Calibrated Regression Against An Adversary Without Regret

Shachi Deshpande[1] **Charles Marx**[2] **Volodymyr Kuleshov**[1]

[1]Computer Science Dept., Cornell University and Cornell Tech, New York, NY , USA
[2]Computer Science Dept., Stanford University, Stanford, CA, USA

Abstract

We are interested in probabilistic prediction in online settings in which data does not follow a probability distribution. Our work seeks to achieve two goals: (1) producing valid probabilities that accurately reflect model confidence; (2) ensuring that traditional notions of performance (e.g., high accuracy) still hold. We introduce online algorithms guaranteed to achieve these goals on arbitrary streams of datapoints, including data chosen by an adversary. Specifically, our algorithms produce forecasts that are (1) calibrated—i.e., an 80% confidence interval contains the true outcome 80% of the time—and (2) have low regret relative to a user-specified baseline model. We implement a post-hoc recalibration strategy that provably achieves these goals in regression; previous algorithms applied to classification or achieved (1) but not (2). In the context of Bayesian optimization, an online model-based decision-making task in which the data distribution shifts over time, our method yields accelerated convergence to improved optima.

1 INTRODUCTION

In applications of machine learning (ML), data can change over time. Online learning algorithms can guarantee good predictive accuracy (e.g., as measured by squared error) on arbitrary data streams, even ones chosen adversarially [Cesa-Bianchi and Lugosi, 2006, Shalev-Shwartz, 2007]. However, we are often interested not only in minimizing predictive error, but also in outputting valid probabilities representative of future outcomes [Vovk et al., 2005b, Kuleshov et al., 2018, Angelopoulos and Bates, 2021]. For example a doctor might wish to estimate the probability of a patient being sick; similarly, a power grid operator might want to know the likelihood that demand for electricity will increase.

In this paper, we are interested in probabilistic predictions in online settings where data does not follow a probability distribution [Shalev-Shwartz, 2007]. This setting is challenging because we need to achieve two goals on data that shifts over time: (1) producing valid probabilities that accurately reflect model confidence; (2) ensuring that traditional notions of performance (e.g., achieving a low squared error) still hold. Additionally, without a data distribution, these goals may not be straightforward to define.

Our approach towards the first goal uses calibration to define valid probabilistic forecasts [Foster and Vohra, 1998, Kuleshov and Ermon, 2017, Gibbs and Candès, 2021]. Intuitively, an algorithm outputs calibrated predictions if the predicted and the empirical probabilities of a predicted outcome match—i.e., an 80% confidence interval contains the true outcome 80% of the time. We formalize the second goal by requiring that calibrated predictions have low regret relative to a baseline uncalibrated forecaster, as measured by a proper score [Gneiting et al., 2007b]. We focus on real-valued outcomes, and define online calibrated regression, a task that seeks to achieve the above two goals.

We propose algorithms for online calibrated regression that output accurate probabilistic predictions via the post-hoc recalibration of a black-box baseline model. Unlike classical recalibration methods [Platt, 1999, Kuleshov et al., 2018], ours work on online non-IID data (even data chosen by an adversary). In contrast to classical online learning [Shalev-Shwartz, 2007], we provide guarantees on not only regret, but also on the validity of probabilistic forecasts. Crucially, unlike many online calibrated and conformal prediction algorithms for classification [Foster and Vohra, 1998] or regression [Gibbs and Candès, 2021], we ensure low regret relative to a baseline forecaster.

Accurate predictive uncertainties can be especially useful in decision-making settings, where an agent uses a model of future outcomes to estimate the results of its actions (e.g., the likelihood of treating a patient) [Malik et al., 2019]. We complement our algorithms with formal guarantees on

expected utility estimation in decision-making applications. We apply our algorithms to several regression tasks, as well in the context of Bayesian optimization, an online model-based decision-making task in which the data distribution shifts over time. We find that improved uncertainties in the Bayesian optimization model yield faster convergence to optimal solutions which are also often of higher quality.

Contributions. First, we formulate a new problem called online calibrated regression, which requires producing calibrated probabilities on potentially adversarial input while retaining the predictive power of a given baseline uncalibrated forecaster. Second, we propose an algorithm for this task that generalizes recalibration in regression to non-IID data. Third, we show that the algorithm can improve the performance of Bayesian optimization, highlighting its potential to improve decision-making.

2 BACKGROUND

We place our work in the framework of online learning [Shalev-Shwartz, 2007]. At each time step $t = 1, 2, ...$, we are given features $x_t \in \mathcal{X}$. We use a forecaster $H : \mathcal{X} \to \mathcal{F}$ to produce a forecast $f_t = H(x_t)$, $f_t \in \mathcal{F}$ in a set of forecasts $\mathcal{F}$ over a target $y \in \mathcal{Y}$. Nature then reveals the true target $y_t \in \mathcal{Y}$ and we incur a loss of $\ell(y_t, f_t)$, where $\ell : \mathcal{Y} \times \mathcal{F} \to \mathbb{R}^+$ is a loss function. Unlike in classical machine learning, we do not assume that the x_t, y_t are i.i.d.: they can be random, deterministic, or even chosen by an adversary. In this regime, online learning algorithms admit strong performance guarantees measured in terms of regret $R_T(g)$ relative to a constant prediction g, $R_T(g) = \sum_{t=1}^{T} \ell(y_t, f_t) - \ell(y_t, g)$. The worst-case regret at time T equals $R_T = \max_{g \in \mathcal{F}} R_T(g)$.

Online forecasting Our work extends the online learning setting to probabilistic predictions. We focus on regression, where $y_t \in \mathbb{R}$ and the prediction f_t can be represented by a cumulative distribution function (CDF), which we denote by $F_t : \mathbb{R} \to [0, 1]$; $F_t(z)$ denotes the predicted probability that y is less than z. The quality of probabilistic forecasts is evaluated using *proper* losses ℓ. Formally, a loss $\ell(y, f)$ is proper if $f \in \arg\min_{g \in \mathcal{F}} \mathbb{E}_{y \sim (f)} \ell(y, g)\ \forall f \in \mathcal{F}$.; i.e., the true data probability minimizes the loss. An important proper loss for CDF predictions is the continuous ranked probability score, defined as $\ell_{\text{CRPS}}(y, F) = \int_{-\infty}^{\infty} (F(z) - \mathbb{I}_{y \leq z})^2 dz$.

Online calibration Proper losses decompose into a calibration and a sharpness component: these quantities precisely define an ideal forecast. Intuitively, calibration means that a 60% prediction should be valid 60% of the time; sharpness means that confidence intervals should be tight.

In the online setting, there exist algorithms guaranteed to produce calibrated forecasts of binary outcomes $y_t \in \{0, 1\}$ even when the y_t is adversarial Foster and Vohra [1998], Cesa-Bianchi and Lugosi [2006], Abernethy et al. [2011]. These algorithms are oftentimes randomized; hence their guarantees hold almost surely (a.s.). Here, and in all other usages going forward, "almost surely" refers to the simulated randomness in the randomized algorithm, and not the data. However, most calibration methods do not account for covariates x_t Foster and Vohra [1998] or assume simple binary y_t Kuleshov and Ermon [2017], Foster and Hart [2023]. We extend this work to regression and add guarantees on regret. We provide a detailed comparison of our work with the broader literature along with some motivating examples in Appendix F.

3 ONLINE CALIBRATED REGRESSION

Next, we define a task in which our goal is to produce calibrated forecasts in a regression setting while maintaining the predictive accuracy of a baseline uncalibrated forecaster.

We start with a forecaster H (e.g., an online learning algorithm) that outputs uncalibrated forecasts F_t at each step; these forecasts are fed into a *recalibrator* such that the resulting forecasts G_t are calibrated and have low regret relative to the baseline forecasts F_t. Formally, we introduce the setup of *online recalibration*, in which at every step $t = 1, 2, ...$ we have:

1: Nature reveals features $x_t \in \mathbb{R}^d$. Forecaster H predicts $F_t = H(x_t)$
2: A recalibration algorithm produces a calibrated forecast G_t based on F_t.
3: Nature reveals continuous label $y_t \in \mathcal{Y} \subseteq \mathcal{R}$ bounded by $|y_t| < B/2$, where $B > 0$.
4: Based on x_t, y_t, we update the recalibration algorithm and optionally update H.

Our task is to produce calibrated forecasts. Intuitively, we say that a forecast F_t is calibrated if for every $y' \in \mathcal{Y}$, the probability $F_t(y')$ on average matches the frequency of the event $\{y \leq y'\}$—in other words the F_t behave like calibrated CDFs. We formalize this intuition by introducing the ratio

$$\rho_T(y, p) = \frac{\sum_{t=1}^{T} \mathbb{I}_{y_t \leq y, F_t(y) = p}}{\sum_{t=1}^{T} \mathbb{I}_{F_t(y) = p}}. \tag{1}$$

Intuitively, we want $\rho_T(y, p) \to p$, as $T \to \infty$ for all y. In other words, out of the times when the predicted probability $F_t(y')$ for $\{y_t \leq y'\}$ to be p, the event $\{y_t \leq y'\}$ holds a fraction p of the time. We define $\rho_T(y, p)$ to be zero when the denominator in Equation (1) is zero. Below, we enforce that $\rho_T(y, p) \to p$ for forecasts p that are played infinitely often, in that $\sum_{t=1}^{T} \mathbb{I}_{F_t(y) = p} \to \infty$; if a forecast ceases to be played, there is no need (or opportunity) to improve calibration for that forecast.

We measure calibration using an extension of the aforementioned calibration error C_T. We define the calibration error of forecasts $\{F_t\}$ as

$$C_T(y) = \sum_{p \in P_T(y)} |\rho_T(y,p) - p| \left(\frac{1}{T} \sum_{t=1}^{T} \mathbb{I}_{\{F_t(y)=p\}} \right), \tag{2}$$

where $P_T(y) = \{F_1(y), F_2(y), ..., F_T(y)\}$ is the set of previous predictions for $\{y_t \leq y\}$. To measure (mis)calibration for the recalibrated forecasts G_t, we replace F_t with G_t in Equation (2).

Definition 1. *A sequence of forecasts G_t is ϵ-calibrated for $y \in \mathcal{Y}$ if $C_T(y) \leq R_T + \epsilon$ for $R_T = o(1)$, where R_T represents the convergence rate.*

The interpretation of ϵ-calibration is simple: for example, if $\epsilon = 0.01$, then of the times when we predict a 90% chance of rain, the observed occurrence of rain will be between 89% and 91%. For most applications, an error tolerance of a few % is acceptable. Note that the use of an error tolerance ϵ mirrors previous works [Foster and Vohra, 1998, Abernethy et al., 2011, Kuleshov and Ermon, 2017].

The goal of recalibration is also to produce forecasts that have high predictive value [Gneiting et al., 2007a]. We enforce this by requiring that the G_t have low regret relative to the baseline F_t in terms of the CRPS proper loss. Since the expected CRPS is a sum of calibration and sharpness terms, by maintaining a good CRPS while being calibrated, we effectively implement Gneitig's principle of maximizing sharpness subject to calibration [Gneiting et al., 2007b]. Formally, this yields the following definition.

Definition 2. *A sequence of forecasts G_t is ϵ-recalibrated relative to forecasts F_t if (a) the forecasts G_t are ϵ-calibrated for all $y \in \mathcal{Y}$ and (b) the regret of G_t with respect to F_t is a.s. small w.r.t. ℓ_{CRPS}:*

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} (\ell_{CRPS}(y_t, G_t) - \ell_{CRPS}(y_t, F_t)) \leq \epsilon.$$

4 ALGORITHMS FOR ONLINE REGRESSION

Next, we propose an algorithm for performing online recalibration (Algorithm 1). This algorithm sequentially observes uncalibrated CDF forecasts F_t and returns forecasts G_t such that $G_t(z)$ is a calibrated estimate for the outcome $y_t \leq z$. This algorithm relies on a classical calibration subroutine (e.g., Foster and Vohra [1998]), which it uses in a black-box manner to construct G_t.

Algorithm 1 can be seen as producing a $[0,1] \to [0,1]$ mapping that remaps the probability of each z into its correct value. More formally, Algorithm 1 partitions $[0,1]$ into M intervals $\mathcal{I} = \{[0, \frac{1}{M}), [\frac{1}{M}, \frac{2}{M}), ..., [\frac{M-1}{M}, 1]\}$; each interval is associated with an instance S^{cal} of a binary calibration algorithm (e.g., Foster and Vohra [1998]; see below). In order to compute $G_t(z)$, we compute $p_{tz} = F_t(z)$ and invoke the subroutine S_j^{cal} associated with interval I_j containing p_{tz}. After observing y_t, each S_j^{cal} observes the binary outcome $o_{tj} = \mathbb{I}_{F_t(y_t) \leq \frac{j}{M}}$ and updates itself.

Algorithm 1 Online Recalibration

Require: Online binary calibration subroutine S^{cal} with resolution N; number of intervals M

Initialize $\mathcal{I} = \{[0, \frac{1}{M}), [\frac{1}{M}, \frac{2}{M}), ..., [\frac{M-1}{M}, 1]\}$, a set of intervals that partition $[0,1]$.
Initialize $\mathcal{S} = \{S_j^{\text{cal}} \mid j = 0, ..., M-1\}$, a set of M instances of S^{cal}, one per $I_j \in \mathcal{I}$.
for $t = 1, 2, ...$: **do**
  Observe uncalibrated forecast F_t.
  Define $G_t(z)$ as the output of $S^{\text{cal}}_{\lfloor F_t(z) \rfloor}$, where $\lfloor F_t(z) \rfloor$ is the index of the subroutine associated with the interval containing $F_t(z)$.
  Output G_t. Observe y_t and update recalibrator:
  for $j = 1, 2, ..., M$: **do**
    $o_{tj} = 1$ if $F(y_t) \leq \frac{j}{M}$ else 0. Pass o_{tj} to S_j^{cal}.

4.1 ONLINE BINARY CALIBRATION SUBROUTINES

A key component of Algorithm 1 is the binary calibration subroutine S^{cal}. This subroutine is treated as a black box, hence can implement a range of known algorithms including regret minimization [Foster and Vohra, 1998, Cesa-Bianchi and Lugosi, 2006], Blackwell approchability [Abernethy et al., 2011] or defensive forecasting [Vovk et al., 2005b]. More formally, let p_{tj} denote the output of the j-th calibration subroutine S_j^{cal} at time t. For any $p \in [0,1]$, we define $\rho_T^{(j)}(p) = (\sum_{t=1}^{T} o_{tj} \mathbb{I}_{p_{tj}=p}) / (\sum_{t=1}^{T} \mathbb{I}_{p_{tj}=p})$ to be the empirical frequency of the event $\{o_{tj} = 1\}$ Online calibration subroutines ensure that $\rho_T^{(j)}(p) \approx p$.

Assumptions. Specifically, a subroutine S_j^{cal} normally outputs a set of discretized probabilities i/N for $i \in \{0, 1, ..., N\}$. We refer to N as their resolution. We define the calibration error of S_j^{cal} at i/N as $C_{T,i}^{(j)} = \left| \rho_T^{(j)}(i/N) - \frac{i}{N} \right| \left(\frac{1}{T} \sum_{t=1}^{T} \mathbb{I}_{t,i}^{(j)} \right)$ where $\mathbb{I}_{t,i}^{(j)} = \mathbb{I}\{p_{tj} = i/N\}$. We may write the calibration loss of S_j^{cal} as $C_T^{(j)} = \sum_{i=0}^{N} C_{T,i}^{(j)}$.

We will assume that the subroutine S^{cal} used in Algorithm 1 is ϵ-calibrated in that $C_T^{(j)} \leq R_T + \epsilon$ uniformly ($R_T = o(1)$ as $T \to \infty$). Recall also that the target y_t is bounded as $|y_t| < B/2$.

4.2 ONLINE RECALIBRATION PRODUCES CALIBRATED FORECASTS

Intuitively, Algorithm 1 produces valid calibrated estimates $G_t(z)$ for each z because each S_j^{cal} is a calibrated subroutine. More formally, we seek to quantify the calibration of Algorithm 1. Since the S^{cal} output discretized probabilities, we may define the calibration loss of Algorithm 1 at y as

$$C_T(y) = \sum_{i=0}^{N} \left| \rho_T(y, i/N) - \frac{i}{N} \right| \left(\frac{1}{T} \sum_{t=1}^{T} \mathbb{I}_{t,i} \right), \quad (3)$$

where $\mathbb{I}_{t,i} = \mathbb{I}\{F(y_t) = i/N\}$. The following lemma establishes that combining the predictions of each S_j^{cal} preserves their calibration. Specifically, the calibration error of Algorithm 1 is bounded by a weighted average of R_{T_j} terms, each is $o(1)$, hence the bound is also $o(1)$ (see next section).

Lemma 1 (Preserving calibration)**.** *Given $y \in \mathcal{Y}$, let $T_j = |\{1 \le t \le T : \lfloor F_t(y) \rfloor = j/M\}|$ denote the number of calls to S_j^{cal} by Algorithm 1. If each S_j^{cal} is ϵ-calibrated, then Algorithm 1 is also ϵ-calibrated and the following bound holds uniformly a.s. over T:*

$$C_T(y) \le \sum_{j=1}^{M} \frac{T_j}{T} R_{T_j} + \epsilon \quad (4)$$

4.3 ONLINE RECALIBRATION PRODUCES FORECASTS WITH VANISHING REGRET

Next, we want to show that the G_t do not decrease the predictive performance of the F_t, as measured by ℓ_{CRPS}. Intuitively, this is true because the ℓ_{CRPS} is a proper loss that is the sum of calibration and sharpness, the former of which improves in G_t.

Establishing this result will rely on the following key technical lemma [Kuleshov and Ermon, 2017] (see Appendix).

Lemma 2. *Each ϵ-calibrated S_j^{cal} a.s. has a small regret w.r.t. the ℓ_2 norm and satisfies uniformly over time T_j the bound* $\max_{i,k} \sum_{t=1}^{T_j} \mathbb{I}_{p_{tj}=i/N} \left(\ell_2(o_{tj}, i/N) - \ell_2(o_{tj}, k/N) \right) \le 2(R_{T_j} + \epsilon)$.

An important consequence of Lemma 2 is that a calibrated algorithm has vanishing regret relative to any fixed prediction (since minimizing internal regret also minimizes external regret). Using this fact, it becomes possible to establish that Algorithm 1 is at least as accurate as the baseline forecaster.

Lemma 3 (Recalibration with low regret accuracy)**.** *Consider Algorithm 1 with parameters $M \ge N > 1/\epsilon$ and let ℓ be the CRPS proper loss. Then the recalibrated G_t a.s. have vanishing ℓ-loss regret relative to F_t and we have a.s.:*

$$\frac{1}{T} \sum_{t=1}^{T} \ell(y_t, G_t) - \frac{1}{T} \sum_{t=1}^{T} \ell(y_t, F_t) < NBR_T + \frac{2B}{N}$$

Proof (sketch). When $p_{tj} = G_t(y)$ is the output of a given binary calibration subroutine S_j^{cal} at some y, we know what $\lfloor F(y) \rfloor = j/M$ (by construction). Additionally, we know from Lemma 2 that S_j^{cal} minimizes regret. Thus, it has vanishing regret in terms of ℓ_2 loss relative to the fixed prediction j/M: $\sum_{t=1}^{T_j} (o_{tj} - p_{tj})^2 \le \sum_{t=1}^{T_j} (o_{tj} - j/M)^2 + o(T_j)$. But $o_{tj} = \mathbb{I}_{F(y_t) \le j/m}$, and during the times t when S_j^{cal} was invoked, during the times t when S_j^{cal} was invoked $p_{tj} = G_t(y)$ and $j/M = F_t(y)$. Aggregating over j and integrating over y yields our result. □

These two lemmas lead to our main claim: that Algorithm 1 solves the online recalibration problem.

Theorem 1. *Let S^{cal} be an $(\epsilon/2B)$-calibrated online subroutine with resolution $N \ge 2B/\epsilon$. Then Algorithm 1 with parameters S^{cal} and $M = N$ outputs ϵ-recalibrated forecasts.*

Proof. By Lemma 1, Algorithm 1 is $(\epsilon/2B)$-calibrated and by Lemma 3, its regret w.r.t. the F_t tends to $< 2B/N < \epsilon$. Hence, Theorem 1 follows. □

General proper losses Throughout our analysis, we have used the CRPS loss to measure the regret of our algorithm. This raises the question: is the CRPS loss necessary? One answer to this question is that if the loss ℓ used to measure regret is not a proper loss, then recalibration is not possible.

Theorem 2. *If ℓ is not proper, then no algorithm achieves recalibration w.r.t. ℓ for all $\epsilon > 0$.*

On the other hand, in Appendix B, we provide a more general analysis that shows that: (1) a calibrated S^{cal} must have vanishing regret relative to a fixed prediction as measured using any proper score; (2) Algorithm 1 achieves vanishing regret relative to any proper score. See Appendix B for a formal statement and proof.

5 APPLICATIONS

5.1 CHOICE OF RECALIBRATION SUBROUTINE

Algorithm 1 is compabible with any binary recalibration subroutine S^{cal}. Two choices of S^{cal} include methods based on **internal regret minimization** [Mannor and Stoltz, 2010] and ones based on **Blackwell approachability** [Abernethy et al., 2011]. These yield different computational costs and convergence rates for Algorithm 1.

Specifically, recall that R_T denotes the rate of convergence of the calibration error C_T of Algorithm 1. For most online calibration subroutines S^{cal}, $R_T \le f(\epsilon)/\sqrt{T}$ for some $f(\epsilon)$.

In such cases, we can further bound the calibration error in Lemma 1 as

$$\sum_{j=1}^{M} \frac{T_j}{T} R_{T_j} \leq \sum_{j=1}^{M} \frac{\sqrt{T_j} f(\epsilon)}{T} \leq \frac{f(\epsilon)}{\sqrt{\epsilon T}}. \quad (5)$$

In the second inequality, we set the T_j to be equal. Thus, our recalibration procedure introduces an overhead of $\frac{1}{\sqrt{\epsilon}}$ in the convergence rate of the calibration error C_T and of the regret in Lemma 3. In addition, we require $\frac{1}{\epsilon}$ times more memory and computation time (we run $1/\epsilon$ instances of S_j^{cal}).

When using an internal regret minimization subroutine, the overall calibration error of Algorithm 1 is bounded as $O(1/\epsilon\sqrt{\epsilon T})$ with $O(1/\epsilon)$ time and $O(1/\epsilon^2)$ space complexity. These numbers improve to $O(\log(1/\epsilon))$ time complexity for a $O(1/\epsilon\sqrt{T})$ calibration bound when using the method of Abernethy et al. [2011] based on Blackwell approachability. The latter choice is what we recommend.

5.2 UNCERTAINTY ESTIMATION

We complement our results with ways in which Algorithm 1 can yield predictions for various confidence intervals.

Theorem 3. *Let G_t for $t = 1, 2, ..., T$ denote a sequence of $(\epsilon/2)$-calibrated forecasts. For any interval $[y_1, y_2]$, we have $\frac{1}{T}\sum_{t=1}^{T}(G_t(y_2) - G_t(y_1)) \to \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}\{y_t \in [y_1, y_2]\}$ as $T \to \infty$ a.s.*

This theorem justifies the use of $F_t(y_2) - F_t(y_1)$ to estimate the probability of the event that y_t falls in the interval $[y_1, y_2]$: on average, predicted probabilities will match true outcomes. The proof follows directly from the definition of ϵ-calibration. This result directly mirrors the construction for calibrated confidence intervals in Kuleshov et al. [2018].

5.3 ONLINE DECISION-MAKING

Consider a doctor seeing a stream of patients. For each patient x_t, they use a model M of an outcome y_t to estimate a loss $\ell(x_t) = \mathbb{E}_{y\sim M(x_t)}\ell(x_t, y, a(x_t))$ for a decision $a(x_t)$ (which could be $a(x_t) = \arg\min_a \mathbb{E}_{y\sim M(x_t)}[\ell(x_t, y, a)]$, e.g., a treatment that optimizes an expected outcome). We want to guarantee that the doctor's predictions will be correct: over time, the estimated expected value will not exceed from the realized loss. Crucially, we want this to hold in non-IID settings.

Our framework enables us to achieve this result with only a weak condition—calibration. The following concentration inequality shows that estimates of v are unlikely to exceed the true v on average (proof in Appendix C). If data was IID, this would be Markov's inequality: surprisingly, a similar statement holds in non-IID settings.

Theorem 4. *Let M be a calibrated model and let $\ell(y, a, x)$ be a monotonically non-increasing or non-decreasing loss in y. Then for any sequence $(x_t, y_t)_{t=1}^T$ and $r > 1$, we have:*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\left[\ell(y_t, a(x_t), x_t) \geq r\ell(x_t))\right] \leq 1/r \quad (6)$$

6 EXPERIMENTS

Next, we evaluate Algorithm 1 on regression tasks as well as on Bayesian optimization, a sequential decision-making process that induces a non-i.i.d. data distribution. We performed all experiments on a laptop, indicating the low overhead of our method.

Baselines We compare our randomized online calibration with two baselines. Calibrated regression is a popular algorithm for the IID setting [Kuleshov et al., 2018] and can be seen as estimating the same mapping as Algorithm 1 using kernel density estimation with a tophat kernel. Non-randomized online calibration uses the same subroutine as Algorithm 1, but outputs the expected probability as opposed to a random sample; we found this to be a strong baseline that outperforms simple density estimation and reveals the value of randomization.

Analysis of calibration. We assess the calibration of the base model and the recalibrated model with calibration scores defined using the probability integral transform [Gneiting et al., 2007a]. We define the calibration score as $\text{cal}(p_1, y_1, .., p_n, y_n) = \sum_{j=1}^{m}((q_j - q_{j-1}) - \hat{p_j})^2$, where $q_0 = 0 < q_1 < q_2 < .. < q_m = 1$ are m confidence levels. The $\hat{p_j}$ is estimated as $\hat{p_j} = |\{y_t | q_{j-1} \leq p_t \leq q_j, t = 1, .., N\}|/N$.

6.1 UCI DATASETS

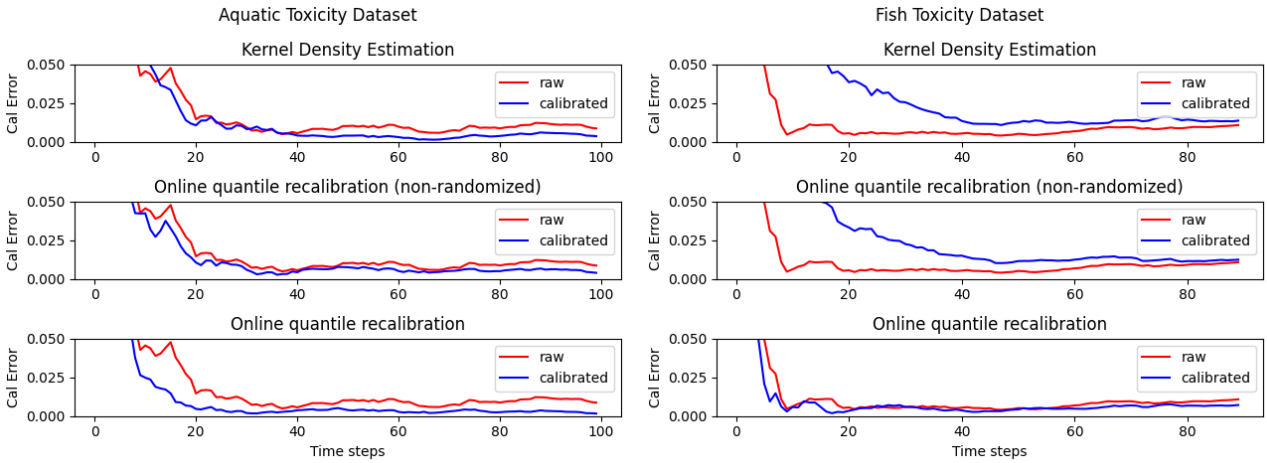
We experiment with four multivariate UCI datasets [Dua and Graff, 2017] to evaluate our online calibration algorithm.

Setup. Our dataset consists of input and output pairs $\{x_t, y_t\}_{t=1}^T$ where T is the size of the dataset. We simulate a stream of data by sending batches of data-points $\{x_t, y_t\}_{t=nt'+1}^{n(t'+1)}$ to our model, where t' is the time-step and n is the batch-size. This simulation is run for $\lceil T/n \rceil$ time-steps. For each batch, Bayesian ridge regression is fit to the data and the recalibrator is trained. We set $N = 20$ in the recalibrator and use a batch size of $n = 10$ unless stated otherwise.

Aquatic toxicity datasets We evaluate our algorithm on the QSAR (Quantitative Structure-Activity Relationship) Aquatic Toxicity Dataset 1(a) (batch size n=5) and Fish Toxicity Dataset 1(b) (batch size n=10), where aquatic toxicity towards two different types of fish is predicted using 8 and

Table 1: Evaluation of Online Calibration on UCI Datasets. We compare the performance of online calibration against non-randomized online calibration, kernel density estimation, and uncalibrated (i.e., raw) baselines. Our method produces the lowest calibration errors in the last time step. Results hold with std error quoted in braces (10 experimental runs, fixed dataset).

Dataset	Uncalibrated (Raw)	Kernel Density Estimation	Online Calibration (Non-randomized)	Online Calibration
Aq. Toxicity (Daphnia Magna)	0.0081 (0.0001)	0.0055 (0.0002)	0.0058 (0.0003)	**0.0027 (0.0001)**
Aq. Toxicity (Fathead Minnow)	0.0111 (0.0000)	0.0097 (0.0005)	0.0084 (0.0005)	**0.0031 (0.0003)**
Energy Efficiency	0.3322 (0.0001)	0.2857 (0.0356)	0.1702 (0.0094)	**0.1156 (0.0061)**
Facebook Comment Volume	0.2510 (0.0000)	0.0589 (0.0050)	0.0623 (0.0000)	**0.0518 (0.0002)**



(a) Aquatic Toxicity (Daphnia Magna)

(b) Aquatic Toxicity (Fathead Minnow)

Figure 1: Performance of Online Calibration on the Aquatic Toxicity Datasets. Aquatic toxicity towards two different types of fish (Daphnia Magna 1(a) and Fathead Minnow 1(b)) is predicted by the base model. In both datasets, online calibration outperforms the baseline methods.

6 molecular descriptors as features respectively. In Figure 1, we can see that the randomized online calibration algorithm produces a lower calibration error than the non-randomized baseline. We also compare the performance of our algorithm against uniform kernel density estimation by maintaining a running average of probabilities in each incoming batch of data-points. For the Fish Toxicity Dataset, we can see that only online calibration improves calibration errors relative to the baseline model. We report all final calibration errors in Table 1.

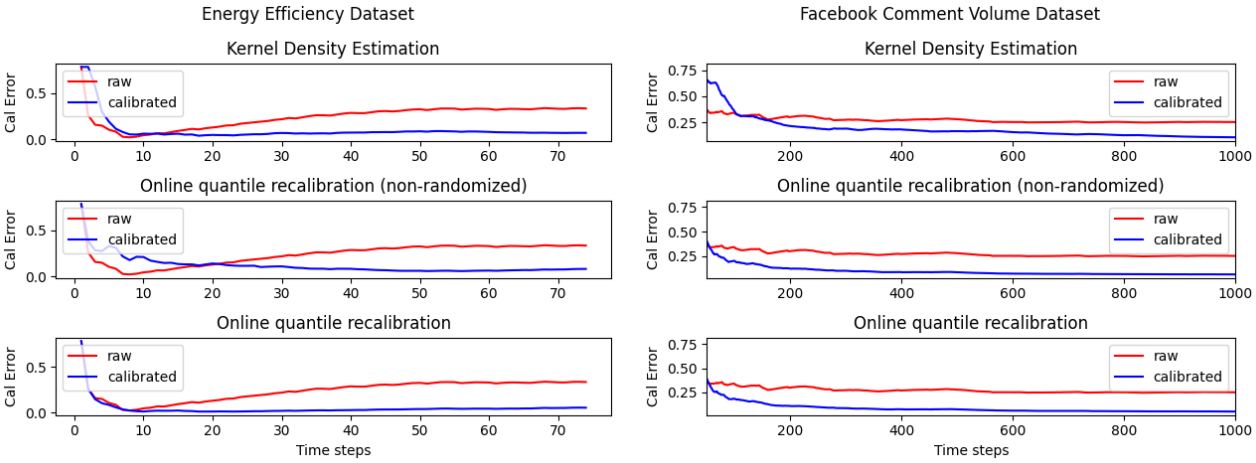
Energy efficiency dataset The heating load and cooling load of a building is predicted using 8 building parameters as features. In Figure 2(a), we see that the calibration errors produced by the online calibration algorithm drop sharply within the initial 10 time-steps. The baselines also produce a drop in calibration scores, but it happens more gradually.

Facebook comment volume dataset In Figure 2(b), the Facebook Comment Volume Dataset is used where the number of comments is to be predicted using 53 attributes associated with a post. We use the initial 10000 data-points from the dataset for this experiment. Here, the non-randomized and randomized online calibration algorithms produce a similar drop in calibration errors, but the randomized online calibration algorithm still dominates both baselines (Table 1).

6.2 BAYESIAN OPTIMIZATION

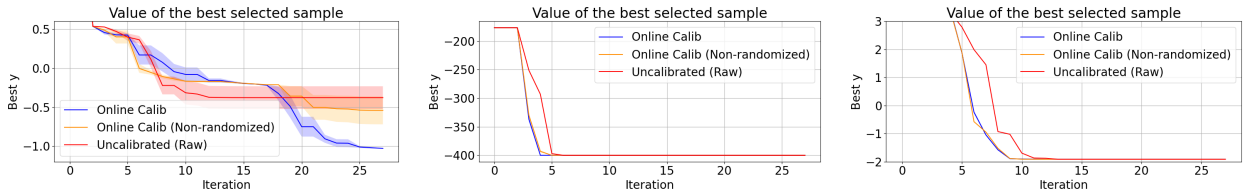
We also apply online recalibration in the context of Bayesian optimization, an online model-based decision-making task in which **the data distribution shifts over time** (it is the result of our actions). We find that improved uncertainties yield faster convergence to higher quality optima.

Setup Bayesian optimization attempts to find the global minimum $x^\star = \arg\min_{x \in \mathcal{X}} f(x)$ of an unknown function $f : \mathcal{X} \to \mathbb{R}$ over an input space $\mathcal{X} \subseteq \mathbb{R}^D$. We are given an initial labeled dataset $x_t, y_t \in \mathcal{X} \times \mathbb{R}$ for $n = 3$. At every time-step t, we use normal and recalibrated uncertainties from the probabilistic model $\mathcal{M} : \mathcal{X} \to (\mathbb{R} \to [0, 1])$ of f



(a) Energy Efficiency (b) Facebook Comment Volume

Figure 2: Performance of Online Calibration on the Energy Efficiency and Facebook Comment Volume Datasets. In both datasets, online recalibration (blue, bottom) attains a lower calibration error at a faster rate than baselines (red and top, middle).



(a) SixHumpCamel (b) Beale (c) Mccormick

Figure 3: Performance of Recalibration Methods on Bayesian Optimization Benchmarks

Table 2: Recalibrated Bayesian Optimization

Benchmark	Uncalibrated	Recalibrated
Ackley (2D)	9.925 (3.502)	**8.313 (3.403)**
SixHump (2D)	-0.378 (0.146)	**-1.029 (0.002)**
Ackley (10D)	14.638 (0.591)	**10.867 (2.343)**
Alpine (10D)	13.911 (1.846)	**12.163 (1.555)**

(here, a Gaussian Process) to select the next data-point x_{next} and iteratively update the model $\mathcal{M}$. We use popular benchmark functions to evaluate the performance of Bayesian optimization. We use the Lower Confidence Bound (LCB) acquisition function to select the data-point x_t. See Appendix E for details.

Table 2 shows that the online recalibration of uncertainties in a Bayesian optimization (BO) model achieves lower minima than an uncalibrated model (results averaged over 5 overall BO runs with fixed initialization). Figure 3 shows that online recalibrated Bayesian optimization can also reach optima in fewer steps. The error bars for the Beale and Mccormick functions are too small to be visible in the plots. All error bars denote standard errors.

7 DISCUSSION

Adversarial calibration methods Table 3 compares our method against its closest alternatives. Unlike previous algorithms aimed at classification that output a binary forecast $p_t \in [0, 1]$ [Foster and Vohra, 1998, Kuleshov and Ermon, 2017], we study marginal quantile calibration in regression. Our work resembles adaptive conformal inference [Gibbs and Candès, 2021], but provides a CDF-like object F_t instead of one confidence interval $q_t \in [0, 1]$ and yields a different notion of calibration. Crucially, we provide regret guarantees relative to a baseline model.

Specifically, our technical goal is marginal CDF calibration: estimating the probability of the event $y_t \leq y$ for all y. Note that these probabilities are marginal over the y_t; this is in contrast to conditional calibration for $y_t = 1|p_t = p$ as in Kuleshov and Ermon [2017]. We call our technical strategy online CDF regression (by analogy to quantile regression): we remap the predicted probabilities $F_t(y)$ (for

Table 3: Comparison to Existing Methods in the Literature

Method	Setting	Output	Calibration	Recalibrator	Regret	Proof Technique
Foster and Vohra [1998]	Class.	$p_t \in [0,1]$	Conditional	n/a	n/a	Int. regret min.
Kuleshov and Ermon [2017]	Class.	$p_t \in [0,1]$	Conditional	p-to-p	L2 loss	Int. regret min.
Gibbs and Candès [2021]	Regr.	$q_t \in [0,1]$	One quantile	q-to-q	n/a	Quantile regr.
Ours	Regr.	CDF F_t	CDF $\forall y$	$F(y)$-to-$F(y)$	CRPS	CDF regr.

any y) to a calibrated probability $R(F_t(y))$. Our proof technique establishes calibration by relating final calibration to the calibration of each subroutine using Jensen's inequality. We establish low regret by aggregating the regret of all the subroutines within one CRPS loss.

Most existing methods in online calibrated classification Foster and Vohra [1998], Vovk et al. [2005b], Abernethy et al. [2011], Okoroafor et al. [2024] or regression Gibbs and Candès [2021] do not provide guarantees for regret, except online recalibrated classification Kuleshov and Ermon [2017] and calibeating Foster and Hart [2023], Lee et al. [2022]. However, these methods are only for binary classification, whereas ours are for regression.When compared with Lee et al. [2022], our work achieves a different calibration definition that is more appropriate for continuous outcomes together with a different notion of regret (See Appendix F.1 for a detailed comparison).

Marginal calibration Our definition of calibration in regression is marginal across all x_t, y_t; this is in contrast to classification [Foster and Vohra, 1998], where calibration is conditional (also known as distributional) on each p. Marginal calibration implies that the true outcome falls below the 90% quantile 90% of times (averaged over all t). Distribution calibration in regression Kuleshov and Deshpande [2022] would be PPAD-hard by reduction from multiclass [Hazan and Kakade, 2012]. Marginal calibration is also currently a common definition of calibration for regression. For example, Kuleshov et al. [2018] in the IID setting or Gibbs and Candès [2021] in the online setting adopt this definition.

Batch vs online calibration Algorithm 1 can be seen as a direct counterpart to the histogram technique, a simple method for density estimation. With the histogram approach, the F_t is split into N bins, and the average y value is estimated for each bin. Because of the i.i.d. assumption, the output probabilities are calibrated, and the bin width determines the sharpness. Note that by Hoeffding's inequality, the average for a specific bin converges at a faster rate of $O(1/\sqrt{T_j})$[Devroye et al., 1996], as opposed to the $O(1/\sqrt{\epsilon T_j})$ rate given by Abernethy et al. [2011]; hence online calibration is harder than batch.

8 PREVIOUS WORK & CONCLUSION

Calibrated probabilities are widely used as confidence measures in the context of binary classification. Such probabilities are obtained via recalibration methods, of which Platt scaling Platt [1999] and isotonic regression Niculescu-Mizil and Caruana [2005] are by far the most popular. Recalibration methods also possess multiclass extensions, which typically involve training multiple one-vs-all predictors Zadrozny and Elkan [2002], as well as extensions to ranking losses Menon et al. [2012], combinations of estimators Zhong and Kwok [2013], and structured prediction Kuleshov and Liang [2015]. Recalibration algorithms have applied to improve reinforcement learning [Malik et al., 2019], Bayesian optimization [Deshpande et al., 2024, Stanton et al., 2023] and deep learning [Kuleshov and Deshpande, 2022]. Crucially, all of these methods implicitly rely on the assumption that data is sampled i.i.d. form an underlying distribution; they can be interpreted as density estimation techniques.

Online calibration was first proposed by [Foster and Vohra, 1998]. Existing algorithms are based on internal regret minimization Cesa-Bianchi and Lugosi [2006] or on Blackwell approachability Foster [1997]; recently, these approaches were shown to be closely related Abernethy et al. [2011], Mannor and Stoltz [2010]. Conformal prediction [Vovk et al., 2005b] is a technique for constructing calibrated predictive sets; it has been extended to handle distribution shifts [Hendrycks et al., 2018, Tibshirani et al., 2019, Barber et al., 2022], and online adversarial data [Gibbs and Candès, 2021].

Conclusion We presented a novel approach to uncertainty estimation that leverages online learning. Our approach extends existing online learning methods to handle predictive uncertainty while ensuring high accuracy, providing formal guarantees on calibration and regret on adversarial input.

We introduced a new problem called online calibrated forecasting, and proposed algorithms that generalize calibrated regression to non-IID settings. Our methods are effective on several predictive tasks and hold potential to improve performance in sequential model-based decision-making settings where we are likely to observe non-stationary data.

References

Jacob Abernethy, Peter L. Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. In *COLT 2011 - The 24th Annual Conference on Learning Theory*, pages 27–46, 2011.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021. URL https://arxiv.org/abs/2107.07511.

The GPyOpt authors. Gpyopt: A bayesian optimization framework in python. http://github.com/SheffieldML/GPyOpt, 2016.

Rina Barber, Emmanuel Candes, Aaditya Ramdas, and Ryan Tibshirani. Conformal prediction beyond exchangeability, 02 2022.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Shachi Deshpande, Charles Marx, and Volodymyr Kuleshov. Online calibrated and conformal prediction improves bayesian optimization, 2024. URL https://arxiv.org/abs/2112.04620.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Applications of mathematics. Springer, New York, Berlin, Heidelberg, 1996. ISBN 978-0-387-94618-4.

Ayya Dheur and Hatem Taleb. A large-scale study of probabilistic calibration in neural network regression. In *Proceedings of the 40th International Conference on Machine Learning*, ICML '23. PMLR, 2023.

Ayya Dheur and Hatem Taleb. Probabilistic calibration by design for neural network regression. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, AISTATS '24. PMLR, 2024.

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

Dean P. Foster. A Proof of Calibration Via Blackwell's Approachability Theorem. Discussion Papers 1182, Northwestern University, February 1997.

Dean P. Foster and Sergiu Hart. "calibeating": Beating forecasters at their own game. *Games and Economic Behavior*, 142:519–536, 2023.

Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration, 1998.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34*, pages 1896–1906. Curran Associates, Inc., 2021.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:243 – 268, 04 2007a. doi: 10.1111/j.1467-9868.2007.00587.x.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69(2):243–268, 2007b.

Elad Hazan and Sham M. Kakade. (weak) calibration is computationally hard. In *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, pages 3.1–3.10, 2012.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1044–1054. Curran Associates, Inc., 2018.

V. Kuleshov and P. Liang. Calibrated structured prediction. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Volodymyr Kuleshov and Shachi Deshpande. Calibrated and sharp uncertainties in deep learning via density estimation. In *Proceedings of the 39th International Conference on Machine Learning*, ICML '22. PMLR, 2022.

Volodymyr Kuleshov and Stefano Ermon. Estimating uncertainty online against an adversary. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI '17, pages 2110–2116, 2017.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18. PMLR, 2018.

Daniel Lee, Georgy Noarov, Mallesh Pai, and Aaron Roth. Online minimax multiobjective optimization: Multicalibeating and other applications. *Advances in Neural Information Processing Systems*, 35:29051–29063, 2022.

Ali Malik, Volodymyr Kuleshov, Jiaming Song, Danny Nemer, Harlan Seymour, and Stefano Ermon. Calibrated model-based deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, ICML '19. PMLR, 2019.

Shie Mannor and Gilles Stoltz. A geometric proof of calibration. *Math. Oper. Res.*, 35(4):721–727, 2010.

Charles Marx, Volodymyr Kuleshov, and Stefano Ermon. Calibrated probabilistic forecasts for arbitrary sequences. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Aditya Krishna Menon, Xiaoqian Jiang, Shankar Vembu, Charles Elkan, and Lucila Ohno-Machado. Predicting accurate probabilities with a ranking loss. In *Proceedings of the 29th International Conference on Machine Learning*, ICML '12, 2012.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, 2005.

Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional prediction for sequential decision making. In *The Twelfth International Conference on Learning Representations*, ICLR '24, 2024.

Princewill Okoroafor, Bobby Kleinberg, and Wen Sun. Faster recalibration of an online predictor via approachability. In *International Conference on Artificial Intelligence and Statistics*, pages 4690–4698. PMLR, 2024.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

Ramya Ramalingam, Shayan Kiyani, and Aaron Roth. The relationship between no-regret learning and online conformal prediction, 2025.

Shai Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. Phd thesis, Hebrew University, 2007.

Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *Proceedings of the 36th International Conference on Machine Learning*, ICML '19. PMLR, 2019.

Samuel Stanton, Wesley Maddox, and Andrew Gordon Wilson. Bayesian optimization with conformal prediction sets. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, AISTATS '23. PMLR, 2023.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9735–9746. Curran Associates, Inc., 2019.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, NY, USA, 2005a.

Vladimir Vovk, Akimichi Takemura, and Glenn Shafer. Defensive forecasting. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS 2005, Bridgetown, Barbados, January 6-8, 2005*, 2005b. URL `http://www.gatsby.ucl.ac.uk/aistats/fullpapers/224.pdf`.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 694–699, 2002.

Leon Wenliang Zhong and James T. Kwok. Accurate probability calibration for multiple classifiers. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 1939–1945. AAAI Press, 2013. ISBN 978-1-57735-633-2.

Calibrated Regression Against An Adversary Without Regret (Supplementary Material)

Shachi Deshpande[1] **Charles Marx**[2] **Volodymyr Kuleshov**[1]

[1]Computer Science Dept., Cornell University and Cornell Tech, New York, NY , USA
[2]Computer Science Dept., Stanford University, Stanford, CA, USA

A CORRECTNESS OF THE RECALIBRATION PROCEDURE

In the appendix, we provide the proofs of the theorems from the main part of the paper.

Notation We use $\mathbb{I}_E$ denote the indicator function of E, $[N]$ and $[N]_0$ to (respectively) denote the sets $\{1, 2, ..., N\}$ and $\{0, 1, 2, ..., N\}$, and Δ_d to denote a d-dimensional simplex.

Setup We place our work in the framework of online learning [Shalev-Shwartz, 2007]. At each time step $t = 1, 2, ...$, we are given features $x_t \in \mathcal{X}$. We use a forecaster $H : \mathcal{X} \to \mathcal{P}$ to produce a prediction $p_t = H(x_t)$, $p_t \in \mathcal{P}$ in the set of distributions $\mathcal{P}$ over a target $y \in \mathcal{Y}$. Nature then reveals the true target $y_t \in \mathcal{Y}$ and we incur a loss of $\ell(y_t, p_t)$, where $\ell : \mathcal{Y} \times \mathcal{P} \to \mathbb{R}^+$ is a loss function. The forecaster H updates itself based on x_t, y_t, and we proceed to time $t + 1$.

Unlike in classical machine learning, we do not assume that the x_t, y_t are i.i.d.: they can be random, deterministic or even chosen by an adversary. Online learning algorithms feature strong performance guarantees in this regime, where performance is usually measured in terms of regret $R_T(q)$ relative to a constant prediction q, $R_T(q) = \sum_{t=1}^T \ell(y_t, p_t) - \ell(y_t, q)$. The worst-case regret at time T equals $R_T = \max_{q \in \mathcal{P}} R_T(q)$.

In this paper, the predictions p_t are probability distributions over the outcome y_t. We focus on regression, where $y_t \in \mathbb{R}$ and the prediction p_t can be represented by a cumulative distribution function (CDF), denoted $F_t : \mathbb{R} \to [0, 1]$ and defined as $F_t(z) = p_t(y \leq z)$.

Learning with expert advice A special case of this framework arises when each x_t represents advice from N *experts*, and H outputs $p_t \in \Delta_{N-1}$, a distribution over experts. Nature reveals an outcome y_t, resulting in an expected loss of $\sum_{i=1}^N p_{ti}\ell(y_t, a_{ti})$, where $\ell(y_t, a_{ti})$ is the loss under expert i's advice a_{ti}. Performance in this setting is measured using two notions of regret.

Definition 3. *The external regret R_T^{ext} and the internal regret R_T^{int} are defined as*

$$R_T^{\text{ext}} = \sum_{t=1}^T \bar{\ell}(y_t, p_t) - \min_{i \in [N]} \sum_{t=1}^T \ell(y_t, a_{it}) \qquad R_T^{\text{int}} = \max_{i,j \in [N]} \sum_{t=1}^T p_{ti} \left(\ell(y_t, a_{it}) - \ell(y_t, a_{jt})\right),$$

where $\bar{\ell}(y, p) = \sum_{i=1}^N p_i \ell(y, a_{it})$ is the expected loss.

Calibration for online binary calibration For now, we focus on the ℓ_1 norm, and we define the calibration error of a forecaster S^{cal} as

$$C_T = \sum_{i=0}^N \left| \rho_T(i/N) - \frac{i}{N} \right| \left(\frac{1}{T} \sum_{t=1}^T \mathbb{I}_{\{p_t = \frac{i}{N}\}} \right), \tag{7}$$

where $\rho_T(p) = \frac{\sum_{t=1}^T y_t \mathbb{I}_{p_t = p}}{\sum_{t=1}^T \mathbb{I}_{p_t = p}}$ denotes the frequency at which event $y = 1$ occurred over the times when we predicted p.

We further define the calibration error when S_j^{cal} predicts i/N as

$$C_{T,i}^{(j)} = \left|\rho_T^{(j)}(i/N) - \frac{i}{N}\right| \left(\frac{1}{T_j}\sum_{t=1}^{T} \mathbb{I}_{t,i}^{(j)}\right)$$

where $\mathbb{I}_{t,i}^{(j)} = \mathbb{I}\{p_t = \frac{i}{N} \cap F_t^H \in [\frac{j-1}{M}, \frac{j}{M})\}$ is an indicator for the event that S_j^{cal} is triggered at time t and predicts i/N. Similarly, $\mathbb{I}_{t,i} = \mathbb{I}\{p_t = i/N\} = \sum_{j=1}^{M} \mathbb{I}_{t,i}^{(j)}$ indicates that i/N was predicted at time t, and $T_j = \sum_{t=1}^{T}\sum_{i=0}^{N} \mathbb{I}_{t,i}^{(j)}$ is the number of calls to S_j^{cal}. Also,

$$\rho_T^{(j)}(i/N) = \frac{\sum_{t=1}^{T} \mathbb{I}_{t,i}^{(j)} y_t}{\sum_{t=1}^{T} \mathbb{I}_{t,i}^{(j)}}$$

is the empirical success rate for S_j^{cal}.

Note that with these definitions, we may write the calibration losses of S_j^{cal} as $C_T^{(j)} = \sum_{i=0}^{N} C_{T,i}^{(j)}$.

Calibration for regression A sequence of forecasts F_t achieves online quantile calibration for all $y \in \mathcal{Y}$ and all $p \in \mathcal{P}$, $\rho_T(y,p) \to p$, a.s. as $T \to \infty$, where

$$\rho_T(y,p) = \frac{\sum_{t=1}^{T} \mathbb{I}_{y_t \le y, F_t(y)=p}}{\sum_{t=1}^{T} \mathbb{I}_{F_t(y)=p}} \tag{8}$$

In other words, out of the times when the predicted probability $F_t(y')$ for $\{y_t \le y'\}$ to be p, the event $\{y_t \le y'\}$ holds a fraction p of the time.

We also seek to quantify more precisely the calibration of Algorithm 1, specifically compare $\rho(y,p)$ with p. We define for this the quantity

$$C_{T,i}(y) = \left|\rho_T(y, i/N) - \frac{i}{N}\right| \left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{t,i}\right),$$

and we define the calibration loss of Algorithm 1 at y as $C_T(y) = \sum_{i=0}^{N} C_{T,i}(y)$.

Proper losses The quality of probabilistic forecasts is evaluated using *proper* losses ℓ. Formally, a loss $\ell(y,p)$ is proper if $p \in \arg\min_{q\in\mathcal{P}} \mathbb{E}_{y\sim(p)}\ell(y,q)\ \forall p \in \mathcal{P}$. An important proper loss for CDF predictions F is the continuous ranked probability score, defined as

$$\ell_{\text{CRPS}}(y,F) = \int_{-\infty}^{\infty} (F(z) - \mathbb{I}_{y\le z})^2 dz. \tag{9}$$

A.1 ASSUMPTIONS

We assume that each subroutine S^{cal} is an instance of a binary calibrated forecasting algorithm (e.g., the methods introduced in Chapter 4 in Cesa-Bianchi and Lugosi [2006]) that produce predictions in $[0,1]$ that are (ϵ, ℓ_2)-calibrated and that $C_T^2 \le R_T + \epsilon$ uniformly ($R_T = o(1)$ as $T \to \infty$; T is the number of calls to instance S_j^{cal}). We also assume that for each t, the target y_t lies in some bounded interval $\mathcal{Y}$ of $\mathbb{R}$ of length at most B.

A.2 ONLINE CALIBRATED REGRESSION

First, we look at algorithms for online calibrated regression (without covariates). Our algorithms leverage classical online binary calibration [Foster and Vohra, 1998] as a subroutine. Formally, Algorithm 1 partitions $[-\frac{B}{2}, \frac{B}{2}]$ into M intervals $\mathcal{I} = \{[\frac{-B}{2}, \frac{-B}{2} + \frac{B}{M}), ..., [\frac{B}{2} - \frac{B}{M}, \frac{B}{2}]\}$; each interval is associated with an instance of an online binary recalibration subroutine S^{cal} [Foster and Vohra, 1998, Cesa-Bianchi and Lugosi, 2006]. In order to compute $G_t(y \le z)$, we invoke the subroutine S_j^{cal} associated with interval I_j containing z. After observing y_t, each S_j^{cal} observes whether y_t falls in its interval and updates its state.

Theorem 5. *Let $\mathcal{Y}_\mathcal{I}$ be the set of upper bounds of the intervals $\mathcal{I}$ and let $\mathcal{P}_S$ be the output space of S^{cal}. Algorithm 1 achieves online calibration and for all $y \in \mathcal{Y}_\mathcal{I}, p \in \mathcal{P}_S$ we have $\rho_T(y,p) \to p$ a.s. as $T \to \infty$.*

Proof. The above theorem follows directly from the construction of Algorithm 1: for each $y \in \mathcal{Y}$, we run an online binary calibration algorithm to target the event $y_t \leq y$.

Specifically, note that for each $y \in \mathcal{Y}_\mathcal{I}$, the empirical frequency $\rho(y,p)$ reduces to the definition of the empirical frequency of a classical binary calibration algorithm targeting probability p and the binary outcome that $y_t \leq y$. The output of the algorithm for $F_t(y)$ is also a prediction for the binary outcome $y_t \leq y$ produced by a classical onlne binary calibration algorithm. Thus, by construction, we have the desired result. □

Algorithms S^{cal} for online binary calibration are randomized. Our procedure needs to be randomized as well and this is a fundamental property of our task.

Theorem 6. *There does not exist a deterministic online calibrated regression algorithm that achieves online calibration.*

Proof. This claim follows because we can encode a standard online binary calibration problem as calibrated regression. Specifically, given a non-randomized online calibrated regression algorithm, we could solve an online binary classification problem. Suppose the adversary chooses a binary $y_t \in \{0,1\} \subseteq [0,1]$ that defines one of two classes. Then we can define an instance of calibrated regression with two buckets $[0, 0.5)$ and $[0.5, 1)$. We use the forecast $F_t(0.5)$ as our prediction for $y_t = 0$ and one minus that as the prediction for 1. Then, the error on the ratio $\rho_T(0.5, p)$ yields the definition of calibration in binary classification. If our deterministic online calibration regression algorithm works, then we have $\rho_T(0.5, p) \to p$, which means that the empirical ratio for the binary algorithm goes to the predicted frequency p as well. But that would yield a deterministic algorithm for online binary calibration, which we know can't exist. □

A.3 PROVING THE CALIBRATION OF ALGORITHM 1

First, we will provide a proof of Lemma 1; this proof holds for any norm ℓ_p.

Lemma 4 (Preserving calibration)**.** *If each S_j^{cal} is (ϵ, ℓ_p)-calibrated, then Algorithm 1 is also (ϵ, ℓ_p)-calibrated and the following bound holds uniformly over T:*

$$C_T \leq \sum_{j=1}^{M} \frac{T_j}{T} R_{T_j} + \epsilon. \tag{10}$$

Proof. Let $\mathbb{I}_i^{(j)} = \sum_{t=1}^{T} \mathbb{I}_{t,i}^{(j)}$ and note that $\sum_{t=1}^{T} \mathbb{I}_{t,i} = \sum_{j=1}^{M} \mathbb{I}_i^{(j)}$. We may write

$$\begin{aligned} C_{T,i}(y) &= \frac{\sum_{t=1}^{T} \mathbb{I}_{t,i}}{T} \left| \rho_T(y, i/N) - \frac{i}{N} \right|^p = \frac{\sum_{j=1}^{M} \mathbb{I}_i^{(j)}}{T} \left| \sum_{j=1}^{M} \frac{\sum_{t=1}^{T} \mathbb{I}_{t,i}^{(j)} o_{tj}}{\sum_{j=1}^{M} \mathbb{I}_i^{(j)}} - \frac{i}{N} \right|^p \\ &= \frac{\sum_{j=1}^{M} \mathbb{I}_i^{(j)}}{T} \left| \sum_{j=1}^{M} \frac{\mathbb{I}_i^{(j)} \rho_T^{(j)}(y, i/N)}{\sum_{j=1}^{M} \mathbb{I}_i^{(j)}} - \frac{i}{N} \right|^p \leq \sum_{j=1}^{M} \frac{\mathbb{I}_i^{(j)}}{T} \left| \rho_T^{(j)}(y, i/N) - \frac{i}{N} \right|^p = \sum_{j=1}^{M} \frac{T_j}{T} C_{T,i}^{(j)}, \end{aligned}$$

where in the last line we used Jensen's inequality. Plugging in this bound in the definition of C_T, we find that

$$C_T = \sum_{i=1}^{N} C_{T,i} \leq \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{T_j}{T} C_{T,i}^{(j)} \leq \sum_{j=1}^{M} \frac{T_j}{T} R_{T_j} + \epsilon,$$

Since each $R_{T_j} \to 0$, Algorithm 1 will be ϵ-calibrated. □

A.4 RECALIBRATED FORECASTS HAVE LOW REGRET UNDER THE CRPS LOSS

Lemma 5 (Recalibration preserves accuracy)**.** *Consider Algorithm 1 with parameters $M \geq N > 1/\epsilon$. Suppose that the S^{cal} are (ϵ, ℓ_2)-calibrated. Then the recalibrated G_t a.s. have vanishing ℓ_{CRPS}-regret relative to F_t:*

$$\frac{1}{T}\sum_{t=1}^{T} \ell_{CRPS}(y_t, G_t) - \frac{1}{T}\sum_{t=1}^{T} \ell_{CRPS}(y_t, F_t) < NBR_T + \frac{2B}{N}. \tag{11}$$

Proof. Our proof will rely on the following fact about any online calibration subroutine S^{cal}. We start by formally establishing this fact.

Fact 1. *Let S^{cal} be an binary online calibration subroutine with actions $0, 1/N, ...1$ whose ℓ_2 calibration error C_T^p is bounded by $R_T = o(T)$. Then the predictions p_t from S^{cal} also minimize external regret relative to any single action i/N:*

$$\sum_{t=1}^{T}(p_t - y_t)^2 - (\frac{i}{N} - y_t)^2 \leq NR_T \text{ for all } i \tag{12}$$

We refer the reader to Lemma 4.4 in Cesa-Bianchi and Lugosi [2006] for a proof.

Next, we prove our main claim. We start with some notation. Let $\mathcal{I} = \{[0, \frac{1}{M}), [\frac{1}{M}, \frac{2}{M}), ..., [\frac{M-1}{M}, 1]\}$ be a set of intervals that partition $[0, 1]$ and let $I_j = [\frac{j-1}{M}, \frac{j}{M})$ be the j-th interval. Also, for each j, we use i_j denote the index $i \in [N]$ that is closest to j in the sense of $|\frac{i_j}{N} - \frac{j}{M}| \leq \frac{1}{N}$. By our assumption that $M \geq N$, this index exists.

We begin our proof by from the definition of the CRPS regret:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \ell_{\text{CRPS}}(y_t, G_t) - \frac{1}{T}\sum_{t=1}^{T} \ell_{\text{CRPS}}(y_t, F_t) \\
&= \frac{1}{T}\sum_{t=1}^{T} \int_{-\infty}^{\infty} (G_t(z) - \mathbb{I}_{y_t \leq z})^2 dz - \frac{1}{T}\sum_{t=1}^{T} \int_{-\infty}^{\infty} (F_t(z) - \mathbb{I}_{y_t \leq z})^2 dz \\
&= \int_{-\infty}^{\infty} \frac{1}{T}\sum_{t=1}^{T} \left[(G_t(z) - \mathbb{I}_{y_t \leq z})^2 - (F_t(z) - \mathbb{I}_{y_t \leq z})^2\right] dz \\
&= \int_{z \in \mathcal{Y}} \frac{1}{T}\sum_{t=1}^{T} \left[(G_t(z) - \mathbb{I}_{y_t \leq z})^2 - (F_t(z) - \mathbb{I}_{y_t \leq z})^2\right] dz \\
&= \int_{z \in \mathcal{Y}} \frac{1}{T}\sum_{t=1}^{T} \left[(G_t(z) - \mathbb{I}_{F_t(y_t) \leq F_t(z)})^2 - (F_t(z) - \mathbb{I}_{F_t(y_t) \leq F_t(z)})^2\right] dz
\end{aligned}$$

In the second-to-last line, we have used the fact that the forecasts have finite support, i.e., the y_t live within a closed bounded set $\mathcal{Y}$. In the last line, we replaced the event $y_t \leq z$ with $F_t(y_t) \leq F_t(z)$, which is valid because F_t is monotonically increasing.

Let's now analyze the above integrand for one fixed value of z:

$$\frac{1}{T}\sum_{t=1}^{T} \left[(G_t(z) - \mathbb{I}_{F_t(y_t) \leq F_t(z)})^2 - (F_t(z) - \mathbb{I}_{F_t(y_t) \leq F_t(z)})^2\right]. \tag{13}$$

Since F_t outputs a finite number of values in the set $\{0, \frac{1}{M}, ..., 1\}$, let j/M denote the value $F_t(z) = j/M$ taken by F_t at z. Additionally, observe that $\mathbb{I}_{F_t(y_t) \leq \frac{j}{M}} = o_{tj}$, where o_{tj} is the binary target variable given to S_j^{cal} at the end of step t. Finally, recall that when $F_t(z) = \frac{j}{M}$, we have defined $G_t(z)$ to be the output of S_j^{cal} at time t, which we denote as G_{tj}. This yields the following expression for the above integrand for a fixed z:

$$\frac{1}{T}\sum_{t=1}^{T} \left[(G_{tj} - o_{tj})^2 - (\frac{j}{M} - o_{tj})^2\right]. \tag{14}$$

Next, recall that i_j is the index $i \in [N]$ that is closest to j in the sense of $|\frac{i_j}{N} - \frac{j}{M}| \leq \frac{1}{N}$. Recall also that $M \geq N$. Note that this implies

$$\ell_2(\frac{j}{M}, o_{tj}) \geq \ell_2(\frac{i_j}{M}, o_{tj}) + \frac{\partial \ell_2}{\partial p}(p, o_{tj})(\frac{j}{M} - \frac{i_j}{M}) \geq \frac{2}{N}. \tag{15}$$

Using this inequality, we obtain the following bound for our earlier integrand:

$$\frac{1}{T}\sum_{t=1}^{T}\left[(G_{tj} - o_{tj})^2 - (\frac{i_j}{N} - o_{tj})^2\right] + \frac{2}{N}. \tag{16}$$

Crucially, this expression is precisely the *external regret* of recalibration subroutine S_j^{cal} relative to the fixed action $\frac{i_j}{N}$ and measured in terms of the L2 loss. By Fact 1, we know that this external regret is bounded by NR_T. Since this bound holds pointwise for any value of z, we can plug it into our original integral to obtain a bound on the CRPS regret:

$$\begin{aligned}
&\int_{z\in\mathcal{Y}} \frac{1}{T}\sum_{t=1}^{T}\left[(G_t(z) - \mathbb{I}_{F_t(y_t)\leq F_t(z)})^2 - (F_t(z) - \mathbb{I}_{F_t(y_t)\leq F_t(z)})^2\right] dz \\
&\leq \int_{z\in\mathcal{Y}} \left[NR_T + \frac{2}{N}\right] dz \\
&\leq NBR_T + \frac{2B}{N}
\end{aligned}$$

In the last line, we used the fact that the integration is over a finite set $\mathcal{Y}$ whose measure is bounded by $B > 0$. This establishes the main claim of this proposition. □

A.5 CORRECTNESS OF ALGORITHM 1

We now prove our main result about the correctness of Algorithm 1.

Theorem 1. *Let S^{cal} be an $(\ell_1, \epsilon/3B)$-calibrated online subroutine with resolution $N \geq 3B/\epsilon$. and let ℓ be a proper loss satisfying the assumptions of Lemma 3. Then Algorithm 1 with parameters S^{cal} and N is an ϵ-accurate online recalibration algorithm for the loss ℓ.*

Proof. It is easy to show that Algorithm 1 is $(\ell_1, \epsilon/3B)$-calibrated by the same argument as Lemma 1 (see the next section for a formal proof). By Lemma 4, its regret w.r.t. the raw F_t^H tends to $< 3B/N < \epsilon$. Hence, the theorem follows. □

A.6 CALIBRATION IMPLIES NO INTERNAL REGRET

Here, we show that a calibrated forecaster also has small internal regret relative to any bounded proper loss [Kuleshov and Ermon, 2017].

Lemma 1. *If ℓ is a bounded proper loss, then an ϵ-calibrated S^{cal} a.s. has a small internal regret w.r.t. ℓ and satisfies uniformly over time T the bound*

$$R_T^{\text{int}} = \max_{ij}\sum_{t=1}^{T}\mathbb{I}_{p_t=i/N}\left(\ell(y_t, i/N) - \ell(y_t, j/N)\right) \leq 2B(R_T + \epsilon). \tag{17}$$

Proof. Let T be fixed for the rest of this proof. Let $\mathbb{I}_{ti} = \mathbb{I}_{p_t=i/N}$ be the indicator of S^{cal} outputting prediction i/N at time t, let $T_i = \sum_{t=1}^{T}\mathbb{I}_{ti}$ denote the number of time i/N was predicted, and let

$$R_{T,ij}^{\text{int}} = \sum_{t=1}^{T}\mathbb{I}_{ti}\left(\ell(y_t, i/N) - \ell(y_t, j/N)\right) \tag{18}$$

denote the gain (measured using the proper loss ℓ) from retrospectively switching all the plays of action i to j. This value forms the basis of the definition of internal regret (Section 2).

Let $T(i,y) = \sum_{t=1}^{T} \mathbb{I}_{ti}\mathbb{I}\{y_t = y\}$ denote the total number of i/N forecasts at times when $y_t = y \in \{0,1\}$. Observe that we have

$$\begin{aligned} T(i,y) &= \sum_{t=1}^{T} \mathbb{I}_{ti}\mathbb{I}\{y_t = y\} = \frac{\sum_{t=1}^{T} \mathbb{I}_{ti}\mathbb{I}\{y_t = y\}}{T_i} T_i = \frac{\sum_{t=1}^{T} \mathbb{I}_{ti}\mathbb{I}\{y_t = y\}}{\sum_{t=1}^{T} \mathbb{I}_{ti}} T_i \\ &= q(i,y)T_i + T_i \left(\frac{\sum_{t=1}^{T} \mathbb{I}_{ti}\mathbb{I}\{y_t = y\}}{\sum_{t=1}^{T} \mathbb{I}_{ti}} - q(i,y) \right) \\ &= q(i,y)T_i + T_i \left(\rho_T(i/N) - i/N \right), \end{aligned}$$

where $q(i,y) = i/N$ if $y = 1$ and $1 - i/N$ if $y = 0$. The last equality follows using some simple algebra after adding and subtracting one inside the parentheses in the second term.

We now use this expression to bound $R^{\text{int}}_{T,ij}$:

$$\begin{aligned} R^{\text{int}}_{T,ij} &= \sum_{t=1}^{T} \mathbb{I}_{ti} \left(\ell(y_t, i/N) - \ell(y_t, j/N) \right) \\ &= \sum_{y \in \{0,1\}} T(i,y) \left(\ell(y, i/N) - \ell(y, j/N) \right) \\ &\leq \sum_{y \in \{0,1\}} q(i,y)T_i \left(\ell(y, i/N) - \ell(y, j/N) \right) + \sum_{y \in \{0,1\}} BT_i \left| \rho_T(i/N) - i/N \right| \\ &\leq 2BT_i \left| \rho_T(i/N) - i/N \right|, \end{aligned}$$

where in the first inequality, we used $\ell(y, i/N) - \ell(y, j/N) \leq \ell(y, i/N) \leq B$, and in the second inequality we used the fact that ℓ is a proper loss.

Since internal regret equals $R^{\text{int}}_T = \max_{i,j} R^{\text{int}}_{T,ij}$, we have

$$R^{\text{int}}_T \leq \sum_{i=1}^{N} \max_j R^{\text{int}}_{T,ij} \leq 2B \sum_{i=0}^{N} T_i \left| \rho(i/N) - i/N \right| \leq 2B(R_T + \epsilon).$$

□

A.7 IMPOSSIBILITY OF RECALIBRATING NON-PROPER LOSSES

We conclude the appendix by explaining why non-proper losses cannot be calibrated [Kuleshov and Ermon, 2017].

Theorem 2. *If ℓ is not proper, then there is no recalibration algorithm w.r.t. ℓ.*

Proof. If ℓ is not proper, there exist a p' and q such that $\mathbb{E}_{y \sim \text{Ber}(p')} \ell(y,q) < \mathbb{E}_{y \sim \text{Ber}(p')} \ell(y,p')$.

Consider a sequence y_t for which $y_t \sim \text{Ber}(p')$ for all t. Clearly the prediction of a calibrated forecaster p_t much converge to p' and the average loss will approach $\ell(y,p')$. This means that we cannot recalibrate the constant predictor $p_t = q$ without making its loss $\ell(y,q)$ higher. We thus have a forecaster that cannot be recalibrated with respect to ℓ. □

B LOW REGRET RELATIVE TO BASELINE CLASSIFIERS

Here, we show that a calibrated forecaster also has small regret relative to any bounded proper loss if we use a certain construction that combines our algorithm with a baseline forecaster. This extends our previous construction to more general settings.

B.1 RECALIBRATION CONSTRUCTION

Setup We start with an online forecaster F that outputs uncalibrated forecasts F_t^H at each step; these forecasts are fed into a *recalibrator* such that the resulting forecasts p_t are calibrated and have low regret relative to the baseline forecasts F_t^H.

Formally, at every step $t = 1, 2, ...$ we have:

1: Forecaster F predicts F_t^H.
2: A recalibration algorithm produces a calibrated forecast p_t based on F_t^H.
3: Nature reveals label y_t
4: Based on x_t, y_t, we update the recalibration algorithm and optionally update H.

Notation We define a discretization V of the space of forecasts. We assume that the forecasts live in a compact set Δ and we define a triangulation of Δ, i.e., a partition into a set of simplices such that any two simplices intersect in either a common face, common vertex, or not at all. Let V be the vertex set of this triangulation, and let $V(p)$ be the set of corners for this simplex.

Note that each distribution p can be uniquely written as a weighted average of its neighboring vertices, $V(p)$. For $v \in V(p)$, we define the test functions $w_v(p)$ to be these linear weights, so they are uniquely defined by the linear equation $p = \sum_{v \in V(p)} w_v(p) v$. We also define the discretization to be sufficiently small: given a target precision $\epsilon > 0$ we define the discretization such that for all f_1, f_2 in the same simplex we have $||f_1 - f_2|| < \epsilon$.

B.2 RECALIBRATION ALGORITHM

We are going to define a general meta-algorithm that follows a construction in which we run multiple instances of our calibrated forecasting algorithms over the inputs of F.

More formally, we take the aforementioned partition of the space of forecasts of Δ of F and we associate each simplex with an instance of our calibration algorithm S^{cal} (using the same Δ and discretization V). In order to compute F_t^H, we invoke the subroutine S_j^{cal} associated with simplex I_j containing F_t^H (with ties broken arbitrarily). After observing y_t, we pass it to S_j^{cal}.

The resulting procedure produces valid calibrated estimates because each S_j^{cal} is a calibrated subroutine. More importantly the new forecasts do not decrease the predictive performance of F, as measured by a proper loss ℓ. In the remainder of this section, we establish these facts formally.

B.3 THEORETICAL ANALYSIS

Notation Our task is to produce calibrated forecasts. Intuitively, we say that a forecast F_t is calibrated if for every $y' \in \mathcal{Y}$, the probability $F_t(y')$ on average matches the frequency of the event $\{y = y'\}$. We formalize this by introducing the ratio

$$\rho_T(p) = \frac{\sum_{t=1}^T y_t \cdot \mathbb{I}_{p_t = p}}{\sum_{t=1}^T \mathbb{I}_{p_t = p}} \tag{19}$$

Intuitively, we want $\rho_T(p) \to p$, a.s. as $T \to \infty$ for all y. In other words, out of the times when the predicted probability for y_t is p, the average y_t look like p.

The quality of probabilistic forecasts is evaluated using *proper* losses ℓ. Formally, a loss $\ell(y, p)$ is proper if $p \in \arg\min_{q \in \mathcal{P}} \mathbb{E}_{y \sim (p)} \ell(y, q) \ \forall p \in \mathcal{P}$. An example in binary classification is the log-loss $\ell_{\log}(y, p) = y \log(p) + (1 - y) \log(1 - p)$. We will assume that the loss is bounded by $B > 0$.

We measure calibration a calibration error C_T. Our algorithms will output discretized probabilities; hence we define the error relative to a set of possible predictions V

$$C_T = \sum_{p \in V} |\rho_T(p) - p| \left(\frac{1}{T} \sum_{t=1}^T \mathbb{I}_{\{p_t = p\}} \right). \tag{20}$$

B.3.1 A Helper Lemma

In order to establish the correctness of our recalibration procedure, we need to start with a helper lemma. This lemma shows that if forecasts are calibrated, then they have small internal regret.

Lemma 2. *If ℓ is a bounded proper loss, then an (ϵ, ℓ_1)-calibrated S^{cal} a.s. has a small internal regret w.r.t. ℓ and satisfies uniformly over time T the bound*

$$R_T^{\mathrm{int}} = \max_{ij} \sum_{t=1}^{T} \mathbb{I}_{p_t=p_i} \left(\ell(y_t, p_i) - \ell(y_t, p_j)\right) \leq 2B(R_T + \epsilon). \tag{21}$$

Proof. Let T be fixed for the rest of this proof. Let $\mathbb{I}_{ti} = \mathbb{I}_{p_t=p_i}$ be the indicator of S^{cal} outputting prediction p_i at time t, let $T_i = \sum_{t=1}^{T} \mathbb{I}_{ti}$ denote the number of time i/N was predicted, and let

$$R_{T,ij}^{\mathrm{int}} = \sum_{t=1}^{T} \mathbb{I}_{ti} \left(\ell(y_t, p_i) - \ell(y_t, p_j)\right) \tag{22}$$

denote the gain (measured using the proper loss ℓ) from retrospectively switching all the plays of action i to j. This value forms the basis of the definition of internal regret.

Let $T(i, y) = \sum_{t=1}^{T} \mathbb{I}_{ti}\mathbb{I}\{y_t = y\}$ denote the total number of p_i forecasts at times when $y_t = y$. Observe that we have

$$\begin{aligned}
T(i,y) &= \sum_{t=1}^{T} \mathbb{I}_{ti}\mathbb{I}\{y_t = y\} = \frac{\sum_{t=1}^{T} \mathbb{I}_{ti}\mathbb{I}\{y_t = y\}}{T_i} T_i = \frac{\sum_{t=1}^{T} \mathbb{I}_{ti}\mathbb{I}\{y_t = y\}}{\sum_{t=1}^{T} \mathbb{I}_{ti}} T_i \\
&= q(i,y)T_i + T_i \left(\frac{\sum_{t=1}^{T} \mathbb{I}_{ti}\mathbb{I}\{y_t = y\}}{\sum_{t=1}^{T} \mathbb{I}_{ti}} - q(i,y) \right) \\
&= q(i,y)T_i + T_i \left(\rho_T(p_i) - p_i\right),
\end{aligned}$$

where $q(i, y) = p_i(y)$. The last equality follows using some simple algebra after adding and subtracting one inside the parentheses in the second term.

We now use this expression to bound $R_{T,ij}^{\mathrm{int}}$:

$$\begin{aligned}
R_{T,ij}^{\mathrm{int}} &= \sum_{t=1}^{T} \mathbb{I}_{ti} \left(\ell(y_t, p_i) - \ell(y_t, p_j)\right) \\
&= \sum_{y} T(i,y) \left(\ell(y, p_i) - \ell(y, p_j)\right) \\
&\leq \sum_{y} q(i,y)T_i \left(\ell(y, p_i) - \ell(y, p_j)\right) + BT_i \left|\rho_T(p_i) - p_i\right| \\
&\leq BT_i \left|\rho_T(p_i) - p_i\right|,
\end{aligned}$$

where in the first inequality, we used $\ell(y, p_i) - \ell(y, p_j) \leq \ell(y, p_i) \leq B$, and in the second inequality we used the fact that ℓ is a proper loss.

Since internal regret equals $R_T^{\mathrm{int}} = \max_{i,j} R_{T,ij}^{\mathrm{int}}$, we have

$$R_T^{\mathrm{int}} \leq \sum_{i=1}^{N} \max_j R_{T,ij}^{\mathrm{int}} \leq 2B \sum_{i=0}^{N} T_i \left|\rho(i/N) - p_i\right| \leq 2B(R_T + \epsilon).$$

□

B.4 RECALIBRATED FORECASTS HAVE LOW REGRET RELATIVE TO UNCALIBRATED FORECASTS

Next, we use the above result to prove that the forecasts recalibrated using the above construction have low regret relative to the baseline uncalibrated forecasts.

Lemma 3 (Recalibration preserves accuracy)**.** *Let ℓ be a bounded proper loss such that $\ell(y_t, p) \leq \ell(y_t, p_j) + B\epsilon$ whenever $||p - p_j|| \leq \epsilon$. Then the recalibrated p_t a.s. have vanishing ℓ-loss regret relative to F_t^H and we have uniformly:*

$$\frac{1}{T}\sum_{t=1}^{T} \ell(y_t, p_t) - \frac{1}{T}\sum_{t=1}^{T} \ell(y_t, F_t^H) < \frac{B}{\epsilon}\sum_{j=1}^{M} \frac{T_j}{T} R_{T_j} + 3B\epsilon. \tag{23}$$

Proof. By the previous lemma, we know that an algorithm whose calibration error is bounded by $R_T = o(1)$ also minimizes internal regret at a rate of $2BR_T$, and thus external regret at a rate of $2BR_T/\epsilon$.

Next, let us use $\mathbb{I}_{j,t}$ to indicate that S_j^{cal} was called at time t. We establish our main claim as follows:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \ell(y_t, p_t) - \frac{1}{T}\sum_{t=1}^{T} \ell(y_t, F_t^H) \\
&= \frac{1}{T}\sum_{t=1}^{T} \left(\sum_{j=1}^{M} \left(\ell(y_t, p_t) - \ell(y_t, F_t^H) \right) \mathbb{I}_{j,t} \right) \\
&< \frac{1}{T}\sum_{t=1}^{T} \left(\sum_{j=1}^{M} \left(\ell(y_t, p_t) - \ell(y_t, p_j) \right) \mathbb{I}_{j,t} + B\epsilon \right) \\
&\leq \frac{1}{\epsilon} B \sum_{j=1}^{M} \frac{T_j}{T} R_{T_j} + 3B\epsilon,
\end{aligned}$$

where R_{T_j} is a bound on the calibration error of S_j^{cal} after T_j plays.

In the first two inequality, we use our assumption on the loss ℓ. The last inequality follows because S_j^{cal} minimizes external regret w.r.t. the constant action p_j at a rate of BR_{T_j}/ϵ. □

B.5 PROVING THAT CALIBRATION HOLDS

We want to also give a proof that the recalibration construction described above yields calibrated forecasts.

Lemma 4. *If each S_j^{cal} is (ϵ, ℓ_p)-calibrated, then the combined algorithm is also (ϵ, ℓ_p)-calibrated and the following bound holds uniformly over T:*

$$C_T \leq \sum_{j=1}^{M} \frac{T_j}{T} R_{T_j} + \epsilon. \tag{24}$$

Proof. Let $M = |V|$. Let $\mathbb{I}_i^{(j)} = \sum_{t=1}^{T} \mathbb{I}_{t,i}^{(j)}$ where $\mathbb{I}_{t,i}^{(j)} = \mathbb{I}\{p_t = p_j \cap F_t^H = p_j\}$ and note that $\sum_{t=1}^{T} \mathbb{I}_{t,i} = \sum_{j=1}^{M} \mathbb{I}_i^{(j)}$. Let also $\rho_T^{(j)}(p_i) = \frac{\sum_{t=1}^{T} \mathbb{I}_{t,i}^{(j)} y_t}{\sum_{t=1}^{T} \mathbb{I}_{t,i}^{(j)}}$. We may write

$$\begin{aligned}
C_{T,i} &= \frac{\sum_{t=1}^{T} \mathbb{I}_{t,i}}{T} \left| \rho_T(p_i) - p_i \right| = \frac{\sum_{j=1}^{M} \mathbb{I}_i^{(j)}}{T} \left| \sum_{j=1}^{M} \frac{\sum_{t=1}^{T} \mathbb{I}_{t,i}^{(j)} y_t}{\sum_{j=1}^{M} \mathbb{I}_i^{(j)}} - p_i \right| \\
&= \frac{\sum_{j=1}^{M} \mathbb{I}_i^{(j)}}{T} \left| \sum_{j=1}^{M} \frac{\mathbb{I}_i^{(j)} \rho_T^{(j)}(p_i)}{\sum_{j=1}^{M} \mathbb{I}_i^{(j)}} - p_i \right| \leq \sum_{j=1}^{M} \frac{\mathbb{I}_i^{(j)}}{T} \left| \rho_T^{(j)}(p_i) - p_i \right| = \sum_{j=1}^{M} \frac{T_j}{T} C_{T,i}^{(j)},
\end{aligned}$$

where $C_{T,i}^{(j)} = \left|\rho_T^{(j)}(p_i) - p_i\right| \left(\frac{1}{T_j}\sum_{t=1}^{T} \mathbb{I}_{t,i}^{(j)}\right)$ and in the last line we used Jensen's inequality. Plugging in this bound in the definition of C_T, we find that

$$C_T = \sum_{i=1}^{N} C_{T,i} \leq \sum_{j=1}^{M}\sum_{i=1}^{N} \frac{T_j}{T} C_{T,i}^{(j)} \leq \sum_{j=1}^{M} \frac{T_j}{T} R_{T_j} + \epsilon,$$

Since each $R_{T_j} \to 0$, the full procedure will be ϵ-calibrated. □

Recall that R_T denotes the rate of convergence of the calibration error C_T. For most online calibration subroutines S^{cal}, $R_T \leq f(\epsilon)/\sqrt{T}$ for some $f(\epsilon)$. In such cases, we can further bound the calibration error in the above lemma as

$$\sum_{j=1}^{M} \frac{T_j}{T} R_{T_j} \leq \sum_{j=1}^{M} \frac{\sqrt{T_j} f(\epsilon)}{T} \leq \frac{f(\epsilon)}{\sqrt{\epsilon T}}. \tag{25}$$

In the second inequality, we set the T_j to be equal. Thus, our recalibration procedure introduces an overhead of $\frac{1}{\sqrt{\epsilon}}$ in the convergence rate of the calibration error C_T and of the regret relative to a baseline forecaster in the earlier lemma.

C APPLICATIONS: DECISION-MAKING

Next, we complement our results with a formal characterization of some benefits of calibration. We are interested in decision-making settings where we wish to estimate the value of a function $v : \mathcal{Y} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ over a set of outcomes $\mathcal{Y}$, actions $\mathcal{A}$, and features $\mathcal{X}$. Note that the function v could be a loss $\ell(y, a, x)$ that quantifies the error of an action $a \in \mathcal{A}$ in a state $x \in \mathcal{X}$ given outcome $y \in \mathcal{Y}$.

We assume that given x, the agent chooses an action $a(x)$ according to a decision-making process. This could be an action $a(x) = \arg\min_a \mathbb{E}_{y\sim H(x)}[\ell(y, a, x)]$ that minimizes a loss that are trying to estimate, but any outcome is possible. The agent then relies on a predictive model H of y to estimate the future values $v(y, a, x)$ for the decision $a(x)$:

$$v(x) = \mathbb{E}_{y\sim H(x)}[v(y, a(x), x)]. \tag{26}$$

We study $v(y, a, x)$ that are monotonically non-increasing or non-decreasing in y. Examples include linear utilities $u(a, x) \cdot y + c(a, x)$ or their monotone transformations.

Expectations under calibrated models If H was a perfect predictive model, we could estimate expected values of outcomes perfectly. In practice, inaccurate models can yield imperfect decisions. Surprisingly, our analysis shows that in many cases, calibration (a much weaker condition that having a perfectly specified model H) is sufficient to correctly estimate the value of various outcomes.

Surprisingly, our guarantees can be obtained with a weak condition—quantile calibration. Additional requirements are the non-negativity and monotonicity of v. Our result is a concentration inequality that shows that estimates of v are unlikely to exceed the true v on average.

Theorem 3. *Let M be a quantile calibrated model as in and let $v(y, a, x)$ be a monotonic value function. Then for any sequence $(x_t, y_t)_{t=1}^T$ and $r > 0$, we have:*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\left[v(y_t, a(x_t), x_t) \geq r v(x_t))\right] \leq 1/r \tag{27}$$

Proof. Recall that $M(x)$ is a distribution over $\mathcal{Y}$, with a density p_x, a quantile function Q_x, and a cdf F_x. Note that for any

x and $s \in (0,1)$ and $y' \leq F_x^{-1}(1-s)$ we have:

$$\begin{aligned} v(x) &= \int v(x,y,a(x))q_x(y)dy \\ &\geq \int_{y \geq y'} v(x,y,a(x))q_x(y)dy \\ &\geq v(x,y',a(x)) \int_{y \geq y'} q_x(y)dy \\ &\geq sv(x,y',a(x)) \end{aligned}$$

The above logic implies that whenever $v(x) \leq sv(x,y,a)$, we have $y \geq F_x^{-1}(1-s)$ or $F_x(y) \geq (1-s)$. Thus, we have for all t,

$$\mathbb{I}\{v(x_t) \leq sv(x_t,y_t,a_t)\} \leq \mathbb{I}\{F_{x_t}(y_t) \geq (1-s)\}.$$

Therefore, we can write

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\{v(x_t) \leq sv(x_t,y_t,a_t)\} \leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\{F_{x_t}(y_t) \geq (1-s)\} = s + o(T),$$

where the last equality follows because M is calibrated. Therefore, the claim holds in the limit as $T \to \infty$ for $r = 1/s$. The argument is similar if v is monotonically non-increasing. In that case, we can show that whenever $y' > F_x^{-1}(s)$, we have $v(x) \geq sv(x,y',a(x))$. Thus, whenever $v(x) \leq sv(x,y,a)$, we have $y \leq F_x^{-1}(s)$ or $F_x(y) \leq s$. Because, F_x is calibrated, we again have that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\{v(x_t) \leq sv(x_t,y_t,a_t)\} \leq \sum_{t=1}^{T} \mathbb{I}\{F_{x_t}(y_t) < s\} = s + o(T),$$

and the claim holds with $r = 1/s$. □

Note that this statement represents an extension of Markov inequality. Note also that this implies the same result for a distribution calibrated model, since distribution calibration implies quantile calibration.

D EXPERIMENTS ON UCI BENCHMARKS

The existing UCI datasets [Dua and Graff, 2017] used in our experiments hold a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Computational resources. Our experiments were conducted on a laptop with 2.3 GHz 8-Core Intel Core i9 processor and 32 GB 2667 MHz DDR4 RAM. The code and datasets take 16MB memory.

Detailed setup. Our dataset consists of input and output pairs $\{x_t, y_t\}_{t=1}^{T}$ where T is the size of the dataset. We simulate a stream of data by sending batches of data-points $\{x_t, y_t\}_{t=nt'+1}^{n(t'+1)}$ to our model, where t' is the time-step and n is the batch-size. This simulation is run for $\lceil T/n \rceil$ time-steps. For each batch, Bayesian ridge regression is fit to the data and the recalibrator is trained. We set $N = 20$ in the recalibrator and use a batch size of $n = 10$ for all experiments except for the Aquatic Toxicity dataset 1(a) where we used $n = 5$. The calibration is evaluated at levels $[0.2, 0.4, 0.5, 0.6, 0.8]$.

E EXPERIMENTS ON BAYESIAN OPTIMIZATION

Bayesian optimization attempts to find the global minimum $x^\star = \arg\min_{x \in \mathcal{X}} f(x)$ of an unknown function $f : \mathcal{X} \to \mathbb{R}$ over an input space $\mathcal{X} \subseteq \mathbb{R}^D$. We are given an initial labeled dataset $x_t, y_t \in \mathcal{X} \times \mathbb{R}$ for $t = 1, 2, ..., N$ of i.i.d. realizations of random variables $X, Y \sim P$. At every time-step t, we use uncertainties from the probabilistic model $\mathcal{M} : \mathcal{X} \to (\mathbb{R} \to [0,1])$ of f to select the next data-point x_{next} and iteratively update the model $\mathcal{M}$. Algorithm 2 outlines this procedure. Since the black-box function evaluation can be expensive, the objective of Bayesian optimization in this context is to find the minima (or maxima) of this function while using a small number of function evaluations.

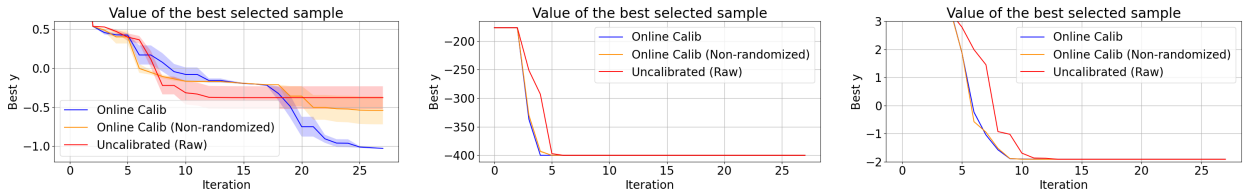
Computational resources. Our experiments were conducted on a laptop with 2.3 GHz 8-Core Intel Core i9 processor and 32 GB 2667 MHz DDR4 RAM. The code and datasets take 16MB memory.

Detailed setup. We use online calibration to improve the uncertainties estimated by the model $\mathcal{M}$. Following Deshpande et al. [2024], we use Algorithm 4 to recalibrate the model $\mathcal{M}$. Since the dataset size is small, we use the CREATESPLITS function to generate leave-one-out cross-validation splits of our dataset $\mathcal{D}$. We train the base model on train-split and use this to obtain probabilistic forecast for data in the test-split. We collect these predictions on all test-splits to form our recalibration dataset and use Algorithm 1 to perform calibration.

Following Deshpande et al. [2024], we perform calibrated Bayesian optimization as detailed in Algorithm 3. Specifically, we recalibrate the base model $\mathcal{M}$ after every step in Bayesian optimization. We build on the GpyOpt library [authors, 2016] for Bayesian optimization that holds the BSD 3-clause license.

We use some popular benchmark functions to evaluate the performance of Bayesian optimization. We initialize the Bayesian optimization with 3 randomly chosen data-points. We use the Lower Confidence Bound (LCB) acquisition function to select the data-point x_t and evaluate a potentially expensive function f as x_t to obtain y_t. At any given time-step T, we have the dataset $\mathcal{D}_T = \{x_t, y_t\}_{t=1}^{T}$ collected iteratively.

In Figure 4, we see that using online calibration of uncertainties from $\mathcal{M}$ allows us to reach a lower minimum or find the same minimum with a smaller number of steps with Bayesian optimization.



(a) SixHumpCamel (b) Beale (c) Mccormick

Figure 4: Online Calibration Improves Bayesian optimization

Algorithm 2 Bayesian Optimization

Initialize base model $\mathcal{M}$ with data $\mathcal{D} = \{x_t, y_t\}_{t=0}^{M}$.
for $n = 1, 2, ..., T$: **do**
  $x_{\text{next}} = \arg\max_{x \in \mathcal{X}} (\text{Acquisition}(x, \mathcal{R} \circ \mathcal{M}))$.
  $y_{\text{next}} = f(x_{\text{next}})$.
  $\mathcal{D} = \mathcal{D} \bigcup \{(x_{\text{next}}, y_{\text{next}})\}$
  Update model $\mathcal{M}$ with data $\mathcal{D}$

Algorithm 3 Calibrated Bayesian Optimization [Deshpande et al., 2024]

Initialize base model $\mathcal{M}$ with data $\mathcal{D} = \{x_t, y_t\}_{t=0}^{M}$.
$\mathcal{R} \leftarrow \text{CALIBRATE}(\mathcal{M}, \mathcal{D})$.
for $n = 1, 2, ..., T$: **do**
  $x_{\text{next}} = \arg\max_{x \in \mathcal{X}} (\text{Acquisition}(x, \mathcal{R} \circ \mathcal{M}))$.
  $y_{\text{next}} = f(x_{\text{next}})$.
  $\mathcal{D} = \mathcal{D} \bigcup \{(x_{\text{next}}, y_{\text{next}})\}$
  Update model $\mathcal{M}$ with data $\mathcal{D}$
  $\mathcal{R} \leftarrow \text{CALIBRATE}(\mathcal{M}, \mathcal{D})$

Algorithm 4 CALIBRATE [Deshpande et al., 2024]

Require: Base model $\mathcal{M}$, Dataset $\mathcal{D} = \{x_t, y_t\}_{t=0}^{N}$
Train a base model $\mathcal{M}$ on training dataset $\{x_t, y_t\}_{t=0}^{N}$.
Initialize recalibration dataset $\mathcal{D}_{\text{recal}} = \phi$
$S = \text{CREATESPLITS(D)}$
for $(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})$ in S: **do**
    $\mathcal{D}_{\text{train}} =$ Train Dataset $\{x_t, y_t\}_{t=0}^{M}$ in split s.
    $\mathcal{D}_{\text{test}} =$ Test Dataset $\{x_t, y_t\}_{t=0}^{L}$ in split s.
    $\mathcal{D}_{\text{train}} = \text{TRAINSPLIT}(s), \mathcal{D}_{\text{test}} = \text{TESTSPLIT}(s)$
    Train base model $\mathcal{M}'$ on dataset $\mathcal{D}_{\text{train}}$
    Compute CDF dataset $\{[M'(x_t)](y_t)\}_{t=1}^{M}$ from dataset $\mathcal{D}_{\text{test}}$
    $\mathcal{D}_{\text{recal}} = \mathcal{D}_{\text{recal}} \bigcup \{[\mathcal{M}'(x_t)](y_t), y_t\}_{t=1}^{M}$
Train recalibrator model $\mathcal{R}$ on the recalibration dataset $\mathcal{D}_{\text{recal}}$ using Algorithm 1
Return $(\mathcal{R})$

F COMPARISON TO PRIOR WORK

Table 4 and Table 5 summarize how our work fits in the broader literature. In brief, we provide calibration with regret guarantees in the setting of quantile regression on adversarial data. By regret guarantees we mean that performance relative to a user-specified baseline classifier is guaranteed not to drop.

Table 4: Summary of literature for IID or Exchangeable Data.

Output Type	**No Regret Guarantees**	**Regret Guarantees**
Classification		
Predicting Sets	Vovk et al. [2005a]	Kuleshov and Deshpande [2022]
Predicting Probabilities	Platt [1999], Niculescu-Mizil and Caruana [2005]	—
Regression		
Quantiles	Kuleshov et al. [2018], Dheur and Taleb [2023, 2024]	—
Distributions	Song et al. [2019]	Kuleshov and Deshpande [2022]

Table 5: Summary of literature for Non-IID ("Adversarial") Data.

Output Type	**No Regret Guarantees**	**Regret Guarantees**
Classification		
Predicting Sets	Vovk et al. [2005b]	—
Predicting Probabilities	Foster and Vohra [1998], Cesa-Bianchi and Lugosi [2006], Abernethy et al. [2011], Okoroafor et al. [2024], Noarov et al. [2024]	Kuleshov and Ermon [2017], Foster and Hart [2023]
Regression		
Marginal	—	Lee et al. [2022]
Quantiles	Gibbs and Candès [2021], Ramalingam et al. [2025]	**This work**
Distributions	Marx et al. [2025]	—

We now summarize the existing literature. We cite a representative paper in each class.

IID DATA

Classification. Many papers on calibration or conformal prediction assume that data is IID or exchangeable. In calibration for classification, representative works include Platt scaling Platt [1999] and isotonic regression Niculescu-Mizil and Caruana [2005]. Both methods output a calibrated probability p of a binary outcome in $\{0, 1\}$, and admit multi-class extensions. On the other hand, conformal prediction outputs confidence sets that contain the outcome with some probabilities. The conformal prediction by Vovk et al. [2005b] and other authors often assumes that data are exchangeable. Kuleshov and Deshpande [2022] proves that these methods admit regret guarantees.

Regression. In regression, the most standard definition is quantile calibration. Kuleshov et al. [2018] extends Platt scaling to this setting. Conformal prediction for continuous outcomes (e.g., Vovk et al. [2005a]) is similar to quantile calibration, but targets one pair of quantiles, while Kuleshov et al. [2018] outputs a full quantile function. Recently, a stronger form of regression called distribution calibration was studied, and it directly extends calibrated classification: of the times one forecasts predictive distribution p, the data looks like it's distributed as p. Song et al. [2019] describes this notion and Kuleshov and Deshpande [2022] shows it has regret guarantees.

NON-IID DATA

Another line of work seeks to extend the above results to settings where data is non-IID and can be even chosen by an adversary. This is the setting that we study.

Classification. The earliest work is by Foster and Vohra [1998], who frame calibration as internal regret minimization. Cesa-Bianchi and Lugosi [2006] provide a modern view on this algorithm based on online learning. Abernethy et al. [2011] presents yet another view based on Blackwell approachability. Most algorithms fall in one of these three approaches (internal regret, online learning, approachability)—ours is a form of internal regret minimization. Similarly, work on conformal prediction establishes comparable results for constructing confidence sets without IID or exchangeability assumptions. There exist many extensions of this work, including extensions for multi-class Ramalingam et al. [2025].

The drawback of these early works is that they only provide calibration results, but not regret. Thus a classifier can predict 50% chance of rain every day and still be calibrated (but not useful). Kuleshov and Ermon [2017] first introduce regret into adversarial online binary recalibration; Foster and Hart [2023] later re-derive the same algorithm. Our work extends these regret guarantees from classification to regression.

Regression. Gibbs and Candès [2021] provides analogous results to Foster and Vohra [1998] in online quantile regression using an approach based on online learning. Ramalingam et al. [2025] further explains the online learning connection. The distribution calibration extension is much more challenging—Marx et al. [2025] provides the first extension. Besides this fully adversarial literature, there exist extensive work prediction under covariate shift (where a data distribution exists, but its shifting an unknown)—Tibshirani et al. [2019] is an example of this long line of literature.

The challenge with methods such as those of Gibbs and Marx is the same as in classification: there is not a guarantee that calibrated predictions will have useful predictive value. Our work provides this no-regret guarantee for a setting that resembles quantile calibration.

F.1 COMPARISON WITH LEE AT AL. [2022]

Below we discuss how the work by Lee et al. [2022] compares and differs with out work. While the framework by Lee et al. [2022] admits a general compact, convex action set A, the calibration definition achieved by their algorithm is different.

- Their definition says: for each time step, draw a sample from the predicted distribution and the true label distribution over $\mathcal{Y}$. Averaged over T, the empirical pdf/CDFs of the samples from both distributions should match. (This definition can be applied to each population group when extending to multi-calibration).
- In contrast, our definition asks for quantile calibration: for any value $p \in [0, 1]$, look at whether the observed outcome $y_t \leq F_t^{-1}(p)$, i.e., whether y_t is below the p-th quantile. The frequency that $y \leq F_t^{-1}(p)$ should approach p as T increases.

Note that these two definitions are not the same. See "Probabilistic forecasts, calibration and sharpness" by Gneiting et al. [2007b] for counter-examples: the above definitions correspond respectively to *marginal* and *probabilistic calibration* (i.e.,

(c) and (a) in Defn. 1 of that work).

For continuous $\mathcal{Y}$, our work provides a more appropriate calibration guarantee for probability distributions over continuous outcomes: our calibration guarantee is closer to the notion of quantile calibration guarantee for regression as defined for the IID case.

Our paper also defines guarantees on regret relative to a baseline forecaster in a different way.

- Lee et al. [2022] define regret relative to a baseline using the average Brier score $(f_t - b_t)^2$, where b_t is a sample drawn from a true (unknown) distribution over the label y chosen by an adversary, and f_t is a forecast coming either from the model or a baseline.
- In our paper, F_t is a CDF over continuous outcomes and we measure its performance relative to a sequence of baseline functions using the Continuous Ranked Probability Score (CRPS), defined as $\int_{y \in \mathcal{Y}} (F_t(y) - G_t(y))^2 \, dy$, which is an integral over losses between the outputs of two CDFs (typically a forecast and an empirical/step-function observed CDF).

These definitions are clearly different: one takes the L_2 loss in the space of outcomes, and the other in the space of probabilities.

F.2 COMPARING TO KULESHOV AND ERMON [2017]

While Kuleshov and Ermon [2017] and the Calibeating technique focus on binary classification, we study regression. We want to emphasize that moving from calibration to regression is non-trivial and significantly more involved than generalizing the scoring rule from CDFs to point forecasts. The regression setting is significantly harder than classification, and requires (1) non-trivial thinking about how to define calibration and (2) algorithms and analyses that are substantially different than in classification.

The classical definition of calibration (of the times when I predict p, binary event holds p % of the time) does not easily carry over to regression. In fact, an "easier" version of regression is multi-class calibration (imagine the continuous label y is discretized), and even that is PPAD-hard (Hazan and Kakade, 2012).

Thus, most work on regression studies marginal notions of calibration: a p-% confidence interval contains the label p-% of the time (note how we omit the "when I predict p" part). Still, maintaining this in a non-IID setting is non-trivial. One well-known method is ACI (Gibbs and Candes, 2021), but it does not admit regret guarantees. We define a novel and slightly stronger notion of marginal calibration (which has elements of conditional calibration; see Eqn 2), and we provide regret guarantees.

Also, quantifying and minimizing regret is itself non-trivial. This requires defining a suitable notion of regret that is compatible with our definition of calibration. We use the CRPS and CDF recalibration as measures of regret and calibration, respectively.

While our method superficially resembles that of Kuleshov and Ermon [2017] (and Calibeating, which is the same algorithm) in that we partition an interval and run simple subroutines in each sub-interval, the analysis is significantly different, especially the part about minimizing regret. Superficially, while that proof takes (1/2)-page in Kuleshov and Ermon, ours is about 2 pages long and is substantially different.

Note also that we provide a significant number of additional results that strengthen our core work: a generalized Markov inequality that guarantees our method is able to accurately estimate losses, an analysis of confidence intervals, and an application to online decision-making and Bayesian optimization.

F.3 COMPARING TO DESHPANDE ET AL. [2024]

Note that the focus and the methods of both papers are different. Our work makes more theoretical contributions around the feasibility of defining and maintaining good calibration and regret in an online non-IID regression setting. The work by Deshpande et al. [2024] is mainly empirical: it applies methods from IID regression (e.g., Kuleshov and Ermon, ICML2018) and additional heuristics to obtain the best possible empirical results on classification.

We adopt a similar setup to Deshpande et al. [2024] in our experiments because the setting is useful and inherently non-IID. However, because our work is more theoretical, our experiments are not as extensive as those of Deshpande et al. [2024]

(whose entire paper is mostly experimental). That said, our non-randomized baseline (orange line) is effectively equivalent to the IID algorithm used in Deshpande et al. [2024] (it simply maintains marginal calibration by counting frequencies in bins), and we outperform that baseline in our experiments by virtue of designing specialized non-IID algorithms.

Lastly, while the paper by Deshpande et al. [2024] has a lemma on online decision-making, ours holds in the online non-IID setting, while theirs is only IID.

F.4 APPLICATIONS

Consider the example of predicting the demand for electricity so that the power grid operator can make decisions. The electricity demand may fluctuate in unpredictable ways depending on changes in variables like weather, special events producing sudden large industrial demands, time of the day, etc. The adversarial setting allows us to accommodate the worst case deviations from i.i.d. data. Having poorly calibrated forecasts in this setting can result in poor decisions (e.g. inadequate electricity supply). For example, an operator might want to provide electricity supply that minimizes a black-out with a target probability: if demand forecast is miscalibrated, then the true probability could be far different from the one inferred from the forecasted demand.

Some other examples include: 1) When assessing patient risk (e.g., sepsis probability) based on streaming vital signs, we require a calibrated forecast to correctly determine the probability of a bad outcome. 2) When market conditions constantly shift, predicting whether a loan is defaulted requires a calibrated probability. These are all examples of temporal data that may become non-IID since the state of the system evolves over time.