Characterizing Risk Regimes for Safe Deployment of Deep Regression Models

Jayaraman J.Thiagarajan [1] **Vivek Narayanaswamy** [1] **Puja Trivedi** [2] **Rushil Anirudh** [1]

Abstract

To ensure the safe deployment of AI models, it is crucial to identify potential failure modes to prevent costly errors. While failure detection in classification problems has received significant attention, characterizing failure or risk in regression is more complex and less explored. In this paper, we propose a new framework to characterize risk regimes in regression models. Our framework leverages the principle of anchoring to estimate both uncertainties and non-conformity scores, that can be used to jointly categorize samples into distinct risk regimes, thus enabling a fine-grained analysis of model failure. Additionally, we introduce a suite of metrics for evaluating such failure detectors in regression settings. Our results on synthetic and real-world benchmarks demonstrate the effectiveness of our framework over existing methods that rely solely on predictive uncertainties or feature inconsistency to assess risk.

1. Introduction

Ensuring the safe deployment of AI models necessitates proactive detection of potential failure modes to prevent costly errors. In classification tasks, this is framed as generalization gap prediction, where the goal is to estimate the expected deviation in model accuracy between an unlabeled test set and a controlled validation set (Guillory et al., 2021; Narayanaswamy et al., 2022; Baek et al., 2022; Chen et al., 2021; Jiang et al., 2019; Deng & Zheng, 2021). In contrast, this paper focuses on detecting failures in deep regression models, motivated by their significance in various critical domains such as healthcare (Luo et al., 2022; Young et al., 2020) and physical sciences (Raissi et al., 2019). Characterizing failure in regression tasks is inherently complex due to the subjective nature of failure and the variation in error tolerances across different use cases.

Traditionally, predictive uncertainty (Lakshminarayanan et al., 2017; Gal & Ghahramani, 2016; He et al., 2020; Amini et al., 2020) is considered as a meaningful surrogate for model risk. However, relying solely on uncertainty for failure detection can be misleading, as low uncertainty regions can still exhibit higher risk due to feature heterogeneity in the training data (Seedat et al., 2022). Furthermore, data regimes outside the training support that have high uncertainty can offer low risk if the model accurately extrapolates. Figure 1 demonstrates the weak correlation between uncertainty and true risk using a simple 1D function with different experimental designs. On the other hand, Seedat *et al.* (Seedat et al., 2022), recently proposed a task-agnostic approach to identify failure modes based on feature inconsistency compared to the training distribution. However, given the task-agnostic nature of this approach, it can be ineffective for arbitrary target functions.

In this paper, we introduce a novel framework for characterizing failure in deep regression models. Our approach organizes samples from a test set into different risk regimes, such as ID (in distribution), Low Risk, Moderate Risk, and High Risk. We use a unified anchoring-based approach to estimate uncertainties as well as non-conformity scores, that measures sample adherence to the training data manifold (Thiagarajan et al., 2022; Netanyahu et al., 2023). Figure 1 highlights the discrepancy between true and predicted risks across different risk regimes. Our framework outperforms state-of-the-art uncertainty-based and inconsistency-based detectors in aligning risk regimes with the true risk. Finally, we introduce novel evaluation metrics and demonstrate the effectiveness of our framework in identifying generalization, out-of-distribution, and out-of-support regimes using synthetic and real-world benchmarks.

[1]Lawrence Livermore National Laboratories, Livermore, CA, USA [2]University of Michigan, USA. Correspondence to: Jayaraman J.Thiagarajan <jjayaram@llnl.gov>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, Lawrence Livermore National Security, LLC.and was supported by the LLNL-LDRD Program under Project No. 22-SI-004 with IM release number LLNL-CONF-850800.

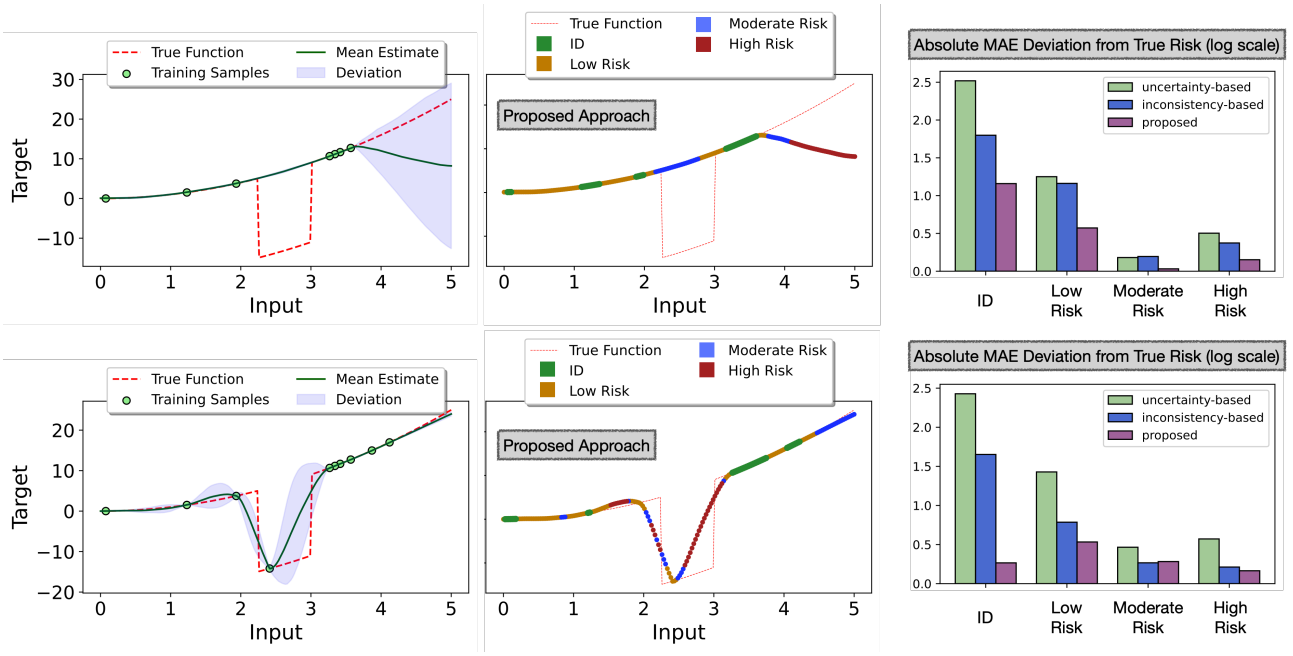

Figure 1. **While predictive uncertainty is necessary to estimate risk, it is insufficient to fully characterize all risk regimes.** Top: Out-of-support (OOS) samples in the range of $[2.2 - 2.7]$ exhibit low uncertainty but moderate model risk due to significant deviation from true function. Bottom: Even with better experiment designs, uncertainty alone in the extrapolating regime $[4.5 - 5]$ is unreliable due to potential drift from the truth. Our proposed framework leverages anchoring (Thiagarajan et al., 2022) to unify prediction uncertainty and non-conformity to the training manifold. It effectively identifies Moderate Risk regimes (highlighted in blue) and outperforms existing baselines in accurately categorizing samples into appropriate risk categories, as indicated by lower MAE.

2. Background

Preliminaries. Let F_θ parameterized by θ be a predictive model trained on a labeled dataset $\mathcal{D} = \{(\mathrm{x}_i, \mathrm{y}_i)\}_{i=1}^{M}$ with M samples. Note, each input $\mathrm{x}_i \in X$ and label $\mathrm{y}_i \in y$ belong to the spaces of inputs X (in $d-$dimensions) and continuous-valued targets y respectively. Given a non-negative loss function $\mathcal{L}$, the sample-level risk of a predictor can be defined as $\mathrm{R}(\mathrm{x}; F_\theta) = \mathbb{E}_{y|\mathrm{x}}\, \mathcal{L}(\mathrm{y}, F_\theta(\mathrm{x}))$. Basically, risk is defined as the cost incurred for incorrect predictions. Estimating true risk is challenging in practice without access to the unknown joint distribution $P(X, y)$. Therefore, developing reliable methods to flag and categorize different risk regimes is crucial for safe model deployment. We now define the different regimes of generalization that we want to characterize: (i) *In-distribution*: This is the scenario where $P(\mathrm{x}_t \in X) > 0$ and $P(\mathrm{x}_t \in \mathcal{D}) > 0$, i.e., there is likelihood for observing the test sample in the training dataset; (ii) *Out-of-Support* (OOS): The scenario where $P(\mathrm{x}_t \in X) > 0$ but $P(\mathrm{x}_t \in \mathcal{D}) = 0$, i.e., the train and test sets have different supports, even though they are drawn from the same space; (iii) *Out-of-Distribution* (OOD): This is the scenario where $P(\mathrm{x}_t \in X) = 0$, i.e., the input spaces for train and test data are disjoint.

Anchoring in Predictive Models. Anchoring is a principle for deep model training that involves reparameterizing an input (query) sample x into a tuple with an anchor r drawn $\mathcal{D}$ and the residual $\Delta\mathrm{x}$, denoted as $[\mathrm{r}, \Delta\mathrm{x}] = [\mathrm{r}, \mathrm{x} - \mathrm{r}]$. Anchoring establishes a relationship between x and r, inducing a joint distribution dependent on both $P(X)$ and the distribution of residuals $P(\Delta)$. During training, anchoring ensures prediction consistency by modeling the combinatorial relationship between every sample in $\mathcal{D}$ and infers the joint distribution $P(X, \Delta)$. During inference, accurate predictions are obtained when $\mathrm{x}_t \in P(X)$ and $[\mathrm{x}_t - \mathrm{r}] \in P(\Delta)$. However, inconsistent predictions occur when the query x_t is OOS/OOD or produces an unseen residual $\Delta\mathrm{x}_t$, which can in turn shed light into model pitfalls. Anchoring has found to be effective for various tasks. For example, Netanyahu et al. (Netanyahu et al., 2023) employed a transduction procedure with anchored models to produce accurate predictions for OOS samples. The key insight was to transform an OOS extrapolation problem into the task of determining an anchor r such that the residual $\Delta\mathrm{x}_t = [\mathrm{r} - \mathrm{x}_t]$ belongs to $P(\Delta)$. On the other hand, Thiagarajan and Anirudh (Thiagarajan et al., 2022) considered anchoring as a means of injecting trivial shifts in the dataset, leading to non-trivial variation in the predictions and enabling uncertainty quantification. In our framework, we consider these anchoring viewpoints and estimate both uncertainty and non-conformity to the

training manifold to characterize failure.

3. Proposed Approach

Generalization gap predictors in the classification setting aim to estimate the correctness of the predicted labels as proxies for failure indication. However, when it comes to regression, defining failure becomes complex due to varying levels of acceptable error tolerances in different scenarios. To address this, we propose a novel framework for systematically characterizing failure in deep regression models. We categorize unlabeled samples from a test set into different risk regimes, namely (*low, moderate, high*) based on their expected levels of risk. Furthermore, we make a significant advancement in the estimation of sample-level risk, which has been a challenging problem. Existing approaches rely on predictive uncertainties or task-agnostic data inconsistency to approximate risk. In contrast, we leverage the principle of anchoring in predictive models (Thiagarajan et al., 2022; Netanyahu et al., 2023), to integrate both predictive uncertainty and non-conformity to the training data manifold which are then used to derive the risk regimes. Notably, our framework eliminates the need for separate estimators for uncertainties and the proposed scores, and it does not require additional calibration data. Finally, we introduce a suite of evaluation metrics to enable a comprehensive evaluation of failure detection methods in deep regression models.

3.1. A Novel Framework for Failure Characterization in Regression Models

Given the challenge of accurately estimating sample-level errors, especially in extrapolation scenarios (OOS or OOD), a more flexible approach is to analyze groups of samples with varying levels of expected risk. Although predictive uncertainties are commonly used to identify sampling deficiencies, Figure 1 shows that uncertainty estimates do not always correlate with true risk. This renders uncertainty-based failure detectors, such as DEUP (Lahlou et al., 2023), ineffective in practice. The reason is that risk can stem from various sources, only some of which are captured by such uncertainties. While uncertainty can to an extent capture failures related to OOS or OOD test samples, larger errors can occur when $(\mathrm{x}_t, \mathrm{y}_t) \notin P(X, y)$, meaning that high risk can arise regardless of the uncertainty on x_t if the test sample (and its unknown target) deviates from the data manifold. Therefore, we advocate for complementary scores that can quantify this non-conformity.

We now describe our failure characterization framework for deep regression models (Figure 2). Broadly, we categorize the set of test samples based on both uncertainty and non-conformity. We accomplish this by dividing the scores into three bins: *low*, *moderate*, and *high*, determined by the conditional quantile ranges $[0, 25]$, $[25, 75]$, and $[75, 100]$ respectively. This categorization allows us to create meaningful partitions of the test data into risk regimes. We assume a typical test set contains samples that are close to the training distribution, as well as OOS and potential OOD samples. Even when this assumption is not valid, and there are no distribution shifts in the test set, our framework can still identify regimes with increasing levels of expected risk.

ID (■): The model generalizes in this regime and is expected to produce the lowest prediction error. In our framework, this corresponds to samples with low uncertainty and low/moderate non-conformity scores;

Low Risk (■): Even when the uncertainty is low, the model can produce higher error than the ID samples, when there is incongruity (e.g., samples within a neighborhood having different target values). Similarly, for OOS samples that are associated with moderate uncertainties, the model can still extrapolate well and produce reduced risk. Hence, we define this regime as the collection of (low uncertainty, high non-conformity) and (moderate uncertainty, low/moderate non-conformity) samples;

Moderate Risk (■): Since predictive uncertainties can be inherently miscalibrated, OOS samples, which the model cannot extrapolate to, can be associated with moderate uncertainties. On the other hand, the model could reasonably generalize to OOD samples that are flagged with high uncertainties. Hence, we define this regime as the collection of (moderate uncertainty, high non-conformity) and (high uncertainty, low/moderate non-conformity) samples;

High Risk (■): When both the uncertainty and non-conformity scores are high, there is no evidence that the model will behave predictably on those samples. In practice, this can correspond to both OOS and OOD samples.

3.2. Uncertainty Estimation via Anchoring

Our failure analysis framework focuses on measuring uncertainties and non-conformity in situations where the true label is unknown. While there are various methods available for estimating predictive uncertainty, accurately measuring non-conformity has proven to be a challenge. In the context of regression problems, existing approaches for characterizing non-conformity include auto-encoding error-based scoring in DataSUITE (Seedat et al., 2022) and feature conformal prediction (Teng et al., 2023). The former utilizes an auto-encoder trained on a calibration dataset to calculate the score, making it applicable to different tasks. However, the latter relies on ground truth labels from a calibration set and applies a conformal interval prediction approach, which is not suitable for our scenario. Both methods exhibit limited performance when the calibration dataset fails to adequately capture the expected shifts during testing. To

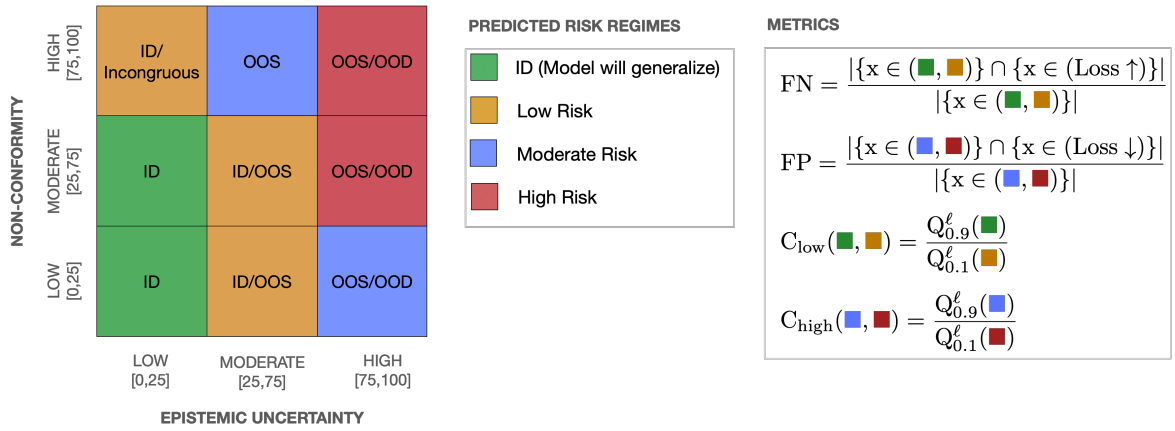

Figure 2. **Overview of our proposed framework.** We organize test examples into bins (*low*, *moderate* and *high*) using both predictive uncertainty and scores of non-conformity to the training manifold. With such a categorization, our framework associates samples into different levels of expected risk (ID, Low Risk, Moderate Risk and High Risk). We also develop a suite of metrics to assess the efficacy of our failure detector.

address these limitations, we propose a unified approach based on anchored neural networks.

As introduced in Section 2, an anchored model is trained by transforming a training sample x into a tuple, $(\mathrm{r}, \mathrm{x} - \mathrm{r})$ based on an anchor r, which is also drawn randomly from the training dataset $\mathcal{D}$. Building upon the findings from (Thiagarajan et al., 2022), at test time, predictions from different anchor choices can be used to obtain the mean and uncertainty estimates as follows:

$$\mu(y_t|\mathrm{x}_t) = \frac{1}{K}\sum_{k=1}^{K} F([\mathrm{r}_k, \mathrm{x}_t - \mathrm{r}_k]);$$
$$\sigma(y_t|\mathrm{x}_t) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(F([\mathrm{r}_k, \mathrm{x}_t - \mathrm{r}_k]) - \mu)^2}, \quad (1)$$

where μ and σ are estimated by marginalizing across K anchors $\{\mathrm{r}_k\}_{k=1}^{K}$ sampled from $\mathcal{D}$.

3.3. Measuring Non-conformity via Inverse Anchoring

Turning our attention to the assessment of non-conformity, we make a noteworthy observation regarding the flexibility of an anchored neural network. It is able to not only capture the relative representation of a test sample in relation to an anchor i.e., (*anchor-centric*), but also the reverse scenario i.e., (*query-centric*). To elaborate, the prediction for an anchor r is given as $F([\mathrm{x}_t, \mathrm{r} - \mathrm{x}_t])$, where x_t represents a test sample. Since the ground truth function value is known for the training samples, we can measure the non-conformity score for a query sample based on its ability to accurately recover the target of the anchor.

From an alternative perspective, the original anchor-centric model (Thiagarajan et al., 2022) ensures accurate predictions for an input $[\mathrm{r}, \Delta]$ only when $\mathrm{r} \in \mathcal{D}$ and $\Delta \in P(\Delta)$. However, for out-of-distribution (OOD) or out-of-sample (OOS) samples, where $\Delta \notin P(\Delta)$, the estimated uncertainty becomes large rendering it unreliable for ranking by expected risk levels. In contrast, our proposed query-centric score addresses this issue by directly quantifying the discrepancy relative to the ground truth target. Specifically, we define our non-conformity score as follows:

$$\texttt{Score}_1(\mathrm{x}) = \max_{r \in \mathcal{D}} \left\| \mathrm{y}_\mathrm{r} - F([\mathrm{x}, \mathrm{r} - \mathrm{x}]) \right\|_1 \quad (2)$$

It is important to note that we measure the largest discrepancy across all training samples in the training dataset. In practice, this can be done for a small batch of randomly selected training samples (e.g., 100). As demonstrated in our results, our proposed non-conformity approach proves highly effective compared to state-of-the-art uncertainty-based and inconsistency-based failure detectors (refer to Figure 1).

Better Resolving Regimes of Medium and High Risk Upon closer examination of Equation (2), it becomes evident that samples located far from the training manifold can exhibit uniformly poor model predictions (i.e., extrapolation), as both $\mathrm{x} \notin \mathcal{D}$ and $\Delta \notin P(\Delta)$. Consequently, distinguishing between samples with moderate risk and those with high risk becomes exceptionally challenging. To address this issue, we propose an approach inspired by the bilinear transduction procedure presented in (Netanyahu et al., 2023). However, a key distinction between the two approaches lies in the fact that, due to our query-centric formulation, both the query x and Δ must be in-distribution to ensure the reliable prediction of the target for the anchor by the anchored model F. We achieve this through the

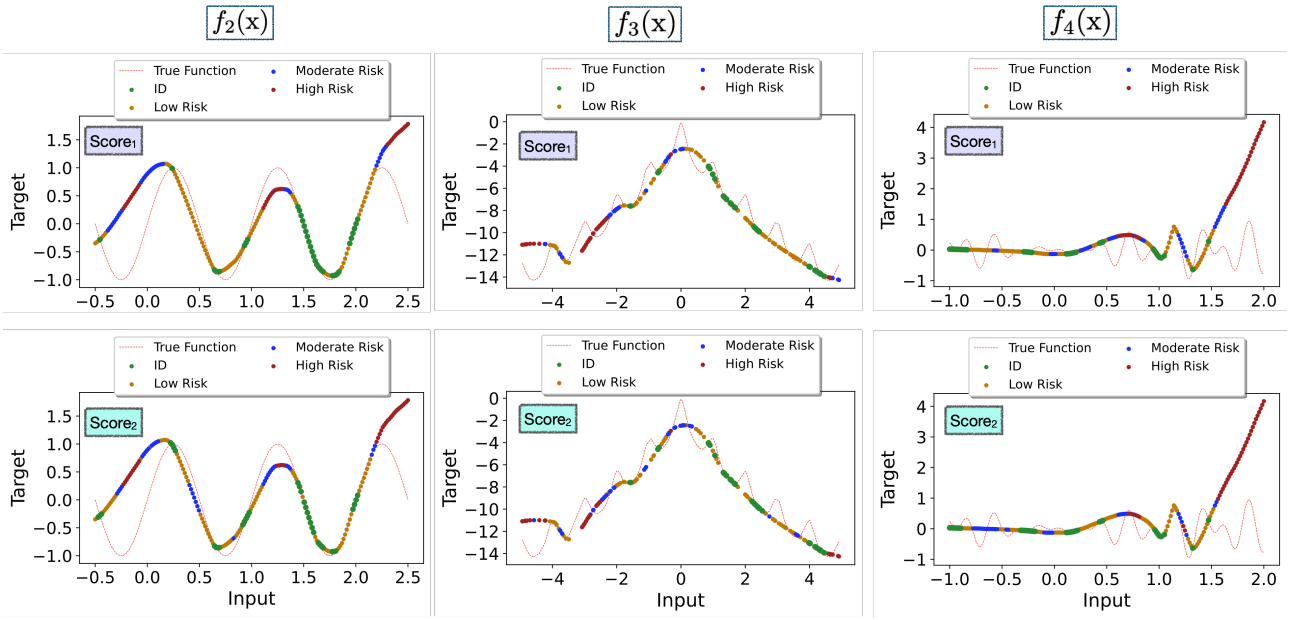

Figure 3. **Risk regimes identified using our framework for 1D regression functions.** By unifying predictive uncertainty and non-conformity to the training manifold, as measured by $\texttt{Score}_1$ or $\texttt{Score}_2$, our approach accurately characterizes the risk regimes and maintains a well-calibrated transition between them across different functions in the input space.

following optimization problem:

$$\texttt{Score}_2(\mathrm{x}) = \max_{r \in \mathcal{D}} \left\| \mathrm{x} - \arg\min_{\bar{\mathrm{x}}} \left(\left\| \mathrm{y_r} - F([\bar{\mathrm{x}}, \mathrm{r} - \bar{\mathrm{x}}]) \right\|_1 + \lambda \mathcal{R}(\bar{\mathrm{x}}) \right) \right\|_2, \tag{3}$$

where $\mathcal{R}(\bar{\mathrm{x}}) = \left\| \bar{\mathrm{x}} - A([\mathrm{x}, \bar{\mathrm{x}} - \mathrm{x}]) \right\|_2 + \left\| \mathrm{x} - A([\bar{\mathrm{x}}, \mathrm{x} - \bar{\mathrm{x}}]) \right\|_2$

In this approach, the score is measured as the discrepancy in the input space to a new fictitious sample that serves as an intermediate anchor, such that its prediction matches the known prediction on the training sample. In other words, we optimize the modification of the query sample x to $\bar{\mathrm{x}}$ in such a way that we accurately match the true target for the anchor r. The non-conformity is then quantified as the distance traversed in the input space to match the target. To ensure that the resulting $\bar{\mathrm{x}}$ remains within the input data manifold, we incorporate a regularizer $\mathcal{R}(\bar{\mathrm{x}})$. Specifically, we train an anchored auto-encoder A on the training dataset $\mathcal{D}$ and enforce cyclical consistency, where A is required to recover x using $\bar{\mathrm{x}}$ as the anchor and vice versa. While $\texttt{Score}_1$ is extremely scalable, $\texttt{Score}_2$ provides better resolution in the medium and high risk regimes at an increased compute cost. In general, the choice of the non-conformity score is determined by the constraints and risk tolerance in different applications.

3.4. Evaluation Metrics

Existing studies (Lahlou et al., 2023) focused on reporting the Spearman correlation between true risk and predicted risk, while DataSUITE measured average error in top inconsistent samples. However, these metrics fail to provide a comprehensive understanding of failure detectors across various risk regimes. To address this limitation, we introduce a new suite of metrics (refer to Figure 2).

False Negatives (FN)(↓) This is the most important metric in applications, where the cost of missing to detect high risk failures is high. Hence, we measure the ratio of samples in the ID or Low Risk regimes that actually have high true risk (top 20^{th} percentile of all test samples).

False Positives (FP)(↓) This reflects the penalty for scenarios where arbitrarily flagging harmless samples as failures. Here, we measure the ratio of samples in the Moderate or High Risk regimes that actually have low true risk (bottom 20^{th} percentile of all test samples).

Confusion in Low Risk Regimes ($\mathbf{C_{low}}$)(↓) A common challenge in fine-grained sample grouping (ID vs Low Risk) is that detection score can confuse samples between neighboring regimes. We define this metric to measure the ratio between the 90^{th} percentile of the ID regime and the 10^{th} percentile of the Low Risk regime.

Confusion in High Risk Regimes ($\mathbf{C_{high}}$)(↓) This is similar to the previous case and instead measures the confusion between the Moderate Risk and High Risk regimes.

4. Experiments

Datasets. We use a suite of regression functions in varying dimensions for evaluating the proposed approach.

(i) 1D Regression Functions: We used the following standard synthetic functions:

1. $f_1(\mathrm{x}) = \begin{cases} \mathrm{x}^2 & \text{if } \mathrm{x} < 2.25 \text{ or } \mathrm{x} > 3.01 \\ \mathrm{x}^2 - 20 & \text{otherwise} \end{cases}$

2. $f_2(\mathrm{x}) = \sin(2\pi\mathrm{x}), \mathrm{x} \in [-0.5, 2.5]$

3. $f_3(\mathrm{x}) = a\exp(-b\mathrm{x}) + \exp(\cos(c\mathrm{x})) - a - \exp(1)$, $\mathrm{x} \in [-5, 5]$, $a = 20, b = 0.2, c = 2\pi$.

4. $f_4(\mathrm{x}) = \sin(\mathrm{x})\cos(5\mathrm{x})\cos(22\mathrm{x}), \mathrm{x} \in [-1, 2]$

In each of these functions, we used 200 test samples drawn from an uniform grid and computed the evaluation metrics.

(ii) HD Regression Benchmarks: (a) Camel (2D), (b) Levy (2D) (ben) characterized by multiple local and global minima, (c) Kinematics (8D), (d) Puma (8D) (del) which are simulated datasets of the forward dynamics of different robotic control arms, (e) Boston Housing (13D) (bh) and (f) Ailerons (39D) (ail) which is a dataset for predicting control action of the ailerons of an F16 aircraft. For each benchmark, we create two variants: Gaps (training exposed to data with targets between $(0 - 30^{th})$ and $(60 - 100^{th})$ percentiles) and Tails (training exposed to $(0 - 70^{th})$ percentiles of the targets) resulting in a total of 12 datasets.

Baselines. (i) DEUP (Lahlou et al., 2023) is the state-of-the-art epistemic uncertainty estimator of deep models. It utilizes a post-hoc error predictor that learns to predict the risk of the underlying model which is considered as a surrogate for uncertainty; (ii) DataSUITE (Seedat et al., 2022) is a task-agnostic approach that estimates the inconsistencies in the data regimes in order to assess data quality. Both baselines rely on the use of additional calibration data to either train the error predictor in case of DEUP and to obtain non-conformity scores that assess the sample level quality in the latter.

Training Protocols. For all experiments, we utilize the open source $\Delta-$UQ (Thiagarajan et al., 2022) which is an efficient and scalable predictive uncertainty estimator based on anchoring. We use an anchored MLP (Bishop & Nasrabadi, 2007) with 4 layers each with a hidden dimension of 128, ReLU activation and batchnorm. We train our models 5000 epochs with learning rate $5e-5$ and ADAM optimizer.Without loss of generality, we utilize the L_1 objective for training.

Table 1. **Metrics for 1D Benchmarks.** We report the FN, FP, C_{low} and C_{high} metrics on evaluation data across the entire target regime (lower the better). Note that for every metric, we identify the first and second best approach across the different benchmarks.

Metrics	Method	$f_1(\mathrm{x})$	$f_2(\mathrm{x})$	$f_3(\mathrm{x})$	$f_4(\mathrm{x})$
FN↓	DEUP	6.19	6.56	16.57	27.13
	DataSUITE	14	8.8	16	7.2
	Ours ($\texttt{Score}_1$)	5.6	0	11.6	2.4
	Ours ($\texttt{Score}_2$)	4.8	5.6	8.4	5.6
FP↓	DEUP	8.91	3.41	8.54	9.09
	DataSUITE	18.67	16	20	5.33
	Ours ($\texttt{Score}_1$)	2.67	0	4.67	6.67
	Ours ($\texttt{Score}_2$)	1.33	2.67	4.33	4
C_{low}↓	DEUP	65.9	57.86	34.13	169.54
	DataSUITE	59.42	24.61	22.44	89.51
	Ours ($\texttt{Score}_1$)	28.08	7.19	19.94	12.05
	Ours ($\texttt{Score}_2$)	20.61	17.8	16.57	19.7
C_{high}↓	DEUP	91.64	4.47	59.46	16.56
	DataSUITE	3.66	46.02	58.32	6.81
	Ours ($\texttt{Score}_1$)	3.09	3.43	8.78	6.9
	Ours ($\texttt{Score}_2$)	3.09	4.67	10.99	5.71

5. Findings

Our Framework Accurately Identifies Risk Regimes. To characterize different risk regimes, it is crucial for a method to align well with the inferred data manifold (ID) and progressively flag regions of low, moderate and high risk as we move away from the training manifold. We achieve this objective effectively, as illustrated in Figure 3. We observe that our framework accurately identifies the training data regimes (Green) as part of the ID. As we traverse further from the training manifold, our approach assigns low risk (Yellow) to unseen examples that are close to the training data. Notably, as we encounter samples that are significantly OOS or OOD, we consistently identify them as Moderate or High risk. We ensure a well-calibrated transition between risk regimes across the entire input space for the regression functions considered. Our unified framework excels in characterizing regimes with moderate or low uncertainty, as demonstrated in Figure 1 (top), where the regime $[2.2, 2.7]$ is correctly identified as moderate risk despite appearing to have lower uncertainty.

Our Framework Produces Lower FN, FP and Confusion Scores. As discussed in Section 3, it is crucial to avoid large prediction errors in regimes identified as ID or low risk, and vice versa, in order to minimize FN and FP. Moreover, a reliable failure detector should effectively delineate risk regimes based on true risk and minimize their overlap. It can be observed from tables 1, 2 and 3 that our framework significantly reduces FN, FP and confusion scores compared to the state-of-the-art baselines. Remarkably, even in higher

Table 2. **Assessing the identified risk regimes for HD Benchmarks (Gaps).** We report the FN, FP, C_{low} and C_{high} metrics on evaluation data across the entire target regime (lower the better). Note that for every metric, we identify the first and second best approach across the different benchmarks.

Metrics	Method	Camel (2D)	Levy (2D)	Kinematics (8D)	Puma (8D)	Housing (13D)	Ailerons (39D)
FN↓	DEUP	15.79	9.25	17.6	13.2	11.46	14.4
	DataSUITE	21.74	19.69	18.4	16.8	17.71	11.2
	Ours ($\texttt{Score}_1$)	12.15	10.9	6.4	10.4	6.25	0.9
	Ours ($\texttt{Score}_2$)	11.39	10.65	6.4	10.8	7.29	1.2
FP↓	DEUP	17.48	10.04	18.67	12.0	10.34	16.0
	DataSUITE	15.74	15.32	10.67	17.33	12.07	8.0
	Ours ($\texttt{Score}_1$)	3.36	5.04	12.0	9.67	8.62	4.0
	Ours ($\texttt{Score}_2$)	7.56	4.2	10.67	8.83	9.07	1.33
C_{low}↓	DEUP	50.59	34.67	10.71	14.82	13.86	15.55
	DataSUITE	42.92	71.06	21.96	15.26	14.8	30.78
	Ours ($\texttt{Score}_1$)	14.05	13.62	12.91	12.44	13.33	12.90
	Ours ($\texttt{Score}_2$)	10.13	10.41	10.93	8.71	10.42	11.18
C_{high}↓	DEUP	15.47	12.42	11.28	6.18	3.36	23.94
	DataSUITE	37.51	36.5	5.97	10.57	22.56	4.2
	Ours ($\texttt{Score}_1$)	8.89	10.39	7.71	8.09	3.19	1.69
	Ours ($\texttt{Score}_2$)	11.03	9.37	7.01	7.30	2.95	1.65

Table 3. **Assessing the identified risk regimes for HD Benchmarks (Tails).** For every metric, we identify the first and second best approach across the different benchmarks.

Metrics	Method	Camel (2D)	Levy (2D)	Kinematics (8D)	Puma (8D)	Housing (13D)	Ailerons (39D)
FN↓	DEUP	10.53	7.34	14.4	16.8	2.11	18.4
	Data SUITE	3.84	9.21	17.6	22.4	17.89	17.6
	Ours ($\texttt{Score}_1$)	0.0	4.56	8	8.8	1.05	9.6
	Ours ($\texttt{Score}_2$)	0.25	4.82	7.2	10.4	2.32	9.6
FP↓	DEUP	9.5	7.35	13.0	14.67	8.77	12.0
	Data SUITE	3.83	6.38	24.0	26.67	19.3	12.0
	Ours ($\texttt{Score}_1$)	0.42	1.68	6.33	13.33	3.51	0.8
	Ours ($\texttt{Score}_2$)	1.68	2.52	6.18	12.2	4.26	0.4
C_{low}↓	DEUP	34.04	52.74	6.36	5.37	13.0	11.07
	Data SUITE	42.08	81.06	7.34	5.67	17.73	16.52
	Ours ($\texttt{Score}_1$)	15.59	26.44	6.58	4.61	5.14	17.19
	Ours ($\texttt{Score}_2$)	14.37	14.04	5.73	5.5	6.67	11.38
C_{high}↓	DEUP	23.69	20.75	6.83	2.63	5.69	7.25
	Data SUITE	17.49	27.32	10.08	6.41	5.15	4.97
	Ours ($\texttt{Score}_1$)	7.5	17.93	7.14	2.46	5.07	2.31
	Ours ($\texttt{Score}_2$)	6.7	15.18	7.09	2.81	4.05	2.43

dimensions and more complex extrapolation scenarios (e.g., Gaps and Tails, as discussed in Section 4), we consistently outperform the baselines. Our non-conformity scores help effectively reduce the overlap between risk regimes producing low (C_{low} and C_{high}), making them reliable for identifying samples that generalize well or those that are completely out-of-distribution/out-of-sample (OOD/OOS). These highlight the limitations of relying solely on predictive uncertainties, such as DEUP, for failure characterization, as they may not be sufficient in practical applications. Additionally, uncertainty methods like DataSUITE, which assess data quality without task-specific considerations, may not accurately identify risk regimes.

$\texttt{Score}_2$ Produces Non-Trivial Improvements Over $\texttt{Score}_1$. As described in Section 3, $\texttt{Score}_2$ quantifies the non-conformities in the input space and requires a test-time optimization strategy to better enhance the identification of moderate and high risk regimes. Our findings across various

benchmarks indicate that $\texttt{Score}_2$ reduces FN and FP and confusion scores over $\texttt{Score}_1$. Notably, $\texttt{Score}_2$ exhibits a more conservative approach in characterizing risk regimes, resulting in comparatively lower confusion scores compared to $\texttt{Score}_1$.

6. Conclusion

In this paper, we propose a novel framework for failure characterization in deep regression models. It leverages the principle of anchoring to integrate predictive uncertainties and novel non-conformity scores, enabling the organization of samples into different risk regimes and facilitating a comprehensive analysis of model errors. We identify two key impacts of our work. First, our framework can enhance the safety of AI model deployment by proactively and preemptively detect failure cases in various high impact scenarios such as scientific simulations. This can prevent costly errors and mitigate risks associated with inaccurate predictions. Second, we contribute to advancing research in failure characterization for deep regression. While we believe that it can improve reliability, its deployment and usage should be accompanied by ethical considerations and human oversight. Decisions and actions based on the detected failure cases should be made responsibly, taking into account potential biases, fairness, and broader societal impact.

References

Ailerons datsets. https://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html. Accessed: 2023-05-11. 6

Virtual library of simulation experiments. https://www.sfu.ca/~ssurjano/index.html. Accessed: 2023-05-01. 6

Boston housing. https://scikit-learn.org/1.0/modules/generated/sklearn.datasets.load_boston.html. Accessed: 2023-05-11. 6

Delve datasets. https://www.cs.toronto.edu/~delve/data/datasets.html. Accessed: 2023-05-11. 6

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020. 1

Baek, C., Jiang, Y., Raghunathan, A., and Kolter, J. Z. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35:19274–19289, 2022. 1

Bishop, C. M. and Nasrabadi, N. M. *Pattern Recognition and Machine Learning*. *J. Electronic Imaging*, 16(4): 049901, 2007. 6

Chen, M., Goel, K., Sohoni, N. S., Poms, F., Fatahalian, K., and Ré, C. Mandoline: Model evaluation under distribution shift. In *International Conference on Machine Learning*, pp. 1617–1629. PMLR, 2021. 1

Deng, W. and Zheng, L. Are labels always necessary for classifier accuracy evaluation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15069–15078, 2021. 1

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016. 1

Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., and Schmidt, L. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1134–1144, 2021. 1

He, B., Lakshminarayanan, B., and Teh, Y. W. Bayesian deep ensembles via the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:1010–1022, 2020. 1

Jiang, Y., Krishnan, D., Mobahi, H., and Bengio, S. Predicting the generalization gap in deep networks with margin distributions. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJlQfnCqKX. 1

Lahlou, S., Jain, M., Nekoei, H., Butoi, V. I., Bertin, P., Rector-Brooks, J., Korablyov, M., and Bengio, Y. DEUP: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=eGLdVRvvfQ. 3, 5, 6

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 1

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), 2022. 1

Narayanaswamy, V., Anirudh, R., Kim, I., Mubarka, Y., Spanias, A., and Thiagarajan, J. J. Predicting the generalization gap in deep models using anchoring. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4393–4397. IEEE, 2022. 1

Netanyahu, A., Gupta, A., Simchowitz, M., Zhang, K., and Agrawal, P. Learning to extrapolate: A transductive approach. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lid14UkLPd4. 1, 2, 3, 4

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019. 1

Seedat, N., Crabbé, J., and van der Schaar, M. Data-SUITE: Data-centric identification of in-distribution incongruous examples. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 19467–19496, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/seedat22a.html. 1, 3, 6

Teng, J., Wen, C., Zhang, D., Bengio, Y., Gao, Y., and Yuan, Y. Predictive inference with feature conformal prediction. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0uRm1YmFTu. 3

Thiagarajan, J. J., Anirudh, R., Narayanaswamy, V., and timo Bremer, P. Single model uncertainty estimation via stochastic data centering. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=j0J9upqN5va. 1, 2, 3, 4, 6

Young, A. T., Xiong, M., Pfau, J., Keiser, M. J., and Wei, M. L. Artificial intelligence in dermatology: a primer. *Journal of Investigative Dermatology*, 140(8):1504–1512, 2020. 1