Sampling-based Multi-dimensional Recalibration

Youngseog Chung[1] **Ian Char**[1] **Jeff Schneider**[1 2]

Abstract

Calibration of probabilistic forecasts in the regression setting has been widely studied in the single dimensional case, where the output variables are assumed to be univariate. In many problem settings, however, the output variables are multi-dimensional, and in the presence of dependence across the output dimensions, measuring calibration and performing recalibration for each dimension separately can be both misleading and detrimental. In this work, we focus on representing predictive uncertainties via samples, and propose a recalibration method which accounts for the joint distribution across output dimensions to produce calibrated samples. Based on the concept of highest density regions (HDR), we define the notion of HDR calibration, and show that our recalibration method produces samples which are HDR calibrated. We demonstrate the performance of our method and the quality of the recalibrated samples on a suite of benchmark datasets in multi-dimensional regression, a real-world dataset in modeling plasma dynamics during nuclear fusion reactions, and on a decision-making application in forecasting demand.

1. Introduction

Calibration in probabilistic forecasting, in general terms, refers to the alignment between the predicted probabilities and empirical frequencies of the true observations. Alongside other quantitative metrics to assess predictive distributions, e.g. negative log-likelihood (NLL) or simply accuracy of the mean, calibration is considered an important and desirable quality of probabilistic forecasts, and many works have appraised the utility of calibration in various application settings (Gneiting et al., 2007; Malik et al., 2019; Deshpande & Kuleshov, 2021; Chung et al., 2023).

Within the general principle of "aligning predicted and empirical probabilities", various notions of calibration exist, and these definitions also vary slightly between classification and regression settings. In this work, we focus on the regression setting where both the inputs, X, and targets, Y, are continuous.

We begin our discussion from the observation that the most widely studied notions of calibration in regression are usually confined to the setting where the targets are single dimensional (Gneiting et al., 2007; Pearce et al., 2018; Kuleshov et al., 2018; Song et al., 2019; Cui et al., 2020; Zhao et al., 2020; Sahoo et al., 2021; Kuleshov & Deshpande, 2022). While multi-dimensional regression models are widely used in machine learning, especially in applications such as model-based control (Chua et al., 2018; Malik et al., 2019; Yu et al., 2020; Kidambi et al., 2020) or modeling in the physical sciences (Sexton et al., 2012; Duraisamy et al., 2019; Abbate et al., 2021; Char et al., 2023a), we find that methods which account for the joint multi-dimensional distribution in *assessing* calibration and *recalibrating* the prediction is generally lacking. In lieu of more sophisticated methods, calibration is often considered for each output dimension independently. However, failing to account for interplay among the output dimensions may be problematic when dependence does exist. In this case, the collection of marginals is not sufficient to provide an accurate assessment of the prediction quality (see Figure 1 for an example).

In this work, we address the problem of calibration in multi-dimensional regression by first formalizing a notion of calibration which *can* account for dependence among the output dimensions and further proposing a recalibration algorithm for the joint predictive distribution. We summarize our main contributions as follows:

- By leveraging existing ideas in highest density regions (HDR), we propose the notion of *HDR calibration*, which accounts for dependence in the output dimensions in defining and evaluating calibration for multi-dimensional distributional predictions.
- We develop a recalibration algorithm for multi-dimensions which produces HDR calibrated predictive distributions via a sampling procedure.
- We provide extensive demonstrations of the merits

[1] Machine Learning Department; [2] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213. Correspondence to: Youngseog Chung <youngsec@cs.cmu.edu>.

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

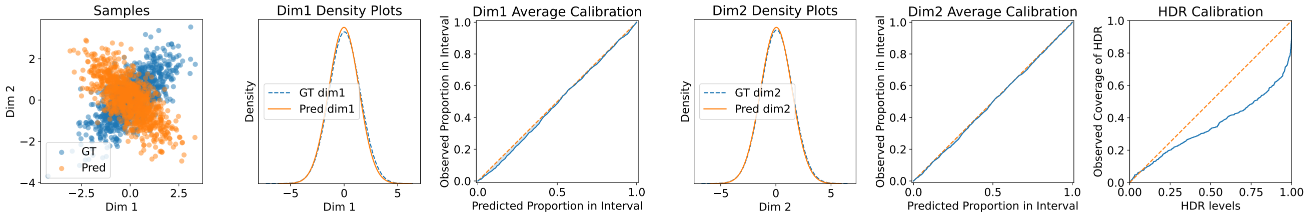

Figure 1. We demonstrate a pitfall of assessing the calibration of each dimension independently for multi-dimensional distributional predictions. **(From Left to Right)** Consider a 2-dimensional target space where samples from the predictive distribution (labeled *Pred*) and ground truth distribution (labeled *GT*) are displayed as a scatter plot. The predictive distribution exhibits the opposite correlation in the output dimensions compared to the ground truth, but each of the marginal distributions are accurate. Assessing calibration of each dimension separately suggests a well-calibrated predictive distribution. Highest density regions (HDRs), on the other hand, are able to account for the dependence in the dimensions. Assessing HDR calibration, which considers the output dimensions *jointly*, reveals the miscalibration of the full joint distribution.

of the notion of HDR calibration and the efficacy of the recalibration algorithm on a suite of benchmark datasets in multi-dimensional regression, and two real-world datasets: a dynamics modeling task in nuclear fusion, and a downstream decision-making application in forecasting customer demand.

We continue our discussion by first describing the problem setting and relevant concepts to motivate the definition of HDR calibration in Section 2. Based on this notion of calibration, we present our proposed HDR recalibration algorithm in Section 3. We provide empirical evaluations in Section 4[1].

2. Preliminaries and Related Works

2.1. Setting and Notation

Upper case letters X, Y denote random variables (r.v.), and lower case letters x, y denote their observed values. We consider the regression setting with an input feature space $\mathcal{X} \subseteq \mathbb{R}^n$ and a target space $\mathcal{Y} \subseteq \mathbb{R}^D$. We use x^d, X^d, y^d, and Y^d to denote the d^{th} dimension of input and target vectors. f and F denote the true probability density function (PDF) and cumulative distribution function (CDF), and when it exists, we denote the true quantile function with F^{-1}. Estimates of these functions are denoted with $\hat{f}$, $\hat{F}$ and $\hat{F}^{-1}$. We use subscripts to indicate the corresponding random variable of the PDFs and CDFs (e.g. f_X and F_X are the marginal PDF and CDF of X, and $f_{Y|X}$ and $F_{Y|X}$ are the PDF and CDF of Y conditioned on X). When conditioning on a specific value $X = x$, we denote the conditional distribution functions as $f_{Y|x}$ and $F_{Y|x}$. Lastly, we assume that new target samples can be drawn from the distribution estimate, and we denote the random variable corresponding to these new target samples as $\hat{Y}$. In particular, this can be done by sampling $X \sim f_X$ from the dataset and subsequently sampling $\hat{Y}|X \sim \hat{f}_{Y|X}$. Importantly, note that the distribution of $\hat{Y}$ is still tied to the distribution of X.

2.2. Calibration in Univariate Regression

Before discussing the multi-dimensional setting, we first provide a brief review of notions of calibration in the univariate setting. A widely accepted notion of calibration in univariate regression is *probabilistic calibration* (Gneiting et al., 2007). A predictive distribution $\hat{F}_{Y|X}$ is probabilistically calibrated if

$$P(Y \leq \hat{F}^{-1}_{Y|X}(p)) = p, \forall p \in (0, 1). \qquad (1)$$

This notion is also referred to as simply *calibration* (Kuleshov et al., 2018), *quantile calibration* (Song et al., 2019), or *average calibration* (Zhao et al., 2020; Chung et al., 2021b) since it focuses on the average validity of the predictive quantile function $\hat{F}^{-1}_{Y|X}$. We henceforth refer to this notion as *average calibration*. Here, we note that the true distribution $F_{Y|X}$ trivially satisfies Eq. 1 since $F_{Y|X}(Y) \sim \mathcal{U}(0, 1)$ by the probability integral transform and $P(\hat{F}_{Y|X}(Y) \leq p) = p$ is the CDF of $\mathcal{U}(0, 1)$.

From this general definition, subsequent works have derived various notions of calibration, usually by placing different conditions in assessing the empirical probability (LHS of Eq. 1). For example, *distribution calibration* (Song et al., 2019) assesses average calibration conditioned on the predictive distribution; *individual calibration* (Zhao et al., 2020) requires average calibration conditioned on each input point, $x \in \mathcal{X}$; and *group calibration* (Kleinberg et al., 2016; Hébert-Johnson et al., 2017) requires average calibration conditioned on specific subsets of the input space with non-zero measure.

In all of the aforementioned notions, Y is assumed to be

[1] Code is available at: https://github.com/YoungseogChung/multi-dimensional-recalibration

univariate (i.e $\mathcal{Y} \subseteq \mathbb{R}$), and predictive conditional quantiles $\hat{F}^{-1}_{Y|X} : \mathcal{X} \times (0, 1) \to \mathcal{Y}$ are utilized to measure the discrepancy between predicted and empirical probabilities (RHS and LHS of Eq. 1).

2.3. The Multi-dimensional Setting

While a naive application of the notions of univariate calibration to multi-dimensional distribution functions may seem plausible, in the multivariate setting, the quantile function is not well-defined (Belloni & Winkler, 2009), and further, $F_Y(Y)$ for $Y \in \mathbb{R}^D$ when $D > 1$ is no longer uniformly distributed (Barbe et al., 1996; Genest & Rivest, 2001). To circumvent these issues, prior works have suggested utilizing projections of the target variable Y in order to *define* and *assess* calibration of multi-dimensional distributional predictions. We formalize such methods as follows.

Consider a mapping $g : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, where $\mathcal{Z} \subseteq \mathbb{R}$. Furthermore, we let Z and $\hat{Z}$ be the r.v.s over the projection outputs when using target labels Y and $\hat{Y}$, respectively. Concretely, $Z := g(X, Y)$ and $\hat{Z} := g(X, \hat{Y})$. Since sampling from the predicted distribution is cheap, we can estimate the CDF $F_{Z|X}$ using the empirical distribution of $\hat{Z}|X$. We refer to this empirical CDF as $\hat{F}_{Z|X}$.

Then, following the definition of average calibration (Eq. 1), we can define calibration in the projected space as satisfying, $\forall p \in (0, 1)$,

$$P(Z \leq \hat{F}^{-1}_{Z|X}(p)) = p \tag{2}$$

$$\text{or equivalently, } P(\hat{F}_{Z|X}(Z) \leq p) = p. \tag{3}$$

We can easily show that the optimal prediction $\hat{F}_{Y|X} = F_{Y|X}$ satisfies this definition of calibration in the projected space.

Proposition 2.1. *The optimal distributional prediction, i.e. $\hat{F}_{Y|X} = F_{Y|X}$, satisfies calibration in the projected space, Eq. 3. (proof in Section A)*

Several prior works have proposed specific versions of Eq. 3 with specific projection functions. Ziegel & Gneiting (2014) introduced *copula calibration* by utilizing the predictive CDF as the projection function, i.e. $g(X, \cdot) = \hat{F}_{Y|X}$. In this specific case, the distribution of the projections is called the Kendall distribution (Nelsen et al., 2003).

One can also utilize the predictive PDF for the projection function such that $g(X, \cdot) = \hat{f}_{Y|X}$, in which case Eq. 3 bears intrinsic relationships to existing concepts of highest predictive density (HPD) values (Harrison et al., 2015; Dalmasso et al., 2020; Zhao et al., 2021) and highest density regions (HDR) (Hyndman, 1996).

While there are several candidates for projection functions, in this work, we choose to focus on using the predictive PDF. In particular, we leverage its connections with HPD and HDR to formalize a notion of calibration in multi-dimensions (Defn. 2.2) and propose a recalibration procedure that achieves this notion of calibration (Section 3). Hence, in the rest of this work, we always assume $Z := \hat{f}_{Y|X}(Y)$ and $\hat{Z} := \hat{f}_{Y|X}(\hat{Y})$.

For any given (x, y), $\texttt{HPD}_x(y)$ is a measure of how plausible y is w.r.t $\hat{f}_{Y|x}$ and is defined as

$$\texttt{HPD}_x(y) = \int_{y' : \hat{f}_{Y|x}(y') \geq \hat{f}_{Y|x}(y)} \hat{f}_{Y|x}(y')dy'. \tag{4}$$

In words, $\texttt{HPD}_x(y)$ is the predicted probability of observing $\hat{Y}$ that is more likely than y, where the likelihood is determined by $\hat{f}_{Y|x}$. Considering the definition of Z and $\hat{Z}$, we see that

$$\begin{aligned} &\texttt{HPD}_x(y) && (5)\\ &= P(\hat{f}_{Y|x}(\hat{Y}) \geq \hat{f}_{Y|x}(y) \mid X = x) && (6)\\ &= 1 - \hat{F}_{Z|x}(\hat{f}_{Y|x}(y)). && (7) \end{aligned}$$

Further, HPD values, which are *probabilities*, have a direct relationship to HDRs, which are *prediction sets*. For clarity, we provide a definition of HDR below using our notation, and we refer the reader to Section C for the original notation by Hyndman (1996). For a fixed x and constant $\lambda \in \mathbb{R}$, we define the λ-density region as $\texttt{DR}_x(\lambda) := \{y : \hat{f}_{Y|x}(y) \geq \lambda\}$. Then for a given coverage level p, the p-HDR is the smallest density region with probability greater than or equal to p. Concretely,

$$\begin{aligned} &\texttt{HDR}_x(p) := \texttt{DR}_x(\lambda^*)\\ \text{where} \quad &\lambda^* = \sup\{\lambda : P(\hat{Y} \in \texttt{DR}_x(\lambda)|X = x) \geq p\}. \end{aligned}$$

By their definitions, the following equivalence holds:

$$\texttt{HPD}_x(y) \leq p \iff y \in \texttt{HDR}_x(p) \tag{8}$$

We note that calibration is generally defined in terms of prediction sets of a distribution, and drawing on the intrinsic relationships between HDR, HPD and Eq. 7, we formalize Eq. 3 with the notion of HDR calibration:

Definition 2.2. A predictive PDF $\hat{f}_{Y|X}$ is *HDR calibrated* if, $\forall p \in (0, 1)$,

$$P(Y \in \texttt{HDR}_X(p)) = p \tag{9}$$

$$\text{or equivalently, } P(\texttt{HPD}_X(Y) \leq p) = p. \tag{10}$$

Proposition 2.3. *HDR calibration holds if and only if Equation 3 holds. (proof in Section A)*

Similar to average calibration (Eq. 1), which requires Y to be contained in the p-prediction set (defined by the p^{th} quantile) with probability p, HDR calibration requires the p-HDR to contain Y with probability p.

Utilizing projections allows one to *define* notions of calibration in the multi-dimensional setting which can account for dependence in the output dimension, granted that the projection function models the dependence. Further, based on the definitions, one can *assess* calibration (or miscalibration) via the discrepancy between the predicted and empirical probabilities. Following the commonly used notion of expected calibration error (ECE) (Guo et al., 2017; Cui et al., 2020; Tran et al., 2020; Chung et al., 2021b) we can measure the (L1-)ECE w.r.t the general notion of calibration defined in Eq. 3 as

$$\mathbb{E}_{p\sim\mathcal{U}(0,1)}\left|P(\hat{F}_{Z|X}(Z)\leq p)-p\right|. \tag{11}$$

Not only do these metrics allow one to evaluate the quality of uncertainty for multi-dimensional predictions, but they can also be used to improve a model's predictive distribution via a recalibration procedure.

2.4. Recalibration

Probabilistic models are usually trained by optimizing a loss function which may not be necessarily aligned with calibration. This can often lead to models being miscalibrated at the end of the training (Guo et al., 2017; Kuleshov et al., 2018; Chung et al., 2021b). A *post-hoc recalibration* step can be applied on top of the trained model to adjust for its level of miscalibration observed on a held-out calibration or validation dataset.

Post-hoc recalibration is well-studied in classification, and there are several methods which have proven to be effective in producing well-calibrated (discrete) class probabilities (Platt et al., 1999; Zadrozny & Elkan, 2001; 2002; Guo et al., 2017; Gupta & Ramdas, 2021).

This problem is not as widely studied in regression, however, and to the best of our knowledge, the most popular method is that of Kuleshov et al. (2018), which learns an isotonic mapping between expected and observed quantile levels. Crucially, this method readily applies only to the case when the targets Y are univariate, and we henceforth refer to this algorithm as "single dimensional (SD) recalibration". In Section 3, we propose a recalibration procedure for the multivariate setting.

While also proposed for the single dimensional setting, it is worth mentioning that Izbicki et al. (2022) proposes a conformal prediction method which bears relevance as their method utilizes HPD values as the conformity score. However, there are key differences: while they are focused on producing prediction *sets* for a fixed coverage level (as is the goal of conformal prediction), we are focused on expressing the full predictive *distribution*. Crucially, since Izbicki et al. (2022) does not consider multi-dimensional target spaces, their method does not account for dependence in the target dimensions, and the algorithm relies on constructing a finite grid of the target space, which is ill-suited for higher dimensions. As we will discuss in Section 3, our recalibration procedure explicitly addresses dependence in the target dimensions and is more scalable as it focuses on *sampling* from a predictive distribution in the multi-dimensional space. We refer the reader to Appendix B for additional details on related works.

2.5. Predictive Uncertainty and Sampling

Ensuring the calibration of predictive uncertainties becomes important when deploying probabilistic models in downstream applications. The application setting will dictate how the uncertainties are utilized. For example, in the context of Bayesian optimization, depending on the acquisition function, the uncertainties may be used to construct confidence bounds (Auer, 2002; Srinivas et al., 2009), compute probabilities or expectations (Jones et al., 1998), or be used to sample from (Thompson, 1933; Kandasamy et al., 2018; Char et al., 2019). For model-based control where a probabilistic dynamics model is learned, the uncertainty-aware model is most often used to sample plausible transitions and trajectories (Chua et al., 2018; Janner et al., 2019; Mehta et al., 2021; Char et al., 2023b).

In this work, we focus on *sampling* to represent and utilize the predictive uncertainties and aim to produce samples from a well-calibrated predictive distribution. Applying SD recalibration (Kuleshov et al., 2018) to multi-dimensional settings will necessitate recalibrating each dimension separately, which renders each dimension independent in the recalibrated samples. However, we note that this, in fact, is how recalibration is utilized in practice to multi-dimensional settings (e.g. Malik et al. (2019)). The algorithm we propose in the next section performs recalibration *jointly* across all output dimensions and is able to consider dependence across the dimensions.

3. Method

In this section, we describe our proposed recalibration procedure which aims to achieve HDR calibration (Defn. 2.2). We describe the procedure in two parts. Section 3.1 details the recalibration algorithm that aims to optimize for Eq. 3, which is equivalent to HDR calibration by Proposition 2.3. Afterwards, Section 3.2 describes a pre-conditioning step that can modify the predictive PDF to account for dependence in the output dimensions when applying the recalibration algorithm.

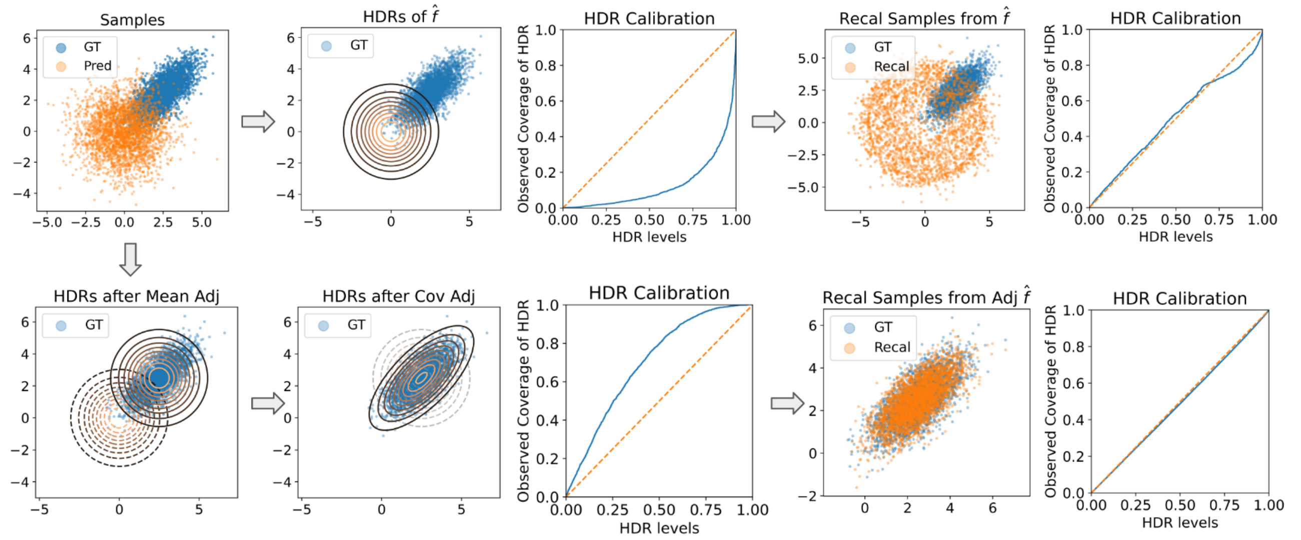

Figure 2. Demonstration of HDR recalibration on a marginal distributional prediction. **(Top Left)** The initial prediction (labeled *Pred*) displays bias in the mean prediction and fails to model the correlation in the ground truth distribution (labeled *GT*). **(Top Row)** Without the PDF adjustment step, we observe that observations (GT points) fall more often in the higher level HDRs (level sets defined by darker boundaries) than lower level HDRs (level sets bounded by lighter colors). HDR recalibration re-samples from each HDR according to the observed frequencies (i.e. the learned recalibration mapping), hence when producing recalibrated samples, the higher level HDRs (i.e. outer level sets of $\hat{f}$) are over-sampled and the lower level HDRs (inner level sets of $\hat{f}$) are under-sampled. The resulting recalibrated samples are HDR calibrated (right-most plot), but we can visually assess that the samples are suboptimal and in particular, fail to model the correlation in the dimensions. **(Bottom Row)** Before the recalibration procedure, we can estimate the bias in the mean on the calibration dataset and correlation in the dimensions with the correlation matrix of the mean prediction error. After applying these two adjustments, HDR calibration reveals that each p-HDR contains more than p proportion of the observations (which also indicates that the level sets are too wide). Hence, HDR recalibration proportionately under-samples from each HDR, which results in well-calibrated samples that also reflect the correlation in the output dimensions.

3.1. HDR Recalibration Algorithm

The proposed recalibration algorithm is comparable to that of Kuleshov et al. (2018) for univariate settings, but with key differences – the recalibration occurs in the projected space $\mathcal{Z}$, and the recalibration output must be translated back into the target space $\mathcal{Y}$.

First, we estimate a recalibration mapping in the projected space by using observations of the random variable $\hat{F}_{Z|X}(Z)$ with a calibration dataset $\{(x_i, y_i)\}_{i=1}^N$, i.e. the observed values are $\{\hat{F}_{Z|x_i}(z_i)\}_{i=1}^N$ where $z_i = \hat{f}_{Y|x_i}(y_i)$. To elaborate more on this procedure, note that z_i is a scalar value produced by evaluating the PDF $\hat{f}_{Y|x_i}$ at y_i, where $\hat{f}_{Y|x_i}$ is the PDF of the predictive distribution. $\hat{F}_{Z|x_i}(z_i)$ is also a scalar value produced by evaluating the CDF $\hat{F}_{Z|x_i}$ at z_i, however, $\hat{F}_{Z|x_i}$ is an empirical CDF over the projected space that is estimated by producing samples from the predictive distribution $\hat{f}_{Y|x_i}$. Again, we note that sampling from the predictive distribution is cheap, thus estimating this empirical CDF is also cheap. Algorithm 3 provides exact details on this estimation step.

Afterwards, we learn the monotonic mapping $R : [0, 1] \rightarrow [0, 1]$ where $R(p) := P(\hat{F}_{Z|X}[Z] \leq p)$. R is then applied to the predictive distribution at each x, $\hat{F}_{Z|x}$, to produce the recalibrated predictive distribution $R \circ \hat{F}_{Z|x}$.

Proposition 3.1. *Consider $R \circ \hat{F}_{Z|X}$ for an invertible mapping R. Then $R \circ \hat{F}_{Z|X}$ satisfies Eq. 3, i.e.*

$$P(R \circ \hat{F}_{Z|X}(Z) \leq p) = p \quad \forall p \in (0, 1).$$

(proof in Section A)

One can therefore use such a recalibration map, R, to draw new, calibrated samples in $\mathcal{Z}$ space. However, it remains unclear how to relate these samples back to their counterparts in $\mathcal{Y}$ space. To address this issue, we present a sampling algorithm that operates over samples of r.v. $\hat{Y}$. The key idea is to re-sample from the set of samples generated from $\hat{f}_{Y|X}$ according to what the distribution should look like in $\mathcal{Z}$ space. In particular, for any fixed x, we can draw many samples from the predictive PDF, $\{\hat{y}_j\}_{j=1}^M \sim \hat{f}_{Y|x}$, then apply the projection $\hat{f}_{Y|x}(\cdot)$ to produce the dataset of tuples

Algorithm 1 HDR Recalibration: Training

```
1: Input: Calibration dataset $\{(x_i, y_i)\}_{i=1}^N$, predictive PDF $\hat{f}_{Y|X}$.
2: $\hat{f}_{Y|X} \leftarrow \texttt{ADJUST}(\hat{f}_{Y|X})$ (Algorithms 5, 6, 7).
3: Construct the dataset $\mathcal{C} = \{\hat{F}_{Z|x_i}(z_i)\}_{i=1}^N$, where $z_i = \hat{f}_{Y|x_i}(y_i)$ (see Algorithm 3).
4: Sort values in $\mathcal{C}$ to construct $\{c_{(i)}\}_{i=1}^N$, construct the recalibration dataset $\mathcal{C}' = \{i/N, c_{(i)}\}_{i=1}^N$.
5: Learn the recalibration mapping $R$ on $\mathcal{C}'$.
6: Output: Recalibration mapping $R$.
```

Algorithm 2 HDR Recalibration: Sampling

```
1: Input: Test point $x$, predictive PDF $\hat{f}_{Y|X}$, recalibration mapping $R$, number of samples $M$.
2: $\hat{f}_{Y|X} \leftarrow \texttt{ADJUST}(\hat{f}_{Y|X})$. (Algorithms 5, 6, 7).
3: Construct $\mathcal{D} = \{(\hat{y}_j, \hat{z}_j)\}_{j=1}^M$ by producing $M$ samples $\hat{y}_j \sim \hat{f}_{Y|x}$ and setting $\hat{z}_j = \hat{f}_{Y|x}(\hat{y}_j)$.
4: Re-sample from $\mathcal{D}$ to construct $\mathcal{D}' = \{(\hat{y}_k, \hat{z}_k)\}_{k=1}^M$ s.t. $\{\hat{z}_k\}_{k=1}^M$ approximately follows $R \circ \hat{F}_{Z|x}$ (see Algorithm 4).
5: Output: Recalibrated samples at $x$, $\{\hat{y}_k\}_{k=1}^M$.
```

$\mathcal{D} = \{(\hat{y}_j, \hat{z}_j)\}$, where $\hat{z}_j = \hat{f}_{Y|x}(\hat{y}_j)$, and note that by definition, $\hat{z}_j \sim \hat{f}_{Z|x}, \hat{F}_{Z|x}$. We then re-sample from $\mathcal{D}$ to produce $\{(y_k, z_k)\} \subseteq \mathcal{D}$ such that the distribution of $\{z_k\}$ is more closely aligned with $R \circ \hat{F}_{Z|x}$. Concretely, this is done by forming an empirical CDF of the $\hat{Z}$ samples $\{\hat{z}_j\}$ using binning, re-weighting each bin to match $R \circ \hat{F}_{Z|x}$, then re-sampling from each bin according to the adjusted weights. The full algorithm is summarized in Algorithms 1 and 2: Algorithm 1 describes the procedure for learning the recalibration map R, and Algorithm 2 describes the test time sampling procedure. We provide more details on each of the steps in Section D. Crucially, the corresponding $\{y_k\}$ are HDR calibrated.

Proposition 3.2. *Suppose that $\hat{Z} \sim R \circ \hat{F}_{Z|X}$ and that R is an invertible mapping. Then the distribution of $\hat{Y}$ is HDR calibrated. (proof in Section A)*

3.2. Adjusting the Predictive PDF

The HDR recalibration algorithm from Section 3.1 produces a predictive distribution (via samples) s.t. the p-HDR contains p proportion of the target observations, on average, $\forall p \in (0, 1)$. However, this predictive distribution can still fail to address dependencies among the output dimensions. This is because, for any fixed x, the HDRs are constructed with *level sets* of $\hat{f}_{Y|x}$, and, if $\hat{f}_{Y|x}$ fails to model dependencies, then the recalibrated samples will also express independence among the output dimensions. We provide an illustration in Figure 2. The top row shows that the pre-hoc predictive distribution assumes independence in the output dimensions, which is reflected in the spherical boundaries of the HDRs. After HDR recalibration, the shape of the recalibrated distribution is still spherical, even though the calibration dataset (i.e. ground truth (GT) observations in blue) displays correlation among the dimensions.

This highlights the importance of the projection function $\hat{f}_{Y|X}$, and ideally, $\hat{f}_{Y|X}$ should better reflect the true distribution in order for the recalibration procedure to produce more accurate samples. Further, if we can estimate the errors in $\hat{f}_{Y|X}$ (e.g. correlation, bias) with a held-out dataset, it can be beneficial to adjust $\hat{f}_{Y|X}$ for these factors prior to recalibration.

As a concrete instantiation of this adjustment, we propose a simple procedure to adjust the PDF of multivariate Gaussian distributions by estimating the bias in the predicted mean (i.e. the *location* of the HDRs), standard deviation (i.e. the *width* of the HDRs in each dimension), and the correlation in output dimensions (i.e. the *shape* of the HDRs) with a held-out dataset and correcting the PDF for each of these aspects. We provide details on each adjustment in Section D, and we suggest applying the composition of these adjustments prior to recalibration, as indicated with the ADJUST step in Line 2 of Algorithms 1 and 2. The bottom row of Figure 2 provides an illustration of the mean adjustment and correlation adjustment. We can observe that the resulting recalibrated samples more closely reflect the ground truth distribution. In our experiments, we always apply the composition of adjustments, and we provide an ablation study of performing HDR recalibration with and without adjustments in Section E.5.

Lastly, we note that the ADJUST steps in Line 2 of Algorithms 1 and 2 are meant to be a general procedure that depends on the type of predictive distribution used, and the adjustments provided in Algorithms 5, 6, and 7 in Section D are examples that are specific to the case when the predictive distribution is Gaussian. When either the predictive or ground truth distributions are complex, these specific adjustment steps may be insufficient to adequately model the relationships among the output dimensions, and more sophisticated adjustments may be necessary.

4. Experiments

We demonstrate the efficacy of the proposed method on two sets of modeling tasks (Section 4.1) and one downstream decision-making task (Section 4.2). Across all experiments, we compare the performance of model predictions with no recalibration (i.e. pre-hoc), with SD recalibration, and with HDR recalibration.

4.1. Modeling Tasks

The two modeling tasks are comprised of 1) a suite of benchmark regression datasets and 2) a real-world dataset from the physical sciences – modeling plasma dynamics in a nuclear fusion device called a tokamak.

Datasets. The "mulan" benchmark (Tsoumakas et al., 2011) is a set of prediction tasks with multi-dimensional targets of up to 16 dimensions. Among these tasks, we take regression datasets with at least 1000 datapoints, which result in the following 5 datasets: **scpf** (3D), **rf1** (8D), **rf2** (8D), **scm1d** (16D), **scm20d** (16D). On each dataset, we make train-validation-test splits of proportions $[65\%, 20\%, 15\%]$, and use the train set to learn a probabilistic neural network (PNN), which is a neural network that predicts a multivariate Gaussian distribution with a diagonal covariance matrix. Section E.2 provides the full set of details on the experiment setup for this benchmark experiment.

In the nuclear fusion experiment, we take three different pre-trained dynamics models of plasma evolution during nuclear fusion reactions as the pre-hoc models. These models were learned from recorded data of nuclear fusion experiments conducted on the DIII-D tokamak (Luxon, 2002), a magnetic confinement nuclear fusion device. Controlling these devices is meticulously difficult, and these dynamics models were used to optimize model-based control policies for deployment on DIII-D. Each model takes in the current plasma state and tokamak actuators, then predicts a multi-dimensional predictive distribution over several key plasma state variables for the next time step. All three models predict a multivariate Gaussian distribution with a diagonal covariance matrix. Two of the models (**Fusion1** and **Fusion2**) predict a 3-dimensional state target, and the third model, (**Fusion3**) predicts one additional state variable to predict a 4-dimensional target. Section E.3 provides the full set of details on the experiment setup for the nuclear fusion experiment.

Evaluation. We perform evaluations for both the benchmark and fusion modeling tasks as follows. For every test input x_i, samples are drawn from the predictive distribution, and we denote this set of samples as $\mathcal{S}_i = \{\hat{y}_j\}_{j=1}^M$. We report evaluation metrics based on the predictions, $\mathcal{S}_i$, and the true target datapoint, y_i, for each (x_i, y_i) in the test set. When applying recalibration (SD or HDR), we use the validation set to learn the recalibration mapping R and apply the mapping in producing $\mathcal{S}_i$. Sections E.2 and E.3 provides full details on the evaluation procedure for each set of experiments.

Metrics. As evaluation metrics, we report one proper scoring rule and two measures of calibration. Proper scoring rules (Gneiting & Raftery, 2007) are summary statistics of overall performance of a distributional prediction, and they serve as both an optimization objective as well as evaluation metrics. Because we represent the predictive distribution via *samples*, we use the energy score as our core evaluation metric. The energy score is defined in terms of expectations w.r.t the predictive distribution and hence, is amenable to estimation with samples.

Given a test datapoint (x_i, y_i) and the predictive distribution at x_i, $\hat{f}_{Y|x_i}$, the (negatively-oriented) energy score, $\text{ES}(\hat{f}_{Y|x_i}, y_i)$, is defined as

$$\text{ES}(\hat{f}_{Y|x_i}, y_i) = \mathbb{E}_{\hat{f}_{Y|x_i}} \left\| \hat{Y} - y_i \right\|_2^\beta - \frac{1}{2} \mathbb{E}_{\hat{f}_{Y|x_i}} \left\| \hat{Y} - \hat{Y}' \right\|_2^\beta,$$

where each of $\hat{Y}$ and $\hat{Y}'$ are independent r.v.s that are both distributed $\sim \hat{f}_{Y|x_i}$, and β is a hyperparameter $\in (0, 2)$.

Additionally, we report two measures of calibration: "HDR Expected Calibration Error (HDR-ECE)" and "single dimensional ECE (SD-ECE)". We estimate both metrics by first drawing K probability values: $0 \leq p_1 < p_2 \cdots < p_K \leq 1$ as the predicted probabilities.

HDR-ECE is computed as ECE (Eq. 11) using the notion of HDR calibration, and this produces one value for the full joint predictive distribution:

$$\widehat{\text{HDR-ECE}} = \frac{1}{K} \sum_{k=1}^{K} |\hat{p}_k - p_k|,$$

where $\hat{p}_k$ is an estimate of the empirical probability term, $P(\hat{F}_{Z|X}(Z) \leq p_k)$.

To compute SD-ECE, we first compute ECE using the notion of average calibration (Eq. 1) for each output dimension separately: $\widehat{\text{SD-ECE}}^d$, $d \in [D]$. Since this produces D values, and we take the average to summarize SD-ECE as a single scalar:

$$\widehat{\text{SD-ECE}} = \frac{1}{D} \sum_{d=1}^{D} \widehat{\text{SD-ECE}}^d$$

We point out that SD-ECE is simply a point of reference since it computes miscalibration as the sum of calibration error from each output dimension. Thus, this metric may not provide an accurate representation of miscalibration of the full joint distribution and may display pathologies, as described in Figure 1.

We refer the reader to Section E.1 for the full set of details on how each metric is estimated. Lastly, note that all three metrics (Energy score, HDR-ECE, SD-ECE) are negatively oriented, i.e. lower values are more desirable.

	Pre-hoc			SD Recalibration			HDR Recalibration		
Dataset	Energy	HDR-ECE	SD-ECE	Energy	HDR-ECE	SD-ECE	Energy	HDR-ECE	SD-ECE
scpf	3.97 (0.37)	0.30 (0.00)	0.15 (0.00)	4.00 (0.37)	**0.04 (0.00)**	**0.02 (0.00)**	**−0.79 (0.42)**	0.07 (0.01)	0.17 (0.00)
rf1	0.11 (0.01)	0.08 (0.00)	0.05 (0.00)	0.11 (0.01)	0.03 (0.00)	**0.01 (0.00)**	**0.08 (0.01)**	**0.01 (0.00)**	0.05 (0.00)
rf2	1.06 (0.28)	0.07 (0.00)	0.05 (0.00)	1.06 (0.28)	**0.04 (0.00)**	**0.01 (0.00)**	**1.04 (0.28)**	0.09 (0.01)	0.06 (0.00)
scm1d	1.13 (0.00)	0.48 (0.00)	0.11 (0.00)	1.13 (0.00)	0.24 (0.00)	0.02 (0.00)	**0.36 (0.05)**	**0.04 (0.00)**	**0.01 (0.00)**
scm20d	1.29 (0.01)	0.48 (0.00)	0.11 (0.00)	1.31 (0.01)	0.25 (0.00)	**0.02 (0.00)**	**0.81 (0.09)**	**0.04 (0.00)**	**0.02 (0.00)**
Fusion1	2.48 (0.03)	0.34 (0.00)	0.11 (0.00)	2.48 (0.03)	0.14 (0.00)	**0.05 (0.00)**	**−3.73 (0.14)**	**0.13 (0.00)**	0.06 (0.00)
Fusion2	1.95 (0.01)	0.45 (0.00)	0.17 (0.00)	1.93 (0.01)	0.31 (0.00)	0.09 (0.00)	**−1.85 (0.05)**	**0.05 (0.00)**	**0.01 (0.00)**
Fusion3	4.79 (0.07)	0.35 (0.00)	0.09 (0.00)	5.03 (0.08)	0.17 (0.00)	**0.03 (0.00)**	**−5.01 (0.39)**	**0.09 (0.00)**	0.05 (0.00)

Table 1. Results from multi-dimensional regression and recalibration experiments. The mean is shown with 1 standard error in parentheses (0.00 indicates that the values were smaller than 2 decimal places). The lowest mean value for each metric is bolded. **(Top)** Results from the benchmark experiments. **(Bottom)** Results from the nuclear fusion dynamics model experiments.

Results. Table 1 provides results on both the benchmark (Top) and nuclear fusion tasks (Bottom). We see that across all 5 benchmark datasets and the 3 nuclear fusion tasks, the energy score indicates that the samples produced by HDR recalibration are the highest quality and has best approximated the ground truth distribution. HDR recalibration also improves HDR-ECE compared to the pre-hoc model on 4 out of 5 benchmark datasets and on all three fusion tasks, which is expected given HDR recalibration aims to minimize HDR-ECE. Likewise, SD recalibration aims to minimize SD-ECE, which it achieves on 4 out of 5 benchmark datasets and 2 out of 3 fusion tasks. However, the fact that SD recalibration does not improve the energy score further supports the argument that SD-ECE is *not an adequate metric* for assessing multi-dimensional predictions.

4.2. Decision-making with Demand Forecasts

We apply the proposed recalibration algorithm in a decision-making setting, where a decision-maker (in this case a grocery store manager) must forecast future demand for store items and stock the items accordingly. Similar to the inventory management experiments in Kuleshov et al. (2018); Malik et al. (2019), we take the “Corporacion Favorita” Kaggle dataset (Favorita et al., 2017), which records the historical sales of items from a supermarket chain in Ecuador. We take the top three most sold items between the dates 2015-01-01 to 2017-08-11 and set up the modeling problem s.t. the grocery store forecasts the demand for the three products in the next business day, given the recent four day history of the sales and variables that indicate the day of the week and week of the year. Given that sales of items in a store may display dependence (e.g. seasonality or cannibalization), modeling the dependencies across the three items will be important.

We set up the decision-making problem s.t. the grocery store attributes very high loss to under-stocking (e.g. loss of reputation and future demand) and also incurs a small loss for over-stocking (e.g. possible spoilage and waste):

$$\text{Loss}_t = 10 * Q_{t,\text{under-stock}} + 1 * Q_{t,\text{over-stock}},$$

where Loss_t denotes the loss on day t, $Q_{t,\text{under-stock}}$ denotes the quantity of under-stocked items on day t, and $Q_{t,\text{over-stock}}$ denotes the quantity of over-stocked items on day t. The demand forecasts are generated as samples from the predictive distribution (i.e. possible realizations of the next day demand), and we use a cautious decision policy that makes the decisions based on the sample that forecasts the highest demand. With a total of 795 days in the dataset, we train the probabilistic model with the first 559 days while using the subsequent 159 days as the validation set, and we use this same validation set for recalibration. With the remaining 77 days, we simulate the decision-making problem and record the accumulated loss. We repeat the simulation across 5 different seeds and report the average loss and standard error. We refer the reader to Section E.4 for more details on the experiment setup.

	Pre-hoc	SD Recal	HDR Recal
Loss	916.81 (4.16)	910.82 (5.19)	**865.34** (7.83)

Table 2. Total loss incurred by each method in sales simulation experiment. The mean loss is shown with 1 standard error in parentheses.

Table 2 displays the accumulated loss based on each method. While SD recalibration marginally outperforms the distribution with no recalibration (Pre-hoc), HDR recalibration significantly improves the loss, demonstrating that the joint calibration of predictions provides utility in this decision-making setting.

5. Discussion

In this work, we addressed the problem of recalibrating multi-dimensional distributional predictions. Prior recalibration methods consider each target dimension separately and fail to take into account dependencies that may exist in the dimensions. Bridging ideas in calibration of projections of multivariate targets, HPD values, and HDRs, we defined the notion of HDR calibration and proposed the HDR recalibration algorithm.

HDR calibration leverages the property that for any distributional prediction setting, the p-HDR of the optimal prediction will contain the true observations with empirical frequency p. As HDRs consider the full joint distribution, it is a more adequate notion for assessment of calibration of multi-dimensional distributional predictions. The proposed HDR recalibration algorithm aims to achieve HDR calibration by performing recalibration in the projected space, sampling from the recalibrated projection distribution, and mapping the projection samples back to the target space. This produces a sampling-based representation of the HDR calibrated distribution. Across the suite of benchmark multi-dimensional regression tasks, plasma dynamics prediction tasks, and decision-making task, HDR recalibration consistently improves the quality of the predictive samples compared to the baseline methods.

We note that HDR calibration is a specific instance of calibration in the projected space with the predictive PDF as the projection function. Other projection functions which capture various aspects of the data distribution can also be used, and we leave for future work exploring such projections.

Acknowledgements

YC is supported by the AI Research Institutes Program funded by the National Science Foundation under AI Institute for Societal Decision Making (AI-SDM), Award No. 2229881 and the Kwanjeong Educational Foundation. IC is supported in part by US Department of Energy grants under contract numbers DE-SC0021414 and DE-AC02-09CH1146 and the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745016 and DGE2140739. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors would also like to thank the anonymous ICML reviewers for their valuable feedback.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning, in particular, probabilistic methods, uncertainty quantification, and calibration. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbate, J., Conlin, R., and Kolemen, E. Data-driven profile prediction for diii-d. *Nuclear Fusion*, 61(4):046027, 2021.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Barbe, P., Genest, C., Ghoudi, K., and Remillard, B. On kendall's process. *journal of multivariate analysis*, 58(2): 197–229, 1996.

Belloni, A. and Winkler, R. L. On multivariate quantiles under partial ordering. Technical report, 2009.

Boyer, M., Wai, J., Clement, M., Kolemen, E., Char, I., Chung, Y., Neiswanger, W., and Schneider, J. Machine learning for tokamak scenario optimization: combining accelerating physics models and empirical models. In *APS Division of Plasma Physics Meeting Abstracts*, volume 2021, pp. PP11–164, 2021.

Char, I., Chung, Y., Neiswanger, W., Kandasamy, K., Nelson, A. O., Boyer, M., Kolemen, E., and Schneider, J. Offline contextual bayesian optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Char, I., Abbate, J., Bardóczi, L., Boyer, M., Chung, Y., Conlin, R., Erickson, K., Mehta, V., Richner, N., Kolemen, E., et al. Offline model-based reinforcement learning for tokamak control. In *Learning for Dynamics and Control Conference*, pp. 1357–1372. PMLR, 2023a.

Char, I., Chung, Y., Shah, R., Neiswanger, W., and Schneider, J. Correlated trajectory uncertainty for adaptive sequential decision making. In *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2023b.

Chua, K., Calandra, R., McAllister, R., and Levine, S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

Chung, Y., Char, I., Guo, H., Schneider, J., and Neiswanger, W. Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification. *arXiv preprint arXiv:2109.10254*, 2021a.

Chung, Y., Neiswanger, W., Char, I., and Schneider, J. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. *Advances in Neural Information Processing Systems*, 34:10971–10984, 2021b.

Chung, Y., Rumack, A., and Gupta, C. Parity calibration. In Evans, R. J. and Shpitser, I. (eds.), *Proceedings of* *the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 413–423. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/chung23a.html.

Cui, P., Hu, W., and Zhu, J. Calibrated reliable regression using maximum mean discrepancy. *Advances in Neural Information Processing Systems*, 33, 2020.

Dalmasso, N., Pospisil, T., Lee, A. B., Izbicki, R., Freeman, P. E., and Malz, A. I. Conditional density estimation tools in python and r with applications to photometric redshifts and likelihood-free cosmological inference. *Astronomy and Computing*, 30:100362, 2020.

Deshpande, S. and Kuleshov, V. Calibration improves bayesian optimization. *arXiv preprint arXiv:2112.04620*, 2021.

Duan, T., Anand, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A., and Schuler, A. Ngboost: Natural gradient boosting for probabilistic prediction. In *International conference on machine learning*, pp. 2690–2700. PMLR, 2020.

Duraisamy, K., Iaccarino, G., and Xiao, H. Turbulence modeling in the age of data. *Annual review of fluid mechanics*, 51:357–377, 2019.

Favorita, C., inversion, Elliot, J., and McDonald, M. Corporación favorita grocery sales forecasting, 2017. URL https://kaggle.com/competitions/favorita-grocery-sales-forecasting.

Feldman, S., Bates, S., and Romano, Y. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023.

Genest, C. and Rivest, L.-P. On the multivariate probability integral transformation. *Statistics & probability letters*, 53(4):391–399, 2001.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69 (2):243–268, 2007.

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-Volume 70*, pp. 1321–1330. JMLR. org, 2017.

Gupta, C. and Ramdas, A. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International Conference on Machine Learning*, pp. 3942–3952. PMLR, 2021.

Harrison, D., Sutton, D., Carvalho, P., and Hobson, M. Validation of bayesian posterior distributions using a multi-dimensional kolmogorov–smirnov test. *Monthly Notices of the Royal Astronomical Society*, 451(3):2610–2624, 2015.

Hébert-Johnson, U., Kim, M. P., Reingold, O., and Rothblum, G. N. Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*, 2017.

Hyndman, R. J. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, 1996.

Izbicki, R., Shimizu, G., and Stern, R. B. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *The Journal of Machine Learning Research*, 23(1):3772–3803, 2022.

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Jones, D. R., Schonlau, M., and Welch, W. J. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

Kandasamy, K., Krishnamurthy, A., Schneider, J., and Póczos, B. Parallelised bayesian optimisation via thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pp. 133–142. PMLR, 2018.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33: 21810–21823, 2020.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Kuleshov, V. and Deshpande, S. Calibrated and sharp uncertainties in deep learning via density estimation. In *International Conference on Machine Learning*, pp. 11683–11693. PMLR, 2022.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. *arXiv preprint arXiv:1807.00263*, 2018.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pp. 6402–6413, 2017.

Luxon, J. L. A design retrospective of the diii-d tokamak. *Nuclear Fusion*, 42(5):614, 2002.

Malik, A., Kuleshov, V., Song, J., Nemer, D., Seymour, H., and Ermon, S. Calibrated model-based deep reinforcement learning. *arXiv preprint arXiv:1906.08312*, 2019.

Mehta, V., Paria, B., Schneider, J., Ermon, S., and Neiswanger, W. An experimental design perspective on model-based reinforcement learning. *arXiv preprint arXiv:2112.05244*, 2021.

Mehta, V., Char, I., Abbate, J., Conlin, R., Boyer, M. D., Ermon, S., Schneider, J., and Neiswanger, W. Exploration via planning for information about the optimal trajectory. *arXiv preprint arXiv:2210.04642*, 2022.

Nelsen, R. B., Quesada-Molina, J. J., Rodríguez-Lallena, J. A., and Úbeda-Flores, M. Kendall distribution functions. *Statistics & probability letters*, 65(3):263–268, 2003.

Nix, D. A. and Weigend, A. S. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994.

Pearce, T., Zaki, M., Brintrup, A., and Neely, A. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. *arXiv preprint arXiv:1802.07167*, 2018.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Rawlings, J. B. Tutorial overview of model predictive control. *IEEE control systems magazine*, 20(3):38–52, 2000.

Sahoo, R., Zhao, S., Chen, A., and Ermon, S. Reliable decisions with threshold calibration. *Advances in Neural Information Processing Systems*, 34:1831–1844, 2021.

Seo, J., Na, Y.-S., Kim, B., Lee, C., Park, M., Park, S., and Lee, Y. Feedforward beta control in the kstar tokamak by deep reinforcement learning. *Nuclear Fusion*, 61(10): 106010, 2021.

Seo, J., Na, Y.-S., Kim, B., Lee, C., Park, M., Park, S., and Lee, Y. Development of an operation trajectory design algorithm for control of multiple 0d parameters using deep reinforcement learning in kstar. *Nuclear Fusion*, 62 (8):086049, 2022.

Sexton, D. M., Murphy, J. M., Collins, M., and Webb, M. J. Multivariate probabilistic projections using imperfect climate models part i: outline of methodology. *Climate dynamics*, 38:2513–2542, 2012.

Song, H., Diethe, T., Kull, M., and Flach, P. Distribution calibration for regression. In *International Conference on Machine Learning*, pp. 5897–5906. PMLR, 2019.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Tran, K., Neiswanger, W., Yoon, J., Zhang, Q., Xing, E., and Ulissi, Z. W. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology*, 1(2):025006, 2020.

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., and Vlahavas, I. Mulan: A java library for multi-label learning. *The Journal of Machine Learning Research*, 12: 2411–2414, 2011.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

Zadrozny, B. and Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pp. 609–616. Citeseer, 2001.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Zhao, D., Dalmasso, N., Izbicki, R., and Lee, A. B. Diagnostics for conditional density models and bayesian inference algorithms. In *Uncertainty in Artificial Intelligence*, pp. 1830–1840. PMLR, 2021.

Zhao, S., Ma, T., and Ermon, S. Individual calibration with randomized forecasting. In *International Conference on Machine Learning*, pp. 11387–11397. PMLR, 2020.

Ziegel, J. F. and Gneiting, T. Copula calibration. *Electronic journal of statistics*, 8(2):2619–2638, 2014.

A. Theoretical Statements

Proposition 2.1 The optimal distributional prediction, i.e. $\hat{F}_{Y|X} = F_{Y|X}$, satisfies calibration in the projected space, Eq. 3.

Proof. Recall the definition of calibration in the projected space, which we restate here:

$$P(Z \leq \hat{F}_{Z|X}^{-1}(p)) = p, \forall p \in (0,1), \tag{12}$$

where $Z \coloneqq g(X,Y), X \sim f_X, Y \sim f_{Y|X}$ and $\hat{F}_{Z|X}$ is the CDF of the r.v. $\hat{Z}|X$.

For any $p \in (0,1)$,

$$\begin{aligned} &P(Z \leq \hat{F}_{Z|X}^{-1}(p)) && (13)\\ &= P(\hat{F}_{Z|X}(Z) \leq p) && (14)\\ &= \int_{\mathcal{X}} P\left(\hat{F}_{Z|x}(Z) \leq p \mid X = x\right) dF_X(x) && (15) \end{aligned}$$

By the condition of the statement, we have $\hat{F}_{Y|X} = F_{Y|X}$, thus $\hat{F}_{Y|x} = F_{Y|x}$.

$$\begin{aligned} &\hat{F}_{Y|x} = F_{Y|x} && (16)\\ &\Rightarrow \hat{Y}|x \overset{d}{=} Y|x \qquad (\hat{Y}|x \sim \hat{F}_{Y|x}, Y|x \sim F_{Y|x}, \text{and “} \overset{d}{=} \text{”denotes “equal in distribution”}) && (17)\\ &\Rightarrow g(x,\hat{Y})|x \overset{d}{=} g(x,Y)|x && (18)\\ &\Rightarrow \hat{Z}|x \overset{d}{=} Z|x \qquad (\hat{Z}|x \coloneqq g(x,\hat{Y})|x \text{ and } Z|x \coloneqq g(x,Y)|x) && (19)\\ &\Rightarrow \hat{F}_{Z|x} = F_{Z|x} \qquad (\hat{Z}|x \sim \hat{F}_{Z|x} \text{ and } Z|x \sim F_{Z|x}) && (20) \end{aligned}$$

Then, by the probability integral transform, $\hat{F}_{Z|x}(Z|x) \sim \mathcal{U}(0,1)$ and $P\left(\hat{F}_{Z|x}(Z) \leq p \mid X = x\right) = p$.

Thus we have

$$\begin{aligned} &\int_{\mathcal{X}} P\left(\hat{F}_{Z|x}(Z) \leq p \mid X = x\right) dF_X(x) && (21)\\ &= \int_{\mathcal{X}} p dF_X(x) && (22)\\ &= \int_{\mathcal{X}} p f_X(x) dx && (23)\\ &= p \int_{\mathcal{X}} f_X(x) dx && (24)\\ &= p. && (25) \end{aligned}$$

□

Proposition 2.3 HDR calibration holds if and only if Equation 3 holds.

Proof. We first prove “HDR calibration holds” $\implies$ Equation 3.

For any given $p \in (0,1)$.

$$
\begin{aligned}
&P(Y \in \texttt{HDR}_X(p)) && (26)\\
&= P(\texttt{HPD}_X(Y) \leq p) && (27)\\
&= \int_{\mathcal{X}} P(\texttt{HPD}_x(Y) \leq p \mid X = x) dF_X(x) && (28)\\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbb{I}\{\texttt{HPD}_x(y) \leq p\} dF_{Y|x}(y) dF_X(x) && (29)\\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbb{I}\{1 - \hat{F}_{Z|x}(\hat{f}_{Y|x}(y)) \leq p\} dF_{Y|x}(y) dF_X(x) && (30)\\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbb{I}\{1 - p \leq \hat{F}_{Z|x}(\hat{f}_{Y|x}(y))\} dF_{Y|x}(y) dF_X(x) && (31)\\
&= \int_{\mathcal{X}} P(1 - p \leq \hat{F}_{Z|x}(\hat{f}_{Y|x}(Y)) \mid X = x) dF_X(x) && (32)\\
&= P(1 - p \leq \hat{F}_{Z|X}(\hat{f}_{Y|X}(Y)) && (33)\\
&= 1 - P(\hat{F}_{Z|X}(\hat{f}_{Y|X}(Y)) \leq 1 - p) && (34)\\
&= p \quad \text{(by condition that HDR calibration holds)} && (35)
\end{aligned}
$$

Thus, $\forall p \in (0,1)$,

$$
\begin{aligned}
&1 - P(\hat{F}_{Z|X}(\hat{f}_{Y|X}(Y)) \leq 1 - p) = p && (36)\\
&\iff P(\hat{F}_{Z|X}(\hat{f}_{Y|X}(Y)) \leq 1 - p) = 1 - p && (37)\\
&\iff P(\hat{F}_{Z|X}(\hat{f}_{Y|X}(Y)) \leq p) = p \quad \text{(Equation 3)}, && (38)
\end{aligned}
$$

proving that "HDR calibration holds" $\implies$ Equation 3.

We can reverse all of the steps to prove the other direction, which completes the proof.

□

Proposition 3.1 Consider $R \circ \hat{F}_{Z|X}$ for an invertible mapping R. Then $R \circ \hat{F}_{Z|X}$ satisfies Eq. 3, i.e.

$$P(R \circ \hat{F}_{Z|X}(Z) \leq p) = p \quad \forall p \in (0,1).$$

Proof. For any fixed $p \in (0,1)$, let $q = R^{-1}(p)$.

Note the following equality

$$
\begin{aligned}
&P(\hat{F}_{Z|X}(Z) \leq p) && (39)\\
&= \int_{\mathcal{X}} P(\hat{F}_Z(Z|x) \leq p \mid X = x) dF(x). && (40)
\end{aligned}
$$

Applying R to $\hat{F}_{Z|X}$, we have

$$
\begin{aligned}
&P(R \circ \hat{F}_{Z|X}(Z) \leq p) && (41)\\
&\int_{\mathcal{X}} P(R \circ \hat{F}_{Z|x}(Z) \leq p \mid X = x) dF(x) && (42)\\
&= \int_{\mathcal{X}} P(\hat{F}_{Z|x}(Z) \leq R^{-1}(p) \mid X = x) dF(x) && (43)\\
&= \int_{\mathcal{X}} P(\hat{F}_{Z|x}(Z) \leq q \mid X = x) dF(x) && (44)\\
&= P(\hat{F}_{Z|X}(Z) \leq q) && (45)\\
&= p \quad \text{(by definition of } R(q)\text{)} && (46)
\end{aligned}
$$

□

Proposition 3.2 Suppose that $\hat{Z} \sim R \circ \hat{F}_{Z|X}$ and that R is an invertible mapping. Then the distribution of $\hat{Y}$ is HDR calibrated.

Proof. To first clarify the notation in the proposition statement, in writing $\hat{Z} \sim R \circ \hat{F}_{Z|X}$, the notation for $\hat{Z}$ has been overloaded as in the main text, we have stated that $\hat{Z}$ is the r.v. of the distribution function $\hat{F}_{Z|X}$.

Following the context of Section 3.1, the $\hat{Y}$ and $\hat{Z}$ in this proposition statement should be taken as any arbitrary r.v. that is generated as follows: $\hat{Y}$ is an arbitrary r.v. in $\mathcal{Y}$, and $\hat{Z}$ is the corresponding r.v. in the projected space induced by the projection function $\hat{f}_{Y|X}$ and the r.v. $\hat{Y}$.

By Proposition 3.1, we know that if $\hat{Z} \sim R \circ \hat{F}_{Z|x}$, then Eq. 3 holds.

By Proposition 2.3, we know that Eq. 3 is equivalent to HDR calibration, i.e. the distribution function of $\hat{Y}$ satisfies HDR calibration.

□

B. Additional Details on Related Works

Izbicki et al. (2022) proposes "HPD-split" as a conformal prediction method that utilizes HPD values (Eq. 4) as the split residuals, and simply by the method of split conformal prediction, given a fixed level $\alpha \in (0, 1)$, the prediction set produced by HPD-split is guaranteed at least $1 - \alpha$ coverage on average over the data distribution, i.e. marginal validity (Definition 1, Izbicki et al. (2022)) holds.

As discussed in Section 2.4, while their method does not consider multi-dimensional target spaces, because of the intrinsic relationship between HPD values and HDRs (Eq. 8), it is interesting to consider the relationship between the prediction set produced by HPD-split, HDR calibration (Defn. 2.2), and our proposed recalibration algorithm (Algorithms 1, 2).

Since HDR calibration is a notion of calibration defined for a predictive *distribution* function $\hat{f}, \hat{F}$, strictly speaking, one cannot state that the prediction *set* by HPD-split is HDR calibrated since there is no notion of a probability distribution function: given a fixed $p \in (0, 1)$, HPD-split produces a set $\mathcal{S} \subseteq \mathcal{Y}$ s.t. $P(Y \in \mathcal{S}) \geq p$, but following Definition 2.2, one cannot construct $\texttt{HDR}(p)$ as the concept of HDR is intrinsically tied to a distribution function. This is not particular to HPD-split, but a key feature of all conformal prediction methods that differentiates it from calibration (or recalibration) methods: conformal prediction methods output prediction *sets* for a specified α-level, while calibration methods output full predictive *distributions*.

However, conversely, one can construct prediction sets from predictive distributions, and we can show that a prediction set constructed from the HDR recalibration procedure satisfies marginal validity that holds for HPD-split, i.e. our recalibration procedure shares the conformal guarantees of HPD-split.

Proof. Assume an invertible interpolation algorithm for the recalibration mapping R in Algorithm 1 s.t. $\forall \alpha \in \{\frac{1}{N}, \frac{2}{N}, \dots \frac{N-1}{N}, 1\}$, $R^{-1}(\alpha) = c_{(N\alpha)}$ (Recall that $c_{(i)}$ is the i^{th} order statistic of the recalibration dataset $\mathcal{C}'$ in Algorithm 1). Assume a fixed level $\alpha \in \{\frac{1}{N}, \frac{2}{N}, \dots \frac{N-1}{N}, 1\}$.

Recall that the CDF of the recalibrated distribution in $\mathcal{Z}$ space is $R \circ \hat{F}_{Z|X}$ (Section 3.1). Given a test input point x, consider constructing a 1-sided prediction interval in $\mathcal{Z}$ with expected coverage equal to $1 - \alpha$, i.e. $\{z : R \circ \hat{F}_{Z|x}(z) \geq \alpha)\}$.

$$\begin{aligned}
&\{z : R \circ \hat{F}_{Z|x}(z) \geq \alpha)\} \\
&= \{z : \hat{F}_{Z|x}(z) \geq R^{-1}(\alpha)\} \\
&= \{z : \hat{F}_{Z|x}(z) \geq c_{(N\alpha)}\}
\end{aligned}$$

Following the definition $Z \coloneqq \hat{f}_{Y|X}(Y)$, we have that the pre-image of this set in $\mathcal{Y}$ space is the following set

$$\mathcal{C}(x) = \{y : \hat{F}_{Z|x}(\hat{f}_{Y|x}(y)) \geq c_{(N\alpha)}\}$$

Note that this set is identical to the conformal prediction set in Definition 15 of Izbicki et al. (2022), and following their Theorem 20, this set satisfies marginal validity, i.e.

$$P(Y \in C(X)) \geq 1 - \alpha$$

□

Feldman et al. (2023) is another conformal prediction method, but which aims to produce prediction sets in multi-dimensional target spaces. Their method relies on training a VAE on the dataset to learn a representation that is amenable to performing quantile regression. In addition to the point that it is a conformal method which produces prediction sets instead of distributions, we note that their work is somewhat orthogonal to our setting as it is a *pre-hoc* method, whereas we are focused on *post-hoc* methods which can be applied on top of pre-hoc trained models (as discussed in Section 2.4).

C. Additional Definitions

Average Calibration. We provide additional notes on average calibration (Eq. 1), which we restate here:

$$P(Y \leq \hat{F}^{-1}_{Y|X}(p)) = p, \forall p \in [0,1].$$

We can rewrite this expression by conditioning on $X = x$ and applying the law of total probability:

$$P(Y \leq \hat{F}^{-1}_{Y|X}(p)) = \mathbb{E}_{x \sim f_X}\left[P\left(Y \leq \hat{F}^{-1}_{Y|x}(p) \middle| X = x\right)\right].$$

Copula Calibration and Kendall Distribution. We first re-state copula calibration using our notation.

Copula calibration requires Eq. 3 with $Z \coloneqq \hat{F}_{Y|X}(Y)$, $\hat{Z} \coloneqq \hat{F}_{Y|X}(\hat{Y})$, and $\hat{F}_{Z|X} : [0,1] \to [0,1]$ is the Kendall distribution: $\hat{F}_{Z|X}(p) = P(\hat{Z} \leq p)$.

Utilizing the notation from Ziegel & Gneiting (2014) the Kendall distribution of any CDF F is denoted as $\mathcal{K}_F : [0,1] \to [0,1]$, where $\mathcal{K}_F(p) = P(F(Y) \leq p)$ for $p \in [0,1]$, and copula calibration is defined as

$$P(\mathcal{K}_{\hat{F}_X}[\hat{F}_X(Y)] \leq p) = p, \forall p \in [0,1].$$

Highest Density Region Hyndman (1996) defines highest density regions as follows. Given a coverage level p, the p-HDR of the predictive PDF $\hat{f}_{Y|x}$, $R_{\hat{f}_{Y|x}}(p)$ is defined as

$$R_{\hat{f}_{Y|x}}(p) = \{y : \hat{f}_{Y|x}(y) \geq \hat{f}_p\}, \tag{47}$$

$$\text{where } \hat{f}_p = \arg\sup_{\hat{f}_p}\{P(\hat{Y} \in R_{\hat{f}_{Y|x}}(p) \mid X = x) \geq p\}. \tag{48}$$

We note that this definition, which is stated here nearly verbatim from the original definitions from Hyndman (1996), is recursive. To clarify, $R_{\hat{f}_{Y|x}}(p)$ is the prediction set which has the highest density values w.r.t. $\hat{f}_{Y|x}$ and which has p integrated probability. We refer the reader to Figure 1 of Hyndman (1996) or Figure 2 of Zhao et al. (2021) for helpful visualizations.

D. Additional Details on Algorithms

D.1. Subroutine Algorithms

This section provides details on the algorithms that are used as subroutines of the main algorithms (Algorithms 1 and 2 from Section 3).

Algorithm 3 is used in Algorithm 1 to construct the recalibration mapping dataset.

Algorithm 4 is used in Algorithm 2 to re-sample from a set of predictive samples to construct a set of recalibrated samples. We point out that Algorithm 4 provides an implementation of using binning to apply the recalibration mapping R during re-sampling. However, there can be other ways of applying R during re-sampling (e.g. isotonic regression).

Algorithms 5, 6, 7 comprise the PDF adjustment step described in Section 3.2, which is used in Line 2 of both Algorithms 1 and 2. While the PDF adjustment step is meant to be a general procedure to correct for biases or correlations that are evident based on the predictions on a held-out dataset, because diagonal Gaussians are commonly used for multi-dimensional distributional predictions, we provide an instantiation of the adjustment procedure specifically for diagonal Gaussian distributions. In our experiments, we used a composition of all three algorithms in order - i.e. the PDF was first adjusted for the mean, then the standard deviation, and lastly the covariance. We note that Algorithm 6 requires a choice of loss function defined for univariate Gaussians, and in our experiments, we used SD-ECE (single dimensional expected calibration error) as the loss function.

Lastly, note that in practice, Algorithms 5, 6, 7 are actually run only once during Training (Line 2 of Algorithm 1). During testing, the learned adjustment functions are simply applied to the distributional predictions (Line 2 of Algorithm 2).

Algorithm 3 Constructing the Recalibration Mapping Dataset

```
**Input**: Calibration dataset $\{(x_i, y_i)\}_{i=1}^{N}$, predictive PDF $\hat{f}_{Y|X}$, number of samples $M$.
$\mathcal{C} \leftarrow \varnothing$
**for** $i \in [N]$ **do**
    Draw $M$ samples from $\hat{f}_{Y|x_i}$ to construct $\{\hat{y}_{i,j}\}_{j=1}^{M}$ and apply $\hat{f}_{Y|x_i}$ to each sample to construct $\{\hat{z}_{i,j}\}_{j=1}^{M}$ where $\hat{z}_{i,j} = \hat{f}_{Y|x_i}(\hat{y}_{i,j})$
    $z_i \leftarrow \hat{f}_{Y|x_i}(y_i)$
    Estimate empirical CDF $\hat{F}_{Z|x_i}$ with $\{\hat{z}_{i,j}\}_{j=1}^{M}$ and evaluate $z_i$: $r_i = \hat{F}_{Z|x_i}(z_i)$
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{r_i\}$
**end for**
**Output**: $\mathcal{C}$
```

Algorithm 4 Producing HDR Recalibrated Samples via Binning

```
**Input**: Dataset of tuples of predictive samples and projections $\mathcal{D} = \{(\hat{y}_j, \hat{z}_j)\}_{j=1}^{M}$, recalibration mapping $R$, number of bins $B$.
Place $\mathcal{D}$ into $B$ equal width bins $\mathcal{M} = \{b_i = [l_i, u_i)\}_{i=1}^{B}$ w.r.t $\{\hat{z}_j\}_{j=1}^{M}$, s.t. $\min_{\hat{z}_j}(\hat{y}_j, \hat{z}_j) \in b_1$, $\max_{\hat{z}_j}(\hat{y}_j, \hat{z}_j) \in b_B$, and the number of elements in each bin $|b_i| = \lfloor \frac{M}{B} \rfloor$.
Recalibrated samples $\mathcal{D}' \leftarrow \varnothing$
**for** $i \in [B]$ **do**
    Sampling rate for bin $s_i = R\left(i * \lfloor \frac{M}{B} \rfloor / M\right) - R\left((i-1) * \lfloor \frac{M}{B} \rfloor / M\right)$.
    Sample from $b_i$ with probability $s_i$ to construct the dataset $\{(\hat{y}_k, \hat{z}_k)\}_{k=1}^{K_i}$ where $K_i = \lfloor (s_i * \lfloor \frac{M}{B} \rfloor) \rfloor$
    $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{\hat{y}_k\}_{k=1}^{K_i}$.
**end for**
**Output**: $\mathcal{D}'$.
```

Algorithm 5 Mean Adjustment for Gaussian Distributions

```
**Input**: Calibration dataset $\{(x_i, y_i)\}_{i=1}^{N}$, predictive Gaussian PDF $\hat{f}_{Y|X} \coloneqq (\hat{\mu}_X, \hat{\sigma}_X)$.
Predict the conditional mean at each $x_i$, $\hat{\mu}_i$, and compute the bias: $\texttt{bias} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mu}_i)$.
Define the mean adjustment function: $\mathcal{A}\left((\hat{\mu}, \hat{\sigma})\right) = (\hat{\mu} + \texttt{bias}, \hat{\sigma})$
**Output**: Mean adjusted Gaussian distribution $\mathcal{A}\left((\hat{\mu}_X, \hat{\sigma}_X)\right)$.
```

Algorithm 6 Standard Deviation Adjustment for Diagonal Gaussian Distributions

```
**Input**: Calibration dataset $\{(x_i, y_i)\}_{i=1}^N$, predictive Gaussian PDF $\hat{f}_{Y|X} := (\hat{\mu}_X, \hat{\sigma}_X)$, loss function for univariate Gaussian distributions $\mathcal{L} : (\mu \times \sigma, \mathcal{Y}) \to \mathbb{R}$
Predict the conditional mean and standard deviation at each $x_i$: $(\hat{\mu}_i, \hat{\sigma}_i)$.
Optimized standard deviation ratios $\mathcal{S} \leftarrow [\,]$.
**for** $d \in [D]$ **do**
    $c^d = \arg\min_{c \in \mathbb{R}_+} \frac{1}{N} \sum_{i=1}^N \mathcal{L}\left((\hat{\mu}_i^d, c * \hat{\sigma}_i^d), y_i^d\right)$.
    Append $c^d$ to $\mathcal{S}$.
**end for**
Concatenate $\mathcal{S}$ into a vector and denote it $\hat{s} \in \mathbb{R}^D$.
Define the standard deviation adjustment function: $\mathcal{A}\left((\hat{\mu}, \hat{\sigma})\right) = (\hat{\mu}, \hat{s} \odot \hat{\sigma})$, where $\odot$ indicates element-wise product.
**Output**: Standard deviation adjusted Gaussian distribution $\mathcal{A}\left((\hat{\mu}_X, \hat{\sigma}_X)\right)$.
```

Algorithm 7 Covariance Adjustment for Diagonal Gaussian Distributions

```
**Input**: Calibration dataset $\{(x_i, y_i)\}_{i=1}^N$, predictive Gaussian PDF $\hat{f}_{Y|X} := (\hat{\mu}_X, \hat{\sigma}_X)$.
Predict the conditional mean and standard deviation at each $x_i$: $(\hat{\mu}_i, \hat{\sigma}_i)$.
Compute the error in mean prediction $\{\epsilon_i\}_{i=1}^N$, where $\epsilon_i = y_i - \hat{\mu}_i$.
Compute the error correlation matrix from $\{\epsilon_i\}_{i=1}^N$: $\hat{\rho} \in \mathbb{R}^{D \times D}$.
Define the covariance adjustment function, $\mathcal{A}\left((\hat{\mu}, \hat{\sigma})\right) = (\hat{\mu}, \hat{\Sigma}(\hat{\sigma}, \hat{\rho}))$, where $\hat{\Sigma}(\hat{\sigma}, \hat{\rho}) = \mathrm{Diag}(\hat{\sigma}) \cdot \hat{\rho} \cdot \mathrm{Diag}(\hat{\sigma})$, $\mathrm{Diag}(\hat{\sigma})$ denotes the $D \times D$ diagonal matrix with elements of $\hat{\sigma}$ in the diagonal, and $\cdot$ is the standard matrix-matrix product operation.
**Output**: Covariance adjusted Gaussian distribution $\mathcal{A}\left((\hat{\mu}_X, \hat{\sigma}_X)\right)$.
```

D.2. Algorithm Analysis

We provide an analysis of the computational complexity of the main algorithms, Algorithms 1 and 2. Because the ADJUST step (Line 2) is an auxiliary procedure, we first analyze the run time of both algorithms excluding this step.

Algorithm 1

- Line 3: When constructing the dataset $\mathcal{C}$, computing the value $\hat{F}_{Z|x_i}(z_i)$ involves drawing M samples from $\hat{f}_{Y|x_i}$ and sorting M values of z_i, hence takes $O(M \log M)$ time for each $i \in [N]$, and thus takes $O(NM \log M)$ time.
- Line 4: Sorting N values takes $O(N \log N)$ time.
- Line 5: Learning R depends on the algorithm used. If one uses binning, this step takes no additional time since the bins are already defined by $\mathcal{C}'$.

Therefore, the whole procedure takes $O(NM \log M)$ time, where N is the number of datapoints in the calibration dataset, and M is the number of samples drawn to estimate the empirical CDF $\hat{F}_{Z|X}$.

Algorithm 2

- Line 3: Drawing M samples and applying $\hat{f}_{Y|x}$ on each sample takes $O(M)$ time.
- Line 4: Following Algorithm 4:
  - Line 2 of Algorithm 4: The most expensive step is sorting M values, which takes $O(M \log M)$ time.
  - Lines 4-8 for-loop of Algorithm 4: In each iteration, we re-sample from each bin of $\lfloor \frac{M}{B} \rfloor$ points, which takes $O(M/B)$ time. Since there are B iterations, the whole for-loop takes $O(M)$ time.

Therefore, the whole procedure takes $O(M \log M)$ time, where M is the number of samples drawn to express the predictive distribution $\hat{f}_{Y|x}$.

PDF Adjustment Step (**ADJUST**)

- Algorithm 5 takes $O(N)$ time.
- The run time of Algorithm 6 depends on the optimization algorithm used in Line 5, but since the optimization is repeated for D dimensions with a dataset of N points, it takes at least $\Omega(ND)$ time.
- In Algorithm 7, correlation estimation (Line 4) takes $O(ND^2)$ time, and the covariance matrix estimation involves two matrix multiplication of size $D \times D$, which takes $O(D^3)$ time, hence the overall complexity is $O(ND^2 + D^3)$.

E. Additional Details on Experiments

E.1. Evaluation Metrics

Energy score Given a test datapoint (x_i, y_i) and the predictive distribution at x_i, $\hat{f}_{Y|x_i}$, the (negatively-oriented) energy score, ES($\hat{f}_{Y|x_i}, y_i$), is defined as

$$\text{ES}(\hat{f}_{Y|x_i}, y_i) = \mathbb{E}_{\hat{f}_{Y|x_i}} \left\| \hat{Y} - y_i \right\|_2^\beta - \frac{1}{2} \mathbb{E}_{\hat{f}_{Y|x_i}} \left\| \hat{Y} - \hat{Y}' \right\|_2^\beta,$$

where each of $\hat{Y}$ and $\hat{Y}'$ are independent r.v.'s following the distribution $\hat{f}_{Y|x_i}$, and β is a hyperparameter $\in (0, 2)$. We estimate this score with

$$\widehat{\text{ES}}(\hat{f}_{Y|x_i}, y_i) = \frac{1}{|\mathcal{S}|} \sum_{\hat{y} \in \mathcal{S}} \|\hat{y} - y_i\|_2^\beta - \frac{1}{2|\mathcal{S}'||\mathcal{S}''|} \sum_{\hat{y}' \in \mathcal{S}', \hat{y}'' \in \mathcal{S}''} \|\hat{y}' - \hat{y}''\|_2^\beta$$

by drawing finite sets of samples $\mathcal{S}, \mathcal{S}'$, and $\mathcal{S}''$ independently from the distribution $\hat{f}_{Y|x_i}$. The exact number of samples drawn for each of $\mathcal{S}, \mathcal{S}', \mathcal{S}''$ is provided in the following sections on each of the experiments. We set $\beta = 1.7$ for all of our experiments, but other values also result in similar trends as reported.

Given the test set, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, we report the mean energy score over $\mathcal{D}$:

$$\widehat{ES} = \frac{1}{N} \sum_{i=1}^N \widehat{\text{ES}}(\hat{f}_{Y|x_i}, y_i).$$

HDR-ECE HDR-ECE (highest density region expected calibration error) is computed following the equation for expected calibration error (ECE) in Eq. 11 with the notion of HDR calibration, specifically Eq. 3, which is equivalent to Definition 2.2 by Proposition 2.3. We estimate HDR-ECE with the test set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ as follows. At a test input x_i, since we do not know the closed form of $\hat{F}_{Z|x_i}$, we estimate $\hat{F}_{Z|x_i}(z_i)$ by first drawing a set of samples from $\hat{f}_{Y|x_i}$: $\{\hat{y}_j\}_{j=1}^M$, then applying $\hat{f}_{Y|x_i}$ to each sample to construct $\mathcal{S}_i = \{\hat{z}_j = \hat{f}_{Y|x_i}(\hat{y}_j)\}_{j=1}^M \sim \hat{F}_{Z|x_i}$, and finally produce the estimate for $\hat{F}_{Z|x_i}(z_i)$ as

$$p_i = \frac{1}{M} \sum_{j=1}^M \mathbb{I}\{\hat{z}_j \leq z_i\} = \frac{1}{M} \sum_{j=1}^M \mathbb{I}\{\hat{f}_{Y|x_i}(\hat{y}_j) \leq \hat{f}_{Y|x_i}(y_i)\}.$$

Then for any $p_k \in [0, 1]$ we estimate the empirical probability term, $P(\hat{F}_{Z|X}(Z) \leq p_k)$, as

$$\hat{p}_k = \frac{1}{N} \sum_{i=1}^N \mathbb{I}\left\{p_i \leq p_k\right\}.$$

Finally, we choose K probability values: $0 \leq p_1 < p_2 \cdots < p_K \leq 1$ and estimate HDR-ECE as the empirical estimate of Eq. 11:

$$\widehat{\text{HDR-ECE}} = \frac{1}{K} \sum_{k=1}^K |\hat{p}_k - p_k|.$$

The exact number of samples drawn to estimate $\hat{F}_{Z|x_i}$ is provided in the following sections on each of the experiments. For the probability values, we set $K = 99$ and used the following grid of probability values: $[0.01, 0.02, 0.03, \ldots 0.98, 0.99]$.

SD-ECE SD-ECE (single dimensional expected calibration error) is computed as the mean of ECE measured for each output dimension independently following the notion of univariate calibration (Eq. 1). For each dimension d, given a test input x_i, we first estimate $\hat{F}_{Y^d|x_i^d}(y_i^d)$ with a set of predictive samples $\mathcal{S}_i = \{\hat{y}_j\}_{j=1}^M \sim \hat{f}_{Y|x_i}$ as

$$p_i^d = \frac{1}{M} \sum_{j=1}^M \mathbb{I}\{\hat{y}_j^d \leq y_i^d\}.$$

Then for a given probability $p_k \in [0, 1]$, we estimate the empirical coverage term, $P(\hat{F}_{Y^d|X^d}(Y^d) \leq p_k)$, as

$$\hat{p}_k^d = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left\{p_i^d \leq p_k\right\}.$$

Thus, we estimate the "SD-ECE at dimension d" (SD-ECE$^{\text{d}}$) as $\widehat{\text{SD-ECE}}^{\text{d}} = \frac{1}{K}\sum_{k=1}^{K} |\hat{p}_k^d - p_k|$, and take the average of $\widehat{\text{SD-ECE}}^{\text{d}}$ over all of the output dimensions to estimate SD-ECE:

$$\widehat{\text{SD-ECE}} = \frac{1}{D} \sum_{d=1}^{D} \widehat{\text{SD-ECE}}^{\text{d}},$$

where D is the number of dimensions of Y.

Again, the exact number of samples drawn to estimate $\hat{F}_{Y|x_i}$ is provided in the following sections on each of the experiments. Just as with HDR-ECE, for the probability values, we set $K = 99$ and used the following grid of probability values: $[0.01, 0.02, 0.03, \ldots 0.98, 0.99]$.

Lastly, we note that calibration plots (a.k.a. reliability diagrams) in Figures 1 and 2 were produced with the *Uncertainty Toolbox* (Chung et al., 2021a).

E.2. Benchmark Regression Tasks

We use the following 5 datasets from the "mulan" benchmark (Tsoumakas et al., 2011): **scpf** (3D), **rf1** (8D), **rf2** (8D), **scm1d** (16D), **scm20d** (16D). With each dataset, we make train-validation-test splits of proportions $[65\%, 20\%, 15\%]$, and train a probabilistic neural network (PNN) (Lakshminarayanan et al., 2017; Nix & Weigend, 1994) that predicts a diagonal Gaussian distribution.

For all of the datasets, the PNN trained has 5 fully connected layers, each with 200 hidden units, and the output parametrizes a diagonal Gaussian with a mean and a log-variance prediction. The Gaussian likelihood loss was used for training, with a learning rate of 0.001 and no weight decay was used. Training was halted early if the validation loss did not improve for more than 100 epochs, for a maximum of 1000 epochs. All of the models early-stopped their training.

After training a PNN on the train set, we learn the recalibration mapping on the validation split and produce 20 independent sets of recalibrated predictive samples on the test split (for methods *SD Recalibration* and *HDR Recalibration*). For the *Pre-hoc* method, samples were drawn from the PNN model without any recalibration. For **scpf**, each set contained 5000 samples of $\hat{y} \sim \hat{f}_{Y|x}$ at each test point x, for **rf1, rf2**, 8000 samples, and for **scm1d, scm20d**, 10000 samples. We compute each evaluation metric (Energy score, HDR-ECE, SD-ECE) on each of these sets of samples, then take the average for each metric across the 20 sets of samples. We then repeat this process (data splits, model training, recalibration and repeated sampling) with 5 different seeds, and report the mean and standard error of the metrics in Table 1 (Top).

E.3. Dynamics Modeling in Nuclear Fusion

We first provide some background information on the problem setup of modeling plasma dynamics for tokamaks. A tokmak is a device that magnetically confines a toroidal plasma, and it is one of the most promising devices for making nuclear fusion energy a reality. With the potential of providing an abundant source of safe and clean power generation, nuclear fusion, which is the physical process during which atomic nuclei combine together to form heavier atomic nuclei, is regarded as the power source of the future. However, fusion reactions are difficult to control, and recently, there has been increasing interest in both learning dynamics models (Boyer et al., 2021; Abbate et al., 2021) and applying those models for control of tokamaks (Mehta et al., 2021; 2022; Char et al., 2023a; Seo et al., 2021; 2022). In model-based control, a model of the system dynamics is learned and used e.g. to optimize a control policy offline, or the model is deployed online for model predictive control (Rawlings, 2000). In either case, the learned model is sampled repeatedly to optimize a control sequence, hence obtaining well-calibrated samples which reflect the intricacies of the true system dynamics is crucial (Malik et al., 2019; Chua et al., 2018).

We take 3 different pre-trained dynamics models that were used to optimize control policies for deployment on the DIII-D tokamak, a nuclear fusion device in San Diego that is operated by General Atomics (Luxon, 2002). All three models were trained with logged data from past experiments (referred to as “shots”) on this device. As input, the models take in the current state of the plasma and the actuators from the tokamak. They then output a multi-dimensional predictive distribution of several key plasma state variables in the next timestep. Two of the models, which we refer to as **Fusion1** and **Fusion2**, predict a 3-dimensional target: β_N (the ratio of plasma pressure over magnetic pressure), *density* (the line-averaged electron density), and *li* (internal inductance). The third model, **Fusion3**, predicts one additional variable, *dr* (differential rotation of the plasma), to predict a 4-dimensional target. For the actuators, the model takes in the amount of power and torque injected from the neutral beams, the current, the magnetic field, and four shape variables (elongation, a_{minor}, triangularity top, and triangularity bottom). This, along with the state space, make for an input dimension of 11 (for **Fusion1** and **Fusion2**) or 12 (for **Fusion3**).

All three pre-trained models have the same model architecture. It is a recurrent probabilistic neural network (RPNN), which features an encoding layer by an RNN with 64 hidden units followed by a fully connected layer with 256 units, and a decoding layer of fully connected layers with [128, 512, 128] units, which finally outputs a diagonal Gaussian parameterized by the mean and a log-variance prediction. The Gaussian likelihood loss was used for training, with a learning rate of 0.0003 and weight decay of 0.0001. In using dynamics models to sample trajectories and train policies, the key metric practitioners are concerned with is explained variance, hence explained variance on a held out validation set of 1000 shots was monitored during training and training was halted early if there was no improvement for more than 250 epochs. **Fusion2** and **Fusion3** were trained with a non-smoothed dataset consisting of 12000 shots in the training dataset, and **Fusion2** explains on average 57% of the variance in the outputs, and **Fusion3** 40%. **Fusion1** was trained with a smoothed version of the dataset and explains on average 63% of variance in the outputs.

For each of these models, we learn the recalibration mapping with a validation dataset consisting of 1000 shots’ worth of data for methods *SD Recalibration* and *HDR Recalibration*. For the *Pre-hoc* method, samples were drawn from the pre-trained models without any recalibration. For all models and methods, we drew 3000 samples at each test datapoint. We compute the average of each evaluation metric across 10 sets of predictive samples on a single set of 20 held-out test shots, and repeat this process for 10 different sets of 20 test shots. We report the mean and standard error across these 10 sets in Table 1 (Bottom).

E.4. Decision-making with Demand Forecasting

The demand forecasting model takes in the recent 4-day history of sales of each of the top three most sold items, and categorical variables which indicate the day of the week and week of the year, which makes for a total of 14 input dimensions. The model is then trained to predict a distribution over how much of each item was sold in the next business day. i.e. over 3-dimensional targets.

To demonstrate versatility of the proposed HDR recalibration method, in this experiment, we used NGBoost (natural gradient boosting (Duan et al., 2020)) for the demand forecasting model. We used the NGBRegressor, which predicts a diagonal Gaussian distribution. The model was trained with the CRPScore with a learning rate of 0.005, and number of estimators set to 1000. Training was stopped early if the CRPScore on the validation set did not improve for 20 iterations. The same validation set was used for recalibration.

During testing, we use the demand forecasting model to produce a set of samples that reflect possible realizations of demand for each of the three products in the next business day. Among these samples, we filter out samples which are over the budget. Then, with the remaining samples, we select the sample with the maximum sum of demand across the three products, and take this sample as the action – i.e. this sample is how much quantity of each product the store will prepare for the next business day. With the true sales data of the next business day, we compute the loss as described in Section 4.2. To set the budget, we took the mean of sales in the validation set.

For each method (Pre-hoc, HDR recalibration, SD recalibration), the simulation across the test set was repeated 10 times and the average of the cumulative loss across these 10 times was recorded. This full pipeline (model training, sampling, 10 repeated simulations) was done with 5 different seeds, and the mean and standard error of the cumulative loss across the 5 seeds is reported in Table 2.

E.5. Ablation Study on the Effects of PDF Adjustment in HDR Recalibration

We present an ablation study which demonstrates the effect of the PDF adjustment step, described in Section 3.2. For the 5 benchmark datasets, we run HDR recalibration *without* the adjustment step, and compare evaluation metrics against running

HDR recalibration *with* the adjustment. The results are shown in the table below. The "With Adjustment" columns have been reproduced here from Table 1 for convenience. The energy score indicates that the adjustment step improves the quality of the predictive samples.

	***Without* Adjustment**			***With* Adjustment**		
Dataset	Energy	HDR-ECE	SD-ECE	Energy	HDR-ECE	SD-ECE
scpf	3.55 (0.37)	**0.03 (0.00)**	**0.15 (0.00)**	**−0.79 (0.42)**	0.07 (0.01)	0.17 (0.00)
rf1	0.09 (0.02)	**0.01 (0.00)**	0.06 (0.00)	**0.08 (0.01)**	**0.01 (0.00)**	**0.05 (0.00)**
rf2	1.05 (0.29)	**0.01 (0.00)**	**0.05 (0.00)**	**1.04 (0.28)**	0.09 (0.01)	0.06 (0.00)
scm1d	1.11 (0.01)	0.40 (0.00)	0.07 (0.00)	**0.36 (0.05)**	**0.04 (0.00)**	**0.01 (0.00)**
scm20d	1.27 (0.01)	0.42 (0.00)	0.07 (0.00)	**0.81 (0.09)**	**0.04 (0.00)**	**0.02 (0.00)**

Table 3. Results from comparing the HDR recalibration method without and with the PDF adjustment step. The energy score indicates that the adjustment step significantly improves the sample quality when performing HDR recalibration.

E.6. Computing Infrastructure

All of the model training was done with 4 NVIDIA GeForce RTX 2080 Ti GPUs.

All of the evaluation was done on a CPU machine with Intel(R) Xeon(R) Gold 6238 CPU @ 2.10GHz.