# On the Role of Model Uncertainties in Bayesian Optimization

**Jonathan Foldager**[1, *] **Mikkel Jordahn**[1,*] **Lars Kai Hansen**[1] **Michael Riis Andersen**[1]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark
[*]Shared first authorship.

## Abstract

Bayesian Optimization (BO) is a popular method for black-box optimization, which relies on uncertainty as part of its decision-making process when deciding which experiment to perform next. However, not much work has addressed the effect of uncertainty on the performance of the BO algorithm and to what extent calibrated uncertainties improve the ability to find the global optimum. In this work, we provide an extensive study of the relationship between the BO performance (regret) and uncertainty calibration for popular surrogate models and acquisition functions, and compare them across both synthetic and real-world experiments. Our results show that Gaussian Processes, and more surprisingly, Deep Ensembles are strong surrogate models. Our results further show a positive association between calibration error and regret, but interestingly, this association disappears when we control for the type of surrogate model in the analysis. We also study the effect of recalibration and demonstrate that it generally does not lead to improved regret. Finally, we provide theoretical justification for why uncertainty calibration might be difficult to combine with BO due to the small sample sizes commonly used.

## 1 INTRODUCTION

Probabilistic machine learning provides a framework in which it is possible to reason about uncertainty for both models and predictions [Ghahramani, 2015]. It is often argued that especially in high-stakes applications (healthcare, robotics, etc.), uncertainty estimates for decisions/predictions should be a central component and that they should be well-calibrated [Kuleshov and Deshpande, 2022]. The intuition behind calibration is that the uncertainty estimates should accurately reflect reality; for example, if a classification model predicts an 80% probability of belonging to class $A$ on 10 datapoints, then (on average) we would expect 8 of those 10 samples actually belong to class $A$. Likewise – but less intuitively – in regression, if a calibrated model generates a prediction $\mu$ and standard deviation $\sigma$, we would expect to see $p$ percent of the data lying inside a $p$ percentile confidence interval of $\mu$ [Busk et al., 2021].

Uncertainty also plays a central role in Bayesian Optimization (BO) [Snoek et al., 2012], which will be the focus of this paper. As a sequential design strategy for global optimization, BO has several applications with perhaps the most popular ones being general experimental design [Shahriari et al., 2015] and model selection for machine learning models [Bergstra et al., 2011].

BO is most often used when the objective function is expensive (e.g. monetary, or time-consuming) or unethical to evaluate, gradients between in- and outputs are not available, noisy, and/or data acquisition is limited to few training samples [Agnihotri and Batra, 2020]. A BO protocol works by iteratively fitting a probabilistic surrogate model to observed values of an objective function, and using a so-called acquisition function (AF) based on the surrogate model, to select where to query the objective function next. In AFs, there is an inherent trade-off between exploring input areas in which the surrogate model is uncertain of the underlying objective function, and exploiting areas where the surrogate model already knows that the objective value is close to optimal. As such, it seems obvious that in order for this exploration-exploitation trade-off to be good, the probabilistic model must be well-calibrated. It is, however, still not well-described how much calibration actually affects BO procedures. One could imagine that if calibration leads to a better model representation of the underlying objective function, as would be the general intuition, it would be natural to expect that improving calibration via so-called *recalibration* [Kuleshov et al., 2018] will aid in finding the global optimum of that same function.

*Accepted for the 39th Conference on Uncertainty in Artificial Intelligence* (UAI 2023).

### 1.1 OUR CONTRIBUTION

In this paper, we set out to investigate how the model uncertainties affect BO performance by means of both numerical and theoretical perspectives. Our work is highly motivated by the general intuition and understanding in the community that BO surrogate models with better / well-calibrated uncertainty estimates will perform better (i.e. reach better final and/or total regret). In particular, our paper is concerned with studying statements such as "BO crucially relying on calibrated uncertainty estimates" [Springenberg et al., 2016] and that methods performing worse "due to their frequentist uncertainty estimates" [Deshwal et al., 2021]. But how well-calibrated do we need to be in order to achieve good BO performance? In order to investigate these questions, we provide four major contributions:

- An extensive study of commonly used surrogate models and acquisition functions, where we study the resulting calibration errors and regrets to assess the relationship between calibration and regret. This includes an intervention study, where we manipulate model calibration and study the effect on regret.
- We show that Deep Ensembles is superior for hyperparameter tuning using BO.
- An investigation of whether recalibration during the BO protocol leads to better BO performance.
- Numerical and theoretical results to substantiate a discussion on the role of calibration in BO. Especially on the relationship between the number of recalibration samples and the variance of the calibration curve.

### 1.2 RELATED WORK

A great deal of work has been carried out for uncertainty calibration for regression models [Kuleshov et al., 2018, Song et al., 2019, Ovadia et al., 2019, Busk et al., 2021, Nado et al., 2021] and the useful uncertainty toolbox [Chung et al., 2021] makes it easy to assess the calibration level of various models. In the very recent work by Deshpande and Kuleshov [2021], a procedure for calibrating Gaussian processes (GPs) during BO was proposed. Given the small sample sizes available in BO, the idea is to use leave-one-out cross-validation and utilize the calibration algorithm proposed in earlier work by Kuleshov et al. [2018]. We note that potential issues might arise from this procedure as the earlier work by Kuleshov et al. [2018] states multiple times their approach produces calibrated forecasts "*given enough i.i.d. data*". However, the data available during BO is rarely large nor independent and identically distributed (i.i.d.), and the goal of our work is to dive deeper into this. Other research on the role of uncertainty calibration includes examples such as the work by Bliznyuk et al. [2008], where the authors propose a way of using Markov Chain Monte Carlo (MCMC) to get calibrated predictions for GPs. In Belakaria et al. [2020], the authors investigate uncertainty-aware multi-objective (multidimensional output) BO and argue that due to the uncertainty incorporating strategy, their model outperforms state-of-the-art procedures.

## 2 BACKGROUND

Bayesian Optimization (BO) is concerned with the optimization task of finding the global minimum $\mathbf{x}^* = [x_1^*, x_2^*, ..., x_D^*]^\top$ of some objective function $f(\mathbf{x})$, where $\mathbf{x}$ is a $D$-dimensional vector, i.e.

$$\mathbf{x}^* = \operatorname{argmin} f(\mathbf{x}). \tag{1}$$

We assume that the optimization objective $f(\mathbf{x}) \in \mathbb{R}$ is contaminated with noise, i.e. we observe $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon$, where $\epsilon$ is additive noise often assumed to follow an isotropic normal distribution. In many scenarios such as hyperparameter tuning of neural networks, the set of input variables $\mathbf{x}$ are rarely all real-valued, and often no closed-form expression for $f$ exists. Hence, BO is well-suited when $f$ is a so-called "black-box" function [Turner et al., 2021]. At least two crucial decisions are to be made when using BO in practice: 1) the choice of surrogate model, which is to learn the underlying objective function $f$, and 2) the acquisition function (AF), which controls the strategy for deciding which input $\mathbf{x}$ to sequentially pick by maximizing the AF. Popular choices for surrogate models include Gaussian Processes (GPs) [Rasmussen, 2003, Snoek et al., 2012] and Random Forests (RFs) [Bergstra et al., 2011], but any model with a probabilistic interpretation, e.g. Deep Ensembles (DEs) [Lakshminarayanan et al., 2017] or mean-field Bayesian Neural Networks (BNNs) [Springenberg et al., 2016], can be used.

**Acquisition Functions** For the choice of AF, *Expected Improvement* (EI) as proposed by Jones et al. [1998] is often used and is defined as follows:

$$\mathrm{EI}(\mathbf{x}) = (\mu(\mathbf{x}) - f(\mathbf{x}^+))\Phi(Z) + \sigma(\mathbf{x})\phi(Z), \tag{2}$$

if $\sigma(\mathbf{x}) > 0$ otherwise $\mathrm{EI}(\mathbf{x}) = 0$, and with $Z(\mathbf{x}) = \frac{\mu(\mathbf{x}) - f(\mathbf{x}^+)}{\sigma(\mathbf{x})}$, where $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ denote the mean and standard deviation, respectively, of the surrogate function at $\mathbf{x}$, $f(\mathbf{x}^+)$ denotes the best function value observed so far, and $\Phi$ and $\phi$ denote the cumulative distribution function (CDF) and probability density function (PDF) of a standard normal distribution, respectively. Another popular AF is the *Upper Confidence Bound* (UCB), proposed in Srinivas et al. [2012] which is defined as:

$$\mathrm{UCB}(\mathbf{x}) = -\mu(\mathbf{x}) + \beta^{1/2}\sigma(\mathbf{x}), \tag{3}$$

for minimization problems, where $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ once again denote the mean and standard deviation of the surrogate function at $\mathbf{x}$ and $\beta$ is a hyperparameter controlling the

Table 1: BO results for experiments with synthetic data. For each of the surrogate and acquisition pairs here, we ran a total of 128 optimization problems, where each problem is repeated with 20 different seeds. For each pair, we report the mean of all $128 \cdot 20 = 2560$ runs and the standard error of the mean for all metrics. The instantaneous and total regret metrics are computed using eq. (8) and (9), respectively. ECE is the expected calibration error and is computed using eq. (7) and sharpness denotes the negatige entropy of the predictive distributions. Rows with Acquisition=Average (AVG) correspond to an average over all three acquisition strategies (EI, UCB, TS), but excluding random sampling (RS). Best performing configurations in each of the three sections (i.e. RS, EI+UCB+TS, AVG) are reported in bold font.

| Surrogate | Acquisition | Inst. Regret | Total Regret | ECE | Sharpness |
|---|---|---|---|---|---|
| GP | RS | **0.496 ± 0.018** | **67.117 ± 2.155** | **0.005 ± 0.000** | -0.183 ± 0.012 |
| DE | RS | 0.508 ± 0.019 | 67.345 ± 2.194 | 0.011 ± 0.000 | 0.030 ± 0.007 |
| RF | RS | 0.511 ± 0.018 | 67.920 ± 2.205 | 0.006 ± 0.000 | -0.478 ± 0.016 |
| BNN Small | RS | 0.519 ± 0.019 | 67.990 ± 2.199 | 0.088 ± 0.001 | 1.253 ± 0.008 |
| BNN | RS | 0.509 ± 0.018 | 67.489 ± 2.165 | 0.105 ± 0.001 | 3.241 ± 0.000 |
| GP | EI | 0.036 ± 0.001 | 13.214 ± 0.325 | 0.016 ± 0.000 | -0.224 ± 0.012 |
| DE | EI | 0.043 ± 0.002 | 21.714 ± 0.524 | 0.029 ± 0.001 | -0.353 ± 0.009 |
| RF | EI | 0.099 ± 0.004 | 33.511 ± 0.994 | 0.025 ± 0.000 | -0.386 ± 0.016 |
| BNN Small | EI | 0.848 ± 0.026 | 91.221 ± 2.719 | 0.113 ± 0.001 | 0.602 ± 0.008 |
| BNN | EI | 0.755 ± 0.024 | 87.944 ± 2.62 | 0.110 ± 0.001 | 3.221 ± 0.000 |
| GP | UCB | **0.027 ± 0.001** | **12.829 ± 0.328** | 0.017 ± 0.000 | -0.322 ± 0.012 |
| DE | UCB | 0.046 ± 0.002 | 21.148 ± 0.508 | 0.028 ± 0.001 | -0.375 ± 0.009 |
| RF | UCB | 0.081 ± 0.003 | 31.173 ± 0.945 | 0.025 ± 0.000 | -0.404 ± 0.016 |
| BNN Small | UCB | 0.480 ± 0.016 | 64.604 ± 1.830 | 0.097 ± 0.001 | 0.861 ± 0.007 |
| BNN | UCB | 0.734 ± 0.023 | 86.777 ± 2.595 | 0.110 ± 0.001 | 3.221 ± 0.000 |
| GP | TS | 0.041 ± 0.003 | 28.729 ± 1.044 | **0.010 ± 0.000** | -0.436 ± 0.011 |
| DE | TS | 0.042 ± 0.002 | 22.116 ± 0.508 | 0.027 ± 0.001 | -0.333 ± 0.009 |
| RF | TS | 0.279 ± 0.013 | 51.166 ± 1.783 | 0.013 ± 0.000 | -0.451 ± 0.015 |
| BNN Small | TS | 0.628 ± 0.021 | 76.086 ± 2.330 | 0.091 ± 0.001 | 0.997 ± 0.007 |
| BNN | TS | 0.519 ± 0.019 | 68.111 ± 2.225 | 0.105 ± 0.001 | 3.242 ± 0.000 |
| GP | AVG | **0.035 ± 0.001** | **18.257 ± 0.390** | **0.015 ± 0.000** | -0.327 ± 0.007 |
| DE | AVG | 0.044 ± 0.001 | 21.659 ± 0.296 | 0.028 ± 0.000 | -0.354 ± 0.005 |
| RF | AVG | 0.153 ± 0.005 | 38.616 ± 0.757 | 0.021 ± 0.000 | -0.414 ± 0.009 |
| BNN Small | AVG | 0.652 ± 0.013 | 77.303 ± 1.346 | 0.100 ± 0.001 | 0.820 ± 0.005 |
| BNN | AVG | 0.669 ± 0.013 | 80.944 ± 1.439 | 0.108 ± 0.000 | 3.228 ± 0.000 |

trade-off between exploitation and exploration. Finally, the acquisition strategy coined *Thompson Sampling* [Thompson, 1933] works by generating a random sample from the posterior of $f$ and then locating the optimal value for the specific sample, i.e. for some sample $f(\mathbf{x}) \sim p(f|\text{Data})$

$$\text{TS}(\mathbf{x}) = -f(\mathbf{x}). \tag{4}$$

For GPs and BNNs this is done by sampling a function from the posterior, whilst for DEs and RFs we sample a neural network or tree, respectively (Elmachtoub et al. [2017]).

**Calibration** Following the work by Kuleshov et al. [2018], a regression model is well-calibrated if approximately $q$ percent of the time test samples fall inside a $q$ percent confidence interval of the predictive distribution. For regression tasks, the model calibration can be assessed using the expected calibration error

$$\text{ECE} = \sum_p w_p (C_y(p) - p)^2, \tag{5}$$

where $C_y(p)$ is defined as

$$C_y(p) = \frac{1}{N_T} \sum_{t=1}^{N_T} \mathbb{I}[y_t \leq F_t^{-1}(p)], \tag{6}$$

where $F_t^{-1}$ is the quantile function, i.e. $F_t^{-1}(p) \equiv \inf_y \{y \mid p \leq F_t(y)\}$, for the $t$'th datapoint evaluated at percentile $p$, $\mathbb{I}$ is an indicator function and $w_p$ can be chosen to adjust the importance of percentiles with fewer datapoints. Throughout this paper, we assume $w_p = 1 \ \forall p$. The closer the ECE is to zero, the better calibrated the model is.

Table 2: BO results for hyperparameter tuning experiments. For each of the surrogate and acquisition pairs here, we ran a total of 6 optimization problems, where each problem is repeated with 100 different seeds. For each pair, we report the mean of all $6 \cdot 100 = 600$ runs and the standard error of the mean for all metrics. The instantaneous and total regret metrics are computed using eq. (8) and (9), respectively. ECE is the expected calibration error and is computed using eq. (7) and sharpness denotes the negative entropy of the predictive distributions. Rows with Acquisition=Average (AVG) correspond to an average over all three acquisition strategies (EI, UCB, TS), but excluding random sampling (RS). Best performing configurations in each of the three sections (i.e. RS, EI+UCB+TS, AVG) are reported in bold font.

| Surrogate | Acquisition | Inst. Regret | Total Regret | ECE | Sharpness |
|---|---|---|---|---|---|
| GP | RS | $0.0151 \pm 0.0006$ | $2.7021 \pm 0.0995$ | $\mathbf{0.0055 \pm 0.0001}$ | $-0.7762 \pm 0.0138$ |
| DE | RS | $0.0161 \pm 0.0007$ | $2.7822 \pm 0.1033$ | $0.0093 \pm 0.0001$ | $-0.2574 \pm 0.0134$ |
| RF | RS | $0.0152 \pm 0.0007$ | $2.6977 \pm 0.1018$ | $0.0072 \pm 0.0002$ | $1.0302 \pm 0.1017$ |
| BNN Small | RS | $\mathbf{0.0150 \pm 0.0007}$ | $\mathbf{2.5948 \pm 0.0942}$ | $0.1015 \pm 0.0005$ | $1.3499 \pm 0.0102$ |
| BNN | RS | $0.0154 \pm 0.0007$ | $2.7820 \pm 0.1009$ | $0.1075 \pm 0.0005$ | $3.2391 \pm 0.0003$ |
| GP | EI | $0.0031 \pm 0.0002$ | $1.5375 \pm 0.0565$ | $0.0153 \pm 0.0004$ | $-0.5433 \pm 0.0155$ |
| DE | EI | $\mathbf{0.0011 \pm 0.0001}$ | $\mathbf{0.9031 \pm 0.0436}$ | $0.0363 \pm 0.0010$ | $-0.2927 \pm 0.0096$ |
| RF | EI | $0.0043 \pm 0.0003$ | $1.0925 \pm 0.0459$ | $0.0146 \pm 0.0004$ | $0.8718 \pm 0.0761$ |
| BNN Small | EI | $0.0332 \pm 0.0018$ | $4.8430 \pm 0.2239$ | $0.1052 \pm 0.0007$ | $0.7928 \pm 0.0136$ |
| BNN | EI | $0.0170 \pm 0.0009$ | $3.1505 \pm 0.1328$ | $0.1092 \pm 0.0005$ | $3.2247 \pm 0.0004$ |
| GP | UCB | $0.0026 \pm 0.0002$ | $1.5156 \pm 0.0560$ | $0.0149 \pm 0.0004$ | $-0.5297 \pm 0.0154$ |
| DE | UCB | $0.0012 \pm 0.0001$ | $0.9159 \pm 0.0437$ | $0.0369 \pm 0.0009$ | $-0.2862 \pm 0.0098$ |
| RF | UCB | $0.0043 \pm 0.0002$ | $1.0979 \pm 0.0455$ | $0.0157 \pm 0.0004$ | $0.9205 \pm 0.0779$ |
| BNN Small | UCB | $0.0104 \pm 0.0007$ | $2.6292 \pm 0.1176$ | $0.1013 \pm 0.0006$ | $1.0458 \pm 0.0088$ |
| BNN | UCB | $0.0152 \pm 0.0008$ | $3.1068 \pm 0.1300$ | $0.1093 \pm 0.0005$ | $3.2244 \pm 0.0004$ |
| GP | TS | $0.0046 \pm 0.0003$ | $1.7544 \pm 0.0643$ | $0.0125 \pm 0.0003$ | $-0.5814 \pm 0.0173$ |
| DE | TS | $0.0016 \pm 0.0002$ | $1.0321 \pm 0.0489$ | $0.0364 \pm 0.0009$ | $-0.2522 \pm 0.0100$ |
| RF | TS | $0.0017 \pm 0.0002$ | $1.3192 \pm 0.0497$ | $\mathbf{0.0101 \pm 0.0002}$ | $0.8893 \pm 0.0859$ |
| BNN Small | TS | $0.0176 \pm 0.0009$ | $2.9900 \pm 0.1231$ | $0.1025 \pm 0.0005$ | $1.0644 \pm 0.0091$ |
| BNN | TS | $0.0150 \pm 0.0007$ | $2.6796 \pm 0.0988$ | $0.1075 \pm 0.0005$ | $3.2405 \pm 0.0003$ |
| GP | AVG | $0.0034 \pm 0.0001$ | $1.6025 \pm 0.0342$ | $0.0142 \pm 0.0002$ | $-0.5515 \pm 0.0093$ |
| DE | AVG | $\mathbf{0.0013 \pm 0.0001}$ | $\mathbf{0.9504 \pm 0.0263}$ | $0.0365 \pm 0.0005$ | $-0.2770 \pm 0.0057$ |
| RF | AVG | $0.0034 \pm 0.0001$ | $1.1699 \pm 0.0273$ | $\mathbf{0.0135 \pm 0.0002}$ | $0.8939 \pm 0.0462$ |
| BNN Small | AVG | $0.0204 \pm 0.0007$ | $3.4874 \pm 0.0965$ | $0.1030 \pm 0.0003$ | $0.9676 \pm 0.0069$ |
| BNN | AVG | $0.0157 \pm 0.0005$ | $2.9790 \pm 0.0703$ | $0.1087 \pm 0.0003$ | $3.2299 \pm 0.0003$ |
| GP (recal.) | AVG | $0.0060 \pm 0.0002$ | $1.8416 \pm 0.0400$ | $0.0149 \pm 0.0002$ | $-0.6552 \pm 0.0058$ |
| DE (recal.) | AVG | $\mathbf{0.0019 \pm 0.0001}$ | $\mathbf{1.1468 \pm 0.0320}$ | $0.0418 \pm 0.0005$ | $-0.3123 \pm 0.0042$ |
| RF (recal.) | AVG | $0.0029 \pm 0.0001$ | $1.1907 \pm 0.0292$ | $\mathbf{0.0112 \pm 0.0001}$ | $-0.5700 \pm 0.0047$ |
| BNN Small (recal.) | AVG | $0.0383 \pm 0.0013$ | $4.9472 \pm 0.1458$ | $0.0937 \pm 0.0003$ | $0.7728 \pm 0.0136$ |
| BNN (recal.) | AVG | $0.0157 \pm 0.0005$ | $3.0210 \pm 0.0721$ | $0.1071 \pm 0.0003$ | $3.1546 \pm 0.0165$ |

**Recalibration** Kuleshov et al. [2018] also propose a general procedure for recalibrating any surrogate model. A so-called recalibrator model $C$ is constructed using an independent and identically distributed (i.i.d.) validation set and subsequently, applied to adjust the CDF of the model's predictive distribution $F_t$ for some observation $y_t$, i.e. the recalibrated predictive distribution is $C \circ F_t$. This is done via learning an isotonic mapping: $C : [0, 1] \rightarrow [0, 1]$ from the predicted probabilities of events of the form $(-\infty, y_t]$ to the corresponding empirical probabilities. In Deshpande and Kuleshov [2021], a recalibration method for BO specifically is proposed, in which the recalibrator model is learnt via leave-one-out CV on the samples gathered during BO. After training the recalibrator model $C$, the relevant summary statistics (e.g. moments and intervals) of the recalibrated distributions can be computed numerically from $C \circ F_t$. See Alg. 1 in Kuleshov et al. [2018] for more details.

## 3 EXPERIMENTS

In this section, we describe a collection of numerical experiments designed to study and investigate the relationship between calibration and regret. We focus our study on four popular models, namely GPs, RFs, DEs, and BNNs. For GPs, DEs, and BNNs, we assume an isotropic Gaussian likelihood and for RFs, we impose a Gaussian predictive distribution, where the mean and variance are estimated as the sample mean and variance of the tree predictions. Our experiments are based on both synthetic and real-world data: for experiments with synthetic data, we use a number of problems from the common benchmark suites for optimization called Sigopt [Jamil and Yang, 2013, Dewancker et al., 2016], and for the real-world data, we apply BO to hyperparameter tuning of various machine learning models including feed-forward Neural Networks, Convolutional Neural Networks and SVMs used on on or more datasets such as MNIST [Lecun et al., 1998], Fashion-MNIST [Xiao et al., 2017], AG News classification [Zhang et al., 2015] and Wine classification [Dua and Graff, 2017]. For all experimental details, see Supplementary Material.

**Experimental Setup** In the synthetic setting, we perform BO experiments on a total of 128 optimisation problems spanning input dimensions ($D \in \{1, 2, .., 10\}$) from the Sigopt benchmark. For each optimisation problem, we repeat the experiment 20 times using different random initialization of both the BO routines and seeds. We do this for all combinations of surrogates and AFs, of which we use the previously mentioned EI, UCB and TS. We consistently use ten initial i.i.d. random samples followed by 90 BO iterations for all experiments. We add Gaussian distributed noise giving a SNR of 100 to all Sigopt objective functions. For reference, we also include a random sampling (RS) acquisition function. In the hyperparameter tuning setting, we perform BO experiments on a total of 6 different hyperparameter tuning problems. The surrogate models and AFs are the same as in the synthetic setting, and we similarly sample 10 i.i.d. points to initiate the BO session, and then run 90 BO iterations. Here we run each experiment 100 times.

Our key performance metrics are regret, calibration error and sharpness as defined in the following. We report the calibration error, ECE, as being the mean squared calibration error evaluated on a large i.i.d. test set ($N_{\text{test}} = 5000$) as

$$\text{ECE} = \frac{1}{P} \sum_{j=1}^{P} (C_y(p_j) - p_j)^2, \tag{7}$$

where $C_y(p_j)$ is defined in eq. (6) and for $0 \leq p_1 \leq p_2 ... \leq p_P \leq 1$ as suggested by Kuleshov et al. [2018]. We use $P = 20$ with equidistant $p_j$ values. The ECE values are reported as averages across all BO iterations. We quantify the BO performance using the regret metric, where we define the instantaneous regret for the last iteration $T$ as follows

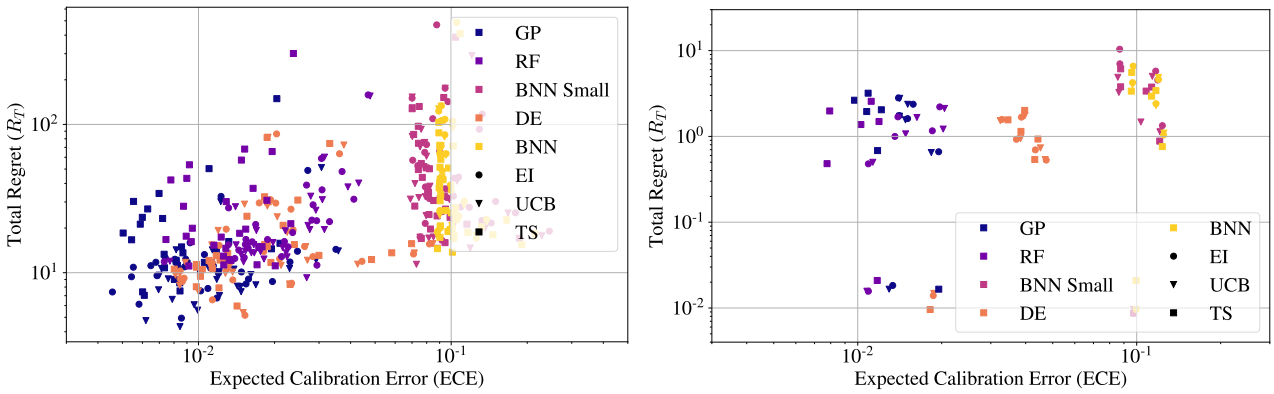
$$\mathcal{R}_I = y_{\text{min}} - y(x_T^*), \tag{8}$$

where $y(x)$ is the objective function value (with added noise in the synthetic case, i.e. $y(x) = f(x) + \sigma$) obtained at point $x$, $y_{\text{min}} \equiv \min_x y(x)$ is function value at the global minimum, and $x_T^* \equiv \arg\min_{x_t} \{y(x_t)\}_{t=1}^T$ is the input value for the best observation after $T$ iterations. Similarly, the total regret is the sum of the instantaneous regret across all iterations

$$\mathcal{R}_T = \sum_{i=1}^{T} \left[ y_{\text{min}} - y(x_i^*) \right]. \tag{9}$$

All regret values are reported after standardizing objective function values. Finally, we report the sharpness as the average negative entropy of the predictive distributions as evaluated on the test-set across all BO iterations. For the choices of surrogate models, we use a GP with an RBF kernel, and optimize hyperparameters of the kernel at every BO iteration using the exact marginal likelihood [Rasmussen, 2003]. We use two different mean-field BNN architectures, a smaller (BNN Small) with a single hidden layer with 10 hidden neurons and a larger (BNN) with two hidden layers with 30 and 10 hiddens nodes respectively. Both are trained using the ELBO loss [Blei et al., 2018]. The DEs consists of 10 neural networks with two hidden layers and are all trained using the MSE loss and Adam optimiser [Kingma and Ba, 2014]. Finally, the RFs have their hyperparameters tuned via CV on a grid of hyperparameters at each BO iteration. With regards to the AFs, we use EI as defined in Eq. 2, UCB with $\beta = 1$, and only sample one posterior function at each BO step when using TS. See detailed experimental details and descriptions in the Supplementary Material. Code is available at `https://github.com/jfold/unibo`.

**Experiment results** The results for the synthetic and real data experiments are summarized in Tables 1 and 2, respectively. We observe that in the synthetic setting, GPs outperform all other models both in terms of instantaneous regret and more importantly, total regret, although closely followed by DEs. RFs perform relatively well (at all times better than random sampling), whilst the BNNs exhibit poor performance and are often outperformed by random sampling. Finally, we see that the GP is best calibrated overall, and that all surrogate models have their lowest ECE when random sampling is used. This is overall not surprising as the ECE is evaluated on a large i.i.d. test set, which is more well-represented by i.i.d. training samples compared to strongly dependent samples acquired iteratively through BO. For the real-data experiments in Table 2, we see that DEs outperform all other models in terms of both regret types, and are closely followed by both GPs and RFs which perform comparatively. Once again, GPs are the best calibrated when random sampling is employed.




(a) Test calibration vs regret of synthetic data experiments.

(b) Test calibration vs regret of real data experiments

Figure 1: Total Regret vs. ECE for synthetic data experiments and hyperparameter tuning experiments. The colors in the scatter plot indicate the type of surrogate model, and the marker indicates the AF used. **OBS:** Each point in the scatter plots corresponds to an average of 20 seeds in the synthetic data experiments and 100 seeds in the hyperparameter tuning experiments for each specific configuration.

**Relationship between calibration and regret** In order to investigate the relationship between BO performance (regret) and calibration (ECE), we first compute the Pearson correlation coefficient between the total regret values and the ECE values, which yield moderate and statistically significant coefficients of $0.28$ and $0.42$ for synthetic and hyperparameter tuning experiments, respectively (see Table 3). The moderate positive association is also visually confirmed by the scatter plots in Figures 1. It is also evident from these plots that the type of surrogate model is important for both ECE and total regret. Therefore, we also compute the partial correlation coefficient controlling for the model type yielding $-0.06$ and $-0.24$ for synthetic and real data, respectively. Interestingly, both correlations become weaker and one statistically insignificant (at level $\alpha = 0.05$) leading to an instance of Simpson's paradox [Wagner, 1982]. To further investigate this, we conducted a multiple linear regression analysis for total regret vs ECE controlling for both the type of model and the specific problem instance. The results for the hyperparameter tuning experiments showed that both the common slope and model-specific slopes for ECE were generally weak and statistically insignificant (see all details in the Supplementary Material). In summary, these results show that models with high ECE are generally associated with high regrets, however, this association diminishes when we control for the type of surrogate model. To further scrutinize these observations, we conduct two additional experiments: an intervention study and a recalibration study.

Table 3: Correlation values between regret and ECE.

| | **Synth. Data** | **Real Data** |
|---|---|---|
| **Correlation** | $0.28$ ($p < 10^{-8}$) | $0.42$ ($p < 10^{-4}$) |
| **Partial Correlation \| Model** | $-0.06$ ($p = 0.19$) | $-0.24$ ($p = 0.026$) |

**Intervention study: Perturbing Predictive Uncertainties** In the intervention study, we explicitly manipulate calibration by perturbing the predictive uncertainty of each model during the BO protocol. Specifically, we multiply the standard deviation of the posterior distribution for all models by a constant $c \in \left[10^{-4}, 10^{2}\right]$ and observe the resulting effect on ECE and total regret. We conduct this experiment for the 6 different hyperparameter tuning problems using the EI acquisition function and repeat the experiment with 40 different seeds. In Figure 2 we show the calibration error (a) and total regret (b) as a function of the multiplicative constant $c$. Several interesting observations can be made from Figure 2. First, all models exhibit the smallest calibration error at $c > 1$, which indicates some degree of overconfidence, and thus, increasing the predictive variance slightly generally improves calibration. Interestingly, DEs and GPs are somewhat robust to these perturbations in their predictive uncertainties with regard to regret, while RFs even seem to benefit from having the uncertainties reduced. Finally, in panel (c) we plot regret vs calibration error for each value of $c$, where each marker is scaled with the size of $c$ and $c = 1$ is marked with black. We have connected the dots for each surrogate function, going from smallest to largest $c$. From this plot, we observe that perturbing by $c > 1$ rapidly increase both regret and ECE, but perturbations with $c < 1$ are less harmful and may actually lead to improved performance. In other words, the results from this experiment suggest that miscalibration caused by models being generally underconfident, i.e. $c > 1$, is more detrimental to BO performance compared to models being overconfident, i.e. $c < 1$.

**Recalibration study: Recalibration during BO** In the recalibration study, we investigate whether recalibrating the models during the BO protocol improves BO performance following the recalibration procedure proposed by

Deshpande and Kuleshov [2021]. We do this by re-running our BO experiments on real data, where we use leave-one out CV on the training data obtained during BO to learn a recalibration model and adjust the resulting predictive distributions accordingly. The results can be seen in the last section in the bottom of Table 2. Except for RFs, it is seen that both regret and ECE are generally worse *after recalbration*. This may seem counter-intuitive, but then recall that we compute the recalibration model using leave-one-out on the training set, but we measure the expected calibration on an independent test set. The recalibration procedure may have improved the calibration metric on the training dataset, but in our experiments, it does not generalize to an independent test set. We note that RFs do benefit from recalibration, but this might be explained by the fact that sharpness is substantially reduced after recalibration. We will shed more light on these observations in the next section.

# 4 DISCUSSION AND SUMMARY

In the previous section, we described and performed a number of numerical experiments to analyze the relationship between calibration and regret for BO. In this section, we will summarize and discuss some of the key take-aways as well as expand the analysis with a theoretical perspective.

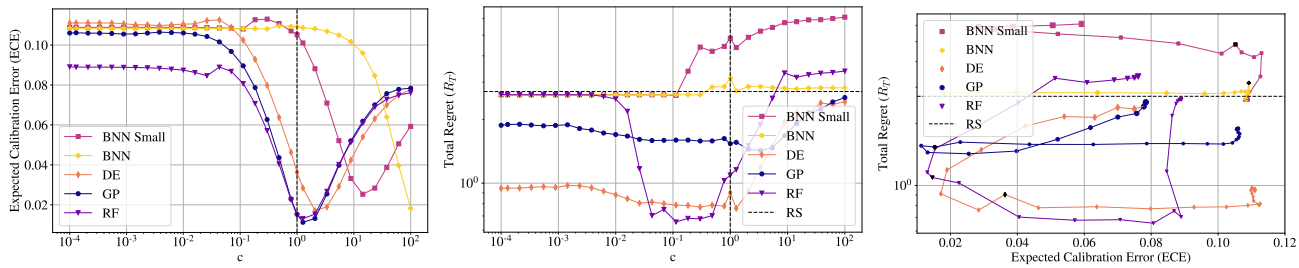
**Take-away 1: Gaussian Processes and Deep Ensembles work well for BO.** Our results for synthetic data is consistent with the apparent consensus that GPs are strong surrogates for BO and that they outperform the competing methods in terms of regret (both total and instant) (see Table 1), with DEs being close followers. Surprisingly, in the hyperparameter tuning experiments, DEs perform exceedingly well, with RFs and GPs performing equally well. One should however note the practical concern that DEs is computationally more expensive to train during the BO procedure, but that this could be rationalized if such compute time is cheap relative to querying the objective function. In both experiments, the mean-field BNNs perform significantly worse than all other methods, including random search. Similar behavior has also been observed in other experimental design settings, e.g. active learning [Foong et al., 2020]. In terms of ECE, the GPs performed slightly better than the RFs and DEs in the synthetic setting, whilst RFs and GPs perform comparably in the hyperparameter tuning setting. Again, we notice that the mean-field BNNs are inferior to the other methods in both experiments.

**Take-away 2: Correlation between BO performance and calibration diminishes when controlling for the type of surrogate model.** For the synthetic and hyperparameter experiments, our analysis showed moderate positive correlations of $0.28$ and $0.42$, respectively, between total regret and ECE, when computed across all problems, seeds, acquisition functions, and surrogate models. However, when we control for the type of surrogate model, the correlation becomes much weaker and one becomes statistically insignificant (see Table 3). That is, within each model family, BO trials with lower calibration errors are generally not linearly associated with lower regret and in turn better BO performance.

**Take-away 3: Under-confidence might be more harmful to BO compared to overconfidence.** In our intervention study, we manipulated all surrogate models to be either under- or overconfident during the BO protocol by multiplying their predictive uncertainties by a constant $c > 0$, where $0 < c < 1$ implies more confident predictions, and $c > 1$ implies less confident predictions. The results showed that all models exhibited some degree of overconfidence, which may not be surprising. However, the results also showed that BO performance decreased (i.e. regret increased) rapidly for all models except for the larger BNN for $c > 1$, whereas BO performance was much more robust to perturbations with $c < 1$, which actually caused an increase in BO performance in some cases. Only for the GP, we observed a slight temporary improvement in regret for $c > 1$. It is also worth emphasizing that the value of $c$ leading to optimal calibration did not coincide with the values for optimal regret. Finally, it is evident from eq. (2) that changing $c$ also affects the effective exploitation-exploration trade-off which, in turn, may also impact the regret (the optimal trade-off is also likely to be intrinsic to the optimisation problem). This can be observed in Figure 2, where both very small and very large values of $c$ caused the methods to behave more like random search.

**Take-away 4: Recalibration does generally not improve BO performance.** We further investigated the potential benefit of recalibrating the surrogate models during the BO process using a leave-one-out procedure. However, in our recalibration experiments on the hyperparameter tuning datasets, the recalibration procedure only lead to improved ECE (measured on a proper independent test set) for two surrogate models, namely the small BNNs and the RFs. In the other cases, it actually worsened the ECE. Moreover, we also noticed that all models got worse total regret performance after employing the recalibration procedure.

**Hypothesis: Calibration curves are not reliable for small sample sizes.** Recent work by Deshpande and Kuleshov [2021] observed that re-calibration might aid BO by yielding smaller total regret in some trials and smaller instant regret for the BO last iteration in fewer trials. However, our experiments suggest that recalibration might actually degrade BO performance. Kuleshov et al. [2018] state that a sufficiently large i.i.d. validation set is a required condition for successful recalibration, which is in stark contrast to the sample collection during BO which is not i.i.d. due to the inherent sequential nature of BO algorithms and is often characterized by small sample sizes.




(a) ECE after BO protocol (b) Regret after BO protocol (c) Regret vs. ECE for various perturbations.

Figure 2: The effect on test calibration and regret when disturbing the posterior predictive uncertainty by $c \cdot \sigma(\mathbf{x})$ during the BO protocol. (a) Shows the overall ECE of each model when a perturbation of $c \cdot \sigma(\mathbf{x})$ is done in each iteration, (b) shows the corresponding total regret, and (c) depicts how regret and calibration varies together for the same experiments. The size of the markers here indicate how large $c$ is, and the plot lines go from smallest to largest $c$. Black points are when $c = 1$.

To investigate this hypothesis, our starting point will be a simple regression setting, where $p_y(y|x)$ denotes the true data generating distribution of $y$ given an input $x$. We further assume a trained model with predictive distribution $p_t(y|x)$ aiming to mimic $p_y$ via training samples. Consider now the task of assessing the calibration of model using a set of i.i.d. validation samples $\{y_1, y_2, ..., y_N\}$. Given the small sample sizes typically used in BO, a natural question to ask is how accurately can we assess the calibration curve as a function of the size of the validation set $N$? To investigate this question, we consider the variance of the estimator in eq. (6) and analyze its decay rate as a function of the sample size $N$. The result is summarized in the following statement:

**Proposition 1.** *Let $F_i$ be the CDF of the predictive distribution for the i'th observation and let $\{y_i\}_{i=1}^N$ be i.i.d. samples $y_i \sim p_y$. For $\mathcal{C}_y(p) = \frac{1}{N}\sum_{i=1}^N \mathbb{I}\left[y_i \leq F_i^{-1}(p)\right]$, then the variance of $C_y(p)$ decays as $\mathbb{V}\left[\mathcal{C}_y(p)\right] = \mathcal{O}(N^{-1})$.*

*Proof.* Let $\mathcal{C}_y(p) = \frac{1}{N}\sum_{i=1}^N z_i$ for $z_i \equiv \mathbb{I}\left[y_i \leq F_i^{-1}(p)\right]$. The variance of $\mathcal{C}_y(p)$ is then given by

$$\mathbb{V}\left[\mathcal{C}_y(p)\right] = \mathbb{V}\left[\frac{1}{N}\sum_{i=1}^N z_i\right]$$

and by independence each of $z_i$,

$$\mathbb{V}\left[\mathcal{C}_y(p)\right] \leq \frac{1}{N^2}\sum_{i=1}^N \sup_i \mathbb{V}\left[z_i\right] = \frac{1}{N^2}\sum_{i=1}^N \frac{1}{2^2} = \frac{1}{N}\frac{1}{2^2}.$$

Hence, it follows the variance of $\mathcal{C}_y(p)$ is bounded by

$$\mathbb{V}\left[\mathcal{C}_y(p)\right] \leq \mathcal{O}\left(N^{-1}\right). \tag{10}$$

See Supplementary Material for detailed proof. □

We also confirmed this result empirically and observe results perfectly consistent with the predictions from Proposition 1 (see in the Supplementary Material), i.e. the maximum standard deviation of the estimator for $C_y(p)$ decays as $\frac{1}{\sqrt{N}}$. Next, we assume our model is perfect, i.e. $p_t(y|x) = p_y(y|x)$, and ask what is the contribution to ECE caused by a small sample size alone. The results are summarized in the next two statements:

**Proposition 2.** *Let $F_i$ be the CDF of the predictive distribution perfect model, i.e. $p_t(y|x) = p_y(y|x)$. If $F_i$ is strictly monotonic, it holds that $\mathbb{V}\left[\mathcal{C}_y(p)\right] = \frac{p(1-p)}{N}$ for all $p$.*

*Proof.* In this setting, we have

$$z_i = \mathbb{I}\left[y_i \leq F_i^{-1}(p)\right] = \mathbb{I}\left[F_i(y_i) \leq p\right] = \mathbb{I}\left[u_i \leq p\right],$$

where $u_i \sim \mathcal{U}[0,1]$ are uniformly distributed on the unit interval due to the probability integral transform. Since $\{u_i\}_{i=1}^N$ are also independent, it follows that $S_n = \sum_{i=1}^N z_i \sim \text{Binomial}(N, p)$. Therefore, we have

$$\mathbb{V}\left[\mathcal{C}_y(p)\right] = \mathbb{V}\left[N^{-1}S_N\right] = N^{-2}\mathbb{V}\left[S_N\right] = N^{-1}p(1-p).$$

This completes the proof. □

**Proposition 3.** *Let $ECE = \sum_{j=1}^P w_j(p_j - \mathcal{C}_y(p_j))^2$ be the weighted mean square calibration error. Assume $w_i \in [0,1]$ and $0 < p_1 < p_2 < ... < p_P < 1$ are fixed, and assume the CDF of the predictive distribution is equal to the true data distribution (almost everywhere), then it holds that $\mathbb{E}\left[ECE\right] = \frac{1}{n}\sum_{j=1}^P w_j p_j(1-p_j) \propto n^{-1}$.*

*Proof.* See Supplementary Material. □

**Take-away 5: Calibration curves may not be reliable for small sample sizes** Proposition 1 and Proposition 2 state that the variance of the estimator of the empirical calibration decreases with $\mathcal{O}\left(N^{-1}\right)$. This suggests that empirical calibration curves may not be reliable for small sample sizes and in the worst case, to improve the accuracy of the estimates by one decimal point, one needs to increase the size of the validation set by a factor of 100, which will often be infeasible in practical BO settings. Furthermore, Proposition 3 states that even for a perfect model, the expected ECE

is proportional to $N^{-1}$. Therefore, for small sample sizes, one should be careful when concluding that a model is miscalibrated, since the observed ECE might as well be caused by the sample size. Even worse, when performing recalibration in this scenario, one might risk adjusting the model in the "wrong direction" causing the model to be more miscalibrated than the original model. In the Supplementary Material, we show several examples of this phenomenon. Although our empirical and theoretical analysis are focused on the i.i.d. setting, we expect the effect to be even more severe in the non-i.i.d. case since the effective sample size is typically smaller for correlated samples [Thiébaux and Zwiers, 1984]. Therefore, we claim that these effects may have profound impact on recalibration in BO protocols.

**Concluding Remarks** In our experiments, we confirm the common knowledge that GPs generally work well in the BO setting, but interestingly, we also find that Deep Ensembles outperform GPs in some cases. There is the computational downside of Deep Ensembles compared to GPs, however, this overhead may be justified if the cost of evaluating the BO objective function is sufficiently expensive. Moreover, we observe that models with high ECEs are generally associated with worse performance in BO, but that this association disappears when we control for the type of surrogate model. However, we still argue that calibration is important for BO because 1) models with lower ECEs are associated with better regrets and 2) when we explicitly intervened on calibration (by manipulating the predictive uncertainty), we observed that the BO performance for all models decrease significantly. Furthermore, our experiments suggest that recalibration during the BO protocol can hurt BO performance. Based on both theoretical and empirical evidence, we attribute this to the fact that it is really difficult to reliably assess calibration using the small (and non-i.i.d.) datasets typically used in BO. Therefore, we advocate cautiousness when using these recalibration methods for small sample sizes in practice.

**Future work** Our study indicates that the common way to diagnose calibration (on a large test set) might not be sensible for BO and that future studies about calibration metrics more relevant to BO are needed.

It will also be of great interest to explore the relationship between calibration and regret from a causal perspective. Lastly, it would be interesting to dig deeper into the effects of under- vs. over-confidence on BO performance.

## References


A. Agnihotri and N. Batra. Exploring bayesian optimization. *Distill*, 5(5):e26, 2020.

S. Belakaria, A. Deshwal, N. K. Jayakodi, and J. R. Doppa. Uncertainty-aware search framework for multi-objective bayesian optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10044–10052, 2020.

J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. 2018.

N. Bliznyuk, D. Ruppert, C. Shoemaker, R. Regis, S. Wild, and P. Mugunthan. Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation. *Journal of Computational and Graphical Statistics*, 17(2): 270–294, 2008.

J. Busk, P. B. Jørgensen, A. Bhowmik, M. N. Schmidt, O. Winther, and T. Vegge. Calibrated uncertainty for molecular property prediction using ensembles of message passing neural networks. *Machine Learning: Science and Technology*, 3(1):015012, 2021.

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

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

A. Deshwal, S. Belakaria, and J. R. Doppa. Bayesian optimization over hybrid spaces. *arXiv preprint arXiv:2106.04682*, 2021.

I. Dewancker, M. McCourt, S. Clark, P. Hayes, A. Johnson, and G. Ke. A stratified analysis of bayesian optimization methods. *arXiv preprint arXiv:1603.09441*, 2016.

D. Dua and C. Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

A. N. Elmachtoub, R. McNellis, S. Oh, and M. Petrik. A practical method for solving contextual bandit problems using decision trees. *CoRR*, abs/1706.04687, 2017. URL `http://arxiv.org/abs/1706.04687`.

A. Foong, D. Burt, Y. Li, and R. Turner. On the expressiveness of approximate inference in bayesian neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15897–15908. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/b6dfd41875bc090bd31d0b1740eb5b1b-Paper.pdf`.

Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

M. Jamil and X.-S. Yang. A literature survey of benchmark functions for global optimization problems. *arXiv preprint arXiv:1308.4008*, 2013.

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

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014.

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

V. Kuleshov, N. Fenner, and S. Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pages 2796–2804. PMLR, 2018.

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

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Z. Nado, N. Band, M. Collier, J. Djolonga, M. W. Dusenberry, S. Farquhar, A. Filos, M. Havasi, R. Jenatton, G. Jerfel, et al. Uncertainty baselines: Benchmarks for uncertainty & robustness in deep learning. *arXiv preprint arXiv:2106.04015*, 2021.

Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019.

C. E. Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104 (1):148–175, 2015.

J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

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

J. T. Springenberg, A. Klein, S. Falkner, and F. Hutter. Bayesian optimization with robust bayesian neural networks. *Advances in neural information processing systems*, 29:4134–4142, 2016.

N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, may 2012. doi: 10.1109/tit.2011.2182033. URL https://doi.org/10.1109%2Ftit.2011.2182033.

H. J. Thiébaux and F. W. Zwiers. The interpretation and estimation of effective sample size. *Journal of Applied Meteorology and Climatology*, 23(5):800–811, 1984.

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

R. Turner, D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. Guyon. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In *NeurIPS 2020 Competition and Demonstration Track*, pages 3–26. PMLR, 2021.

C. H. Wagner. Simpson's paradox in real life. *The American Statistician*, 36:46–48, 1982.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

X. Zhang, J. J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. *CoRR*, abs/1509.01626, 2015. URL http://arxiv.org/abs/1509.01626.