Adaptive Sampling to Reduce Epistemic Uncertainty Using Prediction Interval-Generation Neural Networks

Giorgio Morales, John W. Sheppard

Gianforte School of Computing
Montana State University, Bozeman, MT 59717, USA
giorgiomorales@ieee.org; john.sheppard@montana.edu

Abstract

Obtaining high certainty in predictive models is crucial for making informed and trustworthy decisions in many scientific and engineering domains. However, extensive experimentation required for model accuracy can be both costly and time-consuming. This paper presents an adaptive sampling approach designed to reduce epistemic uncertainty in predictive models. Our primary contribution is the development of a metric that estimates potential epistemic uncertainty leveraging prediction interval-generation neural networks. This estimation relies on the distance between the predicted upper and lower bounds and the observed data at the tested positions and their neighboring points. Our second contribution is the proposal of a batch sampling strategy based on Gaussian processes (GPs). A GP is used as a surrogate model of the networks trained at each iteration of the adaptive sampling process. Using this GP, we design an acquisition function that selects a combination of sampling locations to maximize the reduction of epistemic uncertainty across the domain. We test our approach on three unidimensional synthetic problems and a multi-dimensional dataset based on an agricultural field for selecting experimental fertilizer rates. The results demonstrate that our method consistently converges faster to minimum epistemic uncertainty levels compared to Normalizing Flows Ensembles, MC-Dropout, and simple GPs.

Code — https://github.com/NISL-MSU/AdaptiveSampling

Introduction

In various scientific and engineering fields, the development of accurate predictive models frequently relies on experimentation. Conducting these experiments can be costly and time-consuming, making it important to adopt strategies that extract the most valuable information from each experiment. One notable example is precision agriculture (PA) where experimental results may require an entire growing season to manifest, and only a portion of the field is allocated for such trials (Lawrence, Rew, and Maxwell 2015). This is exacerbated by the fact data can often only be collected every other year, due to crop rotation.

Adaptive sampling (AS) techniques offer a promising solution by selecting samples intelligently that contribute most to improving model accuracy and reducing uncertainty (Di Fiore, Nardelli, and Mainini 2024). This work focuses on sampling techniques designed to reduce uncertainty in the prediction models across the entire input domain. Such techniques are essential for enhancing trust in decision-making systems whose optimization processes rely on accurate prediction models. For instance, in PA, determining optimal fertilizer rates depends on the shape of estimated nitrogen-yield response (N-response) curves (Bullock and Bullock (1994), Morales and Sheppard (2023a)). These curves represent the estimated crop yield values at specific field sites in response to all admissible fertilizer rates. Uncertainty across the domain can severely affect the survey shapes, leading to unreliable recommended fertilizer rates.

We note a distinction between two types of uncertainty: epistemic and aleatoric. Epistemic uncertainty represents the portion of total uncertainty that can be reduced by gathering more information or improving the prediction model. On the other hand, aleatoric uncertainty is the inherent and irreducible component of uncertainty due to the random nature of the data itself (Hüllermeier and Waegeman 2021; Nguyen, Shaker, and Hüllermeier 2022). The total uncertainty associated with a prediction (σ_y^2) encapsulates both the aleatoric (σ_a^2) and epistemic (σ_e^2) components; i.e., $\sigma_y^2 = \sigma_a^2 + \sigma_e^2$. Prediction intervals (PIs) offer a comprehensive representation of this total uncertainty by estimating the upper and lower bounds within which a prediction is expected to fall with a given probability (Khosravi et al. 2011).

Several methods have been proposed to reduce uncertainty through iterative sampling. However, the majority of these methods have been developed within the framework of active learning (AL) (Nguyen, Destercke, and Hüllermeier 2019; Berry and Meger 2023) or in contexts where the primary objective is to identify the location of local or global optima (Hennig and Schuler 2012; Nguyen et al. 2019).

It is important to note that AS and AL fields do not completely overlap (Di Fiore, Nardelli, and Mainini 2024). In AL, the objective is to select training data within a limited budget to maximize model performance. AL can be categorized into population-based AL, where the test input distribution is known, and pool-based AL, where a pool of unlabeled samples is provided. Our problem configuration does not align with those categories as it is not limited to predefined data pools or known distributions. Instead, it aims

to sample from an open domain continuously, focusing on reducing epistemic uncertainty across the entire input space.

We propose a method to reduce epistemic uncertainty through adaptive sampling using PIs generated by neural networks (NNs). Our method, Adaptive Sampling with Prediction-Interval Neural Networks (ASPINN), uses a dual NN architecture comprising a target-estimation network and a PI-generation network. The objective of such NNs is to produce high-quality PIs that reflect both aleatoric and epistemic uncertainties. Our specific contributions are:

1. We introduce a novel metric based on NN-generated PIs to quantify potential levels of epistemic uncertainty.
2. We present an AS method called ASPINN. At each iteration, it builds a Gaussian Process (GP) from calculated potential epistemic uncertainty levels. The GP, a surrogate for the NN models, estimates potential epistemic uncertainty changes across the domain after sampling specific locations. An acquisition function then uses the GP to select sampling locations, aiming to minimize global epistemic uncertainty throughout the input domain.
3. We tackle a real-world application and present an AS benchmark problem that focuses on reducing the epistemic uncertainty of an agricultural field site.
4. Our method is shown to converge faster to minimum epistemic uncertainty levels than the compared methods.

Related Work

The problem addressed in this work shares similarities with Bayesian Optimization (BO), where at each iteration, data points are sampled at locations expected to yield significant improvements in the objective function according to a specified acquisition function. BO methods build a probabilistic model of the objective function, often a GP, to select the most promising points for evaluation (Garnett 2023).

Traditional BO methods explore the domain space sequentially; however, Gonzalez et al. (2016) proposed a batch sampling strategy for BO that accounts for the interactions between different evaluations in the batch using a penalized acquisition function. Some BO strategies focus on maximizing information gain. For instance, Wang and Jegelka (2017) introduced an acquisition function called max-value entropy search (MES), which balances exploration of areas with higher uncertainty in the surrogate model and exploitation towards the believed optimum. In addition, Nguyen et al. (2019) presented the predictive variance reduction search (PVRS) strategy, which reduces uncertainty at perceived optimal locations, leading to convergence when uncertainty at all perceived optimal locations is minimized.

In typical BO applications, the objective is to identify a single location that corresponds to the local or global optimum of an objective function ($\arg\max f(\mathbf{x})$). In contrast, the solution to our problem consists of an augmented dataset that yields minimum epistemic uncertainty across the entire input space. In the fertilizer rate optimization problem discussed in the previous section, finding the rate that produces the higher estimated yield value does not necessarily coincide with the economic optimum nitrogen rate (EONR). The EONR is the N rate beyond which there is no actual profit for the farmers and its calculation depends on the shape of the N-response curves (Bullock and Bullock 1994). Therefore, the epistemic uncertainty across all admissible N rates should be reduced to provide reliable EONR recommendations for future growing seasons.

Similarly, active learning is closely related to this work. The primary distinction is that AL, given known input distributions (population-based AL) or a set of unlabeled points (pool-based AL), aims to select the minimum number of training examples to maximize model performance (Di Fiore, Nardelli, and Mainini 2024). In contrast, our approach is agnostic of the input distribution and is not restricted to a fixed pool of training candidates. Furthermore, our focus being on reducing uncertainty only considers model prediction improvement as a side-effect. What is more, it allows for repetitive sampling at a single location.

Despite the distinction above, some AL techniques can be adapted to our problem. In particular, we are interested in methods that decompose uncertainty into its aleatoric and epistemic components. A common approach is to use Monte-Carlo Dropout (MC-Dropout) (Gal and Ghahramani 2016) to quantify epistemic uncertainty in NNs. MC-Dropout uses dropout repeatedly to select random subsamples of active nodes in the network, turning a single network into an ensemble. Hence, epistemic uncertainty is represented by the sample variance of the ensemble predictions.

Furthermore, Valdenegro-Toro and Mori (2022) used a variance attenuation (VA) loss function to disentangle the epistemic and aleatoric components from the outputs of ensemble models. However, Zhang et al. (2024) pointed out that VA-based methods overestimate aleatoric uncertainty. In response, they presented a denoising approach that involves incorporating a variance approximation module into a trained prediction model to identify the aleatoric uncertainty. Finally, Berry and Meger (2023) proposed using an ensemble of normalizing flows (NFs), created using dropout masks, to estimate both aleatoric and epistemic uncertainty. To demonstrate their results, they suggested an AL framework that compares various uncertainty estimation methods. These methods are used to sample multiple-point candidates and select those with the highest epistemic uncertainty.

Proposed Method

In this work, we examine a system defined by an input vector $\mathbf{x} \in \mathbb{R}^d$ and a scalar response $y \in \mathbb{R}$. The system's underlying function $f : \mathcal{X} \to \mathcal{Y}$ maps the input value space and the response value space such that $y = f(\mathbf{x}) + \varepsilon_a(\mathbf{x})$, where $\varepsilon_a(\mathbf{x})$ is a random variable representing the error term that is a function of the system's aleatoric uncertainty, $\sigma_a^2(\mathbf{x})$.

Let $\mathcal{D}_t = (\mathbf{X}_{obs}^{(t)}, \mathbf{Y}_{obs}^{(t)})$ represent the dataset available at iteration t consisting of n_t observations, where $\mathbf{X}_{obs}^{(t)} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_t}\}$ and $\mathbf{Y}_{obs}^{(t)} = \{y_1, \ldots, y_{n_t}\}$. A prediction model $\hat{f}_t : \mathcal{X} \to \mathcal{Y}$ with parameters $\boldsymbol{\theta}_f$ is trained by minimizing the mean squared error of the estimation:

$$\min_{\boldsymbol{\theta}_f} \frac{1}{n_t} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_t} (\hat{f}_t(\mathbf{x}_i) - y_i)^2.$$

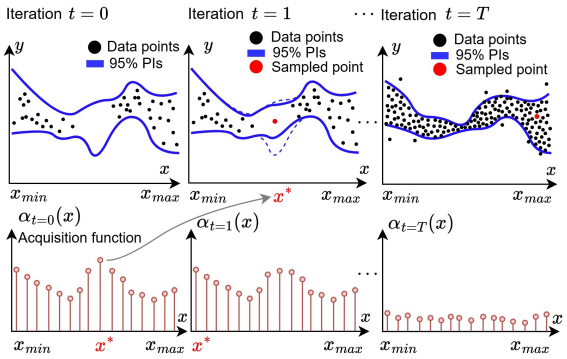

Figure 1: Epistemic uncertainty minimization through AS.

We aim to identify a batch $\mathbf{X}_{acq}^{(t)} = \{\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,B}\}$ of B recommended sampling locations for the next iteration. These locations are chosen to minimize the epistemic uncertainty across the entire input space given a model $\hat{f}_t$ trained on $\mathcal{D}_t$. The epistemic uncertainty, $\sigma_e^2(\mathbf{x}_p)$, arises from the lack of knowledge about f and is due to the limitations of the prediction model trained on the observed dataset.

Preferences over potential sampling locations are encoded by an acquisition function $\alpha_t(\mathbf{x})$. Suppose $J(\mathcal{D}_t)$ is a function that reflects the total potential epistemic uncertainty across the domain. Then $\alpha_t(\mathbf{x})$ is designed to reflect the expected decrease in epistemic uncertainty $\mathbb{E}[J(\mathcal{D}_t) - J(\mathcal{D}_t \cup (\mathbf{x}, y))]$ after making an observation at location $\mathbf{x}$. Fig. 1 depicts an instance of our problem. Here, x^* represents the selected sampling position at each iteration (i.e., $B = 1$). For the general case where $B > 1$, the decision on where to sample the k-th element of the batch, $\mathbf{x}_{t,k}$, depends on the estimated effect of the previous $k-1$ samples of the same batch. This requires a batch sampling strategy, which will be explored in this paper.

In the following, we describe the components of our AS-PINN method. We lay out the steps to derive a metric that reflects the epistemic uncertainty associated with an input value based on PIs. The metric is then used to design an acquisition function that allows for the selection of a batch of sampling locations, which are expected to minimize the global epistemic uncertainty during the next AS iteration.

Prediction Interval Generation

We generate PIs for quantifying the total uncertainty associated with a given sample, thus accounting for both aleatoric and epistemic uncertainty. We employ an NN-based PI generation method called DualAQD (Morales and Sheppard 2023b). This method uses two companion NNs: a target-estimation NN and a PI-generation NN, whose computed functions are denoted as $\hat{f}_t(\cdot)$ and $\hat{g}_t(\cdot)$, respectively. Network $\hat{f}_t(\cdot)$ is trained on $\mathcal{D}_t$ to minimize the target estimation error so that $\hat{y} = \hat{f}_t(\mathbf{x})$ and $\hat{y} \approx y$. Network $\hat{g}_t(\cdot)$ produces two outputs $[\hat{y}^\ell, \hat{y}^u] = \hat{g}_t(\mathbf{x})$, which correspond to the PI lower and upper bounds. Note that $\hat{g}_t(\mathbf{x})$ makes no assumptions about the underlying uncertainty distribution.

Network $\hat{g}_t(\cdot)$ is trained using the DualAQD loss function to produce high-quality PIs that are as narrow as possible while capturing some specified proportion of the predicted data points (e.g., 95%). However, the model should produce wider PIs for out-of-distribution (OOD) samples since these samples are not well-represented in the training set, leading to higher associated epistemic uncertainty. To address this, the bias weights of $\hat{g}_t(\cdot)$ are initialized to generate wide PIs, similar to the approach proposed by Liu et al. (2022). The rationale is that these bias weights will decrease during training for in-distribution samples, resulting in narrower PIs, but will remain high for OOD samples, ensuring appropriately wider PIs to reflect the increased uncertainty.

Potential Epistemic Uncertainty

Let $\sigma_e^2(\mathbf{x}_p)$ represent the epistemic uncertainty at a certain location $\mathbf{x}_p \in \mathcal{X}$. The PI lower and upper bounds generated by NN $\hat{g}_t(\cdot)$ at $\mathbf{x}_p$ are denoted as $\hat{y}_t^\ell(\mathbf{x}_p)$ and $\hat{y}_t^u(\mathbf{x}_p)$, respectively. We claim that using PIs alone does not provide sufficient information to determine $\sigma_e^2(\mathbf{x}_p)$. Consider $\mathbf{x}_p$ as an OOD sample. We may state that the total uncertainty associated with $\mathbf{x}_p$ is primarily due to epistemic uncertainty given the lack of knowledge of the prediction model about the system's behavior in this region of the input domain.

However, we cannot estimate the aleatoric uncertainty around $\mathbf{x}_p$ until we gather observations in such domain region. Alternative methods can be used but they require making assumptions about the noise distribution (Seitzer et al. 2022), training an ensemble of models (Berry and Meger 2023), or using additional trainable modules (Zhang et al. 2024). Therefore, the total uncertainty conveyed by the interval $[\hat{y}_t^\ell(\mathbf{x}_p), \hat{y}_t^u(\mathbf{x}_p)]$ cannot be split effectively into its epistemic and aleatoric components without further information.

Instead of attempting to provide a metric that accurately estimates $\sigma_e^2(\mathbf{x}_p)$ directly, we propose a metric that reflects the potential levels of epistemic uncertainty. Let $\mathcal{N}(\mathbf{x}_p) = \{\mathbf{x} \in \mathbf{X}_{obs}^{(t)} \,|\, \|\mathbf{x} - \mathbf{x}_p\|_2 \leq \theta\}$ denote a neighborhood that considers all samples whose Euclidian distance to $\mathbf{x}_p$ is less than a hyperparameter threshold θ. We create the set of input–response pairs $\mathcal{R}(\mathcal{N}(\mathbf{x}_p)) = \{(\mathbf{x}, y) \,|\, (\mathbf{x}, y) \in \mathcal{D}_t, \mathbf{x} \in \mathcal{N}(\mathbf{x}_p), \hat{y}^\ell(\mathbf{x}) \leq y \leq \hat{y}^u(\mathbf{x})\}$ using the samples in $\mathcal{N}(\mathbf{x}_p)$ whose response values fall within their corresponding PI. Thus, we present the metric $Q_t(\mathbf{x}_p)$, defined as:

$$Q_t(\mathbf{x}_p) = \begin{cases} \min\limits_{(\mathbf{x},y)\in\mathcal{R}(\mathcal{N}(\mathbf{x}_p))} (\hat{y}^u(\mathbf{x}) - y) + \\ \min\limits_{(\mathbf{x},y)\in\mathcal{R}(\mathcal{N}(\mathbf{x}_p))} (y - \hat{y}^\ell(\mathbf{x})) & \text{if } \mathcal{N}(\mathbf{x}_p) \neq \emptyset \\ \hat{y}_t^u(\mathbf{x}_p) - \hat{y}_t^\ell(\mathbf{x}_p) & \text{if } \mathcal{N}(\mathbf{x}_p) = \emptyset \end{cases} \quad (1)$$

The local neighborhood of $\mathbf{x}_p$ may contain important contextual information that an analysis at a single location $\mathbf{x}_p$ cannot capture. For instance, Fig.2a illustrates an interval $\text{PI}(\mathbf{x}_p) = [\hat{y}_t^\ell(\mathbf{x}_p), \hat{y}_t^u(\mathbf{x}_p)]$ generated at a single location. Suppose $Q_t(\mathbf{x}_p)$ is calculated using $\text{PI}(\mathbf{x}_p)$ only (i.e., $\theta = 0$). Since a single point lies within the interval, $Q_t(\mathbf{x}_p)$ is equal to the PI width, indicating that the epistemic uncertainty at $\mathbf{x}_p$ can potentially be completely reduced. Fig.2b depicts a case in which the PI shown in Fig.2a is located in a region of the domain with low data density. As such, there

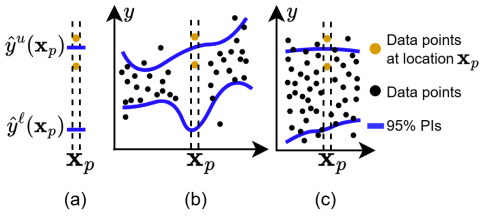

Figure 2: PIs generated at location $\mathbf{x}_p$. **(a)** Data points located at $\mathbf{x}_p$ only. **(b)** PI width is affected by epistemic uncertainty. **(b)** PI width is mainly due to aleatoric uncertainty.

exists an epistemic component that entails that the PI width could be reduced by acquiring more data in this region.

Conversely, Fig.2c shows a similar PI in a high data density context. Here, a reduction in $\text{PI}(\mathbf{x}_p)$ will also lead to a decrease in the PI widths of adjacent locations, provided that the uncertainty at $\mathbf{x}_p$ is not independent of its surroundings. However, model $\hat{g}_t(\cdot)$ is trained to produce narrow PIs while maintaining a nominal coverage (e.g., 95%). Thus, it will not reduce $\text{PI}(\mathbf{x}_p)$ if this reduction would result in several samples near the PI bounds being excluded from their intervals. Notice that if $\theta > 0$, then $Q_t(\mathbf{x}_p) \approx 0$, indicating minimal potential epistemic uncertainty around $\mathbf{x}_p$.

Batch Sampling

When multiple locations are sampled at each iteration, decisions for the entire batch are made based on the current model without observing any data from the batch until the next iteration. Hence, it is necessary to simulate the decisions that would be made under the equivalent sequential policy (i.e., when $B = 1$) (Gonzalez et al. 2016). In other words, the decision of selecting the k-th element of the t-th batch, $\mathbf{x}_{t,k}$, should incorporate the estimates of change in uncertainty after sampling at locations $\mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,k-1}$ (i.e., $\mathbf{x}_{t,1:k-1}$). Following a greedy sampling strategy, we have:

$$\mathbf{x}_{t,k} = \underset{\mathbf{x}_p \in \mathcal{X}}{\text{argmax}}\ \alpha_t(\mathbf{x}_p \,|\, \mathbf{x}_{t,1:k-1}). \tag{2}$$

We consider an acquisition function that estimates the reduction in the total potential epistemic uncertainty across the domain when making an observation at a given location $\mathbf{x}_p$:

$$\alpha_t(\mathbf{x}_p \,|\, \mathbf{x}_{t,1:k-1}) = J\left(\mathcal{D}_{t,k-1}\right) - J(\mathcal{D}_{t,k-1} \cup (\mathbf{x}_p, \hat{f}_t(\mathbf{x}_p))).$$

$\mathcal{D}_{t,k-1}$ is the dataset $\mathcal{D}_t$ augmented with the first $k-1$ samples of the batch and their corresponding estimated response values. The potential epistemic uncertainty at $\mathbf{x}$ during the t-iteration after sampling the first k elements of the batch is denoted as $Q_{t,k}(\mathbf{x})$. Thus, the total potential epistemic uncertainty is calculated as $J(\mathcal{D}_{t,k}) = \sum_{\mathbf{x} \in \mathcal{X}} Q_{t,k}(\mathbf{x})$, where $J(\mathcal{D}_{t,0}) = J(\mathcal{D}_t)$ and $Q_{t,0}(\mathbf{x}) = Q_t(\mathbf{x})$.

Thus, $J(\mathcal{D}_t)$ is computed based on $Q_t(\mathbf{x})$, which is derived from the outputs produced by NNs $\hat{f}_t(\cdot)$ and $\hat{g}_t(\cdot)$ (Eq. 1), trained on $\mathcal{D}_t$. To calculate $J(\mathcal{D}_{t,k-1} \cup (\mathbf{x}_p, \hat{f}_t(\mathbf{x}_p)))$ in a similar manner, it is necessary to train both NNs on the augmented dataset $\mathcal{D}_{t,k-1} \cup (\mathbf{x}_p, \hat{f}_t(\mathbf{x}_p))$. According to Eq. 2, this operation would need to be repeated $\forall \mathbf{x}_p \in \mathcal{X}$ and $\forall k \in [1, \ldots, B]$ and, as such, becomes impractical. Therefore, motivated by most BO-based approaches, we use a GP as a surrogate model. The objective is to simulate, with low computational cost, how the potential epistemic uncertainty would be affected throughout the entire domain after observing a sample at a given position.

Let us define a GP $p(\hat{f}_t) = \mathcal{GP}(\mu_t, \mathbf{K}_t)$ that serves as a surrogate model for $\hat{f}_t(\cdot)$ and its associated epistemic uncertainty during the t-th iteration. This GP is characterized by the mean function μ_t and the positive-definite covariance matrix $\mathbf{K}_t$. Functions μ_t and $\mathbf{K}_t$ are initialized based on the estimations generated by $\hat{f}_t(\cdot)$ and $\hat{g}_t(\cdot)$, trained on $\mathcal{D}_t$.

For the mean function, we consider $\mu_t(\mathbf{x}) = \hat{f}_t(\mathbf{x})$. On the other hand, the diagonal elements of $\mathbf{K}_t$ reflect the uncertainty in the predictions $\hat{f}_t(\mathbf{x})$ due to epistemic uncertainty. Since this uncertainty varies across the domain, it represents heteroscedastic noise. Considering that the uncertainty at a given position may be correlated with nearby positions, $\mathbf{K}_t$ is structured as a matrix with non-zero off-diagonal elements. Thus, the scale of $\mathbf{K}_t$ depends on location and is calculated according to the potential epistemic uncertainty:

$$\mathbf{K}_t(\mathbf{x}, \mathbf{x}') = \begin{cases} Q_t(\mathbf{x}), & \text{if } \mathbf{x} = \mathbf{x}' \\ \rho(\mathbf{x}, \mathbf{x}')\sqrt{Q_t(\mathbf{x})Q_t(\mathbf{x}')}, & \text{otherwise,} \end{cases}$$

where $\rho(\mathbf{x}, \mathbf{x}')$ indicates the correlation between positions $\mathbf{x}$ and $\mathbf{x}'$. We use the radial basis function (RBF) such that $\rho(\mathbf{x}, \mathbf{x}') = e^{-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2r^2}}$, where r is a tunable hyperparameter.

Given we want to assess the impact of observing a data point at a given position $\mathbf{x}_p$, we condition the GP on the data point $(\mathbf{x}_p, \hat{f}_t(\mathbf{x}_p))$, resulting in a GP posterior $p(\hat{f}_t|(\mathbf{x}_p, \hat{f}_t(\mathbf{x}_p)))$ whose covariate matrix is denoted as $\mathbf{K}_t(\mathbf{x}, \mathbf{x}' \,|\, \mathbf{x}_p)$. In general, the covariance matrix when sampling the k-th element of the batch is denoted as $\mathbf{K}_t(\mathbf{x}, \mathbf{x}' \,|\, \mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,k})$ and $Q_{t,k} = \text{diag}\left(\mathbf{K}_t(\mathbf{x}, \mathbf{x}' \,|\, \mathbf{x}_{t,1}, \ldots, \mathbf{x}_{t,k})\right)$.

Given $\mathbf{x}_p$, the covariance matrix is updated as follows:

$$\begin{aligned}\mathbf{K}_t(\mathbf{x}, \mathbf{x}' \,|\, \mathbf{x}_p) =& \mathbf{K}_t(\mathbf{x}, \mathbf{x}') - \\ & \mathbf{K}_t(\mathbf{x}, \mathbf{x}_p)\mathbf{K}_t(\mathbf{x}_p, \mathbf{x}_p)^{-1}\mathbf{K}_t(\mathbf{x}_p, \mathbf{x}').\end{aligned}$$

Hence, the updated GP variance at $\mathbf{x}_p$ collapses to zero after observing a data point at that position. Note that this would only happen when $Q_t(\mathbf{x}_p)$ reflects the level of epistemic uncertainty exclusively. In practice, this assumption may not hold. Nevertheless, it allows us to construct a heuristic that guides the search toward locations where new observations would potentially cause the greatest uncertainty reduction. The next sampling location is selected using Eq. 2 based on the total potential epistemic uncertainty after observing a data point at $\mathbf{x}_p$, which is given by:

$$J\left(\mathcal{D}_t \cup (\mathbf{x}_p, \hat{f}_t(\mathbf{x}_p))\right) = \sum \text{diag}\left(\mathbf{K}_t(\mathbf{x}, \mathbf{x}' \,|\, \mathbf{x}_p)\right).$$

Experimental Results

We compared ASPINN to three methods adapted for AS: Normalizing flows ensembles (NF-Ensemble) (Berry and

Name	Function $f(\mathbf{x})$	Noise $\varepsilon_a(\mathbf{x})$
cos	$10 + 5\cos(\mathbf{x} + 2)$	$\mathcal{N}(0, 2 + 2\cos(1.2\mathbf{x}))$
hetero	$7\sin(\mathbf{x})$	$\mathcal{N}(0, 3\cos(\mathbf{x}/2))$
cosqr	$10 + 5\cos(\frac{\mathbf{x}^2}{5})$	$\mathcal{N}(0, \frac{1}{2}(1 - \frac{\mathbf{x}^2}{100}))$

Table 1: Functions and noise terms of the 1-D problems.

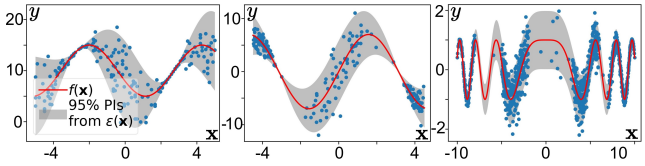

Figure 3: Initial cos, hetero, and cosqr datasets and the ideal 95% PIs calculated from $\varepsilon_a(\mathbf{x})$ across the domain.

Meger 2023), a standard GP (Gardner et al. 2018), and MC-Dropout (Gal and Ghahramani 2016). For our experiments, we considered three synthetic one-dimensional (1-D) regression problems and one multidimensional regression problem based on a real-world problem. We used synthetic problems given that, in AS, we are required to sample at locations with high uncertainty that could not have been observed previously. By utilizing problems with known underlying target and noise functions, which are unknown to the AS methods, we can simulate and evaluate accurately the performance improvements resulting from the decisions made by each method in previous iterations.

Experiments with One-Dimensional Data

We considered three 1-D problems: cos (Morales and Sheppard 2023b), hetero (Depeweg et al. 2018), and cosqr. All three problems are affected by heteroscedastic noise, and their function equations are shown in Table 1. Unlike most AL and AS approaches, we do not initiate the experiments from empty datasets. For each case, we generated incomplete datasets as initial states, as shown in Fig. 3. The motivation for this is to produce areas with low data density, which entails high epistemic uncertainty. Thus, methods that estimate potential epistemic uncertainty more accurately and select sampling locations designed to reduce such uncertainty should require fewer AS iterations to approximate the ground-truth distribution of the problem. Additional implementation details are provided in the Appendix.

For ASPINN, we trained feed-forward NNs with varying depths: two hidden layers with 100 units for problems cos and hetero; and three hidden layers with 500, 100, and 50 units, respectively, for cosqr. The networks $\hat{f}_t$ and $\hat{g}_t$ share the same architecture except for the last layer, as $\hat{f}_t$ uses one output, while $\hat{g}_t$ uses two outputs. Furthermore, ASPINN uses two hyperparameters: the neighbor distance threshold θ and the kernel length r. We performed a grid search with the values $\theta = [0.1, 0.15, 0.2, 0.25]$ and $r = [0.1, 0.15, 0.2, 0.25]$, and selected $\theta = 0.25$ and $r = 0.15$ for all experiments. DualAQD, the PI-generation method used by ASPINN, uses a hyperparameter η as a scale factor to adapt the coefficient that balances the two objectives of the DualAQD loss function. We chose a scale factor $\eta = 0.1$. Other η values (i.e., $\{0.001, 0.005, 0.01, 0.05, 0.1\}$) achieved similar results but with slower convergence rates.

For MC-Dropout, we used the same architecture as the target-estimation NN in ASPINN. For NF-Ensemble, we used flows with 200 hidden units for problems cos and hetero and 300 hidden units for problem cosqr. We employed ensembles consisting of five models trained during 30,000 epochs. For the standard GP, we used the same RBF kernel used by ASPINN. We utilized an inference implementation based on black-box matrix-matrix multiplication (Gardner et al. 2018) that uses 3000 training epochs.

Our objective is to reduce the epistemic uncertainty with as few AS iterations as possible. We define the performance metric $PI_\delta^{(t)}$ to quantify epistemic uncertainty relative to the ground truth at the t-th iteration:

$$PI_\delta^{(t)} = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \left(|y^u(\mathbf{x}) - \hat{y}_t^u(\mathbf{x})| + |y^\ell(\mathbf{x}) - \hat{y}_t^\ell(\mathbf{x})| \right).$$

Here, $y^\ell(\mathbf{x})$ and $y^u(\mathbf{x})$ represent the ideal lower and upper PI bounds, respectively, calculated from the aleatoric noise function: $y^u(\mathbf{x}) = f(\mathbf{x}) + 1.96\,\varepsilon_a(\mathbf{x})$ and $y^\ell(\mathbf{x}) = f(\mathbf{x}) - 1.96\,\varepsilon_a(\mathbf{x})$. This metric is applicable to problems with normally distributed aleatoric noise, which is the case for the problems evaluated in this work. However, none of the tested methods make assumptions about the noise distribution. Note that if $PI_\delta^{(t)} = 0$, the estimated PIs match the ideal intervals, implying that the model's epistemic uncertainty has been minimized, and the total uncertainty is purely aleatoric. A non-zero $PI_\delta^{(t)}$ indicates a discrepancy between the estimated and ideal PIs, suggesting the presence of epistemic uncertainty. The greater the $PI_\delta^{(t)}$, the higher the epistemic uncertainty. To ensure fairness, $\hat{y}_t^\ell(\mathbf{x})$ and $\hat{y}_t^u(\mathbf{x})$ are generated by an independent NN, $\hat{g}_t(\cdot)$, trained on the dataset $\mathcal{D}_t$ produced by each compared method at each iteration. Regardless of the uncertainty estimation model used by each method, we trained an additional PI-generation NN using the DualAQD loss to maintain a consistent uncertainty metric across all comparisons.

It is worth mentioning that other works have used different evaluation approaches. For instance, Berry and Meger (2023) employed an approach where they sampled 50 random locations from the domain. For each location, they generated 1000 samples using the ground-truth distribution and 1000 samples using the distribution predicted by each method. They then calculated the Kullback-Leibler divergence between the ground truth and the model-generated distributions. However, we believe this approach does not provide a consistent basis for evaluation, as each method employs different mechanisms for estimating uncertainty.

For our experiments, the AS process was executed for each problem for 50 iterations. This process is repeated 10 times, initializing the problems with a different seed each time. Figure 4 depicts an initial state of problem cos along with the augmented datasets during iterations $t = 7, 40$. The figure also displays the corresponding calculated potential epistemic uncertainty for all values of the input domain.

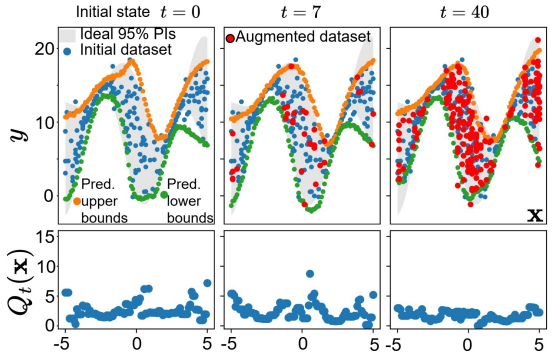

Figure 4: Example of the adaptive sampling process using ASPINN on the `cos` problem.

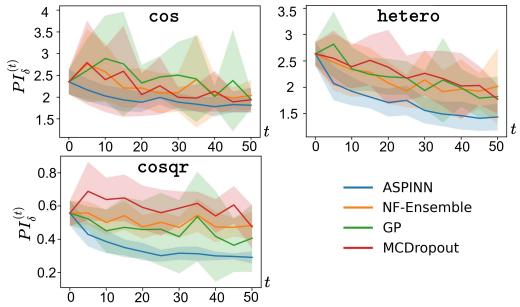

Figure 5: Evolution of the mean $PI_\delta^{(t)}$ value and its corresponding standard deviation for the 1-D problems.

Figure 5 shows the evolution of the mean $PI_\delta^{(t)}$ value and its corresponding standard deviation, calculated across the values obtained from the 10 repetitions at each t. In addition, we calculated the area under the uncertainty curve (AUUC) for each learning curve. For each problem, Table 2 gives the average AUUC for the four methods and corresponding standard deviations. The bold entries indicate the method that achieved the lowest average AUUC value and that its difference with respect to the values obtained by the other methods is statistically significant according to a paired t-test performed at the 0.05 significance level.

Experiments with Simulated Field Data

In this section, we present a multi-dimensional problem that simulates a real-world agricultural field site. A field site is defined as a specific area within a larger field (e.g., a 10×10m). It is used for precise monitoring and management to address local variations in soil and crop conditions.

Note that actual real-world data cannot be considered for a comparative AS study. There are multiple reasons for this. First, a given field site receives a single experimental rate during the fertilization stage and its effects are observed during the harvest season (e.g., five months for winter wheat). Second, additional samples at the same site require collecting data over multiple years. Third, when comparing different AS methods, they may produce different experimental rates, which cannot be implemented simultaneously in a single season. Fourth, real-world conditions, such as unforeseen environmental factors and concept drift, introduce additional complexity, making it difficult to isolate the AS strategies' effects. Therefore, simulations based on the properties of a real field provide a controlled environment where different AS methods can be evaluated under identical conditions, allowing for a fair comparison.

Problem	MCDropout	GP	NF-Ensemble	ASPINN
`cos`	112.57±24.20	123.87±26.22	113.39±19.49	**97.26±7.87**
`hetero`	113.80±13.38	110.21±13.59	106.44±16.26	**85.95±9.11**
`cosqr`	30.39±3.56	23.12±5.55	25.60±2.67	**17.13±1.42**

Table 2: AUUC comparison for the 1-D problems

In previous work, we derived the functional form of N-response curves of different management zones (MZs) from an actual winter wheat field as symbolic skeleton expressions using a Multi-Set Transformer (Morales and Sheppard 2024). An MZ is defined as a distinct sub-region that encompasses sites with relative homogeneity and, thus, similar fertilizer responsivity (i.e., similar response to varying fertilizer rates). A symbolic skeleton expression is a representation of a mathematical expression that captures its structural form without setting specific numerical values. For instance, the relationship between yield, y, and N rate, $\mathbf{x}^{Nr}$, at a given site is given by the skeleton $y = c_1 + c_2 \tanh(c_3 + c_4\, \mathbf{x}^{Nr})$, where c_1–c_4 are placeholder constants. In this work, we propose to use a simulated field site from an MZ whose underlying function is based on the previous skeleton.

In particular, we consider the following yield function:

$$y = f(\mathbf{x}) = \frac{\mathbf{x}^P}{15} + \left(\frac{\mathbf{x}^A}{\pi} + 1\right) \tanh\left(\frac{0.1\,\mathbf{x}^{Nr}}{3\mathbf{x}^{VH} + 2}\right) + \varepsilon_a(\mathbf{x}),$$

where $\mathbf{x} = [\mathbf{x}^P, \mathbf{x}^A, \mathbf{x}^{VH}, \mathbf{x}^{Nr}]$ comprises the following site-specific covariates: annual precipitation (mm), terrain aspect (radians), Sentinel-1 backscattering coefficient from the Vertical Transmit-Horizontal Receive Polarization band, and applied N rate (lbs/ac), respectively. The aleatoric noise is modeled as $\varepsilon_a(\mathbf{x}) = \mathcal{N}(0, (\mathbf{x}^P + \mathbf{x}^{Nr})/150)$. Further details on the selection of these underlying and noise functions are available in the Appendix.

While this yield regression problem considers four explanatory variables, the only one that farmers can control is $\mathbf{x}^{Nr}$. Therefore, the AS search is focused along the $\mathbf{x}^{Nr}$ axis to determine the best experimental N rate for reducing epistemic uncertainty. A field site receives a single fertilizer treatment; thus, we consider $B = 1$. The AS process was conducted over 50 iterations, with each iteration representing a different year or growing season, corresponding to a randomly generated precipitation value $\mathbf{x}^P \sim U(75, 150)$. All compared methods used the same sequence of precipitation values throughout the AS process. $\mathbf{x}^A$ describes topographic information of the field so it is assumed to remain constant throughout all iterations. In contrast, $\mathbf{x}^{VH}$, associated with soil moisture, was modeled as a function of precipitation and topographic aspect. Additional details on data generation are provided in the Appendix.

We applied the AS process ten times. At each iteration, we used a unique initialization seed and evaluated the epistemic uncertainty along the allowed N rates (i.e., 0, 30, 60, 90, 120,

MCDropout	GP	NF-Ensemble	ASPINN
614.68 ± 112.48	593.54 ± 107.42	730.80 ± 74.63	**496.85±71.65**

Table 3: AUUC comparison for the simulated field site

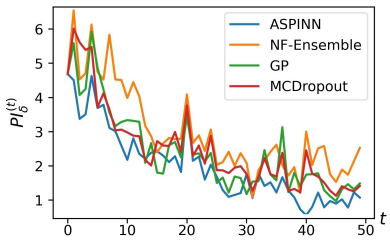

Figure 6: Evolution of the mean $PI_\delta^{(t)}$ value and its corresponding standard deviation for the simulated field site.

and 150 lbs/ac) under the current field conditions. Table 3 presents the average AUUC values and corresponding standard deviations, highlighting the best-performing method in bold. Figure 6 depicts the evolution of the mean $PI_\delta^{(t)}$ values, calculated based on the results from the ten repetitions.

Discussion

The ASPINN method involves training a PI-generation NN, which is used to design a novel potential epistemic uncertainty metric. This metric is then used in our batch sampling strategy to determine the sequence of sampling locations most likely to reduce epistemic uncertainty the greatest across the input domain.

When evaluating ASPINN on the tested 1-D problems, as shown in Fig. 5, we observed that it produced learning curves with faster convergence rates and lower standard deviation than the other methods. Although the confidence bands exhibit some overlap, this is attributed to outliers with high $PI_\delta^{(t)}$ values generated by other methods (e.g., GP), which increase the variance. Nevertheless, it is important to note that the learning curves for ASPINN consistently remain below those of the other methods across all iterations and have narrower confidence bands. Thus, the difference in AUUC values is shown to be statistically significant according to the t-test, as shown in Table 2. Also from Fig. 5, we notice that ASPINN generated constantly decreasing and smoother learning curves. Conversely, other methods, such as MC-Dropout, tend to oversample certain regions of the input domain, leading to imbalanced datasets. This oversampling results in overfitting in those regions while causing a poor fit in others, producing unstable learning curves.

Furthermore, the experiments conducted on the simulated field data exhibit consistent behavior with the results from the 1-D problems In particular, Table 3 demonstrates that ASPINN achieves the lowest AUUC values, and the differences between ASPINN and the compared methods are statistically significant. Given that the precipitation values vary at each iteration, the resulting learning curves are expected to exhibit multiple peaks and valleys rather than a smooth, consistently decreasing trend, as observed in Fig 6. This variability arises because higher precipitation values are associated with increased uncertainty levels, leading to more pronounced fluctuations in the learning curves. Considering that the sequence of precipitation values is not the same for all AS repetitions, Fig 6 reports only the mean curve and not the confidence bands. This is because the $PI_\delta^{(t)}$ values obtained by a method across different iterations are generated from contexts that could correspond to extreme opposites, leading to high variance values that do not necessarily reflect the method's performance. Despite this behavior, we observed that ASPINN consistently produced learning curves that remained below those of the compared methods.

One limitation of our approach is that it does not handle multi-modal aleatoric noise inherently. Multi-modal noise indicates that the data variability comes from different underlying sources, each contributing to a different mode in the noise distribution. In such cases, it would be necessary to use a PI-generation method capable of producing multiple upper and lower bounds based on the identified number of modes. Note, however, that the contributions proposed in this paper are not reliant on a specific PI-generation method. In the presence of multiple PIs, we would need to adapt the epistemic uncertainty metric accordingly and execute the remaining steps similarly. Another limitation, which also applies to the compared methods, is the computational cost when dealing with high-dimensional problems due to the need to evaluate all potential locations in the input space. We plan to address this limitation in future work.

Conclusion

Accurate predictive modeling is essential in many scientific and engineering disciplines, where decisions often rely on data gathered from costly and time-consuming experiments. This is especially true in fields like precision agriculture, where data collection is limited by factors such as growing seasons and crop rotation. In such contexts, reducing uncertainty in prediction models is necessary for optimizing outcomes and ensuring reliable decision-making. Addressing this challenge, our work focuses on minimizing epistemic uncertainty through adaptive sampling techniques.

We introduced ASPINN, an adaptive sampling technique designed to reduce epistemic uncertainty across an input domain using prediction intervals generated by neural networks. The novel potential epistemic uncertainty metric, central to ASPINN, provided a robust basis for guiding the sampling process. The effectiveness of our approach was demonstrated through its consistent ability to achieve faster convergence rates with lower and more stable learning curves compared to other methods. This was observed across all tested scenarios, including 1-D synthetic problems and a multi-dimensional problem that simulates an agricultural field site based on real-world winter wheat data.

In the future, we plan on adapting ASPINN for problems affected by both heteroskedastic and multi-modal noise. In particular, this would involve integrating PI-generation techniques capable of addressing multi-modal noise functions and refining the potential epistemic uncertainty metric to account for multiple PIs at a single location.

Acknowledgments

This research was supported by the Data Intensive Farm Management project (USDA-NIFA-AFRI 2016-68004-24769 and USDA-NRCS NR213A7500013G021). Computational efforts were performed on the Tempest HPC System, operated by University Information Technology Research Cyberinfrastructure at MSU.

References

Berry, L.; and Meger, D. 2023. Normalizing Flow Ensembles for Rich Aleatoric and Epistemic Uncertainty Modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6): 6806–6814.

Bullock, D. G.; and Bullock, D. S. 1994. Quadratic and Quadratic-Plus-Plateau Models for Predicting Optimal Nitrogen Rate of Corn: A Comparison. *Agronomy Journal*, 86(1): 191–195.

Depeweg, S.; Hernandez-Lobato, J.-M.; Doshi-Velez, F.; and Udluft, S. 2018. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1184–1193.

Di Fiore, F.; Nardelli, M.; and Mainini, L. 2024. Active Learning and Bayesian Optimization: A Unified Perspective to Learn with a Goal. *Archives of Computational Methods in Engineering*.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, 1050–1059. New York, USA.

Gardner, J. R.; Pleiss, G.; Bindel, D.; Weinberger, K. Q.; and Wilson, A. G. 2018. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 7587–7597. Red Hook, NY, USA.

Garnett, R. 2023. *Bayesian Optimization*. Cambridge University Press.

Gonzalez, J.; Dai, Z.; Hennig, P.; and Lawrence, N. 2016. Batch Bayesian Optimization via Local Penalization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, 648–657. Cadiz, Spain.

Hennig, P.; and Schuler, C. J. 2012. Entropy search for information-efficient global optimization. *J. Mach. Learn. Res.*, 13: 1809–1837.

Hüllermeier, E.; and Waegeman, W. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3): 457–506.

Khosravi, A.; Nahavandi, S.; Creighton, D. C.; and Atiya, A. F. 2011. Lower Upper Bound Estimation Method for Construction of Neural Network-Based Prediction Intervals. *IEEE Transactions on Neural Networks*, 22(3): 337–346.

Lawrence, P. G.; Rew, L. J.; and Maxwell, B. D. 2015. A probabilistic Bayesian framework for progressively updating site-specific recommendations. *Precision Agriculture*, 16(3): 275–296.

Liu, S.; Zhang, P.; Lu, D.; and Zhang, G. 2022. PI3NN: Out-of-distribution-aware Prediction Intervals from Three Neural Networks. In *International Conference on Learning Representations*.

Morales, G.; and Sheppard, J. W. 2023a. Counterfactual Explanations of Neural Network-Generated Response Curves. In *International Joint Conference on Neural Networks*, 01–08.

Morales, G.; and Sheppard, J. W. 2023b. Dual Accuracy-Quality-Driven Neural Network for Prediction Interval Generation. *IEEE Transactions on Neural Networks and Learning Systems*, 1–11.

Morales, G.; and Sheppard, J. W. 2024. Univariate Skeleton Prediction in Multivariate Systems Using Transformers. In *Machine Learning and Knowledge Discovery in Databases: Research Track. ECML PKDD 2024*.

Nguyen, V.; Gupta, S.; Rana, S.; Thai, M.; Li, C.; and Venkatesh, S. 2019. Efficient Bayesian Optimization for Uncertainty Reduction Over Perceived Optima Locations. In *2019 IEEE International Conference on Data Mining (ICDM)*, 1270–1275.

Nguyen, V.-L.; Destercke, S.; and Hüllermeier, E. 2019. Epistemic Uncertainty Sampling. In Kralj Novak, P.; Šmuc, T.; and Džeroski, S., eds., *Discovery Science*, 72–86. Cham: Springer International Publishing.

Nguyen, V.-L.; Shaker, M. H.; and Hüllermeier, E. 2022. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1): 89–122.

Seitzer, M.; Tavakoli, A.; Antic, D.; and Martius, G. 2022. On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks. In *International Conference on Learning Representations*.

Valdenegro-Toro, M.; and Mori, D. 2022. A Deeper Look into Aleatoric and Epistemic Uncertainty Disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1508–1516. Los Alamitos, CA, USA.

Wang, Z.; and Jegelka, S. 2017. Max-value entropy search for efficient Bayesian Optimization. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 3627–3635. JMLR.org.

Zhang, W.; Ma, Z. M.; Das, S.; Weng, T.-W. L.; Megretski, A.; Daniel, L.; and Nguyen, L. M. 2024. One Step Closer to Unbiased Aleatoric Uncertainty Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15): 16857–16864.