 4.1 (Most powerful discriminator)**.** Consider a regressor $f_{\mathcal{D}_n}$, a discrepancy function d and an error threshold ϵ as in the previous section. Then, for a regressor input $\mathbf{x}$ and for any threshold $\gamma \in [0,\ 1]$ define the decision region:

$$\mathcal{R}_\epsilon(\gamma) \triangleq \{\mathbf{x} \in \mathcal{X} : P_B(\mathbf{x};\epsilon) > \gamma\}\,. \tag{7}$$

This definition allows us to obtain the Neyman-Pearson most powerful test for ϵ-goodness. The proof of the following proposition is presented in Appendix B.1.

Proposition 4.2. *The discriminator $\delta_p(\mathbf{x},\gamma,\epsilon)$, defined as $\delta_p(\mathbf{x},\gamma,\epsilon) = \mathbb{1}\{\mathbf{x} \in \mathcal{R}_\epsilon(\gamma)\}$ is the most powerful statistical test, testing ϵ-goodness against the alternative that $\mathbf{x}$ is ϵ-bad, at a certain false positive probability, which is parameterized by $\gamma \in [0,1]$.*

If the distribution of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ is known, the probability of being ϵ-bad can be computed as:

$$\begin{aligned} P_B(\mathbf{x};\epsilon) &= \mathbb{E}\left[\mathbb{1}\{d(\mathbf{y}, f_{\mathcal{D}_n}(\mathbf{x})) \geq \epsilon\}|\mathbf{X} = \mathbf{x}\right] \\ &= \int_{\mathcal{Y}} \mathbb{1}\{d(\mathbf{y}, f_{\mathcal{D}_n}(\mathbf{x})) \geq \epsilon\}\, dp_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}). \end{aligned} \tag{8}$$

Alternatively, we could define a scalar discrepancy random variable $D(\mathbf{Y},\mathbf{X}) \triangleq d(\mathbf{Y}, f_{\mathcal{D}_n}(\mathbf{X}))$, and compute the probability as follows:

$$\begin{aligned} P_B(\mathbf{x};\epsilon) &= \mathbb{E}\left[\mathbb{1}\{d(\mathbf{Y}, f_{\mathcal{D}_n}(\mathbf{x})) \geq \epsilon\}|\mathbf{X} = \mathbf{x}\right] \\ &= 1 - F_{D(\mathbf{Y},\mathbf{X})|\mathbf{X}=\mathbf{x}}(\epsilon), \end{aligned} \tag{9}$$

where F denotes the cumulative distribution function.

4.2. Vanilla Baseline Algorithms

The most powerful detector offers optimal performance, but cannot be implemented since the required conditional distributions are unknown in general. However, we can define approximate detectors by using the optimal decision rule with an estimate $\hat{P}_B(\mathbf{x};\epsilon)$ of $P_B(\mathbf{x};\epsilon)$, that is, implementing an approximate detector:

$$\hat{\delta}(\mathbf{x},\gamma,\epsilon) = \mathbb{1}\{\hat{P}_B(\mathbf{x};\epsilon) > \gamma\}. \tag{10}$$

Algorithm 1 Baseline based on the estimation of the conditional distribution of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$

Input: a trained regressor $f_{\mathcal{D}_n}$, trained on the dataset $\mathcal{D}_n = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$. A detection threshold $\gamma \in [0,\ 1]$.

Training: Reuse $\mathcal{D}_n$ to estimate the conditional distribution of $\mathbf{Y}|\mathbf{X}$.

After training: Given an input test sample $\mathbf{x}$:
Step 1: Estimate $\hat{P}_B(\mathbf{x};\epsilon)$ using (8) [involves the regressor $f_{\mathcal{D}_n}(\mathbf{x})$]
Step 2: Decide that the regressor will be ϵ-bad at $\mathbf{x}$ if $\hat{P}_B(\mathbf{x};\epsilon) > \gamma$.

Possible estimates of $P_B(\mathbf{x};\epsilon)$ can be obtained from either (8) or (9), by estimating the respective conditional distribution. These two approaches yield the baseline Algorithm 1, based on estimating the distribution of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$, and Algorithm 3, based on the estimation of the distribution of $D(\mathbf{Y},\mathbf{X})|\mathbf{X} = \mathbf{x}$. Due to the similarities between both algorithms the latter is presented in Appendix C.

It is noteworthy that theoretically, both (8) and (9) are equivalent, as the distribution of $D(\mathbf{Y},\mathbf{x})|\mathbf{X} = \mathbf{x}$ can be computed from the distribution of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$. However, in practice, the estimation problems differ. $D(\mathbf{Y},\mathbf{X})$ is a scalar non-negative random variable, while $\mathbf{Y}$ could be multidimensional and assume arbitrary values. In Algorithm 1, the regressor is not explicitly utilized in the training phase of the estimation of the conditional distribution. It is only employed after training to compute the probability $P_B(\mathbf{x};\epsilon)$ for a given test sample $\mathbf{x}$. Thus the training of the regressor and the estimator can be done in parallel. In Algorithm 3, the regressor is employed during training of the estimator of the conditional distribution, but is not used in the operation of the detector in test. In this case the training of the regressor and the estimator has to be done sequentially.

5. Error Diversity-Based Detectors for Unreliable Regression Behavior

A limitation of the baseline Algorithms 1 and 3 is that in practice the estimation of the conditional distributions may be difficult, leading to inaccuracies in the estimation of $P_B(\mathbf{x};\epsilon)$. To mitigate this issue we propose a new approach based on the notion of diversity coefficients (Rao, 1982).

Definition 5.1 (Diversity coefficients)**.** Consider a population of samples $\pi = \{V_1, V_2, \dots\}$ whose elements are drawn independently from a probability distribution p_V and

a non-negative symmetric function $h(v_1, v_2)$ which measures the difference between two elements in the population. The diversity coefficient $\mathbb{H}$ of the population is defined as the expected difference, as measured by h, of two elements of π drawn independently, that is (Rao, 1982):

$$\mathbb{H} \triangleq \mathbb{E}\left[h(V_1, V_2)\right], \tag{11}$$

where V_1 and V_2 are independent, $V_1 \sim p_V$ and $V_2 \sim p_V$. This coefficient gives a characterization of the variability within a population as measured by the chosen function h, i.e., the expected difference between two independently chosen random elements of the same population. To see this, consider, e.g., $h(u, v) = (u - v)^2$, which results in $\mathbb{H} = 2\text{var}(V)$.

For each input $\mathbf{x}$, we will consider the discrepancy random variable $D = d(\mathbf{Y}, f_{\mathcal{D}_n}(\mathbf{x}))$, where $\mathbf{Y}$ is distributed according to $p_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$, i.e., conditioned on $\mathbf{X} = \mathbf{x}$. For each $\mathbf{x}$, we define the diversity coefficient.

Definition 5.2 (Diversity metric for regressor)**.** Given a symmetric non-negative bounded function $h(y_1, y_2)$ and a trained regressor $f_{\mathcal{D}_n}(\cdot)$ we define the diversity of the discrepancy variables associated to $\mathbf{X} = \mathbf{x}$ as:

$$\mathbb{H}(\mathbf{x}) \triangleq \mathbb{E}\left[h(d(\mathbf{Y}_1, f_{\mathcal{D}_n}(\mathbf{x})), d(\mathbf{Y}_2, f_{\mathcal{D}_n}(\mathbf{x})))|\mathbf{X} = \mathbf{x}\right], \tag{12}$$

where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are drawn independently according to $p_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$. The diversity coefficient $\mathbb{H}(\mathbf{x})$ can be computed for any input value $\mathbf{x} \in \mathcal{X}$.

Given a diversity metric we could construct a discriminator which separates the points in $\mathcal{X}$ according to their diversity. That is, we can implement a discriminator as follows.

Definition 5.3 (Diversity discriminator)**.** Given a diversity metric $\mathbb{H}$ for a regressor, as in Definition 5.2, for a threshold $\gamma > 0$ we define a diversity discriminator (DV) as:

$$\delta_{\mathbb{H}}(\mathbf{x}, \gamma) = \mathbb{1}\{\mathbb{H}(\mathbf{x}) > \gamma\}. \tag{13}$$

Changing the function h that defines the diversity metric allows for the creation of different discriminators. Our objective is to implement a diversity discriminator that is sensitive to points where the regressor is ϵ-bad. In particular, we explore the possibility of implementing the optimal detector presented in Section 4.1 as a diversity discriminator. For this purpose, we present the following proposition. For the proof, refer to Appendix B.2.

Proposition 5.4. *The most powerful detector from Definition 4.1 can be implemented as a diversity discriminator, using the non-negative symmetric function*

$$h_p(u, v) \triangleq \mathbb{1}\{u > \epsilon\}\mathbb{1}\{v > \epsilon\}. \tag{14}$$

Algorithm 2 – DV-Y: Diversity discriminator based on the estimates of the distribution of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$.

Input: A regressor $f_{\mathcal{D}_n}$ trained on the dataset $\mathcal{D}_n$, dissimilarity threshold γ

Training:

Step 1: Use $\mathcal{D}_n$ to obtain an estimate $\hat{F}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}$ of the distribution of $\mathbf{Y}|\mathbf{X}$.

Step 2: separate ϵ-good and bad samples in $\mathcal{D}_n$:

$$\mathcal{G} \triangleq \{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_n : d(\mathbf{y}, f_{\mathcal{D}_n}(\mathbf{x})) \leq \epsilon\}$$
$$\mathcal{B} \triangleq \{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_n : d(\mathbf{y}, f_{\mathcal{D}_n}(\mathbf{x})) > \epsilon\}$$

Step 3: learn the function h that defined the diversity metric $\mathbb{H}$ for the discriminator:

for i in $1, .., N_{\text{epochs}}$ **do**

for each $(\mathbf{x}, y) \in \mathcal{D}_n$ **do**

Generate two extended datasets of N_u samples:

$$\mathcal{D}_1(\mathbf{x}) = \{\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_{N_u}\} \quad \mathcal{D}_2(\mathbf{x}) = \{\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{N_u}\}.$$

where each $\bar{y}_i, \tilde{y}_i$ is generated i.i.d. with $\hat{F}_{Y|\mathbf{X}=\mathbf{x}}$ from Step 1.

Estimate the diversity metric for $\mathbf{x}$:

$$\mathbb{H}(\mathbf{x}) = \frac{1}{N_u}\sum_{i=1}^{N_u} h(d(\bar{\mathbf{y}}_i, f_{\mathcal{D}_n}(\mathbf{x})), d(\tilde{\mathbf{y}}_i, f_{\mathcal{D}_n}(\mathbf{x}))),$$

end for

Update h to minimize:

$$\mathcal{L} = \frac{1}{2}\frac{1}{|\mathcal{G}|}\sum_{\mathbf{x}\in\mathcal{G}} \mathbb{H}(\mathbf{x}) - \frac{1}{2}\frac{1}{|\mathcal{B}|}\sum_{\mathbf{x}\in\mathcal{B}} \mathbb{H}(\mathbf{x})$$

such that h is symmetric and $h : \mathbb{R}^2_{\geq 0} \to [0, 1]$.

end for

After training: Given an input test sample $\mathbf{x}$:

Step 1: compute $\mathbb{H}(\mathbf{x})$ via a Monte Carlo experiment like in Step 3 of training.

Step 2: Decide that the regressor will be ϵ-bad at $\mathbf{x}$ if $\mathbb{H}(\mathbf{x}) > \gamma$.

This proposition indicates that if the true data distributions are known, it is possible to construct the optimal detector using a diversity detector. In real-world scenarios, the true distribution is not known, and thus, discriminators can only operate on estimates of the distributions. It can be verified that constructing a diversity detector based on these estimates, and using the function (14) of the optimal detector, will recover the baseline algorithms outlined in Section 4.2. However, since practical situations involve estimates of these distributions, it may be possible to find a different h function that improves the performance of the baselines by compensating for the estimation errors in the distributions.

We propose to use the data to obtain an empirical discrepancy metric as an approximate solution to an optimization problem which induces the desired behavior:

$$h^* \triangleq \underset{h:\mathbb{R}^2_{\geq 0}\to[0,1]}{\arg\min} \; \mathbb{E}_{\mathbf{X}}\left[\mathbb{H}(\mathbf{X})\mathbb{1}\{\mathbf{X}\in\mathcal{R}_\epsilon(\gamma)\}\right] - \mathbb{E}_{\mathbf{X}}\left[\mathbb{H}(\mathbf{X})\mathbb{1}\{\mathbf{X}\notin\mathcal{R}_\epsilon(\gamma)\}\right], \quad (15)$$

where $\mathbb{H}(\cdot)$ is given by (12) and $\mathcal{R}_\epsilon(\gamma)$ is the optimal decision region for the ideal detector (7). This optimization problem searches for a diversity metric $\mathbb{H}$ which is minimal on the good points and maximal on the bad points, as defined by the optimal detector. We adapt this optimization problem and derive two algorithms, Algorithms 2 (DV-Y) and 4 (DV-D). In the former, the expectation that defines $\mathbb{H}$ is approximated by using the estimates of the conditional distribution of $\mathbf{Y}|\mathbf{X}=\mathbf{x}$, and in the latter, we use estimates of the conditional distribution of the discrepancy variable $D(\mathbf{Y},\mathbf{X})$, also conditioned on $\mathbf{X}=\mathbf{x}$. Since the algorithms are similar in nature, DV-D is relegated to the Appendix C. The key aspects of the algorithms are as follows:

- The goal is to learn the symmetric bounded non-negative function h that is the argument of the expectation in $\mathbb{H}$ (see (12)) and defines the diversity metric.
- The expectation in $\mathbb{H}$ is estimated through a Monte Carlo experiment using the empirical distribution estimates as a generative model to sample the data.
- The two terms in the loss function in (15) are approximated by empirically separating the training set into the samples that happen to be good or bad.

Thus, the algorithms will utilize the diversity in the distribution estimates and the empirical information of good/bad points in the training dataset to define a diversity detector which is tailored to the underlying data estimates.

6. Experimental Results

We compare the diversity discriminators DV-Y (Algorithm 2) and DV-D (Algorithm 4) with the baseline Algorithms 1 and 3. To do this we require the estimation of the conditional distributions $\mathbf{Y}|\mathbf{X}=\mathbf{x}$ and $D(\mathbf{Y},\mathbf{X})|\mathbf{X}=\mathbf{x}$. To obtain these estimates, we consider several different distribution estimation methods, namely, *Conditional Gaussian* (CG) (Lakshminarayanan et al., 2017), *Simultaneous Quantile Regression* (SQR) (Tagasovska & Lopez-Paz, 2019), KNIFE (Pichler et al., 2022) and RIO (Qiu et al., 2020). CG fits a conditional Gaussian distribution whose mean and variance depend on the input $\mathbf{x}$. SQR estimates the conditional quantile function of the data using the *pinball loss* (Koenker, 2005; Koenker & Bassett, 1978) and KNIFE fits a conditional Gaussian mixture model to the data. Finally, RIO models the errors as a Gaussian processes. Each distribution estimation algorithm yields a different version of the baseline algorithms and diversity detectors. We denote as B1-* and B2-*, where * can be SQR, CG, KNIFE the baseline Algorithms 1 and 2 obtained with the corresponding distribution estimators. Since RIO models the errors directly, it is only suitable for implementing baseline B2. Likewise we denote as DV-Y-* and DV-D-*, where * can be SQR, CG, KNIFE the corresponding diversity discriminators implemented with the aforementioned distribution estimators. For fairness, the same estimates will be used for B1 and DV-Y, and for B2 and DV-D. Thus, in our extensive numerical evaluation, we consider six variations of the baselines and six variations of the algorithms based on the diversity metrics (code as available supplementary data).

6.1. Setup

We consider 8 well-known UCI (Kelly et al.) regression datasets that have been extensively used in uncertainty quantification (Chung et al., 2021; Gal & Ghahramani, 2016; Hernandez-Lobato & Adams, 2015; Tagasovska & Lopez-Paz, 2019). For each dataset, we consider two discrepancies, namely, the absolute error: $d_a(Y, f_{\mathcal{D}_n}(\mathbf{x})) = |y - f_{\mathcal{D}_n}(\mathbf{x})|$; and the relative error: $d_r(y, f_{\mathcal{D}_n}(\mathbf{x})) = \frac{|y-f_{\mathcal{D}_n}(\mathbf{x}))|}{|f_{\mathcal{D}_n}(\mathbf{x})|}$, measuring relative discrepancy to the amplitude of predictions.

For each dataset and discrepancy function, we explore three values of ϵ and conduct our experiments across 10 different seeds. When considering the relative error, the chosen ϵ values signify a relative magnitude, independent of the data range. However, for the absolute error, suitable ϵ values may depend on the dataset and could have significantly different ranges. To mitigate this, we report the value of ϵ for the absolute error normalized by the empirical standard deviation of the training set. This normalization ensures resulting values of normalized ϵ are less than 1. It is important to mention that the representative values of ϵ, for the relative and absolute metrics, depend on each specific datasets. It is not possible to use the same values of ϵ for all algorithms because each regressor achieves different levels of approximation to the data. It is also important to mention that for each value of ϵ we obtain a fundamentally different detection problem because the rejection region of the optimal detector changes completely. The best performance achievable for a value of ϵ depends on the statistical dispersion of the data around the regressor, and on the train/test splits that are drawn from each seed. For this reason the performance within the same value of ϵ and the same dataset may change from seed to seed, especially for the smaller datasets.

For each dataset and seed, we initially train a regressor using a consistent neural architecture featuring three hidden layers and 64 neurons. This involves employing 5-fold cross-validation on the learning rate and weight decay, with

Table 1. AUROC (mean and standard deviation, normalized to 100) for the absolute error discrepancy metric. B1, B2 correspond to baseline Algorithms 1 and 3 resp. DV-Y and DV-D correspond to the proposed Algorithms 2 and 4 resp. For each dataset we report the baselines that are the best for at least one ϵ, and the same for the proposed algorithms. The full tables are in Appendix G.

naval	ϵ	**0.175**	**0.2**	**0.225**
	Avg. % bad	2.6 [0.9]	1.4 [0.7]	0.9 [0.4]
	B2-SQR	93.9 [3.8]	93.9 [4.2]	88.8 [7.2]
	B2-KNIFE	91.6 [3.5]	93.4 [3.6]	95.2 [2.7]
	DV-D-SQR	**96.7** [1.3]	**97.3** [1.8]	**96.2** [3.6]
power	ϵ	**0.3**	**0.35**	**0.4**
	Avg. % bad	14.1 [1.0]	9.2 [1.0]	6.4 [0.7]
	B1-SQR	70.3 [2.5]	71.4 [2.5]	71.1 [1.6]
	B2-SQR	70.4 [2.6]	70.9 [2.7]	71.1 [3.8]
	DV-Y-SQR	70.8 [2.7]	**72.2** [2.5]	**72.8** [3.3]
	DV-D-SQR	**71.5** [2.5]	**72.2** [2.7]	72.0 [3.4]
kin8nm	ϵ	**0.6**	**0.7**	**0.8**
	Avg. % bad	4.2 [0.6]	2.2 [0.5]	1.1 [0.3]
	B1-CG	**78.6** [2.7]	**81.9** [3.7]	**86.9** [5.4]
	DV-Y-CG	**78.6** [2.7]	81.7 [3.8]	86.7 [5.4]
wine	ϵ	**0.1**	**0.3**	**0.5**
	Avg. % bad	87.4 [4.5]	65.2 [6.1]	46.8 [5.0]
	B1-SQR	91.9 [5.1]	84.2 [5.2]	71.3 [3.6]
	B1-KNIFE	94.5 [2.5]	86.0 [4.7]	70.3 [4.9]
	DV-Y-SQR	94.3 [4.3]	86.1 [4.3]	**72.2** [3.6]
	DV-Y-KNIFE	**95.2** [2.1]	**87.3** [3.8]	**72.2** [2.9]
concrete	ϵ	**0.2**	**0.5**	**0.6**
	Avg. % bad	36.7 [5.6]	9.6 [2.8]	6.4 [2.0]
	B1-SQR	73.1 [3.8]	81.1 [8.3]	76.3 [14.2]
	B2-CG	69.8 [7.1]	78.7 [9.6]	81.5 [9.1]
	DV-Y-SQR	**74.4** [3.4]	**83.5** [8.3]	**82.1** [8.4]
energy	ϵ	**0.025**	**0.0625**	**0.1**
	Avg. % bad	61.2 [5.0]	25.8 [5.2]	9.5 [3.4]
	B1-CG	72.5 [6.8]	81.2 [5.5]	81.7 [5.4]
	DV-Y-CG	**75.1** [6.0]	**82.0** [6.9]	**82.5** [7.7]
boston	ϵ	**0.25**	**0.3**	**0.35**
	Avg. % bad	31.8 [2.7]	24.1 [5.8]	21.2 [5.4]
	B1-SQR	72.8 [7.2]	77.1 [6.0]	76.8 [5.6]
	DV-Y-SQR	**73.7** [7.7]	**77.6** [6.7]	**78.6** [5.4]
yacht	ϵ	**0.025**	**0.05**	**0.075**
	Avg. % bad	62.6 [9.6]	35.8 [11.3]	25.5 [10.6]
	B1-CG	92.4 [4.4]	94.1 [5.7]	96.3 [3.0]
	DV-Y-CG	**94.3** [3.3]	**94.4** [5.7]	**96.6** [2.9]

Table 2. FPR at TPR 90% (mean and standard deviation, lower is better) for the absolute error discrepancy metric. B1, B2 correspond to baseline Algorithms 1 and 3 resp. DV-Y and DV-D correspond to our Algorithms 2 and 4 resp. For each dataset we report the baselines that are the best for at least one ϵ, and the same for the proposed algorithms. The full tables are in Appendix G.

naval	ϵ	**0.175**	**0.2**	**0.225**
	B2-KNIFE	0.22 [0.11]	0.19 [0.13]	**0.12** [0.06]
	B2-CG	0.20 [0.04]	0.19 [0.05]	0.15 [0.07]
	DV-D-SQR	**0.11** [0.05]	**0.09** [0.05]	0.15 [0.18]
	DV-D-KNIFE	0.22 [0.08]	0.19 [0.10]	**0.12** [0.10]
power	ϵ	**0.3**	**0.35**	**0.4**
	B1-SQR	0.69 [0.06]	0.69 [0.08]	0.74 [0.11]
	DV-Y-SQR	**0.68** [0.05]	**0.67** [0.07]	**0.67** [0.09]
	DV-D-SQR	**0.68** [0.05]	0.71 [0.06]	0.75 [0.07]
kin8nm	ϵ	**0.6**	**0.7**	**0.8**
	B1-CG	**0.48** [0.09]	**0.44** [0.12]	**0.33** [0.20]
	DV-Y-CG	**0.48** [0.10]	**0.44** [0.12]	**0.33** [0.21]
wine	ϵ	**0.1**	**0.3**	**0.5**
	B1-KNIFE	0.14 [0.14]	0.49 [0.14]	**0.73** [0.08]
	DV-Y-SQR	**0.11** [0.13]	0.51 [0.15]	**0.72** [0.06]
	DV-Y-KNIFE	0.13 [0.12]	**0.46** [0.15]	**0.72** [0.07]
concrete	ϵ	**0.2**	**0.5**	**0.6**
	B1-SQR	**0.67** [0.10]	0.77 [0.39]	0.90 [0.30]
	B2-KNIFE	0.73 [0.13]	0.54 [0.18]	**0.45** [0.27]
	B1-CG	0.69 [0.10]	**0.44** [0.13]	0.49 [0.18]
	DV-Y-SQR	**0.67** [0.10]	**0.43** [0.21]	0.51 [0.22]
	DV-D-CG	**0.67** [0.12]	0.51 [0.22]	0.49 [0.23]
energy	ϵ	**0.025**	**0.0625**	**0.1**
	B1-CG	0.61 [0.11]	**0.42** [0.14]	**0.39** [0.14]
	DV-Y-CG	**0.59** [0.10]	**0.42** [0.13]	**0.39** [0.18]
boston	ϵ	**0.25**	**0.3**	**0.35**
	B1-SQR	0.69 [0.17]	0.60 [0.17]	0.64 [0.25]
	B1-KNIFE	0.70 [0.14]	0.64 [0.16]	**0.57** [0.12]
	DV-Y-SQR	**0.64** [0.21]	**0.53** [0.20]	**0.56** [0.24]
yacht	ϵ	**0.025**	**0.05**	**0.075**
	B1-CG	**0.16** [0.12]	**0.11** [0.12]	0.16 [0.30]
	DV-Y-CG	**0.16** [0.10]	**0.11** [0.12]	**0.10** [0.07]

90% of the data serving as the training set and the remaining 10% as a test set for both the regressors and the detectors. Subsequently, the training set of the regressor is reused to train estimators for the conditional distributions of $Y|\mathbf{X} = \mathbf{x}$ and $D|\mathbf{X} = \mathbf{x}$. For the mean and variance of CG, we utilize a neural network with a single hidden layer comprising 64 neurons and average over 10 models. The learning rate is optimized through 5-fold cross-validation, and the training continues for 150 epochs. Adhering to the setup in (Tagasovska & Lopez-Paz, 2019), SQR is trained using a neural network with 3 hidden layers and 64 neurons. The learning rate and weight decay are optimized through 5-fold cross-validation, and the training proceeds for 1000 epochs. For RIO we use the default setup in the reference code provided. Finally, for KNIFE, we employ 256 modes and train for 1000 epochs, cross-validating the learning rate and weight decay. All algorithms except RIO utilize the Adam optimizer (Kingma & Ba, 2015) and batch normalization (Ioffe & Szegedy, 2015). Additional details can be found in Appendix D. With the conditional estimators we implement the baselines following the procedure outlined in Appendix D.2. For the diversity detectors DV-Y and DV-D we use a neural network with four hidden layers of 64 neurons. We train for 25 epochs and choose the optimal learning rate using only a validation set of 20% of the data. The best parameters are chosen to maximize the AUROC of the validation set. The symmetry in the output is achieved by averaging two replicas of the output of the network with exchanged inputs.

Considered metrics We measure the performance of the detectors using the mean and standard deviation over 10

Table 3. AUROC (mean and standard deviation, normalized to 100) for the relative error discrepancy metric. B1, B2 correspond to baseline Algorithms 1 and 3 resp. DV-Y and DV-D correspond to the proposed Algorithms 2 and 4 resp. For each dataset we report the baselines that are the best for at least one ϵ, and the same for the proposed algorithms. The full tables are in Appendix G.

	ϵ [%]	**0.15%**	**0.175%**	**0.2%**
naval	**Avg. % bad**	1.5 [0.6]	0.8 [0.4]	0.5 [0.3]
	B2-SQR	94.8 [4.1]	87.6 [7.2]	90.2 [9.7]
	B2-KNIFE	92.4 [3.8]	95.1 [3.2]	96.7 [2.7]
	DV-D-SQR	**97.3** [1.8]	**97.0** [2.3]	**98.0** [1.8]
	ϵ [%]	**1%**	**1.25%**	**1.5%**
power	**Avg. % bad**	18.2 [0.8]	10.9 [0.9]	6.6 [0.8]
	B2-SQR	71.4 [2.9]	72.3 [2.9]	71.6 [2.6]
	DV-Y-SQR	70.4 [2.3]	72.5 [2.9]	**72.6** [3.2]
	DV-D-SQR	**71.6** [2.7]	**72.8** [2.5]	72.3 [2.3]
	ϵ [%]	**10%**	**20%**	**30%**
kin8nm	**Avg. % bad**	34.4 [1.3]	12.2 [1.2]	5.2 [0.6]
	B1-CG	76.8 [1.8]	**85.8** [1.7]	**91.2** [1.9]
	DV-Y-CG	**77.0** [1.9]	**85.8** [1.8]	**91.2** [2.0]
	ϵ [%]	**5%**	**10%**	**15%**
wine	**Avg. % bad**	62.6 [4.8]	32.8 [3.6]	17.0 [2.1]
	B1-SQR	82.1 [5.2]	**61.8** [3.8]	65.1 [6.2]
	B1-KNIFE	82.7 [5.9]	60.6 [6.3]	69.9 [6.0]
	DV-Y-SQR	84.2 [4.8]	59.6 [3.4]	64.4 [8.4]
	DV-Y-KNIFE	**84.4** [5.0]	59.3 [8.5]	**70.2** [7.0]
	ϵ [%]	**10%**	**15%**	**20%**
concrete	**Avg. % bad**	35.4 [6.2]	22.1 [5.1]	13.9 [3.4]
	B1-SQR	71.3 [5.3]	74.8 [5.1]	74.6 [6.3]
	DV-Y-SQR	**72.2** [5.4]	**75.3** [5.0]	**77.4** [4.6]
	ϵ [%]	**2%**	**3.5%**	**5%**
energy	**Avg. % bad**	46.8 [3.7]	21.0 [7.0]	10.5 [5.2]
	B1-SQR	71.1 [8.2]	80.5 [6.4]	91.3 [5.9]
	B1-CG	74.9 [3.9]	78.3 [6.8]	81.9 [12.1]
	DV-Y-SQR	73.1 [7.9]	81.0 [5.9]	**91.5** [5.7]
	DV-Y-KNIFE	73.5 [6.9]	**81.9** [5.2]	87.4 [8.4]
	DV-Y-CG	**76.3** [3.2]	80.1 [6.6]	89.1 [7.9]
	ϵ [%]	**10%**	**15%**	**20%**
boston	**Avg. % bad**	33.7 [5.2]	20.8 [5.0]	14.5 [6.3]
	B1-SQR	72.2 [8.0]	78.3 [7.5]	81.9 [11.9]
	DV-Y-SQR	**74.0** [7.4]	**79.5** [8.4]	**85.6** [8.2]
	ϵ [%]	**10%**	**15%**	**20%**
yacht	**Avg. % bad**	59.4 [9.0]	46.1 [9.3]	39.0 [8.9]
	B1-CG	**92.7** [4.4]	**95.7** [3.0]	**97.2** [2.1]
	DV-Y-CG	92.1 [5.0]	94.6 [3.5]	96.2 [2.9]

seeds of the metrics: **False Positive Rate** (FPR, fraction of samples classified as ϵ-bad that are ϵ-good) at 90% of **True Positive Rate** (TPR, fraction of samples classified as ϵ-bad but are ϵ-bad), labeled as FPR at TPR 90% (lower is better); **Area under the Receiver Operating Characteristic curve (AUROC)** which relates the FPR to the TPR (perfect detector = 100%) (Bewick et al., 2004).

Connection to Conformal Estimates In general, conformal approaches cannot be directly applied to solve the issue we address. Specifically, the guarantees provided by conformal predictions, in terms of marginal coverage, do not offer any insight into the conditional probability required for constructing a detector. Nevertheless, to provide more insight, we discuss the main differences and propose a detection approach based on conformalized quantile regression (Romano et al., 2019). This analysis is relegated to Appendix F due to space limitations.

Table 4. FPR at TPR 90% (mean and standard deviation, lower is better) for the relative error discrepancy metric. B1, B2 correspond to baseline Algorithms 1 and 3 resp. DV-Y and DV-D correspond to the proposed Algorithms 2 and 4 resp. For each dataset we report the baselines that are the best for at least one ϵ, and the same for the proposed algorithms. The full tables are in Appendix G.

	ϵ [%]	**0.15%**	**0.175%**	**0.2%**
naval	B2-KNIFE	0.22 [0.13]	0.12 [0.07]	**0.09** [0.08]
	B2-CG	0.16 [0.03]	0.14 [0.07]	0.11 [0.09]
	DV-D-SQR	**0.09** [0.08]	**0.08** [0.07]	**0.08** [0.09]
	ϵ [%]	**1%**	**1.25%**	**1.5%**
power	B1-SQR	**0.68** [0.06]	**0.67**[0.07]	0.77 [0.13]
	B1-CG	0.82 [0.03]	0.79 [0.05]	0.76 [0.06]
	DV-Y-SQR	**0.67** [0.05]	**0.66** [0.05]	**0.67** [0.10]
	ϵ [%]	**10%**	**20%**	**30%**
kin8nm	B1-CG	**0.57** [0.04]	**0.35** [0.04]	**0.19** [0.05]
	DV-Y-CG	**0.57** [0.04]	**0.35** [0.04]	**0.19** [0.05]
	ϵ [%]	**5%**	**10%**	**15%**
wine	B1-KNIFE	0.59 [0.13]	0.79 [0.08]	**0.68** [0.13]
	DV-Y-KNIFE	**0.55** [0.15]	**0.76** [0.11]	**0.68** [0.15]
	ϵ [%]	**10%**	**15%**	**20%**
concrete	B1-SQR	**0.63** [0.11]	0.62 [0.21]	0.76 [0.32]
	B2-CG	0.73 [0.09]	0.64 [0.17]	0.62 [0.23]
	DV-Y-SQR	**0.63** [0.11]	**0.56** [0.09]	**0.47** [0.14]
	ϵ [%]	**2%**	**3.5%**	**5%**
energy	B1-SQR	0.76 [0.14]	0.64 [0.20]	0.35 [0.31]
	B1-CG	**0.62** [0.09]	**0.53** [0.19]	0.40 [0.27]
	DV-Y-CG	**0.62** [0.11]	0.54 [0.18]	**0.29** [0.25]
	ϵ [%]	**10%**	**15%**	**20%**
boston	B1-KNIFE	0.65 [0.11]	0.61 [0.20]	0.46 [0.30]
	B1-CG	0.65 [0.22]	0.56 [0.28]	0.48 [0.33]
	DV-Y-SQR	**0.63** [0.18]	**0.52** [0.18]	**0.34** [0.15]
	ϵ [%]	**10%**	**15%**	**20%**
yacht	B1-CG	**0.25** [0.21]	**0.17** [0.17]	**0.10** [0.12]
	DV-Y-SQR	0.38 [0.25]	**0.17** [0.10]	0.16 [0.10]
	DV-Y-CG	**0.26** [0.21]	0.22 [0.21]	**0.10** [0.11]

6.2. Discussion

In Tables 1 and 3, we present the AUROC (mean and standard deviation, normalized to 100) for the absolute and relative dispersion metrics, respectively. Tables 2 and 4 showcase the FPR at 90% (mean and standard deviation, where lower values are better). Due to the large number of considered algorithms (6 baselines and 6 competing algorithms), we highlight the baselines that perform best for at least one value of ϵ and do the same for the proposed algorithms. This way, for each ϵ, we compare the best baseline with the best among the proposed algorithms. The complete tables are available in Appendix G. In the tables with the AUROC values (Tables 1 and 3) we also include the mean and standard deviation of the percentage of samples

in the test set that are ϵ-bad for each value of ϵ as reference. Overall, it is observed that, in most cases, using the **DV algorithms attains better performance in terms of AUROC**, specifically up to 3.4/100 and 3.7/100 for the absolute and relative error metrics, respectively, compared with the best baselines. On the other hand, DV algorithms perform, at worst, comparable or only marginally worse than competing algorithms. For some datasets, the **DV algorithms achieve a reduction of up to 50% in FPR at TPR 90%**, while in the worst case the performance is comparable.

The ultimate measure of a detector's performance is determined by the conditional distribution of $\mathbf{Y}|\mathbf{X}$. Particularly for small datasets, the fairness in data splits plays a crucial role. Therefore, in Tables 1-4 we present our results by ablating across different values of the tolerated error epsilon. This approach mirrors the actions of a practitioner who must adjust the tolerance level to suit a specific task and dataset. The key observation from the tables provided is that the proposed method (DVIC+SQR...) consistently outperforms the considered baseline (SQR, ...) across various metrics.

The algorithms computed using the distribution of $Y|\mathbf{x}$ (B1-* and DV-Y-*) tend to outperform their counterparts using the distribution of $D|\mathbf{x}$, possibly due to how the regressor is used during training in the second case. However, there are cases where the algorithms based on the estimates of $D|\mathbf{X} = \mathbf{x}$ perform better. A detailed examination of the complete tables in Appendix G reveals significant variations in baseline performance across different methods for each dataset, likely due to the alignment of underlying assumptions of each algorithm with respect to the inherent distributions in the data. The parameter γ plays a pivotal role, serving as a parameter that governs the balance between TPR and FPR, i.e., the trade-off between accurately identifying positives and minimizing false alarms. Thus, the choice of this parameter should be tailored to the requirements of each specific application. To select γ, we recommend utilizing a calibration set, consisting of representative data to guide this selection process. A similar rationale applies to the choice of ϵ, which denotes the tolerance for error. Generally, the selection of ϵ depends on the task at hand, and practitioners should choose it based on the level of error they deem acceptable for their specific use-case.

The proposed framework is inherently versatile, applicable to any type of regressor. This flexibility is one of its primary advantages. As demonstrated in our experimental results, its overall performance is generally comparable to, and often significantly better than, the baseline. However, the extent of performance improvement depends on the specific data distribution inherent in the task under consideration. A distinct advantage of our detector is its data-driven approach, which accounts for the data distribution.

Bayesian analysis and related methods, such as the mixture of experts' paradigms, provide uncertainty estimates that can inform our proposed solution. Utilizing these estimates, like error distribution estimates, is of great interest and warrants future research. We plan to investigate the performance of Bayesian methods as a baseline and integrate them with our regression error detection method. Specifically, we aim to explore how the uncertainty estimates generated by Bayesian techniques can complement our approach, potentially enhancing the accuracy and robustness of our detection framework.

Finally, we add an analysis concerning the computational time requirements of the algorithms, and the specific of the hardware we utilized in the experiments in Appendix E.

7. Summary and Concluding Remarks

We proposed a novel framework to analyze and detect potential anomalies in the operation of regression algorithms. We introduced the notion of unreliability in regression, i.e., when the output of the regressor exceeds a specified discrepancy. We estimate the discrepancy density and measure its statistical diversity using our proposed data-driven score. Our framework demonstrates its effectiveness in compensating for estimation inaccuracies of the density of a target variable given the input, showcasing its potential for improving the reliability of the regressor's predictions in diverse applications, across various datasets and error metrics, which emphasizes its robustness and versatility. Many existing rejection-based algorithms require simultaneous training with the regressor, which differs from our framework. As a future direction, it would be interesting to explore different setups and evaluation frameworks to compare them with our proposed solution. Furthermore, extensions of our framework to other types of data (e.g. medical images, biometric, and multidimensional data), as well as other types of networks is a promising avenue for future research.

Acknowledgements

This work has been supported by the project PSPC AIDA: 2019-PSPC-09 funded by BPI-France. This work was granted access to the HPC/AI resources of IDRIS under the allocation 2022 - AD011012803R1 made by GENCI.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., and Mané, D. Concrete problems in AI safety. *CoRR*, 2016.

Antun, V., Renna, F., Poon, C., Adcock, B., and Hansen, A. C. On instabilities of deep learning in image reconstruction and the potential costs of ai. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020. doi: 10.1073/pnas.1907377117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1907377117.

Bewick, V., Cheek, L., and Ball, J. Statistics review 13: receiver operating characteristic curves. *Crit Care.*, 2004. doi: 10.1186/cc3000.

Chung, Y., Neiswanger, W., Char, I., and Schneider, J. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. In *Advances in Neural Information Processing Systems*, 2021.

Corbière, C., Thome, N., Bar-Hen, A., Cord, M., and Pérez, P. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, 2019.

Dadalto, E., Alberge, F., Duhamel, P., and Piantanida, P. Igeood: An information geometry approach to out-of-distribution detection. In *International Conference on Learning Representations*, 2022.

Fan, J. and Yao, Q. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 1998.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., and Goude, Y. Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, 2021.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

Gomes, E. D. C., Romanelli, M., Pichler, G., and Piantanida, P. A data-driven measure of relative uncertainty for misclassification detection. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Granese, F., Romanelli, M., Gorla, D., Palamidessi, C., and Piantanida, P. DOCTOR: A simple method for detecting misclassification errors. In *Advances in Neural Information Processing Systems*, 2021.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Hernandez-Lobato, J. M. and Adams, R. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Huang, L., Zhang, C., and Zhang, H. Self-adaptive training: beyond empirical risk minimization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ioffe15.html.

Kelly, M., Longjohn, R., and Nottingham, K. Uci machine learning repository. URL https://archive.ics.uci.edu.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

Kiureghian, A. D. and Ditlevsen, O. Aleatory or epistemic? does it matter? *Structural Safety*, 2009.

Koenker, R. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005.

Koenker, R. W. and Bassett, Gilbert, J. Regression Quantiles. *Econometrica*, 1978.

Kotelevskii, N., Artemenkov, A., Fedyanin, K., Noskov, F., Fishkov, A., Shelmanov, A., Vazhentsev, A., Petiushko, A., and Panov, M. Nonparametric uncertainty quantification for single deterministic neural network. In *Advances in Neural Information Processing Systems*, 2022.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, 2017.

Lehmann, E. L., Romano, J. P., and Casella, G. *Testing statistical hypotheses*, volume 3. Springer, 2005.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., and Zhou, B. Trustworthy ai: From principles to practices. *ACM Comput. Surv.*, 2023.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Liu, J. Z., Lin, Z., Padhy, S., Tran, D., Bedrax-Weiss, T., and Lakshminarayanan, B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems*, 2020.

Ming, Y., Sun, Y., Dia, O., and Li, Y. How to exploit hyperspherical embeddings for out-of-distribution detection? In *The Eleventh International Conference on Learning Representations*, 2023.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In Elomaa, T., Mannila, H., and Toivonen, H. (eds.), *Machine Learning: ECML 2002*, pp. 345–356, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-36755-0.

Pearce, T., Brintrup, A., Zaki, M., and Neely, A. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Pichler, G., Colombo, P. J. A., Boudiaf, M., Koliander, G., and Piantanida, P. A differential entropy estimator for training neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Picot, M., Messina, F., Boudiaf, M., Labeau, F., Ayed, I. B., and Piantanida, P. Adversarial robustness via fisher-rao regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2698–2710, 2023. doi: 10.1109/TPAMI.2022.3174724.

Qiu, X., Meyerson, E., and Miikkulainen, R. Quantifying point-prediction uncertainty in neural networks via residual estimation with an I/O kernel. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rkxNh1Stvr.

Rao, C. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 1982.

Ritter, H., Botev, A., and Barber, D. A scalable laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skdvd2xAZ.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, 2019.

Sharma, A., Azizan, N., and Pavone, M. Sketching curvature for efficient out-of-distribution detection for deep neural networks. In *Proc. Conf. on Uncertainty in Artificial Intelligence*, 2021.

Tagasovska, N. and Lopez-Paz, D. Single-model uncertainties for deep learning. In *Advances in Neural Information Processing Systems*, 2019.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

Zaoui, A., Denis, C., and Hebiri, M. Regression with reject option and application to knn. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Zhao, S., Ma, T., and Ermon, S. Individual calibration with randomized forecasting. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

A. Analysis of the Example and Error Distributions of Real Datasets

In our example we consider a regressor whose difference random variable $Y - f_{\mathcal{D}_n}(x)|X = x$ is distributed as $\mathcal{N}(b(x), \sigma(x))$. Thus, defining the error variable $E = |Y - f_{\mathcal{D}_n}(x)|$ we may compute the probability that the error exceed the predefined threshold as:

$$\mathbb{P}\left(E > \epsilon | X = x\right) = 1 - \left[\Phi\left(\frac{\epsilon - b(x)}{\sigma(x)}\right) - \Phi\left(\frac{-\epsilon - b(x)}{\sigma(x)}\right)\right], \tag{16}$$

where Φ is the cumulative distribution function of a standard normal random variable. From there we can compute the density of the error distribution from the difference variable $Y - f_{\mathcal{D}_n}(x)$ as:

$$p_{E|X}(e|x) = p_{Y - f_{\mathcal{D}_n}(x)}(e) + p_{Y - f_{\mathcal{D}_n}(x)}(-e), \tag{17}$$

where, as mentioned before, $Y - f_{\mathcal{D}_n}(x)|X = x \sim \mathcal{N}(b(x), \sigma(x))$. For each possible $x \in \mathcal{X}$, comparing the probability in (16) with the detection threshold γ, we can identify the critical region of points where the regressor is deemed unreliable by the optimal detector.

For the specific example in Fig. 1, we have taken $\sigma(x) = 0.05(1 + (x + 0.2)^2)$ and $b(x) = 0.1(x - 0.2)^3$ and chosen $\gamma = 0.4$. With this we can compute the rejection region which consists on the input points x where the regressor is deemed unreliable (see Fig. 2). To compute the plots in Figs. 1(a) and (b), we randomly choose 50 points for $X \sim \mathcal{N}(0, 1)$, classify

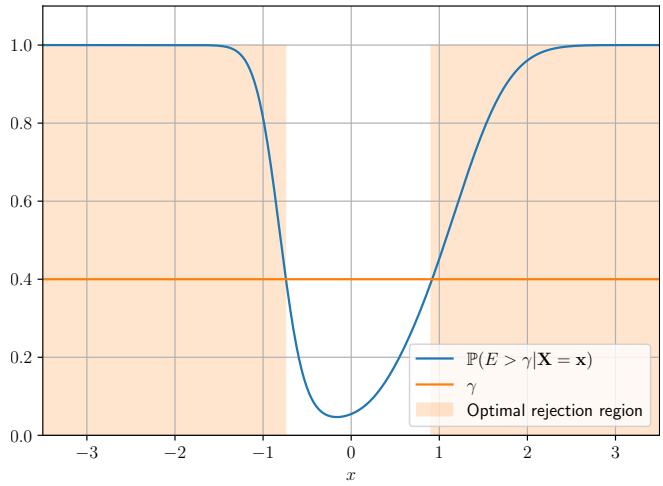


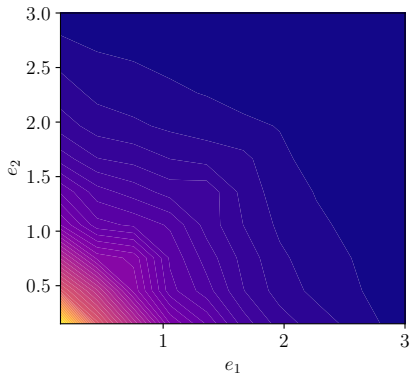
Figure 2. Optimal rejection region for the example in Section 2.

them as good and bad and compute the theoretical density (17). We then compute the error product density (3) for each point in each class and average them within the class.

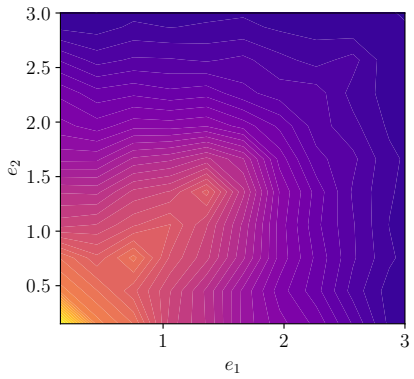
To find the estimates in Figs. 1 (c) and (d), we generate a dataset $\{(x_i, e_i)\}_{i=1}^{1000}$ of input data and error samples. We then train a simple neural network architecture to learn the conditional quantile function of $E|X = x$ from this dataset. This network is used as a generative model to obtain an histogram of the conditional error distribution for the 50 input points used to generate Fig. 1(a) and (b). We then estimate the error product distributions (3) and compute the average of the estimates of the joint error distributions.

We can follow a similar approach with real datasets, although establishing a ground truth is not possible in this scenario. In Fig. 3, we employ the regressors and distribution estimation algorithms (specifically, SQR, CG, KNIFE) trained in Section 6 to estimate the conditional error distribution, focusing on the absolute error metric $|Y - f_{\mathcal{D}_n}(\mathbf{x})|$. The distribution estimators are trained on the training set of the regressor, and subsequently, the error product distributions (3) are estimated on its test set. As observed in the simple example of Section 2, the classes of good and bad points exhibit distinct behaviors in terms of the error product distributions.

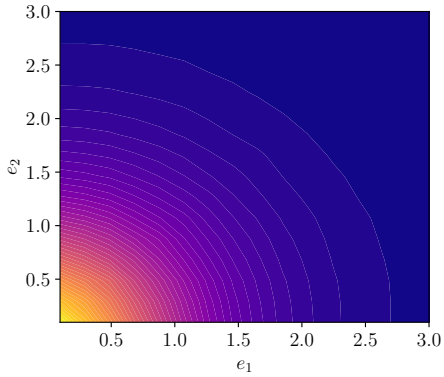
In Fig. 4, a similar procedure is applied to two datasets, namely *energy* and *Boston*, for the relative error metric $|Y - f_{\mathcal{D}_n}(\mathbf{x})|/|f_{\mathcal{D}_n}(\mathbf{x})|$ described in Section 6. Once again, different patterns for each class can be observed in this case.



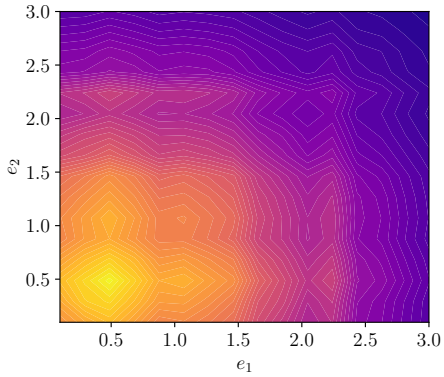
(a) Naval dataset. Average joint error distribution for points that are ϵ-good in the test set. $\epsilon = 0.1$. Estimated using KNIFE.



(b) Naval dataset. Average joint error distribution for points that are ϵ-bad in the test set. $\epsilon = 0.1$. Estimated using KNIFE.



(c) Kin8nm dataset. Average joint error distribution for points that are ϵ-good in the test set. $\epsilon = 0.8$. Estimated using CG.



(d) Kin8nm dataset. Average joint error distribution for points that are ϵ-bad in the test set. $\epsilon = 0.8$. Estimated using CG.

Figure 3. Examples of average joint error distributions given by (3) for the absolute error discrepancy function $|Y - f_{\mathcal{D}_n}(x)|$ using the regressors trained in Section 6. The patterns are obtained from the corresponding test set and the distributions of the discrepancy variables $D(Y, f_{\mathcal{D}_n}(x))$ are used.

B. Proofs of Propositions

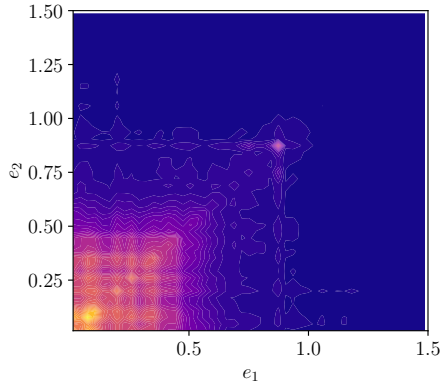
B.1. Proof of Proposition 4.2

Define the random variable $J \triangleq A_\epsilon(\mathbf{X}, \mathbf{Y})$ using (5). The null hypothesis of ϵ-goodness is then simply $J = 0$ and the alternative is $J = 1$. We wish to apply the fundamental lemma of Neyman and Pearson (Thm. 3.2.1 in (Lehmann et al., 2005)) and choose the measure $\mu \triangleq P_{\mathbf{X}} \times \delta_{\{0,1\}}$ on $\mathcal{X} \times \{0, 1\}$, where $P_{\mathbf{X}}$ is the distribution of $\mathbf{X}$ and $\delta_{\{0,1\}}$ is the counting measure.

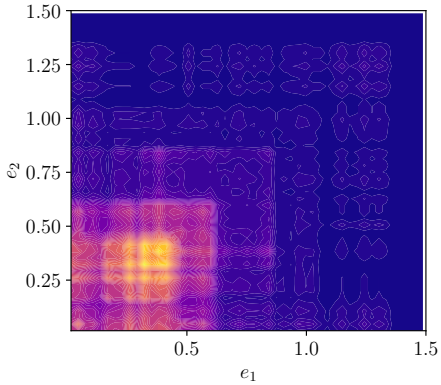
For $\mathbb{P}(J = j) > 0$, we can write the density $p_{\mathbf{X}}(\mathbf{x}|J = j)$ of $\mathbf{X}$ given $J = j$ w.r.t. $P_{\mathbf{X}}$ as

$$p_{\mathbf{X}}(\mathbf{x}|J = j) = \frac{1}{\mathbb{P}(J = j)} \frac{d\mathbb{P}(\mathbf{X}, J)}{d\mu}(\mathbf{x}, j) = \frac{1}{\mathbb{P}(J = j)} \mathbb{P}(J = j|\mathbf{X} = \mathbf{x}). \tag{18}$$

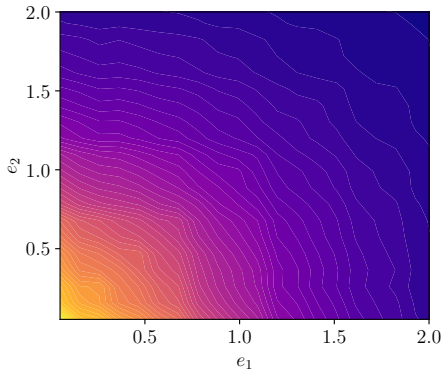
The null hypothesis is rejected if and only if $\mathbb{P}(J = 1|\mathbf{X} = \mathbf{x}) = P_B(\mathbf{x}; \epsilon) > \gamma$. This condition can now be rewritten as $p_{\mathbf{X}}(\mathbf{x}|J = 1) > k\, p_{\mathbf{X}}(\mathbf{x}|J = 0)$ using (18), where $k = \frac{\gamma P\{J=0\}}{(1-\gamma)P\{J=1\}}$. However, this is exactly the sufficient condition for the most powerful test by Thm. 3.2.1 in (Lehmann et al., 2005).



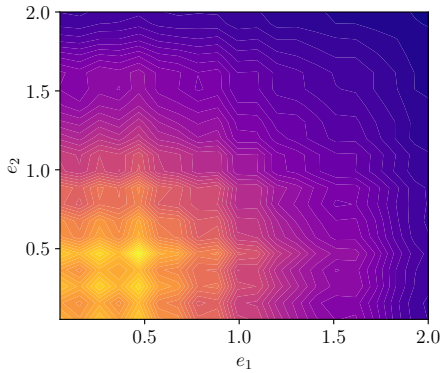
(a) Boston dataset. Average joint error distribution for points that are ϵ-good in the test set. $\epsilon = 0.25$. Estimated using SQR.



(b) Boston dataset. Average joint error distribution for points that are ϵ-bad in the test set. $\epsilon = 0.25$. Estimated using SQR.



(c) Energy dataset. Average joint error distribution for points that are ϵ-good in the test set. $\epsilon = 0.05$. Estimated using CG.



(d) Energy dataset. Average joint error distribution for points that are ϵ-bad in the test set. $\epsilon = 0.05$. Estimated using CG.

Figure 4. Examples of average joint error distributions given by (3) for the relative error discrepancy function $|Y - f_{\mathcal{D}_n}|/|f_{\mathcal{D}_n}|$ using the regressors trained in Section 6. The patterns are obtained from the corresponding test set and the distributions of the discrepancy variables $D(Y, f_{\mathcal{D}_n})$ are used.

B.2. Proof of Proposition 5.4

We can consider the diversity for the regressor obtained from the symmetric function:

$$h_p(u, v) \triangleq \mathbb{1}\{u > \epsilon\}\mathbb{1}\{v > \epsilon\}. \tag{19}$$

Then, since in the definition of $\mathbb{H}(\mathbf{x})$ in Eq. (12) $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are drawn independently according to $\mathbf{Y}|\mathbf{X} = \mathbf{x}$, we have:

$$\begin{aligned}\mathbb{H}(\mathbf{x}) &= \mathbb{E}\left[\mathbb{1}\{d(\mathbf{Y}_1, f_{\mathcal{D}_n}(\mathbf{x})) > \epsilon\} \times \mathbb{1}\{d(\mathbf{Y}_2, f_{\mathcal{D}_n}(\mathbf{x})) > \epsilon\}|\mathbf{X} = \mathbf{x}\right] \\ &= \mathbb{E}\left[\mathbb{1}\{d(\mathbf{Y}_1, f_{\mathcal{D}_n}(\mathbf{x})) > \epsilon\}|\mathbf{X} = \mathbf{x}\right] \times \mathbb{E}\left[\mathbb{1}\{d(\mathbf{Y}_2, f_{\mathcal{D}_n}(\mathbf{x})) > \epsilon\}|\mathbf{X} = \mathbf{x}\right] \\ &= P_B^2(\mathbf{x}; \epsilon).\end{aligned} \tag{20}$$

Since $P_B^2(\mathbf{x}; \epsilon)$ is an increasing function of $P_B(\mathbf{x}; \epsilon)$, this detector is equivalent to the optimal one.

C. Algorithms based on the discrepancy variable

In this Appendix, we present the baseline Algorithm 3 which is similar to Algorithm 1, but based on the estimation of the conditional distribution of $D|\mathbf{X}=\mathbf{x}$. We also present Algorithm 4 (DV-D), which is similar to Algorithm 2 (DV-Y) but also based on the estimation of the conditional distribution of $D|\mathbf{X}=\mathbf{x}$.

Algorithm 3 Baseline based on the estimation of the conditional distribution of $D(\mathbf{Y},\mathbf{X})|\mathbf{X}=\mathbf{x}$

Input: a trained regressor $f_{\mathcal{D}_n}$, trained on the dataset $\mathcal{D}_n=\{(\mathbf{x}_i,y_i)\}_{i=1}^n$. A detection threshold $\gamma\in[0,\ 1]$.

Training:

Step 1: Using $\mathcal{D}_n$ define a new discrepancy samples dataset:

$$\tilde{\mathcal{D}}_n \triangleq \{(\mathbf{x}_i, d(\mathbf{y}_i, f_{\mathcal{D}_n}(\mathbf{x}_i))\}_{i=1}^n.$$

Step 2: Use $\tilde{\mathcal{D}}_n$ to estimate the conditional distribution of $D(\mathbf{Y},\mathbf{X})|\mathbf{X}$.

After training: Given an input test sample $\mathbf{x}$:

Step 1: Estimate $\hat{P}_B(\mathbf{x};\epsilon)$ using (9)

Step 2: Decide that the regressor is ϵ-bad at $\mathbf{x}$ if $\hat{P}_B(\mathbf{x};\epsilon)>\gamma$.

Algorithm 4 – DV-Y-D: Diversity discriminator based on the estimation of the conditional distribution of $D(\mathbf{Y},\mathbf{X})|\mathbf{X}=\mathbf{x}$

Input: A regressor $f_{\mathcal{D}_n}$ trained on dataset $\mathcal{D}_n$, dissimilarity threshold γ.

Training:

Step 1: Using $\mathcal{D}_n$ define a new discrepancy dataset:

$$\tilde{\mathcal{D}}_n \triangleq \{(\mathbf{x}_i, d(\mathbf{y}_i, f_{\mathcal{D}_n}(\mathbf{x}_i))\}_{i=1}^n.$$

Step 2: Use $\tilde{\mathcal{D}}_n$ to obtain an estimate of the distribution of the discrepancy variable $D(\mathbf{Y},\mathbf{X})|\mathbf{X}$.

Step 3: separate good and bad samples in $\mathcal{D}_n$:

$$\mathcal{G} \triangleq \{(\mathbf{x},\mathbf{y})\in\mathcal{D}_n : d(\mathbf{y}, f_{\mathcal{D}_n}(\mathbf{x}))\le\epsilon\},$$
$$\mathcal{B} \triangleq \{(\mathbf{x},\mathbf{y})\in\mathcal{D}_n : d(\mathbf{y}, f_{\mathcal{D}_n}(\mathbf{x}))>\epsilon\}.$$

Step 4: learn the function h that defined the diversity metric $\mathbb{H}$ for the discriminator:

for i in $1,..,N_{\text{epochs}}$ **do**

for each $(\mathbf{x}, d(\mathbf{y}, f_{\mathcal{D}_n}(\mathbf{x})))\in\tilde{\mathcal{D}}_n$ **do**

Generate two extended datasets of N_u samples:

$$\tilde{\mathcal{D}}_1(\mathbf{x})=\{\bar{d}_1,\dots,\bar{d}_{N_u}\}\quad \tilde{\mathcal{D}}_2(\mathbf{x})=\{\tilde{d}_1,\dots,\tilde{d}_{N_u}\}$$

of discrepancy samples, where each $\bar{d}_i,\tilde{d}_i$ is generated i.i.d. with $\hat{F}_{\mathbf{D}|\mathbf{X}=\mathbf{x}}$ from Step 2.

Estimate the diversity metric for $\mathbf{x}$:

$$\mathbb{H}(\mathbf{x})=\frac{1}{N_u}\sum_{i=1}^{N_u} h(\tilde{d}_i,\bar{d}_i),$$

end for

Update h to minimize:

$$\mathcal{L}=\frac{1}{2}\frac{1}{|\mathcal{G}|}\sum_{\mathbf{x}\in\mathcal{G}}\mathbb{H}(\mathbf{x})-\frac{1}{2}\frac{1}{|\mathcal{B}|}\sum_{\mathbf{x}\in\mathcal{B}}\mathbb{H}(\mathbf{x})$$

such that h is symmetric and $h:\mathbb{R}^2_{\ge 0}\to[0,1]$.

end for

After training: Given an input test sample $\mathbf{x}$:

Step 1: compute $\mathbb{H}(\mathbf{x})$ via a Monte Carlo experiment like in Step 4 of training.

Step 2: Decide that the regressor will be ϵ-bad at $\mathbf{x}$ if $\mathbb{H}(\mathbf{x})>\gamma$.

D. Additional Simulation Details

D.1. Hyperparameters

For the CG, SQR and KNIFE the following setups were used:

- CG: we averaged 10 models and performed 5-fold cross-validation over the learning rate in $\{0.01, 0.001, 0.0001\}$. Each model was trained for 150 epochs. For the mean and variance a neural network with a single hidden layer of 64 neurons was used.
- SQR: we used a neural network with 3 hidden layers of 64 neurons. We performed 5-fold cross-validation over the learning rate in $\{10^{-3}, 5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-5}, 10^{-5}\}$ and weight decay in $\{0, 0.025, 0.05, 0.075, 0.1\}$. The models were trained for 1000 epochs.
- KNIFE: we used 256 modes for the Gaussian mixture, and performed 5-fold cross-validation over the learning rate in $\{0.001, 0.0005, 0.0001\}$ and weight decay in $\{0, 0.0125, 0.025\}$. The models were trained for 1000 epochs.

For the implementation of the DV-Y and DV-D algorithms, a neural network with 4 hidden layers, each containing 64 neurons, was employed. The network's symmetry was achieved by averaging the output of two copies of the network fed with interchanged inputs. In all Monte Carlo experiments to compute $\mathbb{H}$, 20000 uniform random variables were used. For learning the h function in DV-Y and DV-D algorithms, the learning rates were chosen based on a 20% validation set, within the following ranges:

- *yacht*, *energy*: $\{5 \times 10^{-3}, 10^{-3}, 5 \times 10^{-4}\}$;
- *concrete*, *kin8nm*, *naval*: $\{10^{-4}, 5 \times 10^{-5}, 10^{-5}\}$;
- *power*, *wine*: $\{10^{-3}, 5 \times 10^{-4}, 10^{-4}\}$;
- *Boston*: $\{5 \times 10^{-4}, 5 \times 10^{-5}, 10^{-5}\}$.

D.2. Computations of $\hat{P}_B(\mathbf{x}; \epsilon)$ for the baseline algorithms

For the computations of the baselines in Section 6 where the output variable $\mathbf{Y}$ is scalar, $\mathcal{Y} \subseteq \mathbb{R}$, we computed $\hat{P}_B(\mathbf{x}; \epsilon)$ in the simulations using the following expressions:

- For the absolute error and a scalar output Y we have:

$$\begin{aligned} P_B(\mathbf{x}; \epsilon) &= \int_{\mathcal{Y}} \mathbb{1}\{d_a(y, f_{\mathcal{D}_n}(\mathbf{x})) \geq \epsilon\} \, dF_{Y|\mathbf{X}=\mathbf{x}}(y) \\ &= 1 - \mathbb{P}\left(f_{\mathcal{D}_n}(\mathbf{x}) - \epsilon \leq Y \leq f_{\mathcal{D}_n}(\mathbf{x}) + \epsilon | \mathbf{X} = \mathbf{x}\right) \\ &= 1 - F_{Y|\mathbf{X}=\mathbf{x}}(f_{\mathcal{D}_n}(\mathbf{x}) + \epsilon) + F_{Y|\mathbf{X}=\mathbf{x}}(f_{\mathcal{D}_n}(\mathbf{x}) - \epsilon). \end{aligned} \tag{21}$$

- For the relative error and a scalar output Y similar computations show that:

$$P_B(\mathbf{x}; \epsilon) = 1 - F_{Y|\mathbf{X}=\mathbf{x}}(f_{\mathcal{D}_n}(\mathbf{x}) + \epsilon |f_{\mathcal{D}_n}(\mathbf{x})|) + F_{Y|\mathbf{X}=\mathbf{x}}(f_{\mathcal{D}_n}(\mathbf{x}) - \epsilon |f_{\mathcal{D}_n}(\mathbf{x})|). \tag{22}$$

In the case of CG and KNIFE, which are parametric models involving Gaussian random variables, we can compute these probabilities in closed form. For SQR, the algorithm provides the quantile function (inverse cumulative distribution function). To compute the probability, we then need to numerically invert this function to find the required probabilities, which is straightforward to do since the quantile function is monotonous and defined on $[0, 1]$.

E. Analysis of computational cost

The computational cost of training the DV detectors can be separated in two components:

1. In the first step, an algorithm that predicts the conditional distribution either of $\mathbf{Y}|\mathbf{X}$ (for DV-Y) or of the discrepancy random variables (for DV-D) has to be considered. The cost associated with this stage comes from the algorithm of choice. For instance, the cost of SQR, KNIFE and the conditional Gaussian method are different, and they depend on the complexity of the chosen model for each case. This training has to be performed in an equal manner for both the baseline algorithms and the algorithm with the diversity coefficients.

2. In the second step, the diversity coefficient algorithm needs to be trained. This process involves training a model that learns the function h, which is essentially a mapping from $\mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$. In other words, h operates as a system with two real inputs and one output. For our experiments, we employed a neural network with 4 hidden layers, each containing 64 neurons. To obtain the Monte Carlo estimation of the diversity coefficients $\mathbb{H}$ given by (12) in the paper, we utilize the model trained in the first step as a generative model. This model is used to generate a large number of realizations of the discrepancy random variables, typically 20,000 in our experiments. Subsequently, these realizations serve as inputs to the neural network responsible for generating the samples of h. Since the network that learns h only has two inputs, the computational complexity of this training process does not surpass that of training a network with inputs equivalent to a square image of approximately 141 pixels per side, with the added step of generating the scalar samples.

In Table 5 we show some examples of training time for the diversity coefficients in different scenarios.

Table 5: Average training time for one epoch of training for each type of diversity coefficient using a Nvidia V100 GPU with 16GB of RAM. This time considers only the training of the diversity coefficients. The time differences are due to the different times required to generate the samples with each conditional model.

Dataset	Sample size	DV type	Conditional estimator	Discrepancy	Avg. time per it (sec.)
yacht	308	DV-Y	SQR	abs	1.0
		DV-Y	KNIFE	abs	1.1
		DV-Y	CG	abs	2.5
		DV-Y	SQR	rel	1.0
		DV-Y	CG	rel	2.5
		DV-Y	KNIFE	rel	1.1
		DV-D	SQR	abs	1.0
		DV-D	CG	abs	2.5
		DV-D	KNIFE	abs	1.1
		DV-D	KNIFE	rel	1.1
		DV-D	CG	rel	2.5
		DV-D	SQR	rel	1.0
boston	506	DV-Y	KNIFE	abs	1.9
		DV-Y	CG	abs	4.1
		DV-Y	CG	rel	4.1
		DV-Y	KNIFE	rel	1.9
		DV-D	SQR	abs	1.7
		DV-D	CG	abs	4.1
		DV-D	KNIFE	abs	1.8
		DV-D	KNIFE	rel	1.8
		DV-D	CG	rel	4.1
		DV-D	SQR	rel	1.7
energy	768	DV-Y	KNIFE	abs	2.7
		DV-Y	CG	abs	6.2
		DV-Y	CG	rel	6.3
		DV-Y	KNIFE	rel	2.7
		DV-D	SQR	abs	2.6
		DV-D	CG	abs	6.2
		DV-D	KNIFE	abs	2.7
		DV-D	KNIFE	rel	2.7
		DV-D	CG	rel	6.2
		DV-D	SQR	rel	2.6
concrete	1030	DV-Y	KNIFE	abs	3.6
		DV-Y	CG	abs	8.3
		DV-Y	CG	rel	8.3
		DV-Y	KNIFE	rel	3.6
		DV-D	SQR	abs	3.4

		DV-D	CG	abs	8.3
		DV-D	KNIFE	abs	3.6
		DV-D	KNIFE	rel	3.6
		DV-D	CG	rel	8.3
		DV-D	SQR	rel	3.4
wine	4898	DV-Y	KNIFE	abs	5.5
		DV-Y	SQR	abs	5.3
		DV-Y	CG	abs	13.0
		DV-Y	KNIFE	rel	5.5
		DV-Y	CG	rel	13.6
		DV-Y	SQR	rel	5.3
		DV-D	SQR	abs	5.3
		DV-D	CG	abs	12.8
		DV-D	KNIFE	abs	5.5
		DV-D	CG	rel	12.8
		DV-D	SQR	rel	5.3
		DV-D	KNIFE	rel	5.5
kin8nm	8192	DV-Y	KNIFE	abs	28.9
		DV-Y	CG	abs	66.1
		DV-Y	CG	rel	66.1
		DV-Y	KNIFE	rel	28.7
		DV-D	SQR	abs	27.2
		DV-D	CG	abs	65.1
		DV-D	KNIFE	abs	28.0
		DV-D	KNIFE	rel	28.4
		DV-D	CG	rel	65.4
		DV-D	SQR	rel	27.1
power	9568	DV-Y	KNIFE	abs	33.7
		DV-Y	CG	abs	76.9
		DV-Y	CG	rel	77.0
		DV-Y	KNIFE	rel	33.8
		DV-D	SQR	abs	31.6
		DV-D	CG	abs	76.5
		DV-D	KNIFE	abs	33.0
		DV-D	KNIFE	rel	33.1
		DV-D	CG	rel	77.0
		DV-D	SQR	rel	31.6
naval	11934	DV-Y	KNIFE	abs	41.6
		DV-Y	CG	abs	96.5
		DV-Y	CG	rel	96.1
		DV-Y	KNIFE	rel	42.2
		DV-D	SQR	abs	39.6
		DV-D	CG	abs	95.3
		DV-D	KNIFE	abs	41.7
		DV-D	KNIFE	rel	41.7
		DV-D	CG	rel	95.3
		DV-D	SQR	rel	39.5

F. Connection to conformal prediction

While there are some common elements with conformal prediction, the objective of this work differs. In contrast, conformal approaches cannot be directly applied to solve the problem we tackle, although we acknowledge and verify the useful properties of marginal coverage they possess. The key differences are as follows:

1. Our goal is to detect unreliable regressor behavior as measured by an application dependent error margin (ϵ) and a discrepancy metric. For our baselines we use a probabilistic model, such as quantile regression (SQR) or others (conditional gaussian, KNIFE), to estimate the conditional probability of the unreliable behavior event:

$$P_B(\mathbf{x}) = \mathbb{P}(d(\mathbf{Y}, f_{\mathcal{D}_n}(\mathbf{x})) > \epsilon | \mathbf{X} = \mathbf{x}).$$

2. Using this probability, we approximate the optimal detector. Conformal approaches provide a prediction interval that contains the desired variables with certain marginal guarantees. To achieve this, a desired value for the marginal

coverage probability (denoted as $1-\alpha$) is selected, and then an interval (similar in spirit to the interval defined by ϵ) that satisfies this marginal coverage condition is obtained. In other words, a value of ϵ is determined such that:

$$\mathbb{P}(d(\mathbf{X}, f_{\mathcal{D}_n}(\mathbf{X})) \leq \epsilon) \geq 1-\alpha.$$

Thus, in conformal learning the marginal probability is fixed and the width ϵ is found, to guarantee marginal coverage. The analysis above demonstrates that the problem we are tackling differs from the problem addressed by conformal predictions. Moreover, the guarantees provided by conformal predictions, in terms of marginal coverage, do not offer any insight into the conditional probability required for constructing our baselines.

Nevertheless, we explore the use of conformal approaches within our framework, by constructing the following approach: given value of ϵ and an input $\mathbf{x}$:

a. Construct a conformal prediction such that the interval matches the condition of being ϵ-good, i.e. $d\left(\mathbf{Y}, f_{\mathcal{D}_n}(\mathbf{x})\right) \leq \epsilon$.

b. Obtain the marginal coverage probability for this event, denoted as $P_c(\mathbf{x})$.

c. Use $1-P_c(\mathbf{x})$ as a proxy for the conditional probability $P_B(\mathbf{x})$.

This approach does not have any theoretical guarantees since we approximate the conditional probability via a marginal coverage probability of an interval of the chosen width provided by a specific conformal algorithm. To test this algorithm we need a conformal algorithm which satisfies two conditions:

1. The conformal interval has to be possibly non-symmetric (allow for different tails) ;

2. The interval has to depend on $\mathbf{x}$. The fact that not all conformal techniques satisfy this(Vovk et al., 2005), shows in fact that this construction is not inherent to these approaches.

In this case, we rely on the well-known method introduced in (Romano et al., 2019), specifically Theorem 2, which provides an algorithm that satisfies the desired properties and allows us to construct the conformal intervals based on any distribution of the coverage of its tails. This algorithm starts from a preconstructed confidence interval and then applies a conformal correction. This means that for each algorithm used in our baseline (i.e., SQR, KNIFE, conditional Gaussian) we can construct a "conformal" version, which consists of computing the baseline with the procedure described in the previous paragraph. We compare the performance of these algorithms with the ones already present in the paper, which are indicated with the suffix CF (e.g., B1-SQR-CF denotes Baseline algorithm 1, with quantiles computed with SQR and the conformal correction). The results can be seen in the tables below. In particular, for the tables in this section, we report the original results on the left and the conformal correction results on the right. As we see in the results, both for the relative error and absolute error metric we performed 144 experiments with the 8 datasets, 3 epsilons and the 6 baselines. In terms of AUROC, the conformal modification of the baseline only results in improvement in 52% of the experiments. This shows that the effects of this type of correction are uncertain since none of the theoretical guarantees afforded by conformal intervals apply.

Table 6. AUROC for the relative error discrepancy metric for the small datasets. B1, B2 correspond to baseline Algorithms. DV-Y and DV-D correspond to the proposed algorithms. CF corresponds to the baseline with the conformal correction.

	ϵ [%]	10%	15%	20%
concrete	**Avg. % bad**	35.4 [6.2]	22.1 [5.1]	13.9 [3.4]
	B1-SQR	71.3 [5.3]	74.8 [5.1]	74.6 [6.3]
	B2-SQR	66.5 [4.3]	65.9 [3.8]	62.9 [6.7]
	B1-KNIFE	62.6 [4.2]	63.2 [7.4]	66.0 [6.1]
	B2-KNIFE	61.4 [3.4]	63.8 [5.1]	64.9 [8.4]
	B1-CG	62.9 [5.8]	66.9 [8.0]	70.6 [7.7]
	B2-CG	67.2 [5.6]	69.6 [8.2]	71.3 [11.4]
	DV-Y-SQR	**72.2** [5.4]	**75.3** [5.0]	**77.4** [4.6]
	DV-D-SQR	67.8 [4.4]	68.9 [5.9]	69.5 [7.8]
	DV-Y-KNIFE	62.9 [5.0]	63.7 [5.4]	63.6 [6.1]
	DV-D-KNIFE	61.5 [3.7]	62.1 [5.8]	61.6 [6.2]
	DV-Y-CG	62.2 [4.6]	67.4 [7.7]	70.2 [7.5]
	DV-D-CG	67.3 [6.0]	69.5 [8.2]	71.2 [11.3]
energy	ϵ [%]	2%	3.5%	5%
	Avg. % bad	46.8 [3.7]	21.0 [7.0]	10.5 [5.2]
	B1-SQR	71.1 [8.2]	80.5 [6.4]	91.3 [5.9]
	B2-SQR	69.4 [7.6]	74.0 [6.7]	78.4 [9.2]
	B1-KNIFE	73.5 [6.4]	77.2 [6.9]	83.5 [11.6]
	B2-KNIFE	65.4 [5.4]	71.7 [6.3]	78.7 [7.1]
	B1-CG	74.9 [3.9]	78.3 [6.8]	81.9 [12.1]
	B2-CG	68.5 [6.6]	73.6 [7.3]	82.2 [6.9]
	DV-Y-SQR	73.1 [7.9]	81.0 [5.9]	**91.5** [5.7]
	DV-D-SQR	68.4 [8.9]	75.3 [6.4]	85.3 [7.4]
	DV-Y-KNIFE	73.5 [6.9]	**81.9** [5.2]	87.4 [8.4]
	DV-D-KNIFE	65.6 [5.4]	72.3 [6.4]	79.3 [7.9]
	DV-Y-CG	**76.3** [3.2]	80.1 [6.6]	89.1 [7.9]
	DV-D-CG	68.3 [6.7]	74.1 [6.8]	82.5 [6.9]
boston	ϵ [%]	10%	15%	20%
	Avg. % bad	33.7 [5.2]	20.8 [5.0]	14.5 [6.3]
	B1-SQR	72.2 [8.0]	78.3 [7.5]	81.9 [11.9]
	B2-SQR	65.1 [10.7]	66.6 [12.8]	65.5 [16.5]
	B1-KNIFE	70.2 [8.1]	75.4 [8.5]	81.2 [12.4]
	B2-KNIFE	67.4 [12.3]	71.5 [13.9]	76.0 [14.6]
	B1-CG	70.8 [9.3]	75.9 [10.3]	80.7 [12.1]
	B2-CG	69.3 [13.2]	74.0 [14.7]	78.1 [15.8]
	DV-Y-SQR	**74.0** [7.4]	**79.5** [8.4]	**85.6** [8.2]
	DV-D-SQR	68.5 [7.6]	70.8 [12.3]	76.4 [13.6]
	DV-Y-KNIFE	70.1 [8.2]	75.0 [10.1]	80.9 [14.0]
	DV-D-KNIFE	69.5 [10.4]	72.1 [13.7]	77.7 [13.7]
	DV-Y-CG	69.0 [10.0]	74.4 [10.7]	79.9 [12.6]
	DV-D-CG	69.6 [12.3]	74.2 [13.4]	77.7 [15.5]
yacht	ϵ [%]	10%	15%	20%
	Avg. % bad	59.4 [9.0]	46.1 [9.3]	39.0 [8.9]
	B1-SQR	83.4 [9.9]	87.8 [7.2]	91.9 [5.4]
	B2-SQR	85.7 [7.3]	89.2 [7.0]	92.6 [5.7]
	B1-KNIFE	63.1 [11.5]	67.0 [11.4]	71.9 [10.5]
	B2-KNIFE	83.5 [7.1]	87.1 [5.7]	90.9 [6.3]
	B1-CG	**92.7** [4.4]	**95.7** [3.0]	**97.2** [2.1]
	B2-CG	77.5 [13.3]	82.2 [13.7]	87.1 [15.1]
	DV-Y-SQR	86.7 [7.9]	91.5 [5.9]	93.5 [5.3]
	DV-D-SQR	87.2 [6.5]	90.6 [5.8]	93.6 [5.3]
	DV-Y-KNIFE	68.9 [11.1]	68.9 [10.5]	75.5 [9.8]
	DV-D-KNIFE	84.4 [7.0]	87.3 [5.7]	91.1 [6.4]
	DV-Y-CG	92.1 [5.0]	94.6 [3.5]	96.2 [2.9]
	DV-D-CG	79.5 [13.5]	81.5 [17.6]	86.4 [16.9]

	ϵ [%]	10%	15%	20%
concrete	**Avg. % bad**	35.4 [6.2]	22.1 [5.1]	13.9 [3.4]
	B1-SQR-CF	71.1 [5.5]	74.8 [5.2]	76.4 [5.0]
	B2-SQR-CF	67.6 [4.4]	69.1 [6.5]	70.7 [8.1]
	B1-KNIFE-CF	62.8 [3.5]	64.6 [7.2]	66.0 [5.7]
	B2-KNIFE-CF	59.2 [4.2]	62.7 [5.1]	63.7 [8.5]
	B1-CG-CF	62.7 [6.6]	66.9 [7.9]	70.3 [8.1]
	B2-CG-CF	67.4 [6.1]	69.5 [8.0]	71.1 [10.7]
energy	ϵ [%]	2%	3.5%	5%
	Avg. % bad	46.8 [3.7]	21.0 [7.0]	10.5 [5.2]
	B1-SQR-CF	74.2 [6.2]	82.2 [8.5]	91.3 [5.4]
	B2-SQR-CF	68.6 [8.3]	75.6 [5.4]	84.5 [8.5]
	B1-KNIFE-CF	73.3 [6.8]	77.0 [7.3]	82.3 [12.6]
	B2-KNIFE-CF	65.5 [5.7]	71.8 [7.7]	79.4 [7.7]
	B1-CG-CF	72.6 [4.3]	77.3 [6.2]	81.7 [11.3]
	B2-CG-CF	67.4 [6.9]	74.1 [6.8]	82.7 [6.8]
boston	ϵ [%]	10%	15%	20%
	Avg. % bad	33.7 [5.2]	20.8 [5.0]	14.5 [6.3]
	B1-SQR-CF	72.7 [7.7]	79.0 [7.8]	83.0 [9.3]
	B2-SQR-CF	64.5 [12.4]	67.4 [15.3]	72.1 [18.6]
	B1-KNIFE-CF	70.8 [8.0]	76.7 [7.1]	81.3 [12.5]
	B2-KNIFE-CF	68.6 [9.5]	71.4 [14.1]	74.4 [17.6]
	B1-CG-CF	70.8 [9.3]	76.7 [9.5]	82.7 [11.6]
	B2-CG-CF	69.7 [12.8]	74.2 [14.8]	78.7 [15.8]
yacht	ϵ [%]	10%	15%	20%
	Avg. % bad	59.4 [9.0]	46.1 [9.3]	39.0 [8.9]
	B1-SQR-CF	79.5 [6.8]	88.5 [7.1]	92.9 [4.4]
	B2-SQR-CF	75.8 [17.7]	76.4 [22.4]	82.0 [23.8]
	B1-KNIFE-CF	61.5 [13.8]	69.3 [10.3]	75.2 [10.9]
	B2-KNIFE-CF	74.6 [10.5]	77.7 [12.4]	80.9 [14.4]
	B1-CG-CF	92.2 [3.5]	95.7 [3.2]	**97.2** [2.0]
	B2-CG-CF	72.0 [12.6]	77.6 [15.6]	83.9 [16.5]

Table 7. FPR at TPR 90% for the relative error discrepancy metric for the small datasets. B1, B2 correspond to baseline Algorithms. DV-Y and DV-D correspond to the proposed algorithms. CF corresponds to the baseline with the conformal correction.

	ϵ [%]	10%	15%	20%
concrete	B1-SQR	**0.63** [0.11]	0.62 [0.21]	0.76 [0.32]
	B2-SQR	0.86 [0.15]	0.95 [0.16]	1.00 [0.00]
	B1-KNIFE	0.76 [0.08]	0.79 [0.11]	0.73 [0.10]
	B2-KNIFE	0.75 [0.04]	0.73 [0.08]	0.67 [0.12]
	B1-CG	0.80 [0.04]	0.69 [0.14]	0.63 [0.19]
	B2-CG	0.73 [0.09]	0.64 [0.17]	0.62 [0.23]
	DV-Y-SQR	**0.63** [0.11]	**0.56** [0.09]	**0.47** [0.14]
	DV-D-SQR	0.71 [0.14]	0.71 [0.12]	0.69 [0.17]
	DV-Y-KNIFE	0.79 [0.08]	0.76 [0.13]	0.74 [0.14]
	DV-D-KNIFE	0.76 [0.04]	0.74 [0.07]	0.74 [0.10]
	DV-Y-CG	0.79 [0.04]	0.70 [0.12]	0.63 [0.20]
	DV-D-CG	0.72 [0.11]	0.65 [0.16]	0.62 [0.23]
energy	ϵ [%]	2%	3.5%	5%
	B1-SQR	0.76 [0.14]	0.64 [0.20]	0.35 [0.31]
	B2-SQR	0.73 [0.11]	0.69 [0.24]	0.82 [0.38]
	B1-KNIFE	0.68 [0.13]	0.61 [0.14]	0.42 [0.26]
	B2-KNIFE	0.73 [0.13]	0.65 [0.19]	0.48 [0.18]
	B1-CG	**0.62** [0.09]	**0.53** [0.19]	0.40 [0.27]
	B2-CG	0.69 [0.12]	0.60 [0.15]	0.44 [0.25]
	DV-Y-SQR	0.68 [0.18]	0.62 [0.21]	**0.29** [0.22]
	DV-D-SQR	0.72 [0.11]	0.60 [0.19]	0.35 [0.22]
	DV-Y-KNIFE	0.67 [0.15]	0.54 [0.19]	0.32 [0.22]
	DV-D-KNIFE	0.75 [0.12]	0.63 [0.19]	0.51 [0.24]
	DV-Y-CG	**0.62** [0.11]	0.54 [0.18]	**0.29** [0.25]
	DV-D-CG	0.71 [0.13]	0.58 [0.16]	0.44 [0.25]
boston	ϵ [%]	10%	15%	20%
	B1-SQR	0.69 [0.21]	0.63 [0.15]	0.62 [0.41]
	B2-SQR	0.88 [0.20]	0.91 [0.21]	0.90 [0.31]
	B1-KNIFE	0.65 [0.11]	0.61 [0.20]	0.46 [0.30]
	B2-KNIFE	0.76 [0.20]	0.66 [0.24]	0.56 [0.26]
	B1-CG	0.65 [0.22]	0.56 [0.28]	0.48 [0.33]
	B2-CG	0.68 [0.22]	0.59 [0.24]	0.48 [0.28]
	DV-Y-SQR	**0.63** [0.18]	**0.52** [0.18]	**0.34** [0.15]
	DV-D-SQR	0.73 [0.15]	0.70 [0.15]	0.54 [0.26]
	DV-Y-KNIFE	0.67 [0.14]	0.65 [0.24]	0.50 [0.35]
	DV-D-KNIFE	0.68 [0.17]	0.65 [0.19]	0.51 [0.21]
	DV-Y-CG	0.68 [0.20]	0.61 [0.27]	0.49 [0.33]
	DV-D-CG	0.69 [0.23]	0.61 [0.22]	0.48 [0.25]
yacht	ϵ [%]	10%	15%	20%
	B1-SQR	0.46 [0.27]	0.37 [0.17]	0.23 [0.15]
	B2-SQR	0.41 [0.29]	0.39 [0.37]	0.27 [0.38]
	B1-KNIFE	0.82 [0.13]	0.82 [0.14]	0.77 [0.22]
	B2-KNIFE	0.49 [0.23]	0.32 [0.16]	0.26 [0.25]
	B1-CG	**0.25** [0.21]	**0.17** [0.17]	**0.10** [0.12]
	B2-CG	0.63 [0.27]	0.55 [0.35]	0.30 [0.31]
	DV-Y-SQR	0.38 [0.25]	**0.17** [0.10]	0.16 [0.10]
	DV-D-SQR	0.33 [0.19]	0.27 [0.20]	0.20 [0.21]
	DV-Y-KNIFE	0.74 [0.20]	0.84 [0.15]	0.73 [0.25]
	DV-D-KNIFE	0.40 [0.18]	0.37 [0.20]	0.29 [0.30]
	DV-Y-CG	**0.26** [0.21]	0.22 [0.21]	**0.10** [0.11]
	DV-D-CG	0.56 [0.22]	0.35 [0.30]	0.32 [0.36]

	ϵ [%]	10%	15%	20%
concrete	B1-SQR-CF	0.66 [0.11]	0.59 [0.13]	0.50 [0.17]
	B2-SQR-CF	0.74 [0.13]	0.71 [0.11]	0.66 [0.19]
	B1-KNIFE-CF	0.76 [0.07]	0.76 [0.09]	0.70 [0.14]
	B2-KNIFE-CF	0.79 [0.05]	0.76 [0.07]	0.71 [0.10]
	B1-CG-CF	0.80 [0.06]	0.70 [0.15]	0.64 [0.21]
	B2-CG-CF	0.74 [0.12]	0.67 [0.16]	0.65 [0.21]
energy	ϵ [%]	2%	3.5%	5%
	B1-SQR-CF	0.71 [0.14]	0.65 [0.28]	**0.23** [0.18]
	B2-SQR-CF	0.74 [0.14]	0.59 [0.17]	0.38 [0.22]
	B1-KNIFE-CF	0.69 [0.13]	0.59 [0.15]	0.42 [0.29]
	B2-KNIFE-CF	0.79 [0.09]	0.64 [0.19]	0.51 [0.23]
	B1-CG-CF	0.64 [0.08]	0.54 [0.17]	0.40 [0.31]
	B2-CG-CF	0.77 [0.10]	0.58 [0.16]	0.45 [0.23]
boston	ϵ [%]	10%	15%	20%
	B1-SQR-CF	0.70 [0.14]	0.55 [0.20]	0.42 [0.20]
	B2-SQR-CF	0.77 [0.24]	0.76 [0.23]	0.71 [0.33]
	B1-KNIFE-CF	**0.63** [0.12]	0.60 [0.22]	0.47 [0.31]
	B2-KNIFE-CF	0.73 [0.17]	0.69 [0.23]	0.61 [0.27]
	B1-CG-CF	0.64 [0.23]	0.61 [0.30]	0.44 [0.32]
	B2-CG-CF	0.72 [0.21]	0.66 [0.25]	0.49 [0.28]
yacht	ϵ [%]	10%	15%	20%
	B1-SQR-CF	0.64 [0.24]	0.37 [0.23]	0.25 [0.26]
	B2-SQR-CF	0.61 [0.29]	0.54 [0.35]	0.54 [0.36]
	B1-KNIFE-CF	0.81 [0.14]	0.82 [0.18]	0.71 [0.26]
	B2-KNIFE-CF	0.69 [0.29]	0.71 [0.32]	0.52 [0.37]
	B1-CG-CF	0.25 [0.16]	0.13 [0.12]	**0.10** [0.11]
	B2-CG-CF	0.72 [0.21]	0.57 [0.29]	0.48 [0.36]

Table 8. AUROC for the absolute error discrepancy metric for the small datasets. B1, B2 correspond to baseline Algorithms. DV-Y and DV-D correspond to the proposed Algorithms. CF corresponds to the baseline with the conformal correction.

	ϵ	**0.2**	**0.5**	**0.6**
concrete	**Avg. % bad**	36.7 [5.6]	9.6 [2.8]	6.4 [2.0]
	B1-SQR	73.1 [3.8]	81.1 [8.3]	76.3 [14.2]
	B2-SQR	69.8 [7.3]	59.2 [9.0]	57.3 [9.3]
	B1-KNIFE	63.8 [7.4]	68.7 [5.9]	71.9 [9.6]
	B2-KNIFE	67.7 [6.9]	76.9 [8.9]	80.9 [12.2]
	B1-CG	66.6 [4.7]	77.6 [5.6]	79.8 [7.2]
	B2-CG	69.8 [7.1]	78.7 [9.6]	81.5 [9.1]
	DV-Y-SQR	**74.4** [3.4]	**83.5** [8.3]	**82.1** [8.4]
	DV-D-SQR	70.2 [6.9]	79.0 [9.5]	80.5 [9.2]
	DV-Y-KNIFE	63.3 [5.7]	67.2 [12.4]	69.3 [14.8]
	DV-D-KNIFE	67.9 [7.8]	75.2 [8.5]	79.0 [11.3]
	DV-Y-CG	64.2 [7.4]	77.3 [6.2]	77.4 [10.7]
	DV-D-CG	69.9 [6.8]	78.1 [10.2]	81.9 [8.7]
energy	ϵ	**0.025**	**0.0625**	**0.1**
	Avg. % bad	61.2 [5.0]	25.8 [5.2]	9.5 [3.4]
	B1-SQR	68.2 [8.3]	75.8 [7.3]	81.5 [6.6]
	B2-SQR	63.8 [4.8]	73.9 [7.7]	66.5 [15.5]
	B1-KNIFE	67.7 [3.6]	74.7 [6.8]	72.6 [10.2]
	B2-KNIFE	64.8 [7.0]	71.4 [8.3]	72.3 [12.5]
	B1-CG	72.5 [6.8]	81.2 [5.5]	81.7 [5.4]
	B2-CG	64.3 [7.2]	72.1 [7.8]	74.0 [11.3]
	DV-Y-SQR	70.3 [6.7]	76.5 [7.6]	81.5 [6.6]
	DV-D-SQR	65.5 [5.6]	73.1 [7.6]	75.4 [11.8]
	DV-Y-KNIFE	68.9 [4.1]	78.3 [7.1]	79.4 [9.3]
	DV-D-KNIFE	65.2 [6.5]	72.1 [8.9]	74.0 [13.3]
	DV-Y-CG	**75.1** [6.0]	**82.0** [6.9]	**82.5** [7.7]
	DV-D-CG	65.0 [7.1]	72.0 [7.8]	74.1 [11.3]
boston	ϵ	**0.25**	**0.3**	**0.35**
	Avg. % bad	31.8 [2.7]	24.1 [5.8]	21.2 [5.4]
	B1-SQR	72.8 [7.2]	77.1 [6.0]	76.8 [5.6]
	B2-SQR	61.0 [7.1]	60.8 [7.4]	57.4 [8.1]
	B1-KNIFE	69.9 [8.7]	73.1 [7.7]	75.7 [6.6]
	B2-KNIFE	62.0 [5.5]	63.5 [6.4]	65.7 [6.5]
	B1-CG	67.5 [6.9]	72.4 [8.6]	73.7 [8.9]
	B2-CG	63.7 [5.0]	63.7 [8.3]	66.0 [8.6]
	DV-Y-SQR	**73.7** [7.7]	**77.6** [6.7]	**78.6** [5.4]
	DV-D-SQR	63.3 [7.6]	62.8 [8.7]	64.0 [7.7]
	DV-Y-KNIFE	69.0 [8.5]	72.1 [7.5]	75.5 [6.4]
	DV-D-KNIFE	63.5 [6.8]	63.4 [11.3]	64.4 [11.4]
	DV-Y-CG	65.7 [8.6]	69.7 [11.5]	70.2 [12.3]
	DV-D-CG	63.5 [6.4]	64.2 [9.1]	65.8 [9.4]
yacht	ϵ	**0.025**	**0.05**	**0.075**
	Avg. % bad	62.6 [9.6]	35.8 [11.3]	25.5 [10.6]
	B1-SQR	84.2 [6.8]	90.1 [6.2]	90.9 [5.1]
	B2-SQR	79.1 [10.8]	86.8 [8.3]	89.2 [4.9]
	B1-KNIFE	86.4 [11.4]	89.0 [6.6]	88.7 [6.2]
	B2-KNIFE	73.3 [8.9]	82.0 [9.1]	85.4 [7.9]
	B1-CG	92.4 [4.4]	94.1 [5.7]	96.3 [3.0]
	B2-CG	80.9 [10.3]	84.7 [9.7]	87.8 [7.0]
	DV-Y-SQR	86.4 [6.8]	90.4 [6.8]	92.2 [5.3]
	DV-D-SQR	81.6 [7.6]	87.3 [7.4]	90.9 [4.7]
	DV-Y-KNIFE	89.8 [7.9]	87.9 [6.5]	89.0 [5.7]
	DV-D-KNIFE	77.0 [9.3]	83.7 [7.4]	86.4 [6.7]
	DV-Y-CG	**94.3** [3.3]	**94.4** [5.7]	**96.6** [2.9]
	DV-D-CG	79.9 [9.2]	84.6 [9.4]	87.5 [6.8]

	ϵ	**0.2**	**0.5**	**0.6**
concrete	**Avg. % bad**	36.7 [5.6]	9.6 [2.8]	6.4 [2.0]
	B1-SQR-CF	73.3 [3.3]	81.4 [10.3]	80.0 [13.1]
	B2-SQR-CF	70.1 [6.2]	76.1 [9.4]	72.4 [10.6]
	B1-KNIFE-CF	64.7 [6.9]	69.9 [5.8]	72.8 [8.8]
	B2-KNIFE-CF	67.8 [7.4]	76.6 [9.3]	77.3 [14.5]
	B1-CG-CF	66.9 [4.8]	77.8 [5.5]	79.3 [7.7]
	B2-CG-CF	70.0 [6.9]	74.5 [13.7]	74.2 [16.0]
energy	ϵ	**0.025**	**0.0625**	**0.1**
	Avg. % bad	61.2 [5.0]	25.8 [5.2]	9.5 [3.4]
	B1-SQR-CF	71.1 [6.5]	76.2 [7.8]	82.2 [6.7]
	B2-SQR-CF	64.5 [6.4]	73.2 [7.3]	69.4 [18.3]
	B1-KNIFE-CF	67.6 [4.7]	74.9 [6.5]	72.9 [9.9]
	B2-KNIFE-CF	64.2 [6.3]	71.0 [8.1]	61.8 [20.6]
	B1-CG-CF	72.2 [6.5]	80.9 [5.5]	80.4 [6.3]
	B2-CG-CF	64.7 [7.1]	72.1 [7.6]	66.1 [19.2]
boston	ϵ	**0.25**	**0.3**	**0.35**
	Avg. % bad	31.8 [2.7]	24.1 [5.8]	21.2 [5.4]
	B1-SQR-CF	71.8 [8.7]	77.4 [6.6]	77.3 [7.3]
	B2-SQR-CF	63.5 [7.8]	61.7 [7.5]	61.7 [7.2]
	B1-KNIFE-CF	70.3 [8.3]	72.9 [7.6]	75.0 [6.6]
	B2-KNIFE-CF	62.2 [5.8]	62.4 [7.4]	65.1 [6.5]
	B1-CG-CF	67.7 [6.0]	74.2 [7.4]	75.3 [7.4]
	B2-CG-CF	63.1 [4.9]	64.0 [7.7]	64.6 [7.7]
yacht	ϵ	**0.025**	**0.05**	**0.075**
	Avg. % bad	62.6 [9.6]	35.8 [11.3]	25.5 [10.6]
	B1-SQR-CF	84.1 [5.6]	87.2 [11.5]	87.9 [7.5]
	B2-SQR-CF	80.0 [8.0]	87.4 [8.2]	88.3 [6.1]
	B1-KNIFE-CF	86.4 [11.1]	87.3 [6.4]	87.2 [5.7]
	B2-KNIFE-CF	73.5 [9.1]	81.2 [9.5]	83.2 [8.7]
	B1-CG-CF	94.2 [4.7]	95.7 [5.1]	96.5 [2.5]
	B2-CG-CF	79.6 [11.1]	84.8 [9.6]	84.9 [8.9]

Table 9. FPR at TPR 90% for the absolute error discrepancy metric for the small datasets. B1, B2 correspond to baseline Algorithms. DV-Y and DV-D correspond to the proposed Algorithms. CF corresponds to the baseline with the conformal correction.

	ϵ	**0.2**	**0.5**	**0.6**
concrete	B1-SQR	**0.67** [0.10]	0.77 [0.39]	0.90 [0.30]
	B2-SQR	0.74 [0.13]	1.00 [0.00]	1.00 [0.00]
	B1-KNIFE	0.76 [0.08]	0.68 [0.16]	0.61 [0.21]
	B2-KNIFE	0.73 [0.13]	0.54 [0.18]	**0.45** [0.27]
	B1-CG	0.69 [0.10]	**0.44** [0.13]	0.49 [0.18]
	B2-CG	**0.68** [0.12]	0.50 [0.22]	0.49 [0.23]
	DV-Y-SQR	**0.67** [0.10]	**0.43** [0.21]	0.51 [0.22]
	DV-D-SQR	0.69 [0.13]	0.54 [0.24]	0.52 [0.24]
	DV-Y-KNIFE	0.79 [0.05]	0.66 [0.17]	0.62 [0.19]
	DV-D-KNIFE	0.70 [0.12]	0.64 [0.21]	0.52 [0.25]
	DV-Y-CG	0.73 [0.14]	0.45 [0.14]	0.53 [0.24]
	DV-D-CG	**0.67** [0.12]	0.51 [0.22]	0.49 [0.23]
energy	ϵ	**0.025**	**0.0625**	**0.1**
	B1-SQR	0.74 [0.15]	0.56 [0.18]	0.48 [0.14]
	B2-SQR	0.79 [0.10]	0.63 [0.22]	0.85 [0.32]
	B1-KNIFE	0.69 [0.06]	0.56 [0.13]	0.55 [0.18]
	B2-KNIFE	0.79 [0.07]	0.65 [0.23]	0.54 [0.22]
	B1-CG	0.61 [0.11]	**0.42** [0.14]	**0.39** [0.14]
	B2-CG	0.84 [0.08]	0.63 [0.20]	0.56 [0.21]
	DV-Y-SQR	0.72 [0.18]	0.49 [0.17]	0.49 [0.16]
	DV-D-SQR	0.76 [0.08]	0.61 [0.18]	0.57 [0.25]
	DV-Y-KNIFE	0.67 [0.06]	0.49 [0.14]	0.46 [0.20]
	DV-D-KNIFE	0.79 [0.07]	0.64 [0.19]	0.53 [0.24]
	DV-Y-CG	**0.59** [0.10]	**0.42** [0.13]	**0.39** [0.18]
	DV-D-CG	0.83 [0.07]	0.64 [0.19]	0.56 [0.22]
boston	ϵ	**0.25**	**0.3**	**0.35**
	B1-SQR	0.69 [0.17]	0.60 [0.17]	0.64 [0.25]
	B2-SQR	0.94 [0.11]	0.97 [0.09]	1.00 [0.00]
	B1-KNIFE	0.70 [0.14]	0.64 [0.16]	**0.57** [0.12]
	B2-KNIFE	0.79 [0.12]	0.73 [0.17]	0.75 [0.19]
	B1-CG	0.72 [0.14]	0.64 [0.23]	0.66 [0.23]
	B2-CG	0.74 [0.13]	0.69 [0.15]	0.68 [0.16]
	DV-Y-SQR	**0.64** [0.21]	**0.53** [0.20]	**0.56** [0.24]
	DV-D-SQR	0.78 [0.09]	0.78 [0.10]	0.73 [0.12]
	DV-Y-KNIFE	0.74 [0.18]	0.68 [0.17]	0.60 [0.16]
	DV-D-KNIFE	0.77 [0.13]	0.69 [0.18]	0.72 [0.20]
	DV-Y-CG	0.75 [0.16]	0.66 [0.26]	0.67 [0.26]
	DV-D-CG	0.77 [0.12]	0.68 [0.17]	0.70 [0.17]
yacht	ϵ	**0.025**	**0.05**	**0.075**
	B1-SQR	0.44 [0.19]	0.21 [0.13]	0.24 [0.16]
	B2-SQR	0.52 [0.21]	0.37 [0.33]	0.35 [0.27]
	B1-KNIFE	0.32 [0.16]	0.27 [0.20]	0.27 [0.16]
	B2-KNIFE	0.69 [0.22]	0.42 [0.30]	0.43 [0.29]
	B1-CG	**0.16** [0.12]	**0.11** [0.12]	0.16 [0.30]
	B2-CG	0.53 [0.22]	0.33 [0.22]	0.32 [0.17]
	DV-Y-SQR	0.44 [0.23]	0.19 [0.14]	0.22 [0.16]
	DV-D-SQR	0.49 [0.18]	0.33 [0.23]	0.26 [0.14]
	DV-Y-KNIFE	0.23 [0.16]	0.30 [0.19]	0.29 [0.15]
	DV-D-KNIFE	0.66 [0.27]	0.35 [0.23]	0.36 [0.20]
	DV-Y-CG	**0.16** [0.10]	**0.11** [0.12]	**0.10** [0.07]
	DV-D-CG	0.56 [0.25]	0.32 [0.20]	0.34 [0.17]

	ϵ	**0.2**	**0.5**	**0.6**
concrete	B1-SQR-CF	0.69 [0.13]	0.44 [0.25]	0.61 [0.30]
	B2-SQR-CF	0.67 [0.12]	0.79 [0.26]	0.87 [0.27]
	B1-KNIFE-CF	0.74 [0.05]	0.68 [0.15]	0.64 [0.21]
	B2-KNIFE-CF	0.74 [0.12]	0.66 [0.28]	0.64 [0.40]
	B1-CG-CF	0.66 [0.08]	0.47 [0.14]	0.51 [0.19]
	B2-CG-CF	0.69 [0.11]	0.70 [0.33]	0.79 [0.32]
energy	ϵ	**0.025**	**0.0625**	**0.1**
	B1-SQR-CF	0.67 [0.10]	0.56 [0.19]	0.49 [0.16]
	B2-SQR-CF	0.82 [0.11]	0.61 [0.18]	0.71 [0.32]
	B1-KNIFE-CF	0.68 [0.05]	0.56 [0.13]	0.54 [0.17]
	B2-KNIFE-CF	0.83 [0.07]	0.74 [0.20]	0.83 [0.31]
	B1-CG-CF	0.61 [0.12]	0.40 [0.13]	0.45 [0.25]
	B2-CG-CF	0.86 [0.07]	0.67 [0.18]	0.80 [0.26]
boston	ϵ	**0.25**	**0.3**	**0.35**
	B1-SQR-CF	0.71 [0.19]	0.59 [0.19]	0.54 [0.24]
	B2-SQR-CF	0.80 [0.14]	0.86 [0.09]	0.89 [0.14]
	B1-KNIFE-CF	0.72 [0.11]	0.65 [0.18]	0.61 [0.13]
	B2-KNIFE-CF	0.80 [0.13]	0.74 [0.16]	0.77 [0.19]
	B1-CG-CF	0.75 [0.15]	0.61 [0.21]	0.61 [0.20]
	B2-CG-CF	0.79 [0.14]	0.73 [0.16]	0.76 [0.17]
yacht	ϵ	**0.025**	**0.05**	**0.075**
	B1-SQR-CF	0.44 [0.10]	0.25 [0.23]	0.34 [0.21]
	B2-SQR-CF	0.62 [0.19]	0.37 [0.29]	0.41 [0.27]
	B1-KNIFE-CF	0.31 [0.14]	0.30 [0.22]	0.28 [0.15]
	B2-KNIFE-CF	0.72 [0.24]	0.48 [0.35]	0.50 [0.35]
	B1-CG-CF	0.14 [0.15]	**0.10** [0.11]	**0.10** [0.06]
	B2-CG-CF	0.58 [0.25]	0.33 [0.23]	0.49 [0.32]

Table 10. AUROC for the relative error discrepancy metric for the large datasets. B1, B2 correspond to baseline Algorithms. DV-Y and DV-D correspond to the proposed Algorithms. CF corresponds to the baseline with the conformal correction.

	ϵ [%]	**0.15%**	**0.175%**	**0.2%**
naval	**Avg. % bad**	1.5 [0.6]	0.8 [0.4]	0.5 [0.3]
	B1-SQR	83.7 [5.0]	90.4 [6.1]	89.7 [11.2]
	B2-SQR	94.8 [4.1]	87.6 [7.2]	90.2 [9.7]
	B1-KNIFE	59.0 [10.0]	68.7 [13.5]	62.1 [19.8]
	B2-KNIFE	92.4 [3.8]	95.1 [3.2]	96.7 [2.7]
	B1-CG	89.3 [3.1]	91.2 [2.3]	89.7 [3.2]
	B2-CG	93.0 [1.6]	94.1 [4.1]	95.2 [4.6]
	DV-Y-SQR	83.9 [6.6]	90.4 [7.3]	92.5 [7.3]
	DV-D-SQR	**97.3** [1.8]	**97.0** [2.3]	**98.0** [1.8]
	DV-Y-KNIFE	80.5 [8.8]	83.4 [15.5]	79.9 [17.8]
	DV-D-KNIFE	91.7 [3.8]	93.8 [4.5]	94.2 [6.7]
	DV-Y-CG	91.0 [2.7]	92.0 [1.9]	90.4 [3.4]
	DV-D-CG	92.8 [1.5]	93.6 [4.4]	94.8 [5.6]
power	ϵ [%]	**1%**	**1.25%**	**1.5%**
	Avg. % bad	18.2 [0.8]	10.9 [0.9]	6.6 [0.8]
	B1-SQR	70.1 [2.3]	72.0 [2.7]	70.7 [2.3]
	B2-SQR	71.4 [2.9]	72.3 [2.9]	71.6 [2.6]
	B1-KNIFE	61.5 [2.6]	60.2 [3.2]	59.9 [3.5]
	B2-KNIFE	61.6 [1.9]	62.8 [2.9]	63.1 [3.8]
	B1-CG	59.5 [3.0]	61.0 [4.3]	63.0 [5.1]
	B2-CG	60.4 [1.8]	60.4 [2.1]	60.7 [3.0]
	DV-Y-SQR	70.4 [2.3]	72.5 [2.9]	**72.6** [3.2]
	DV-D-SQR	**71.6** [2.7]	**72.8** [2.5]	72.3 [2.3]
	DV-Y-KNIFE	64.8 [2.3]	65.3 [2.8]	65.4 [3.1]
	DV-D-KNIFE	61.7 [2.1]	62.5 [3.0]	61.6 [4.2]
	DV-Y-CG	59.5 [3.1]	60.9 [4.3]	62.9 [5.2]
	DV-D-CG	61.4 [1.9]	62.1 [2.5]	61.7 [3.4]
kin8nm	ϵ [%]	**10%**	**20%**	**30%**
	Avg. % bad	34.4 [1.3]	12.2 [1.2]	5.2 [0.6]
	B1-SQR	76.1 [2.6]	85.2 [1.3]	89.7 [2.3]
	B2-SQR	74.1 [1.7]	79.5 [2.9]	73.8 [8.8]
	B1-KNIFE	72.2 [2.1]	81.6 [2.2]	87.7 [2.6]
	B2-KNIFE	72.4 [1.8]	80.1 [2.6]	85.6 [4.0]
	B1-CG	76.8 [1.8]	**85.8** [1.7]	**91.2** [1.9]
	B2-CG	75.6 [1.5]	84.6 [1.2]	90.4 [1.8]
	DV-Y-SQR	76.4 [2.5]	85.4 [1.4]	90.6 [1.4]
	DV-D-SQR	74.4 [1.6]	83.8 [1.9]	89.5 [2.5]
	DV-Y-KNIFE	73.4 [1.8]	82.9 [2.2]	88.6 [2.7]
	DV-D-KNIFE	72.6 [1.8]	81.0 [2.1]	86.9 [3.3]
	DV-Y-CG	**77.0** [1.9]	**85.8** [1.8]	**91.2** [2.0]
	DV-D-CG	75.7 [1.5]	84.5 [1.2]	90.2 [1.8]
wine	ϵ [%]	**5%**	**10%**	**15%**
	Avg. % bad	62.6 [4.8]	32.8 [3.6]	17.0 [2.1]
	B1-SQR	82.1 [5.2]	**61.8** [3.8]	65.1 [6.2]
	B2-SQR	62.7 [6.4]	57.2 [3.7]	57.6 [7.5]
	B1-KNIFE	82.7 [5.9]	60.6 [6.3]	69.9 [6.0]
	B2-KNIFE	57.8 [7.8]	56.8 [8.0]	58.8 [8.1]
	B1-CG	51.6 [5.1]	58.3 [5.0]	62.2 [5.1]
	B2-CG	56.3 [7.7]	55.6 [7.4]	57.7 [9.3]
	DV-Y-SQR	84.2 [4.8]	59.6 [3.4]	64.4 [8.4]
	DV-D-SQR	63.0 [6.0]	57.4 [3.5]	56.0 [7.5]
	DV-Y-KNIFE	**84.4** [5.0]	59.3 [8.5]	**70.2** [7.0]
	DV-D-KNIFE	58.5 [7.0]	57.1 [7.9]	60.0 [8.3]
	DV-Y-CG	51.6 [5.1]	58.4 [5.0]	62.2 [5.1]
	DV-D-CG	55.5 [7.5]	55.6 [7.3]	57.7 [9.2]

	ϵ [%]	**0.15%**	**0.175%**	**0.2%**
naval	**Avg. % bad**	1.5 [0.6]	0.8 [0.4]	0.5 [0.3]
	B1-SQR-CF	83.3 [6.5]	90.1 [6.8]	90.7 [8.7]
	B2-SQR-CF	97.2 [1.6]	94.8 [5.9]	93.9 [7.1]
	B1-KNIFE-CF	67.8 [9.2]	72.8 [13.7]	70.1 [15.1]
	B2-KNIFE-CF	91.0 [5.2]	89.4 [11.7]	89.4 [13.7]
	B1-CG-CF	88.6 [4.4]	90.3 [2.6]	88.6 [2.8]
	B2-CG-CF	92.4 [1.9]	92.0 [8.2]	91.4 [10.5]
power	ϵ [%]	**1%**	**1.25%**	**1.5%**
	Avg. % bad	18.2 [0.8]	10.9 [0.9]	6.6 [0.8]
	B1-SQR-CF	70.0 [2.1]	72.1 [2.6]	71.2 [2.1]
	B2-SQR-CF	71.6 [2.8]	72.9 [2.7]	72.2 [2.7]
	B1-KNIFE-CF	61.3 [2.8]	60.3 [3.1]	60.0 [3.5]
	B2-KNIFE-CF	61.5 [2.2]	61.9 [3.2]	60.9 [4.4]
	B1-CG-CF	59.2 [3.1]	60.7 [4.2]	62.8 [5.1]
	B2-CG-CF	60.7 [1.9]	60.4 [2.1]	60.1 [3.0]
kin8nm	ϵ [%]	**10%**	**20%**	**30%**
	Avg. % bad	34.4 [1.3]	12.2 [1.2]	5.2 [0.6]
	B1-SQR-CF	75.9 [2.6]	85.2 [1.2]	90.9 [1.5]
	B2-SQR-CF	73.4 [1.3]	83.5 [2.0]	89.2 [2.5]
	B1-KNIFE-CF	71.8 [2.2]	81.4 [2.2]	87.6 [2.6]
	B2-KNIFE-CF	71.8 [2.1]	80.7 [1.8]	86.0 [3.8]
	B1-CG-CF	76.8 [1.8]	85.7 [1.7]	91.0 [2.0]
	B2-CG-CF	75.6 [1.5]	84.6 [1.2]	90.3 [1.8]
wine	ϵ [%]	**5%**	**10%**	**15%**
	Avg. % bad	62.6 [4.8]	32.8 [3.6]	17.0 [2.1]
	B1-SQR-CF	81.5 [4.7]	60.7 [4.8]	59.5 [7.9]
	B2-SQR-CF	62.9 [6.3]	57.6 [4.2]	56.9 [6.9]
	B1-KNIFE-CF	81.4 [4.3]	60.1 [6.5]	69.4 [5.9]
	B2-KNIFE-CF	58.1 [5.9]	56.0 [7.1]	59.7 [8.6]
	B1-CG-CF	54.8 [5.4]	58.4 [5.2]	61.6 [4.0]
	B2-CG-CF	56.6 [6.0]	55.5 [7.2]	57.6 [9.3]

Table 11. FPR at TPR 90% for the relative error discrepancy metric for the large datasets. B1, B2 correspond to baseline Algorithms. DV-Y and DV-D correspond to the proposed Algorithms. CF corresponds to the baseline with the conformal correction.

	ϵ [%]	**0.15%**	**0.175%**	**0.2%**
naval	B1-SQR	0.43 [0.13]	0.26 [0.17]	0.38 [0.44]
	B2-SQR	0.33 [0.46]	0.80 [0.41]	0.60 [0.51]
	B1-KNIFE	0.88 [0.11]	0.63 [0.28]	0.70 [0.26]
	B2-KNIFE	0.22 [0.13]	0.12 [0.07]	**0.09** [0.08]
	B1-CG	0.23 [0.09]	0.14 [0.04]	0.15 [0.05]
	B2-CG	0.16 [0.03]	0.14 [0.07]	0.11 [0.09]
	DV-Y-SQR	0.45 [0.24]	0.29 [0.25]	0.20 [0.20]
	DV-D-SQR	**0.09** [0.08]	**0.08** [0.07]	**0.08** [0.09]
	DV-Y-KNIFE	0.46 [0.23]	0.35 [0.32]	0.37 [0.28]
	DV-D-KNIFE	0.21 [0.09]	0.14 [0.08]	0.16 [0.18]
	DV-Y-CG	0.19 [0.07]	0.14 [0.03]	0.15 [0.05]
	DV-D-CG	0.17 [0.03]	0.14 [0.08]	0.12 [0.11]
	ϵ [%]	**1%**	**1.25%**	**1.5%**
power	B1-SQR	**0.68** [0.06]	0.67 [0.07]	0.77 [0.13]
	B2-SQR	0.71 [0.06]	0.72 [0.05]	0.94 [0.12]
	B1-KNIFE	0.79 [0.03]	0.81 [0.06]	0.81 [0.05]
	B2-KNIFE	0.80 [0.03]	0.83 [0.04]	0.82 [0.05]
	B1-CG	0.82 [0.03]	0.79 [0.05]	0.76 [0.06]
	B2-CG	0.79 [0.03]	0.80 [0.04]	0.81 [0.05]
	DV-Y-SQR	**0.67** [0.05]	**0.66** [0.05]	**0.67** [0.10]
	DV-D-SQR	0.69 [0.05]	0.69 [0.06]	0.75 [0.05]
	DV-Y-KNIFE	0.76 [0.04]	0.75 [0.07]	0.73 [0.08]
	DV-D-KNIFE	0.79 [0.02]	0.82 [0.03]	0.83 [0.04]
	DV-Y-CG	0.82 [0.02]	0.79 [0.05]	0.77 [0.07]
	DV-D-CG	0.79 [0.02]	0.80 [0.05]	0.82 [0.05]
	ϵ [%]	**10%**	**20%**	**30%**
kin8nm	B1-SQR	0.58 [0.06]	**0.36** [0.04]	0.22 [0.05]
	B2-SQR	0.65 [0.03]	1.00 [0.00]	1.00 [0.00]
	B1-KNIFE	0.64 [0.06]	0.43 [0.06]	0.29 [0.05]
	B2-KNIFE	0.67 [0.03]	0.50 [0.09]	0.35 [0.15]
	B1-CG	**0.57** [0.04]	**0.35** [0.04]	**0.19** [0.05]
	B2-CG	0.60 [0.04]	0.37 [0.04]	0.21 [0.05]
	DV-Y-SQR	**0.57** [0.07]	**0.35** [0.04]	0.22 [0.05]
	DV-D-SQR	0.63 [0.03]	0.40 [0.06]	0.25 [0.08]
	DV-Y-KNIFE	0.63 [0.05]	0.41 [0.07]	0.27 [0.06]
	DV-D-KNIFE	0.67 [0.04]	0.48 [0.08]	0.32 [0.14]
	DV-Y-CG	**0.57** [0.04]	**0.35** [0.04]	**0.19** [0.05]
	DV-D-CG	0.60 [0.04]	0.38 [0.04]	0.21 [0.05]
	ϵ [%]	**5%**	**10%**	**15%**
wine	B1-SQR	0.61 [0.14]	0.91 [0.12]	1.00 [0.00]
	B2-SQR	0.80 [0.09]	0.88 [0.07]	0.93 [0.11]
	B1-KNIFE	0.59 [0.13]	0.79 [0.08]	**0.68** [0.13]
	B2-KNIFE	0.81 [0.08]	0.79 [0.07]	0.81 [0.12]
	B1-CG	0.84 [0.05]	0.83 [0.06]	0.80 [0.09]
	B2-CG	0.82 [0.05]	0.81 [0.07]	0.79 [0.12]
	DV-Y-SQR	0.57 [0.14]	0.85 [0.06]	0.80 [0.11]
	DV-D-SQR	0.81 [0.10]	0.86 [0.03]	0.87 [0.06]
	DV-Y-KNIFE	**0.55** [0.15]	**0.76** [0.11]	**0.68** [0.15]
	DV-D-KNIFE	0.82 [0.08]	0.80 [0.06]	0.78 [0.11]
	DV-Y-CG	0.84 [0.04]	0.83 [0.06]	0.80 [0.08]
	DV-D-CG	0.83 [0.06]	0.82 [0.08]	0.78 [0.11]

	ϵ [%]	**0.15%**	**0.175%**	**0.2%**
naval	B1-SQR-CF	0.49 [0.21]	0.30 [0.23]	0.35 [0.39]
	B2-SQR-CF	**0.10** [0.08]	0.24 [0.40]	0.41 [0.51]
	B1-KNIFE-CF	0.73 [0.19]	0.54 [0.33]	0.52 [0.29]
	B2-KNIFE-CF	0.31 [0.29]	0.33 [0.37]	0.37 [0.45]
	B1-CG-CF	0.23 [0.11]	0.15 [0.04]	0.16 [0.05]
	B2-CG-CF	0.16 [0.04]	0.29 [0.38]	0.35 [0.45]
	ϵ [%]	**1%**	**1.25%**	**1.5%**
power	B1-SQR-CF	0.69 [0.06]	0.69 [0.07]	0.73 [0.08]
	B2-SQR-CF	0.69 [0.06]	0.71 [0.05]	0.76 [0.04]
	B1-KNIFE-CF	0.79 [0.04]	0.82 [0.06]	0.82 [0.05]
	B2-KNIFE-CF	0.79 [0.03]	0.85 [0.04]	0.88 [0.08]
	B1-CG-CF	0.82 [0.02]	0.80 [0.05]	0.78 [0.06]
	B2-CG-CF	0.80 [0.02]	0.83 [0.04]	0.92 [0.08]
	ϵ [%]	**10%**	**20%**	**30%**
kin8nm	B1-SQR-CF	0.58 [0.07]	**0.36** [0.04]	0.21 [0.04]
	B2-SQR-CF	0.66 [0.05]	0.41 [0.06]	0.27 [0.08]
	B1-KNIFE-CF	0.65 [0.06]	0.44 [0.06]	0.29 [0.05]
	B2-KNIFE-CF	0.70 [0.03]	0.49 [0.08]	0.34 [0.14]
	B1-CG-CF	**0.57** [0.04]	0.36 [0.04]	0.20 [0.05]
	B2-CG-CF	0.61 [0.04]	0.38 [0.04]	0.21 [0.05]
	ϵ [%]	**5%**	**10%**	**15%**
wine	B1-SQR-CF	0.63 [0.14]	0.84 [0.09]	0.87 [0.11]
	B2-SQR-CF	0.79 [0.11]	0.85 [0.04]	0.89 [0.09]
	B1-KNIFE-CF	0.61 [0.11]	0.79 [0.09]	0.70 [0.12]
	B2-KNIFE-CF	0.83 [0.07]	0.79 [0.08]	0.84 [0.12]
	B1-CG-CF	0.80 [0.06]	0.83 [0.06]	0.80 [0.08]
	B2-CG-CF	0.86 [0.07]	0.83 [0.07]	0.84 [0.12]

Table 12. AUROC for the absolute error discrepancy metric for the large datasets. B1, B2 correspond to baseline Algorithms. DV-Y and DV-D correspond to the proposed Algorithms.CF corresponds to the baseline with the conformal correction.

	ϵ	**0.175**	**0.2**	**0.225**
	Avg. % bad	2.6 [0.9]	1.4 [0.7]	0.9 [0.4]
naval	B1-SQR	78.2 [8.5]	84.7 [5.4]	89.0 [5.7]
	B2-SQR	93.9 [3.8]	93.9 [4.2]	88.8 [7.2]
	B1-KNIFE	56.9 [8.4]	59.1 [10.7]	66.4 [13.4]
	B2-KNIFE	91.6 [3.5]	93.4 [3.6]	95.2 [2.7]
	B1-CG	88.3 [3.8]	89.2 [3.0]	90.5 [2.0]
	B2-CG	91.7 [1.1]	93.0 [2.0]	94.4 [2.6]
	DV-Y-SQR	79.8 [8.6]	85.1 [6.7]	89.7 [5.3]
	DV-D-SQR	**96.7** [1.3]	**97.3** [1.8]	**96.2** [3.6]
	DV-Y-KNIFE	79.6 [6.6]	80.5 [8.7]	81.6 [13.8]
	DV-D-KNIFE	91.2 [3.8]	93.0 [3.5]	94.4 [3.6]
	DV-Y-CG	90.8 [3.3]	91.1 [2.5]	92.1 [2.3]
	DV-D-CG	91.2 [1.0]	92.7 [2.0]	94.1 [2.7]
	ϵ	**0.3**	**0.35**	**0.4**
	Avg. % bad	14.1 [1.0]	9.2 [1.0]	6.4 [0.7]
power	B1-SQR	70.3 [2.5]	71.4 [2.5]	71.1 [1.6]
	B2-SQR	70.4 [2.6]	70.9 [2.7]	71.1 [3.8]
	B1-KNIFE	60.6 [3.3]	58.8 [3.1]	59.4 [3.3]
	B2-KNIFE	61.8 [2.2]	62.5 [3.0]	63.0 [3.2]
	B1-CG	59.9 [3.9]	60.6 [4.2]	62.5 [5.2]
	B2-CG	59.2 [2.4]	59.5 [2.3]	59.9 [2.6]
	DV-Y-SQR	70.8 [2.7]	**72.2** [2.5]	**72.8** [3.3]
	DV-D-SQR	**71.5** [2.5]	**72.2** [2.7]	72.0 [3.4]
	DV-Y-KNIFE	65.0 [3.2]	64.7 [3.0]	65.3 [3.7]
	DV-D-KNIFE	62.1 [2.8]	62.0 [3.5]	62.4 [3.4]
	DV-Y-CG	59.9 [3.9]	60.6 [4.2]	62.4 [5.2]
	DV-D-CG	60.7 [2.1]	61.1 [2.6]	60.8 [2.6]
	ϵ	**0.6**	**0.7**	**0.8**
	Avg. % bad	4.2 [0.6]	2.2 [0.5]	1.1 [0.3]
kin8nm	B1-SQR	70.5 [3.6]	73.3 [3.5]	71.4 [9.9]
	B2-SQR	52.2 [2.2]	49.9 [0.1]	50.0 [0.0]
	B1-KNIFE	68.9 [5.1]	72.9 [4.8]	78.2 [5.5]
	B2-KNIFE	71.3 [3.7]	70.9 [4.3]	71.1 [8.0]
	B1-CG	**78.6** [2.7]	**81.9** [3.7]	**86.9** [5.4]
	B2-CG	75.9 [3.7]	79.5 [5.2]	83.1 [8.7]
	DV-Y-SQR	71.6 [2.6]	75.1 [4.6]	80.3 [3.8]
	DV-D-SQR	73.5 [3.9]	74.1 [5.6]	77.4 [9.9]
	DV-Y-KNIFE	68.7 [4.9]	74.0 [4.2]	74.9 [5.5]
	DV-D-KNIFE	72.8 [3.9]	74.3 [4.9]	76.0 [9.9]
	DV-Y-CG	**78.6** [2.7]	81.7 [3.8]	86.7 [5.4]
	DV-D-CG	75.8 [3.7]	79.4 [5.4]	82.2 [9.0]
	ϵ	**0.1**	**0.3**	**0.5**
	Avg. % bad	87.4 [4.5]	65.2 [6.1]	46.8 [5.0]
wine	B1-SQR	91.9 [5.1]	84.2 [5.2]	71.3 [3.6]
	B2-SQR	63.5 [7.6]	63.2 [5.9]	59.5 [4.3]
	B1-KNIFE	94.5 [2.5]	86.0 [4.7]	70.3 [4.9]
	B2-KNIFE	62.0 [7.6]	59.0 [7.1]	56.6 [5.8]
	B1-CG	56.0 [8.0]	55.3 [3.5]	57.0 [4.0]
	B2-CG	60.9 [8.8]	56.8 [7.3]	56.6 [6.4]
	DV-Y-SQR	94.3 [4.3]	86.1 [4.3]	**72.2** [3.6]
	DV-D-SQR	66.0 [7.2]	63.2 [5.8]	60.2 [4.7]
	DV-Y-KNIFE	**95.2** [2.1]	**87.3** [3.8]	**72.2** [2.9]
	DV-D-KNIFE	62.2 [6.1]	59.5 [5.5]	57.7 [5.7]
	DV-Y-CG	54.0 [9.1]	55.2 [3.5]	57.0 [4.1]
	DV-D-CG	58.8 [9.9]	56.1 [7.4]	56.8 [6.2]

	ϵ	**0.175**	**0.2**	**0.225**
	Avg. % bad	2.6 [0.9]	1.4 [0.7]	0.9 [0.4]
naval	B1-SQR-CF	77.9 [7.9]	84.3 [7.2]	88.3 [5.9]
	B2-SQR-CF	96.5 [1.5]	**97.3** [1.8]	96.1 [4.7]
	B1-KNIFE-CF	67.3 [8.6]	67.4 [8.7]	70.0 [11.1]
	B2-KNIFE-CF	86.5 [7.5]	87.9 [6.4]	85.3 [10.8]
	B1-CG-CF	87.9 [5.1]	88.6 [4.4]	89.5 [2.0]
	B2-CG-CF	91.1 [1.6]	92.0 [2.5]	93.3 [3.9]
	ϵ	**0.3**	**0.35**	**0.4**
	Avg. % bad	14.1 [1.0]	9.2 [1.0]	6.4 [0.7]
power	B1-SQR-CF	70.3 [2.3]	71.4 [2.1]	71.6 [1.7]
	B2-SQR-CF	71.5 [2.5]	71.7 [2.7]	71.8 [3.3]
	B1-KNIFE-CF	60.5 [3.4]	58.9 [3.0]	59.5 [3.2]
	B2-KNIFE-CF	61.4 [2.5]	59.9 [3.1]	57.6 [4.2]
	B1-CG-CF	59.8 [3.8]	60.5 [4.2]	62.4 [5.2]
	B2-CG-CF	59.1 [2.5]	59.2 [2.3]	58.8 [2.9]
	ϵ	**0.6**	**0.7**	**0.8**
	Avg. % bad	4.2 [0.6]	2.2 [0.5]	1.1 [0.3]
kin8nm	B1-SQR-CF	71.9 [3.0]	76.0 [4.1]	80.7 [3.8]
	B2-SQR-CF	61.4 [7.7]	54.0 [3.9]	49.8 [0.3]
	B1-KNIFE-CF	69.0 [5.1]	73.0 [4.8]	78.8 [4.7]
	B2-KNIFE-CF	64.5 [6.9]	56.9 [8.4]	51.8 [5.9]
	B1-CG-CF	78.5 [2.8]	81.5 [4.1]	86.3 [6.1]
	B2-CG-CF	72.6 [4.8]	68.8 [9.7]	60.3 [9.4]
	ϵ	**0.1**	**0.3**	**0.5**
	Avg. % bad	87.4 [4.5]	65.2 [6.1]	46.8 [5.0]
wine	B1-SQR-CF	91.3 [6.2]	83.8 [5.3]	71.1 [3.2]
	B2-SQR-CF	65.2 [8.4]	62.8 [5.5]	60.8 [5.1]
	B1-KNIFE-CF	92.9 [3.5]	84.5 [4.3]	69.4 [3.6]
	B2-KNIFE-CF	61.8 [5.9]	58.5 [6.3]	57.2 [6.1]
	B1-CG-CF	59.4 [6.9]	57.4 [4.4]	58.3 [4.2]
	B2-CG-CF	60.3 [10.0]	55.8 [6.6]	57.1 [6.2]

Table 13. FPR at TPR 90% for the absolute error discrepancy metric for the large datasets. B1, B2 correspond to baseline Algorithms. DV-Y and DV-D correspond to the proposed Algorithms. CF corresponds to the baseline with the conformal correction.

	ϵ	**0.175**	**0.2**	**0.225**
naval	B1-SQR	0.58 [0.18]	0.45 [0.20]	0.31 [0.12]
	B2-SQR	0.45 [0.48]	0.53 [0.50]	0.80 [0.41]
	B1-KNIFE	0.88 [0.06]	0.87 [0.13]	0.70 [0.25]
	B2-KNIFE	0.22 [0.11]	0.19 [0.13]	**0.12** [0.06]
	B1-CG	0.25 [0.07]	0.23 [0.09]	0.18 [0.07]
	B2-CG	0.20 [0.04]	0.19 [0.05]	0.15 [0.07]
	DV-Y-SQR	0.53 [0.19]	0.45 [0.26]	0.30 [0.13]
	DV-D-SQR	**0.11** [0.05]	**0.09** [0.05]	0.15 [0.18]
	DV-Y-KNIFE	0.44 [0.15]	0.50 [0.28]	0.43 [0.31]
	DV-D-KNIFE	0.22 [0.08]	0.19 [0.10]	**0.12** [0.10]
	DV-Y-CG	0.21 [0.08]	0.19 [0.06]	0.14 [0.04]
	DV-D-CG	0.21 [0.04]	0.20 [0.06]	0.15 [0.07]
	ϵ	**0.3**	**0.35**	**0.4**
power	B1-SQR	0.69 [0.06]	0.69 [0.08]	0.74 [0.11]
	B2-SQR	0.73 [0.06]	0.74 [0.07]	0.94 [0.12]
	B1-KNIFE	0.81 [0.04]	0.82 [0.05]	0.81 [0.05]
	B2-KNIFE	0.81 [0.03]	0.81 [0.05]	0.78 [0.06]
	B1-CG	0.80 [0.03]	0.79 [0.04]	0.77 [0.06]
	B2-CG	0.81 [0.05]	0.80 [0.06]	0.80 [0.05]
	DV-Y-SQR	**0.68** [0.05]	**0.67** [0.07]	**0.67** [0.09]
	DV-D-SQR	**0.68** [0.05]	0.71 [0.06]	0.75 [0.07]
	DV-Y-KNIFE	0.75 [0.07]	0.74 [0.08]	0.75 [0.07]
	DV-D-KNIFE	0.79 [0.03]	0.80 [0.03]	0.82 [0.03]
	DV-Y-CG	0.80 [0.02]	0.79 [0.04]	0.77 [0.06]
	DV-D-CG	0.80 [0.04]	0.80 [0.06]	0.81 [0.05]
	ϵ	**0.6**	**0.7**	**0.8**
kin8nm	B1-SQR	0.83 [0.22]	1.00 [0.00]	0.92 [0.24]
	B2-SQR	1.00 [0.00]	1.00 [0.00]	1.00 [0.00]
	B1-KNIFE	0.64 [0.10]	0.58 [0.13]	0.48 [0.13]
	B2-KNIFE	0.68 [0.11]	0.69 [0.14]	0.64 [0.17]
	B1-CG	**0.48** [0.09]	**0.44** [0.12]	**0.33** [0.20]
	B2-CG	0.58 [0.07]	0.54 [0.17]	0.43 [0.19]
	DV-Y-SQR	0.64 [0.10]	0.61 [0.15]	0.45 [0.12]
	DV-D-SQR	0.62 [0.11]	0.64 [0.18]	0.58 [0.24]
	DV-Y-KNIFE	0.65 [0.11]	0.60 [0.09]	0.58 [0.16]
	DV-D-KNIFE	0.66 [0.07]	0.62 [0.17]	0.51 [0.19]
	DV-Y-CG	**0.48** [0.10]	**0.44** [0.12]	**0.33** [0.21]
	DV-D-CG	0.57 [0.07]	0.55 [0.17]	0.46 [0.22]
	ϵ	**0.1**	**0.3**	**0.5**
wine	B1-SQR	0.16 [0.13]	0.52 [0.11]	0.75 [0.07]
	B2-SQR	0.79 [0.13]	0.78 [0.11]	0.83 [0.08]
	B1-KNIFE	0.14 [0.14]	0.49 [0.14]	**0.73** [0.08]
	B2-KNIFE	0.82 [0.13]	0.78 [0.10]	0.79 [0.06]
	B1-CG	0.82 [0.10]	0.81 [0.05]	0.80 [0.06]
	B2-CG	0.77 [0.13]	0.78 [0.10]	0.80 [0.05]
	DV-Y-SQR	**0.11** [0.13]	0.51 [0.15]	**0.72** [0.06]
	DV-D-SQR	0.78 [0.09]	0.80 [0.08]	0.83 [0.05]
	DV-Y-KNIFE	0.13 [0.12]	**0.46** [0.15]	**0.72** [0.07]
	DV-D-KNIFE	0.80 [0.13]	0.78 [0.10]	0.80 [0.06]
	DV-Y-CG	0.84 [0.10]	0.80 [0.04]	0.80 [0.06]
	DV-D-CG	0.80 [0.10]	0.78 [0.11]	0.80 [0.05]

	ϵ	**0.175**	**0.2**	**0.225**
naval	B1-SQR-CF	0.63 [0.15]	0.49 [0.26]	0.38 [0.22]
	B2-SQR-CF	0.12 [0.05]	**0.09** [0.04]	0.25 [0.40]
	B1-KNIFE-CF	0.72 [0.20]	0.79 [0.16]	0.66 [0.27]
	B2-KNIFE-CF	0.29 [0.12]	0.34 [0.27]	0.44 [0.41]
	B1-CG-CF	0.26 [0.09]	0.24 [0.11]	0.18 [0.06]
	B2-CG-CF	0.21 [0.04]	0.34 [0.35]	0.23 [0.28]
	ϵ	**0.3**	**0.35**	**0.4**
power	B1-SQR-CF	0.70 [0.06]	0.70 [0.08]	0.73 [0.06]
	B2-SQR-CF	**0.69** [0.06]	0.72 [0.06]	0.78 [0.08]
	B1-KNIFE-CF	0.81 [0.03]	0.82 [0.04]	0.83 [0.05]
	B2-KNIFE-CF	0.83 [0.04]	0.86 [0.06]	0.91 [0.07]
	B1-CG-CF	0.80 [0.02]	0.79 [0.05]	0.77 [0.06]
	B2-CG-CF	0.82 [0.05]	0.85 [0.08]	0.96 [0.09]
	ϵ	**0.6**	**0.7**	**0.8**
kin8nm	B1-SQR-CF	0.65 [0.10]	0.61 [0.15]	0.51 [0.18]
	B2-SQR-CF	1.00 [0.00]	1.00 [0.00]	1.00 [0.00]
	B1-KNIFE-CF	0.64 [0.11]	0.58 [0.14]	0.47 [0.09]
	B2-KNIFE-CF	0.96 [0.14]	1.00 [0.00]	1.00 [0.00]
	B1-CG-CF	0.48 [0.10]	0.48 [0.20]	0.41 [0.33]
	B2-CG-CF	1.00 [0.00]	1.00 [0.00]	1.00 [0.00]
	ϵ	**0.1**	**0.3**	**0.5**
wine	B1-SQR-CF	0.16 [0.15]	0.51 [0.13]	0.74 [0.10]
	B2-SQR-CF	0.76 [0.11]	0.78 [0.09]	0.82 [0.05]
	B1-KNIFE-CF	0.22 [0.15]	0.54 [0.10]	0.73 [0.07]
	B2-KNIFE-CF	0.83 [0.15]	0.81 [0.08]	0.81 [0.07]
	B1-CG-CF	0.76 [0.07]	0.77 [0.05]	0.79 [0.06]
	B2-CG-CF	0.78 [0.14]	0.84 [0.09]	0.80 [0.05]

G. Complete Tables of Results

This Appendix contains the full tables of simulations results over all the baseline algorithms and the proposed approaches. There are several interesting observations that can be made from the tables:

- Looking at the baselines we see that there is not a clear winner in terms of the distribution estimation algorithms over the datasets. This is probably because the underlying assumptions of each method may be better adjusted to model specific datasets. In most cases the baselines based on the estimation of the distribution of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ perform better than the counterparts based on $D|\mathbf{X} = \mathbf{x}$ but in some cases this is not the case.
- The DV algorithms in general perform better than their counterpart based on the same conditional estimators. In some cases achieving gains of up to 18.5/100 in average AUROC. This means that DV is capable of substantially improving the detection results in many cases and reduce the variability across dataset splits.

Table 14. AUROC for the absolute error discrepancy metric for the small datasets. B1, B2 correspond to baseline Algorithms 1 and 3 resp. DV-Y and DV-D correspond to the proposed Algorithms 2 and 4 resp.

	ϵ	**0.2**	**0.5**	**0.6**
concrete	**Avg. % bad**	36.7 [5.6]	9.6 [2.8]	6.4 [2.0]
	B1-SQR	73.1 [3.8]	81.1 [8.3]	76.3 [14.2]
	B2-SQR	69.8 [7.3]	59.2 [9.0]	57.3 [9.3]
	B1-KNIFE	63.8 [7.4]	68.7 [5.9]	71.9 [9.6]
	B2-KNIFE	67.7 [6.9]	76.9 [8.9]	80.9 [12.2]
	B1-CG	66.6 [4.7]	77.6 [5.6]	79.8 [7.2]
	B2-CG	69.8 [7.1]	78.7 [9.6]	81.5 [9.1]
	B2-RIO	70.4 [6.9]	75.2 [10.1]	77.2 [9.1]
	DV-Y-SQR	**74.4** [3.4]	**83.5** [8.3]	**82.1** [8.4]
	DV-D-SQR	70.2 [6.9]	79.0 [9.5]	80.5 [9.2]
	DV-Y-KNIFE	63.3 [5.7]	67.2 [12.4]	69.3 [14.8]
	DV-D-KNIFE	67.9 [7.8]	75.2 [8.5]	79.0 [11.3]
	DV-Y-CG	64.2 [7.4]	77.3 [6.2]	77.4 [10.7]
	DV-D-CG	69.9 [6.8]	78.1 [10.2]	81.9 [8.7]
energy	ϵ	**0.025**	**0.0625**	**0.1**
	Avg. % bad	61.2 [5.0]	25.8 [5.2]	9.5 [3.4]
	B1-SQR	68.2 [8.3]	75.8 [7.3]	81.5 [6.6]
	B2-SQR	63.8 [4.8]	73.9 [7.7]	66.5 [15.5]
	B1-KNIFE	67.7 [3.6]	74.7 [6.8]	72.6 [10.2]
	B2-KNIFE	64.8 [7.0]	71.4 [8.3]	72.3 [12.5]
	B1-CG	72.5 [6.8]	81.2 [5.5]	81.7 [5.4]
	B2-CG	64.3 [7.2]	72.1 [7.8]	74.0 [11.3]
	B2-RIO	66.0 [6.1]	69.6 [11.0]	70.8 [12.6]
	DV-Y-SQR	70.3 [6.7]	76.5 [7.6]	81.5 [6.6]
	DV-D-SQR	65.5 [5.6]	73.1 [7.6]	75.4 [11.8]
	DV-Y-KNIFE	68.9 [4.1]	78.3 [7.1]	79.4 [9.3]
	DV-D-KNIFE	65.2 [6.5]	72.1 [8.9]	74.0 [13.3]
	DV-Y-CG	**75.1** [6.0]	**82.0** [6.9]	**82.5** [7.7]
	DV-D-CG	65.0 [7.1]	72.0 [7.8]	74.1 [11.3]
boston	ϵ	**0.25**	**0.3**	**0.35**
	Avg. % bad	31.8 [2.7]	24.1 [5.8]	21.2 [5.4]
	B1-SQR	72.8 [7.2]	77.1 [6.0]	76.8 [5.6]
	B2-SQR	61.0 [7.1]	60.8 [7.4]	57.4 [8.1]
	B1-KNIFE	69.9 [8.7]	73.1 [7.7]	75.7 [6.6]
	B2-KNIFE	62.0 [5.5]	63.5 [6.4]	65.7 [6.5]
	B1-CG	67.5 [6.9]	72.4 [8.6]	73.7 [8.9]
	B2-CG	63.7 [5.0]	63.7 [8.3]	66.0 [8.6]
	B2-RIO	63.8 [8.6]	63.2 [9.0]	64.8 [9.8]
	DV-Y-SQR	**73.7** [7.7]	**77.6** [6.7]	**78.6** [5.4]
	DV-D-SQR	63.3 [7.6]	62.8 [8.7]	64.0 [7.7]
	DV-Y-KNIFE	69.0 [8.5]	72.1 [7.5]	75.5 [6.4]
	DV-D-KNIFE	63.5 [6.8]	63.4 [11.3]	64.4 [11.4]
	DV-Y-CG	65.7 [8.6]	69.7 [11.5]	70.2 [12.3]
	DV-D-CG	63.5 [6.4]	64.2 [9.1]	65.8 [9.4]
yacht	ϵ	**0.025**	**0.05**	**0.075**
	Avg. % bad	62.6 [9.6]	35.8 [11.3]	25.5 [10.6]
	B1-SQR	84.2 [6.8]	90.1 [6.2]	90.9 [5.1]
	B2-SQR	79.1 [10.8]	86.8 [8.3]	89.2 [4.9]
	B1-KNIFE	86.4 [11.4]	89.0 [6.6]	88.7 [6.2]
	B2-KNIFE	73.3 [8.9]	82.0 [9.1]	85.4 [7.9]
	B1-CG	92.4 [4.4]	94.1 [5.7]	96.3 [3.0]
	B2-CG	80.9 [10.3]	84.7 [9.7]	87.8 [7.0]
	B2-RIO	78.5 [7.2]	85.9 [8.0]	89.4 [5.7]
	DV-Y-SQR	86.4 [6.8]	90.4 [6.8]	92.2 [5.3]
	DV-D-SQR	81.6 [7.6]	87.3 [7.4]	90.9 [4.7]
	DV-Y-KNIFE	89.8 [7.9]	87.9 [6.5]	89.0 [5.7]
	DV-D-KNIFE	77.0 [9.3]	83.7 [7.4]	86.4 [6.7]
	DV-Y-CG	**94.3** [3.3]	**94.4** [5.7]	**96.6** [2.9]
	DV-D-CG	79.9 [9.2]	84.6 [9.4]	87.5 [6.8]

Table 15. FPR at TPR 90% for the absolute error discrepancy metric for the small datasets. B1, B2 correspond to baseline Algorithms 1 and 3 resp. DV-Y and DV-D correspond to the proposed Algorithms 2 and 4 resp.

	ϵ	**0.2**	**0.5**	**0.6**
concrete	B1-SQR	**0.67** [0.10]	0.77 [0.39]	0.90 [0.30]
	B2-SQR	0.74 [0.13]	1.00 [0.00]	1.00 [0.00]
	B1-KNIFE	0.76 [0.08]	0.68 [0.16]	0.61 [0.21]
	B2-KNIFE	0.73 [0.13]	0.54 [0.18]	**0.45** [0.27]
	B1-CG	0.69 [0.10]	**0.44** [0.13]	0.49 [0.18]
	B2-CG	**0.68** [0.12]	0.50 [0.22]	0.49 [0.23]
	B2-RIO	0.73 [0.15]	0.62 [0.22]	0.55 [0.25]
	DV-Y-SQR	**0.67** [0.10]	**0.43** [0.21]	0.51 [0.22]
	DV-D-SQR	0.69 [0.13]	0.54 [0.24]	0.52 [0.24]
	DV-Y-KNIFE	0.79 [0.05]	0.66 [0.17]	0.62 [0.19]
	DV-D-KNIFE	0.70 [0.12]	0.64 [0.21]	0.52 [0.25]
	DV-Y-CG	0.73 [0.14]	0.45 [0.14]	0.53 [0.24]
	DV-D-CG	**0.67** [0.12]	0.51 [0.22]	0.49 [0.23]
energy	ϵ	**0.025**	**0.0625**	**0.1**
	B1-SQR	0.74 [0.15]	0.56 [0.18]	0.48 [0.14]
	B2-SQR	0.79 [0.10]	0.63 [0.22]	0.85 [0.32]
	B1-KNIFE	0.69 [0.06]	0.56 [0.13]	0.55 [0.18]
	B2-KNIFE	0.79 [0.07]	0.65 [0.23]	0.54 [0.22]
	B1-CG	0.61 [0.11]	**0.42** [0.14]	**0.39** [0.14]
	B2-CG	0.84 [0.08]	0.63 [0.20]	0.56 [0.21]
	B2-RIO	0.85 [0.11]	0.72 [0.16]	0.67 [0.22]
	DV-Y-SQR	0.72 [0.18]	0.49 [0.17]	0.49 [0.16]
	DV-D-SQR	0.76 [0.08]	0.61 [0.18]	0.57 [0.25]
	DV-Y-KNIFE	0.67 [0.06]	0.49 [0.14]	0.46 [0.20]
	DV-D-KNIFE	0.79 [0.07]	0.64 [0.19]	0.53 [0.24]
	DV-Y-CG	**0.59** [0.10]	**0.42** [0.13]	**0.39** [0.18]
	DV-D-CG	0.83 [0.07]	0.64 [0.19]	0.56 [0.22]
boston	ϵ	**0.25**	**0.3**	**0.35**
	B1-SQR	0.69 [0.17]	0.60 [0.17]	0.64 [0.25]
	B2-SQR	0.94 [0.11]	0.97 [0.09]	1.00 [0.00]
	B1-KNIFE	0.70 [0.14]	0.64 [0.16]	**0.57** [0.12]
	B2-KNIFE	0.79 [0.12]	0.73 [0.17]	0.75 [0.19]
	B1-CG	0.72 [0.14]	0.64 [0.23]	0.66 [0.23]
	B2-CG	0.74 [0.13]	0.69 [0.15]	0.68 [0.16]
	B2-RIO	0.76 [0.12]	0.72 [0.16]	0.73 [0.14]
	DV-Y-SQR	**0.64** [0.21]	**0.53** [0.20]	**0.56** [0.24]
	DV-D-SQR	0.78 [0.09]	0.78 [0.10]	0.73 [0.12]
	DV-Y-KNIFE	0.74 [0.18]	0.68 [0.17]	0.60 [0.16]
	DV-D-KNIFE	0.77 [0.13]	0.69 [0.18]	0.72 [0.20]
	DV-Y-CG	0.75 [0.16]	0.66 [0.26]	0.67 [0.26]
	DV-D-CG	0.77 [0.12]	0.68 [0.17]	0.70 [0.17]
yacht	ϵ	**0.025**	**0.05**	**0.075**
	B1-SQR	0.44 [0.19]	0.21 [0.13]	0.24 [0.16]
	B2-SQR	0.52 [0.21]	0.37 [0.33]	0.35 [0.27]
	B1-KNIFE	0.32 [0.16]	0.27 [0.20]	0.27 [0.16]
	B2-KNIFE	0.69 [0.22]	0.42 [0.30]	0.43 [0.29]
	B1-CG	**0.16** [0.12]	**0.11** [0.12]	0.16 [0.30]
	B2-CG	0.53 [0.22]	0.33 [0.22]	0.32 [0.17]
	B2-RIO	0.60 [0.27]	0.34 [0.27]	0.29 [0.19]
	DV-Y-SQR	0.44 [0.23]	0.19 [0.14]	0.22 [0.16]
	DV-D-SQR	0.49 [0.18]	0.33 [0.23]	0.26 [0.14]
	DV-Y-KNIFE	0.23 [0.16]	0.30 [0.19]	0.29 [0.15]
	DV-D-KNIFE	0.66 [0.27]	0.35 [0.23]	0.36 [0.20]
	DV-Y-CG	**0.16** [0.10]	**0.11** [0.12]	**0.10** [0.07]
	DV-D-CG	0.56 [0.25]	0.32 [0.20]	0.34 [0.17]

Table 16. AUROC for the absolute error discrepancy metric for the large datasets. B1, B2 correspond to baseline Algorithms 1 and 3 resp. DV-Y and DV-D correspond to the proposed Algorithms 2 and 4 resp.

	ϵ	**0.175**	**0.2**	**0.225**
naval	**Avg. % bad**	2.6 [0.9]	1.4 [0.7]	0.9 [0.4]
	B1-SQR	78.2 [8.5]	84.7 [5.4]	89.0 [5.7]
	B2-SQR	93.9 [3.8]	93.9 [4.2]	88.8 [7.2]
	B1-KNIFE	56.9 [8.4]	59.1 [10.7]	66.4 [13.4]
	B2-KNIFE	91.6 [3.5]	93.4 [3.6]	95.2 [2.7]
	B1-CG	88.3 [3.8]	89.2 [3.0]	90.5 [2.0]
	B2-CG	91.7 [1.1]	93.0 [2.0]	94.4 [2.6]
	B2-RIO	90.3 [3.8]	89.6 [3.8]	89.1 [5.1]
	DV-Y-SQR	79.8 [8.6]	85.1 [6.7]	89.7 [5.3]
	DV-D-SQR	**96.7** [1.3]	**97.3** [1.8]	**96.2** [3.6]
	DV-Y-KNIFE	79.6 [6.6]	80.5 [8.7]	81.6 [13.8]
	DV-D-KNIFE	91.2 [3.8]	93.0 [3.5]	94.4 [3.6]
	DV-Y-CG	90.8 [3.3]	91.1 [2.5]	92.1 [2.3]
	DV-D-CG	91.2 [1.0]	92.7 [2.0]	94.1 [2.7]
power	ϵ	**0.3**	**0.35**	**0.4**
	Avg. % bad	14.1 [1.0]	9.2 [1.0]	6.4 [0.7]
	B1-SQR	70.3 [2.5]	71.4 [2.5]	71.1 [1.6]
	B2-SQR	70.4 [2.6]	70.9 [2.7]	71.1 [3.8]
	B1-KNIFE	60.6 [3.3]	58.8 [3.1]	59.4 [3.3]
	B2-KNIFE	61.8 [2.2]	62.5 [3.0]	63.0 [3.2]
	B1-CG	59.9 [3.9]	60.6 [4.2]	62.5 [5.2]
	B2-CG	59.2 [2.4]	59.5 [2.3]	59.9 [2.6]
	B2-RIO	66.6 [5.1]	66.4 [5.9]	66.0 [7.1]
	DV-Y-SQR	70.8 [2.7]	**72.2** [2.5]	**72.8** [3.3]
	DV-D-SQR	**71.5** [2.5]	**72.2** [2.7]	72.0 [3.4]
	DV-Y-KNIFE	65.0 [3.2]	64.7 [3.0]	65.3 [3.7]
	DV-D-KNIFE	62.1 [2.8]	62.0 [3.5]	62.4 [3.4]
	DV-Y-CG	59.9 [3.9]	60.6 [4.2]	62.4 [5.2]
	DV-D-CG	60.7 [2.1]	61.1 [2.6]	60.8 [2.6]
kin8nm	ϵ	**0.6**	**0.7**	**0.8**
	Avg. % bad	4.2 [0.6]	2.2 [0.5]	1.1 [0.3]
	B1-SQR	70.5 [3.6]	73.3 [3.5]	71.4 [9.9]
	B2-SQR	52.2 [2.2]	49.9 [0.1]	50.0 [0.0]
	B1-KNIFE	68.9 [5.1]	72.9 [4.8]	78.2 [5.5]
	B2-KNIFE	71.3 [3.7]	70.9 [4.3]	71.1 [8.0]
	B1-CG	**78.6** [2.7]	**81.9** [3.7]	**86.9** [5.4]
	B2-CG	75.9 [3.7]	79.5 [5.2]	83.1 [8.7]
	B2-RIO	74.7 [3.4]	76.5 [6.0]	77.9 [12.2]
	DV-Y-SQR	71.6 [2.6]	75.1 [4.6]	80.3 [3.8]
	DV-D-SQR	73.5 [3.9]	74.1 [5.6]	77.4 [9.9]
	DV-Y-KNIFE	68.7 [4.9]	74.0 [4.2]	74.9 [5.5]
	DV-D-KNIFE	72.8 [3.9]	74.3 [4.9]	76.0 [9.9]
	DV-Y-CG	**78.6** [2.7]	81.7 [3.8]	86.7 [5.4]
	DV-D-CG	75.8 [3.7]	79.4 [5.4]	82.2 [9.0]
wine	ϵ	**0.1**	**0.3**	**0.5**
	Avg. % bad	87.4 [4.5]	65.2 [6.1]	46.8 [5.0]
	B1-SQR	91.9 [5.1]	84.2 [5.2]	71.3 [3.6]
	B2-SQR	63.5 [7.6]	63.2 [5.9]	59.5 [4.3]
	B1-KNIFE	94.5 [2.5]	86.0 [4.7]	70.3 [4.9]
	B2-KNIFE	62.0 [7.6]	59.0 [7.1]	56.6 [5.8]
	B1-CG	56.0 [8.0]	55.3 [3.5]	57.0 [4.0]
	B2-CG	60.9 [8.8]	56.8 [7.3]	56.6 [6.4]
	B2-RIO	75.1 [10.0]	76.4 [7.9]	71.6 [6.4]
	DV-Y-SQR	94.3 [4.3]	86.1 [4.3]	**72.2** [3.6]
	DV-D-SQR	66.0 [7.2]	63.2 [5.8]	60.2 [4.7]
	DV-Y-KNIFE	**95.2** [2.1]	**87.3** [3.8]	**72.2** [2.9]
	DV-D-KNIFE	62.2 [6.1]	59.5 [5.5]	57.7 [5.7]
	DV-Y-CG	54.0 [9.1]	55.2 [3.5]	57.0 [4.1]
	DV-D-CG	58.8 [9.9]	56.1 [7.4]	56.8 [6.2]

Table 17. FPR at TPR 90% for the absolute error discrepancy metric for the large datasets. B1, B2 correspond to baseline Algorithms 1 and 3 resp. DV-Y and DV-D correspond to the proposed Algorithms 2 and 4 resp.

	ϵ	**0.175**	**0.2**	**0.225**
naval	B1-SQR	0.58 [0.18]	0.45 [0.20]	0.31 [0.12]
	B2-SQR	0.45 [0.48]	0.53 [0.50]	0.80 [0.41]
	B1-KNIFE	0.88 [0.06]	0.87 [0.13]	0.70 [0.25]
	B2-KNIFE	0.22 [0.11]	0.19 [0.13]	**0.12** [0.06]
	B1-CG	0.25 [0.07]	0.23 [0.09]	0.18 [0.07]
	B2-CG	0.20 [0.04]	0.19 [0.05]	0.15 [0.07]
	B2-RIO	0.25 [0.07]	0.27 [0.10]	0.24 [0.12]
	DV-Y-SQR	0.53 [0.19]	0.45 [0.26]	0.30 [0.13]
	DV-D-SQR	**0.11** [0.05]	**0.09** [0.05]	0.15 [0.18]
	DV-Y-KNIFE	0.44 [0.15]	0.50 [0.28]	0.43 [0.31]
	DV-D-KNIFE	0.22 [0.08]	0.19 [0.10]	**0.12** [0.10]
	DV-Y-CG	0.21 [0.08]	0.19 [0.06]	0.14 [0.04]
	DV-D-CG	0.21 [0.04]	0.20 [0.06]	0.15 [0.07]
	ϵ	**0.3**	**0.35**	**0.4**
power	B1-SQR	0.69 [0.06]	0.69 [0.08]	0.74 [0.11]
	B2-SQR	0.73 [0.06]	0.74 [0.07]	0.94 [0.12]
	B1-KNIFE	0.81 [0.04]	0.82 [0.05]	0.81 [0.05]
	B2-KNIFE	0.81 [0.03]	0.81 [0.05]	0.78 [0.06]
	B1-CG	0.80 [0.03]	0.79 [0.04]	0.77 [0.06]
	B2-CG	0.81 [0.05]	0.80 [0.06]	0.80 [0.05]
	B2-RIO	0.75 [0.07]	0.76 [0.08]	0.78 [0.09]
	DV-Y-SQR	**0.68** [0.05]	**0.67** [0.07]	**0.67** [0.09]
	DV-D-SQR	**0.68** [0.05]	0.71 [0.06]	0.75 [0.07]
	DV-Y-KNIFE	0.75 [0.07]	0.74 [0.08]	0.75 [0.07]
	DV-D-KNIFE	0.79 [0.03]	0.80 [0.03]	0.82 [0.03]
	DV-Y-CG	0.80 [0.02]	0.79 [0.04]	0.77 [0.06]
	DV-D-CG	0.80 [0.04]	0.80 [0.06]	0.81 [0.05]
	ϵ	**0.6**	**0.7**	**0.8**
kin8nm	B1-SQR	0.83 [0.22]	1.00 [0.00]	0.92 [0.24]
	B2-SQR	1.00 [0.00]	1.00 [0.00]	1.00 [0.00]
	B1-KNIFE	0.64 [0.10]	0.58 [0.13]	0.48 [0.13]
	B2-KNIFE	0.68 [0.11]	0.69 [0.14]	0.64 [0.17]
	B1-CG	**0.48** [0.09]	**0.44** [0.12]	**0.33** [0.20]
	B2-CG	0.58 [0.07]	0.54 [0.17]	0.43 [0.19]
	B2-RIO	0.60 [0.08]	0.63 [0.17]	0.47 [0.22]
	DV-Y-SQR	0.64 [0.10]	0.61 [0.15]	0.45 [0.12]
	DV-D-SQR	0.62 [0.11]	0.64 [0.18]	0.58 [0.24]
	DV-Y-KNIFE	0.65 [0.11]	0.60 [0.09]	0.58 [0.16]
	DV-D-KNIFE	0.66 [0.07]	0.62 [0.17]	0.51 [0.19]
	DV-Y-CG	**0.48** [0.10]	**0.44** [0.12]	**0.33** [0.21]
	DV-D-CG	0.57 [0.07]	0.55 [0.17]	0.46 [0.22]
	ϵ	**0.1**	**0.3**	**0.5**
wine	B1-SQR	0.16 [0.13]	0.52 [0.11]	0.75 [0.07]
	B2-SQR	0.79 [0.13]	0.78 [0.11]	0.83 [0.08]
	B1-KNIFE	0.14 [0.14]	0.49 [0.14]	**0.73** [0.08]
	B2-KNIFE	0.82 [0.13]	0.78 [0.10]	0.79 [0.06]
	B1-CG	0.82 [0.10]	0.81 [0.05]	0.80 [0.06]
	B2-CG	0.77 [0.13]	0.78 [0.10]	0.80 [0.05]
	B2-RIO	0.55 [0.19]	0.51 [0.14]	0.60 [0.10]
	DV-Y-SQR	**0.11** [0.13]	0.51 [0.15]	**0.72** [0.06]
	DV-D-SQR	0.78 [0.09]	0.80 [0.08]	0.83 [0.05]
	DV-Y-KNIFE	0.13 [0.12]	**0.46** [0.15]	**0.72** [0.07]
	DV-D-KNIFE	0.80 [0.13]	0.78 [0.10]	0.80 [0.06]
	DV-Y-CG	0.84 [0.10]	0.80 [0.04]	0.80 [0.06]
	DV-D-CG	0.80 [0.10]	0.78 [0.11]	0.80 [0.05]

Table 18. AUROC for the relative error discrepancy metric for the small datasets. B1, B2 correspond to baseline Algorithms 1 and 3 resp. DV-Y and DV-D correspond to the proposed Algorithms 2 and 4 resp.

	ϵ [%]	10%	15%	20%
concrete	**Avg. % bad**	35.4 [6.2]	22.1 [5.1]	13.9 [3.4]
	B1-SQR	71.3 [5.3]	74.8 [5.1]	74.6 [6.3]
	B2-SQR	66.5 [4.3]	65.9 [3.8]	62.9 [6.7]
	B1-KNIFE	62.6 [4.2]	63.2 [7.4]	66.0 [6.1]
	B2-KNIFE	61.4 [3.4]	63.8 [5.1]	64.9 [8.4]
	B1-CG	62.9 [5.8]	66.9 [8.0]	70.6 [7.7]
	B2-CG	67.2 [5.6]	69.6 [8.2]	71.3 [11.4]
	B2-RIO	68.0 [4.3]	67.2 [5.3]	69.5 [7.0]
	DV-Y-SQR	**72.2** [5.4]	**75.3** [5.0]	**77.4** [4.6]
	DV-D-SQR	67.8 [4.4]	68.9 [5.9]	69.5 [7.8]
	DV-Y-KNIFE	62.9 [5.0]	63.7 [5.4]	63.6 [6.1]
	DV-D-KNIFE	61.5 [3.7]	62.1 [5.8]	61.6 [6.2]
	DV-Y-CG	62.2 [4.6]	67.4 [7.7]	70.2 [7.5]
	DV-D-CG	67.3 [6.0]	69.5 [8.2]	71.2 [11.3]
energy	ϵ [%]	2%	3.5%	5%
	Avg. % bad	46.8 [3.7]	21.0 [7.0]	10.5 [5.2]
	B1-SQR	71.1 [8.2]	80.5 [6.4]	91.3 [5.9]
	B2-SQR	69.4 [7.6]	74.0 [6.7]	78.4 [9.2]
	B1-KNIFE	73.5 [6.4]	77.2 [6.9]	83.5 [11.6]
	B2-KNIFE	65.4 [5.4]	71.7 [6.3]	78.7 [7.1]
	B1-CG	74.9 [3.9]	78.3 [6.8]	81.9 [12.1]
	B2-CG	68.5 [6.6]	73.6 [7.3]	82.2 [6.9]
	B2-RIO	66.6 [8.2]	74.1 [7.8]	80.9 [9.2]
	DV-Y-SQR	73.1 [7.9]	81.0 [5.9]	**91.5** [5.7]
	DV-D-SQR	68.4 [8.9]	75.3 [6.4]	85.3 [7.4]
	DV-Y-KNIFE	73.5 [6.9]	**81.9** [5.2]	87.4 [8.4]
	DV-D-KNIFE	65.6 [5.4]	72.3 [6.4]	79.3 [7.9]
	DV-Y-CG	**76.3** [3.2]	80.1 [6.6]	89.1 [7.9]
	DV-D-CG	68.3 [6.7]	74.1 [6.8]	82.5 [6.9]
boston	ϵ [%]	10%	15%	20%
	Avg. % bad	33.7 [5.2]	20.8 [5.0]	14.5 [6.3]
	B1-SQR	72.2 [8.0]	78.3 [7.5]	81.9 [11.9]
	B2-SQR	65.1 [10.7]	66.6 [12.8]	65.5 [16.5]
	B1-KNIFE	70.2 [8.1]	75.4 [8.5]	81.2 [12.4]
	B2-KNIFE	67.4 [12.3]	71.5 [13.9]	76.0 [14.6]
	B1-CG	70.8 [9.3]	75.9 [10.3]	80.7 [12.1]
	B2-CG	69.3 [13.2]	74.0 [14.7]	78.1 [15.8]
	B2-RIO	69.3 [8.0]	72.7 [12.4]	79.0 [12.8]
	DV-Y-SQR	**74.0** [7.4]	**79.5** [8.4]	**85.6** [8.2]
	DV-D-SQR	68.5 [7.6]	70.8 [12.3]	76.4 [13.6]
	DV-Y-KNIFE	70.1 [8.2]	75.0 [10.1]	80.9 [14.0]
	DV-D-KNIFE	69.5 [10.4]	72.1 [13.7]	77.7 [13.7]
	DV-Y-CG	69.0 [10.0]	74.4 [10.7]	79.9 [12.6]
	DV-D-CG	69.6 [12.3]	74.2 [13.4]	77.7 [15.5]
yacht	ϵ [%]	10%	15%	20%
	Avg. % bad	59.4 [9.0]	46.1 [9.3]	39.0 [8.9]
	B1-SQR	83.4 [9.9]	87.8 [7.2]	91.9 [5.4]
	B2-SQR	85.7 [7.3]	89.2 [7.0]	92.6 [5.7]
	B1-KNIFE	63.1 [11.5]	67.0 [11.4]	71.9 [10.5]
	B2-KNIFE	83.5 [7.1]	87.1 [5.7]	90.9 [6.3]
	B1-CG	**92.7** [4.4]	**95.7** [3.0]	**97.2** [2.1]
	B2-CG	77.5 [13.3]	82.2 [13.7]	87.1 [15.1]
	B2-RIO	83.1 [5.2]	87.7 [7.5]	92.0 [6.4]
	DV-Y-SQR	86.7 [7.9]	91.5 [5.9]	93.5 [5.3]
	DV-D-SQR	87.2 [6.5]	90.6 [5.8]	93.6 [5.3]
	DV-Y-KNIFE	68.9 [11.1]	68.9 [10.5]	75.5 [9.8]
	DV-D-KNIFE	84.4 [7.0]	87.3 [5.7]	91.1 [6.4]
	DV-Y-CG	92.1 [5.0]	94.6 [3.5]	96.2 [2.9]
	DV-D-CG	79.5 [13.5]	81.5 [17.6]	86.4 [16.9]

Table 19. FPR at TPR 90% for the relative error discrepancy metric for the small datasets. B1, B2 correspond to baseline Algorithms 1 and 3 resp. DV-Y and DV-D correspond to the proposed Algorithms 2 and 4 resp.

	ϵ [%]	10%	15%	20%
concrete	B1-SQR	**0.63** [0.11]	0.62 [0.21]	0.76 [0.32]
	B2-SQR	0.86 [0.15]	0.95 [0.16]	1.00 [0.00]
	B1-KNIFE	0.76 [0.08]	0.79 [0.11]	0.73 [0.10]
	B2-KNIFE	0.75 [0.04]	0.73 [0.08]	0.67 [0.12]
	B1-CG	0.80 [0.04]	0.69 [0.14]	0.63 [0.19]
	B2-CG	0.73 [0.09]	0.64 [0.17]	0.62 [0.23]
	B2-RIO	0.71 [0.13]	0.72 [0.14]	0.70 [0.24]
	DV-Y-SQR	**0.63** [0.11]	**0.56** [0.09]	**0.47** [0.14]
	DV-D-SQR	0.71 [0.14]	0.71 [0.12]	0.69 [0.17]
	DV-Y-KNIFE	0.79 [0.08]	0.76 [0.13]	0.74 [0.14]
	DV-D-KNIFE	0.76 [0.04]	0.74 [0.07]	0.74 [0.10]
	DV-Y-CG	0.79 [0.04]	0.70 [0.12]	0.63 [0.20]
	DV-D-CG	0.72 [0.11]	0.65 [0.16]	0.62 [0.23]
energy	ϵ [%]	2%	3.5%	5%
	B1-SQR	0.76 [0.14]	0.64 [0.20]	0.35 [0.31]
	B2-SQR	0.73 [0.11]	0.69 [0.24]	0.82 [0.38]
	B1-KNIFE	0.68 [0.13]	0.61 [0.14]	0.42 [0.26]
	B2-KNIFE	0.73 [0.13]	0.65 [0.19]	0.48 [0.18]
	B1-CG	**0.62** [0.09]	**0.53** [0.19]	0.40 [0.27]
	B2-CG	0.69 [0.12]	0.60 [0.15]	0.44 [0.25]
	B2-RIO	0.75 [0.16]	0.62 [0.17]	0.48 [0.28]
	DV-Y-SQR	0.68 [0.18]	0.62 [0.21]	**0.29** [0.22]
	DV-D-SQR	0.72 [0.11]	0.60 [0.19]	0.35 [0.22]
	DV-Y-KNIFE	0.67 [0.15]	0.54 [0.19]	0.32 [0.22]
	DV-D-KNIFE	0.75 [0.12]	0.63 [0.19]	0.51 [0.24]
	DV-Y-CG	**0.62** [0.11]	0.54 [0.18]	**0.29** [0.25]
	DV-D-CG	0.71 [0.13]	0.58 [0.16]	0.44 [0.25]
boston	ϵ [%]	10%	15%	20%
	B1-SQR	0.69 [0.21]	0.63 [0.15]	0.62 [0.41]
	B2-SQR	0.88 [0.20]	0.91 [0.21]	0.90 [0.31]
	B1-KNIFE	0.65 [0.11]	0.61 [0.20]	0.46 [0.30]
	B2-KNIFE	0.76 [0.20]	0.66 [0.24]	0.56 [0.26]
	B1-CG	0.65 [0.22]	0.56 [0.28]	0.48 [0.33]
	B2-CG	0.68 [0.22]	0.59 [0.24]	0.48 [0.28]
	B2-RIO	0.71 [0.11]	0.64 [0.24]	0.49 [0.24]
	DV-Y-SQR	**0.63** [0.18]	**0.52** [0.18]	**0.34** [0.15]
	DV-D-SQR	0.73 [0.15]	0.70 [0.15]	0.54 [0.26]
	DV-Y-KNIFE	0.67 [0.14]	0.65 [0.24]	0.50 [0.35]
	DV-D-KNIFE	0.68 [0.17]	0.65 [0.19]	0.51 [0.21]
	DV-Y-CG	0.68 [0.20]	0.61 [0.27]	0.49 [0.33]
	DV-D-CG	0.69 [0.23]	0.61 [0.22]	0.48 [0.25]
yacht	ϵ [%]	10%	15%	20%
	B1-SQR	0.46 [0.27]	0.37 [0.17]	0.23 [0.15]
	B2-SQR	0.41 [0.29]	0.39 [0.37]	0.27 [0.38]
	B1-KNIFE	0.82 [0.13]	0.82 [0.14]	0.77 [0.22]
	B2-KNIFE	0.49 [0.23]	0.32 [0.16]	0.26 [0.25]
	B1-CG	**0.25** [0.21]	**0.17** [0.17]	**0.10** [0.12]
	B2-CG	0.63 [0.27]	0.55 [0.35]	0.30 [0.31]
	B2-RIO	0.48 [0.26]	0.35 [0.22]	0.23 [0.21]
	DV-Y-SQR	0.38 [0.25]	**0.17** [0.10]	0.16 [0.10]
	DV-D-SQR	0.33 [0.19]	0.27 [0.20]	0.20 [0.21]
	DV-Y-KNIFE	0.74 [0.20]	0.84 [0.15]	0.73 [0.25]
	DV-D-KNIFE	0.40 [0.18]	0.37 [0.20]	0.29 [0.30]
	DV-Y-CG	**0.26** [0.21]	0.22 [0.21]	**0.10** [0.11]
	DV-D-CG	0.56 [0.22]	0.35 [0.30]	0.32 [0.36]

Table 20. AUROC for the relative error discrepancy metric for the large datasets. B1, B2 correspond to baseline Algorithms 1 and 3 resp. DV-Y and DV-D correspond to the proposed Algorithms 2 and 4 resp.

	ϵ [%]	**0.15%**	**0.175%**	**0.2%**
naval	**Avg. % bad**	1.5 [0.6]	0.8 [0.4]	0.5 [0.3]
	B1-SQR	83.7 [5.0]	90.4 [6.1]	89.7 [11.2]
	B2-SQR	94.8 [4.1]	87.6 [7.2]	90.2 [9.7]
	B1-KNIFE	59.0 [10.0]	68.7 [13.5]	62.1 [19.8]
	B2-KNIFE	92.4 [3.8]	95.1 [3.2]	96.7 [2.7]
	B1-CG	89.3 [3.1]	91.2 [2.3]	89.7 [3.2]
	B2-CG	93.0 [1.6]	94.1 [4.1]	95.2 [4.6]
	B2-RIO	90.1 [3.9]	88.7 [6.9]	89.8 [5.9]
	DV-Y-SQR	83.9 [6.6]	90.4 [7.3]	92.5 [7.3]
	DV-D-SQR	**97.3** [1.8]	**97.0** [2.3]	**98.0** [1.8]
	DV-Y-KNIFE	80.5 [8.8]	83.4 [15.5]	79.9 [17.8]
	DV-D-KNIFE	91.7 [3.8]	93.8 [4.5]	94.2 [6.7]
	DV-Y-CG	91.0 [2.7]	92.0 [1.9]	90.4 [3.4]
	DV-D-CG	92.8 [1.5]	93.6 [4.4]	94.8 [5.6]
power	ϵ [%]	**1%**	**1.25%**	**1.5%**
	Avg. % bad	18.2 [0.8]	10.9 [0.9]	6.6 [0.8]
	B1-SQR	70.1 [2.3]	72.0 [2.7]	70.7 [2.3]
	B2-SQR	71.4 [2.9]	72.3 [2.9]	71.6 [2.6]
	B1-KNIFE	61.5 [2.6]	60.2 [3.2]	59.9 [3.5]
	B2-KNIFE	61.6 [1.9]	62.8 [2.9]	63.1 [3.8]
	B1-CG	59.5 [3.0]	61.0 [4.3]	63.0 [5.1]
	B2-CG	60.4 [1.8]	60.4 [2.1]	60.7 [3.0]
	B2-RIO	60.9 [2.3]	61.6 [3.0]	60.1 [4.3]
	DV-Y-SQR	70.4 [2.3]	72.5 [2.9]	**72.6** [3.2]
	DV-D-SQR	**71.6** [2.7]	**72.8** [2.5]	72.3 [2.3]
	DV-Y-KNIFE	64.8 [2.3]	65.3 [2.8]	65.4 [3.1]
	DV-D-KNIFE	61.7 [2.1]	62.5 [3.0]	61.6 [4.2]
	DV-Y-CG	59.5 [3.1]	60.9 [4.3]	62.9 [5.2]
	DV-D-CG	61.4 [1.9]	62.1 [2.5]	61.7 [3.4]
kin8nm	ϵ [%]	**10%**	**20%**	**30%**
	Avg. % bad	34.4 [1.3]	12.2 [1.2]	5.2 [0.6]
	B1-SQR	76.1 [2.6]	85.2 [1.3]	89.7 [2.3]
	B2-SQR	74.1 [1.7]	79.5 [2.9]	73.8 [8.8]
	B1-KNIFE	72.2 [2.1]	81.6 [2.2]	87.7 [2.6]
	B2-KNIFE	72.4 [1.8]	80.1 [2.6]	85.6 [4.0]
	B1-CG	76.8 [1.8]	**85.8** [1.7]	**91.2** [1.9]
	B2-CG	75.6 [1.5]	84.6 [1.2]	90.4 [1.8]
	B2-RIO	75.9 [1.7]	**85.8** [1.5]	**91.2** [1.9]
	DV-Y-SQR	76.4 [2.5]	85.4 [1.4]	90.6 [1.4]
	DV-D-SQR	74.4 [1.6]	83.8 [1.9]	89.5 [2.5]
	DV-Y-KNIFE	73.4 [1.8]	82.9 [2.2]	88.6 [2.7]
	DV-D-KNIFE	72.6 [1.8]	81.0 [2.1]	86.9 [3.3]
	DV-Y-CG	**77.0** [1.9]	**85.8** [1.8]	**91.2** [2.0]
	DV-D-CG	75.7 [1.5]	84.5 [1.2]	90.2 [1.8]
wine	ϵ [%]	**5%**	**10%**	**15%**
	Avg. % bad	62.6 [4.8]	32.8 [3.6]	17.0 [2.1]
	B1-SQR	82.1 [5.2]	**61.8** [3.8]	65.1 [6.2]
	B2-SQR	62.7 [6.4]	57.2 [3.7]	57.6 [7.5]
	B1-KNIFE	82.7 [5.9]	60.6 [6.3]	69.9 [6.0]
	B2-KNIFE	57.8 [7.8]	56.8 [8.0]	58.8 [8.1]
	B1-CG	51.6 [5.1]	58.3 [5.0]	62.2 [5.1]
	B2-CG	56.3 [7.7]	55.6 [7.4]	57.7 [9.3]
	B2-RIO	77.5 [5.8]	70.0 [5.9]	67.6 [6.1]
	DV-Y-SQR	84.2 [4.8]	59.6 [3.4]	64.4 [8.4]
	DV-D-SQR	63.0 [6.0]	57.4 [3.5]	56.0 [7.5]
	DV-Y-KNIFE	**84.4** [5.0]	59.3 [8.5]	**70.2** [7.0]
	DV-D-KNIFE	58.5 [7.0]	57.1 [7.9]	60.0 [8.3]
	DV-Y-CG	51.6 [5.1]	58.4 [5.0]	62.2 [5.1]
	DV-D-CG	55.5 [7.5]	55.6 [7.3]	57.7 [9.2]

Table 21. FPR at TPR 90% for the relative error discrepancy metric for the large datasets. B1, B2 correspond to baseline Algorithms 1 and 3 resp. DV-Y and DV-D correspond to the proposed Algorithms 2 and 4 resp.

	ϵ [%]	**0.15%**	**0.175%**	**0.2%**
naval	B1-SQR	0.43 [0.13]	0.26 [0.17]	0.38 [0.44]
	B2-SQR	0.33 [0.46]	0.80 [0.41]	0.60 [0.51]
	B1-KNIFE	0.88 [0.11]	0.63 [0.28]	0.70 [0.26]
	B2-KNIFE	0.22 [0.13]	0.12 [0.07]	**0.09** [0.08]
	B1-CG	0.23 [0.09]	0.14 [0.04]	0.15 [0.05]
	B2-CG	0.16 [0.03]	0.14 [0.07]	0.11 [0.09]
	B2-RIO	0.25 [0.08]	0.24 [0.16]	0.22 [0.14]
	DV-Y-SQR	0.45 [0.24]	0.29 [0.25]	0.20 [0.20]
	DV-D-SQR	**0.09** [0.08]	**0.08** [0.07]	**0.08** [0.09]
	DV-Y-KNIFE	0.46 [0.23]	0.35 [0.32]	0.37 [0.28]
	DV-D-KNIFE	0.21 [0.09]	0.14 [0.08]	0.16 [0.18]
	DV-Y-CG	0.19 [0.07]	0.14 [0.03]	0.15 [0.05]
	DV-D-CG	0.17 [0.03]	0.14 [0.08]	0.12 [0.11]
	ϵ [%]	**1%**	**1.25%**	**1.5%**
power	B1-SQR	**0.68** [0.06]	0.67 [0.07]	0.77 [0.13]
	B2-SQR	0.71 [0.06]	0.72 [0.05]	0.94 [0.12]
	B1-KNIFE	0.79 [0.03]	0.81 [0.06]	0.81 [0.05]
	B2-KNIFE	0.80 [0.03]	0.83 [0.04]	0.82 [0.05]
	B1-CG	0.82 [0.03]	0.79 [0.05]	0.76 [0.06]
	B2-CG	0.79 [0.03]	0.80 [0.04]	0.81 [0.05]
	B2-RIO	0.79 [0.04]	0.81 [0.03]	0.83 [0.04]
	DV-Y-SQR	**0.67** [0.05]	**0.66** [0.05]	**0.67** [0.10]
	DV-D-SQR	0.69 [0.05]	0.69 [0.06]	0.75 [0.05]
	DV-Y-KNIFE	0.76 [0.04]	0.75 [0.07]	0.73 [0.08]
	DV-D-KNIFE	0.79 [0.02]	0.82 [0.03]	0.83 [0.04]
	DV-Y-CG	0.82 [0.02]	0.79 [0.05]	0.77 [0.07]
	DV-D-CG	0.79 [0.02]	0.80 [0.05]	0.82 [0.05]
	ϵ [%]	**10%**	**20%**	**30%**
kin8nm	B1-SQR	0.58 [0.06]	**0.36** [0.04]	0.22 [0.05]
	B2-SQR	0.65 [0.03]	1.00 [0.00]	1.00 [0.00]
	B1-KNIFE	0.64 [0.06]	0.43 [0.06]	0.29 [0.05]
	B2-KNIFE	0.67 [0.03]	0.50 [0.09]	0.35 [0.15]
	B1-CG	**0.57** [0.04]	**0.35** [0.04]	**0.19** [0.05]
	B2-CG	0.60 [0.04]	0.37 [0.04]	0.21 [0.05]
	B2-RIO	0.62 [0.06]	0.36 [0.04]	0.21 [0.05]
	DV-Y-SQR	**0.57** [0.07]	**0.35** [0.04]	0.22 [0.05]
	DV-D-SQR	0.63 [0.03]	0.40 [0.06]	0.25 [0.08]
	DV-Y-KNIFE	0.63 [0.05]	0.41 [0.07]	0.27 [0.06]
	DV-D-KNIFE	0.67 [0.04]	0.48 [0.08]	0.32 [0.14]
	DV-Y-CG	**0.57** [0.04]	**0.35** [0.04]	**0.19** [0.05]
	DV-D-CG	0.60 [0.04]	0.38 [0.04]	0.21 [0.05]
	ϵ [%]	**5%**	**10%**	**15%**
wine	B1-SQR	0.61 [0.14]	0.91 [0.12]	1.00 [0.00]
	B2-SQR	0.80 [0.09]	0.88 [0.07]	0.93 [0.11]
	B1-KNIFE	0.59 [0.13]	0.79 [0.08]	**0.68** [0.13]
	B2-KNIFE	0.81 [0.08]	0.79 [0.07]	0.81 [0.12]
	B1-CG	0.84 [0.05]	0.83 [0.06]	0.80 [0.09]
	B2-CG	0.82 [0.05]	0.81 [0.07]	0.79 [0.12]
	B2-RIO	0.51 [0.12]	0.62 [0.09]	0.62 [0.10]
	DV-Y-SQR	0.57 [0.14]	0.85 [0.06]	0.80 [0.11]
	DV-D-SQR	0.81 [0.10]	0.86 [0.03]	0.87 [0.06]
	DV-Y-KNIFE	**0.55** [0.15]	**0.76** [0.11]	**0.68** [0.15]
	DV-D-KNIFE	0.82 [0.08]	0.80 [0.06]	0.78 [0.11]
	DV-Y-CG	0.84 [0.04]	0.83 [0.06]	0.80 [0.08]
	DV-D-CG	0.83 [0.06]	0.82 [0.08]	0.78 [0.11]