ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10889532

Online scalable Gaussian processes with conformal prediction for guaranteed coverage

Jinwen Xu
School of ECE
University of Georgia
Athens, GA 30602, USA
jinwen.xu@uga.edu

Qin Lu
School of ECE
University of Georgia
Athens, GA 30602, USA
qin.lu@uga.edu

Georgios B. Giannakis
Dept. of ECE
University of Minnesota
Minneapolis, MN 55455, USA
georgios@umn.edu

***Abstract*—The Gaussian process (GP) is a Bayesian nonparametric paradigm that is widely adopted for uncertainty quantification (UQ) in a number of safety-critical applications, including robotics, healthcare, as well as surveillance. The consistency of the resulting uncertainty values however, hinges on the premise that the learning function conforms to the properties specified by the GP model, such as smoothness, periodicity and more, which may not be satisfied in practice, especially with data arriving on the fly. To combat against such model mis-specification, we propose to wed the GP with the prevailing conformal prediction (CP), a distribution-free post-processing framework that produces *prediction sets* with a provably valid coverage under the sole assumption of data exchangeability. However, this assumption is usually violated in the online setting, where a prediction set is sought before revealing the true label. To ensure long-term coverage guarantee, we will adpatively set the key threshold parameter based on the feedback whether the true label falls inside the prediction set. Numerical results demonstrate the merits of the online GP-CP approach relative to existing alternatives in the long-term coverage performance.**

***Index Terms*—Uncertainty Quantification, Conformal Prediction, Online Gaussian Processes, Model Mis-specification**

I. Introduction

Uncertainty quantification (UQ) is of great importance for safety-critical applications, including robotics, healthcare, as well as surveillance. The Gaussian process (GP) is a well-established Bayesian nonparametric paradigm that provides well-calibrated uncertainty values as long as the learning function conforms to the properties specified by the GP model, such as smoothness, periodicity and more [1]. In practise though, the modeling assumptions in GPs may not be satisfied, especially with data arriving on the fly. To combat against such model mis-specification, we propose to wed the online GPs with the prevailing conformal prediction (CP) framework, a *distribution-free* post-processing approach that produces prediction sets with provably valid coverage for any pre-specified miscoverage rate as long as data are (pseudo-)exchangeable [2], [3].

Related Works. Bayesian methods are predominant for UQ in machine learning. Thanks to its sample efficiency and closed-form expression of the posterior probability density function (pdf) of the learning function, the GP is a well-received Bayesian nonparametric framework for UQ [1]. In spite of this, strong stationarity and Gaussianity assumptions in the GP limits its capability to cope with the general nonstationary functions. Alternatively, Bayesian deep learning can model arbitrary nonstationary functions owing to the representational power of deep neural networks (DNNs), but incurs high computational complexity in model training and posterior inference [4]. Also, a prior pdf has to be specified for the weights of the DNN. Aiming to combining the merits of these two frameworks, deep kernel learning leverages a DNN to obtain a feature mapping of the original input before sending it to the GP model [5]. Still, it is highly susceptible to model mis-specification.

CP, on the other hand, makes minimal statistical assumption of the learning function. As a post-processing method, CP can be coupled with any learning model and can provide valid coverage set as long as data are (pseudo)exchangeable [2], [3], [6], [7]. Specifically, CP relies on the so-termed *score* function that assesses the degree of conformality a candidate label is to its prediction, so as to determine whether the label should be assigned to the prediction set. With the score function being the negative predictive log-likelihood, conformal Bayes has also emerged that relies on the CP to adjust the prediction sets produced by Bayesian methods so as to yield the pre-determined coverage probability [8]. Recently, CP has been adapted to the GP framework [9], [10] to mitigate the strong modeling assumptions. To ensure valid coverage guarantees, the condition of data exchangeability is necessitated, which however, is easily violated with data arriving on the fly. While CP with distributional shift has been investigated [11]–[15], how to enable online conformalized GP with scalability and robustness to distributional shift has not been touched.

Contributions. Building on the aforementioned existing works, the present paper puts forth an online GP-based conformal predictor that mitigates the issue of model mis-specification in the vanilla GP. To ensure computational scalability with online continuous arrival of data, we rely on the spectral features of the GP's kernel function to obtain a linear parametric model. To further encourage valid coverage guarantees with robustness to distributional shift, the feedback

J. Xu and Q. Lu are supported by NSF CAREER #2340049; G. B. Giannakis is supported by NSF grants # 2102312, 2103256, 2126052, 2128593, 2212318, 2220292, and 2312547.

that whether the true label falls into the prediction set will be leveraged to adaptively tune the key parameter in the CP. Numerical results demonstrate that the enhanced coverage performances of the resulting online GP-CP predictor relative to the plain GP-based Bayes credible set predictor and the standard CP.

II. Preliminaries

A. Gaussian processes

A plethora of tasks in ML boil down to learning a function f that connects the d-dimensional input $\mathbf{x}$ with real-valued output y, visualized as $\mathbf{x} \to f(\mathbf{x}) \to y$. As a well-established framework to learn functions with UQ, the GP assumes a GP prior for f, denoted as $f \sim \mathcal{GP}(0, \kappa(\mathbf{x}, \mathbf{x}'))$, where $\kappa(\cdot, \cdot)$ is a positive-definite kernel measuring pairwise similarity between any two inputs, $\mathbf{x}$ and $\mathbf{x}'$. For any number of inputs $\mathbf{X}_t := [\mathbf{x}_1, \ldots, \mathbf{x}_t]$, the joint prior pdf of the function evaluations $\mathbf{f}_t := [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_t)]^\top$ is $p(\mathbf{f}_t|\mathbf{X}_t) = \mathcal{N}(\mathbf{f}_t; \mathbf{0}_t, \mathbf{K}_t)$, where $\mathbf{K}_t$ is a $t \times t$ covariance matrix with (i, j)th entry $[\mathbf{K}_t]_{ij} = \mathrm{cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)) := \kappa(\mathbf{x}_i, \mathbf{x}_j)$.

To estimate f, we rely on the labels $\mathbf{y}_t := [y_1, \ldots, y_t]^\top$ that are linked with $\mathbf{f}_t$ via the conditional likelihood $p(\mathbf{y}_t|\mathbf{f}_t, \mathbf{X}_t) = \prod_{\tau=1}^t p(y_\tau|f(\mathbf{x}_\tau))$. For regression, the per-datum likelihood is $p(y_\tau|f(\mathbf{x}_\tau)) = \mathcal{N}(y_\tau; f(\mathbf{x}_\tau), \sigma_n^2)$, which corresponds to the observation model $y_\tau = f(\mathbf{x}_\tau) + n_\tau$, where n_τ is the additive white Gaussian noise with variance σ_n^2.

Given $\mathcal{D}_t := \{\mathbf{X}_t, \mathbf{y}_t\}$, and a new test input $\mathbf{x}$, the goal is to find the predictive pdf for the test output y. Towards this, we will first write the joint pdf based on the GP prior and the Gaussian likelihood ($\mathbf{k}_t(\mathbf{x}) := [\kappa(\mathbf{x}_1, \mathbf{x}), \ldots, \kappa(\mathbf{x}_t, \mathbf{x})]^\top$)

$$\begin{bmatrix} \mathbf{y}_t \\ y \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}_{t+1}, \begin{bmatrix} \mathbf{K}_t & \mathbf{k}_t(\mathbf{x}) \\ \mathbf{k}_t^\top(\mathbf{x}) & \kappa(\mathbf{x}, \mathbf{x}) + \sigma_n^2 \end{bmatrix}\right). \quad (1)$$

Upon conditioning on $\mathcal{D}_t$, one can readily obtain the predictive pdf for y as

$$p(y|\mathcal{D}_t, \mathbf{x}) = \mathcal{N}(y;\ \hat{y}_t(\mathbf{x}), \sigma_t^2(\mathbf{x})) \quad (2)$$

where the mean and variance are given as [1]

$$\hat{y}_t(\mathbf{x}) = \mathbf{k}_t^\top(\mathbf{x})(\mathbf{K}_t + \sigma_n^2\mathbf{I}_t)^{-1}\mathbf{y}_t \quad (3a)$$

$$\sigma_t^2(\mathbf{x}) = \kappa(\mathbf{x}, \mathbf{x}) - \mathbf{k}_t^\top(\mathbf{x})(\mathbf{K}_t + \sigma_n^2\mathbf{I}_t)^{-1}\mathbf{k}_t(\mathbf{x}) + \sigma_n^2. \quad (3b)$$

Having available the predictive pdf (2) for y, one can not only obtain a *point* prediction $\hat{y}_t(\mathbf{x})$ for y, but also the *Bayes β-credible prediction set* $\mathcal{K}_t^\beta(\mathbf{x})$ that self-assesses the quality of the prediction. When $\beta = 95\%$, the Bayes credible set based on the GP model is given by $\mathcal{K}_t^\beta(\mathbf{x}) := [\hat{y}_t(\mathbf{x}) - 2\sigma_t(\mathbf{x}), \hat{y}_t(\mathbf{x}) + 2\sigma_t(\mathbf{x})]$. Notably, larger prediction sets indicates highly uncertain predictions, and the vice versa. However, the consistency of the Bayes credible set $\mathcal{K}_t^\beta(\mathbf{x})$ is contingent on the model fit. If the data do not match well with the GP assumptions, the coverage of the prediction set will not be consistent [9]. To combat against such model misspecification, we will rely on the CP framework.

B. Conformal prediction

CP, on the other hand, is a distribution-free framework [16] for UQ that is compatible with any prediction model [2], [3], [6], [17]. Suppose that we are given a prediction model $p(y|\mathcal{D}_t, \mathbf{x})$ trained on $\mathcal{D}_t$, which could be the aforementioned GP model (2) or any other (non-)Bayesian model [18], [19]. To proceed, CP will rely on a negatively-oriented *conformity* score $s_t(\mathbf{x}, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which measures how well the prediction produced by the fitted model based on $\mathcal{D}_t$ conforms with the true value of y [20]. Specifically, a *larger* score indicates significant *disagreement* of the prediction with the true label y. Upon inverting the score function, one can obtain the conformal prediction set for y as

$$\mathcal{C}_t(\mathbf{x}) = \{y \in \mathcal{Y} : s_t(\mathbf{x}, y) \leq q_t\} \quad (4)$$

where q_t is an estimated $1 - \alpha$ quantile for the distribution of the score $s_t(\mathbf{x}, y)$. In standard CP, q_t is set as $\lceil (1 - \alpha)(t + 1) \rceil$ smallest of $\{s_t(\mathbf{x}_\tau, y_\tau)\}_{\tau=1}^t$ [2]. If $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_t, y_t)$ are exchangeable, the prediction set (4) enjoys the coverage guarantee: $\mathbb{P}(y \in \mathcal{C}_t(\mathbf{x})) \geq 1 - \alpha$.

In spite of the appealing coverage guarantee, the exchangeability assumption is often violated in practise, especially in the online setting where data samples arrive sequentially. In addition, since the prediction model is trained on $\mathcal{D}_t$, the conformity score evaluated at the training sample $(\mathbf{x}_\tau, y_\tau)(\tau \leq t)$ is usually smaller than that at any test point $\{\mathbf{x}, y\}$, thus invalidating the exchangeability of the scores and further leading to undercoverage [6]. Towards addressing these issues, we will devise a mechanism to adjust q_t adaptively in the ensuing section.

III. Online scalable GP with CP

Our goal here is to develop a conformalized GP predictor that can yield prediction sets with coverage guarantees in the online setting. In this context, the GP prediction model will suffer from inscalability as inverting the $t \times t$ kernel matrix in (3) incurs the complexity of $\mathcal{O}(t^3)$, which will become prohibitively high as t grows. To effect scalability, we will first leverage a parametric GP approximant, as delineated next.

A. Scalable GP with the random feature approximation

Various attempts have been made to effect scalability in GP-based learning; see, e.g., [21]. Most existing approaches amount to summarizing the training data via a much smaller number (m) of pseudo data with inducing inputs, thereby obtaining a training-set-dependent low-rank approximant of $\mathbf{K}_t$ [22]. Targeting a low-rank approximant that is not dependent on the training set, we rely here on a standardized shift-invariant $\bar{\kappa}(\cdot)$, whose inverse Fourier transform is

$$\bar{\kappa}(\mathbf{x}, \mathbf{x}') = \bar{\kappa}(\mathbf{x} - \mathbf{x}') = \int \pi_{\bar{\kappa}}(\mathbf{v}) e^{j\mathbf{v}^\top(\mathbf{x} - \mathbf{x}')} d\mathbf{v}$$
$$:= \mathbb{E}_{\pi_{\bar{\kappa}}}\left[e^{j\mathbf{v}^\top(\mathbf{x} - \mathbf{x}')}\right] \quad (5)$$

where $\pi_{\bar{\kappa}}$ is the power spectral density (PSD), and the last equality follows after normalizing so that $\pi_{\bar{\kappa}}(\mathbf{v})$ integrates to 1, what allows one to view it as a pdf.

Upon drawing a sufficient number, say D, of independent and identically distributed (i.i.d.) samples (features) $\{\mathbf{v}_i\}_{i=1}^D$ from $\pi_{\bar{\kappa}}(\mathbf{v})$, the ensemble mean in (5) can be approximated by the sample average $\check{\bar{\kappa}}_c(\mathbf{x},\mathbf{x}') := \frac{1}{D}\sum_{i=1}^D \cos(\mathbf{v}_i^\top(\mathbf{x}-\mathbf{x}'))$.

Define the real $2D\times 1$ random feature (RF) vector as [23]

$$\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}) := \frac{1}{\sqrt{D}}\left[\sin(\mathbf{v}_1^\top\mathbf{x}),\cos(\mathbf{v}_1^\top\mathbf{x}),\ldots,\sin(\mathbf{v}_D^\top\mathbf{x}),\cos(\mathbf{v}_D^\top\mathbf{x})\right]^\top \tag{6}$$

which allows us to replace $\check{\bar{\kappa}}_c$ with $\check{\bar{\kappa}}(\mathbf{x},\mathbf{x}') = \boldsymbol{\phi}_{\mathbf{v}}^\top(\mathbf{x})\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}')$; and thus, the parametric approximant

$$\check{f}(\mathbf{x}) = \boldsymbol{\phi}_{\mathbf{v}}^\top(\mathbf{x})\boldsymbol{\theta}, \quad \boldsymbol{\theta}\sim\mathcal{N}(\mathbf{0}_{2D},\sigma_\theta^2\mathbf{I}_{2D}) \tag{7}$$

can be viewed as coming from a realization of the Gaussian $\boldsymbol{\theta}$ combined with $\boldsymbol{\phi}_{\mathbf{v}}$ to yield the GP prior with $\kappa = \sigma_\theta^2\bar{\kappa}$, where σ_θ^2 is the magnitude of κ.

With the parametric form of $\check{f}(\mathbf{x})$ in (7), the likelihood is also parametrized by $\boldsymbol{\theta}$ as $p(y_\tau|\boldsymbol{\theta},\mathbf{x}_\tau) = \mathcal{N}(y_\tau;\boldsymbol{\phi}_{\mathbf{v}}^\top(\mathbf{x}_\tau)\boldsymbol{\theta},\sigma_n^2)$. This together with the Gaussian prior of $\boldsymbol{\theta}$ (cf. (7)), yields the Gaussian posterior $p(\boldsymbol{\theta}|\mathcal{D}_t) = \mathcal{N}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_t,\boldsymbol{\Sigma}_t)$ with mean $\hat{\boldsymbol{\theta}}_t$ and covariance matrix $\boldsymbol{\Sigma}_t$, based on which we can predict y at each test input $\mathbf{x}$. In particular, this linear parametric model readily accommodates *online* operation, where prediction of y_{t+1} is due upon receiving $\mathbf{x}_{t+1}$ at the beginning of slot $t+1$, and the pdf of $\boldsymbol{\theta}$ is then updated after receiving y_{t+1} at the end of slot $t+1$.

Given $\mathbf{x}_{t+1}$, the predictive pdf for y_{t+1} is obtained as

$$p(y_{t+1}|\mathcal{D}_t,\mathbf{x}_{t+1}) = \mathcal{N}(y_{t+1};\hat{y}_{t+1|t},\sigma_{t+1|t}^2) \tag{8}$$

where the predictive mean and variance are $\hat{y}_{t+1|t} = \boldsymbol{\phi}_{\mathbf{v}}^\top(\mathbf{x}_{t+1})\hat{\boldsymbol{\theta}}_t$ and $\sigma_{t+1|t}^2 = \boldsymbol{\phi}_{\mathbf{v}}^\top(\mathbf{x}_{t+1})\boldsymbol{\Sigma}_t\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}_{t+1}) + \sigma_n^2$, respectively. Thus, one can readily obtain the following Bayes credible set for y_{t+1}

$$\mathcal{K}_t^\beta(\mathbf{x}_{t+1}) = [\hat{y}_{t+1|t} - c_\beta\sigma_{t+1|t}, \hat{y}_{t+1|t} + c_\beta\sigma_{t+1|t}] \tag{9}$$

where c_β is chosen such that $\int_{y\in\mathcal{K}_t^\beta(\mathbf{x}_{t+1})} p(y|\mathcal{D}_t,\mathbf{x}_{t+1})dy = \beta$.

Upon receiving y_{t+1}, the posterior pdf of $\boldsymbol{\theta}$ will be propagated using Bayes' rule as

$$p(\boldsymbol{\theta}|\mathcal{D}_{t+1}) = \frac{p(\boldsymbol{\theta}|\mathcal{D}_t)p(y_{t+1}|\boldsymbol{\theta},\mathbf{x}_{t+1})}{p(y_{t+1}|\mathcal{D}_t,\mathbf{x}_{t+1})} = \mathcal{N}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_{t+1},\boldsymbol{\Sigma}_{t+1}) \tag{10}$$

whose mean and covariance are updated across slots as [24]–[26]

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t + \sigma_{t+1|t}^{-2}\boldsymbol{\Sigma}_t\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}_{t+1})(y_{t+1}-\hat{y}_{t+1|t})$$
$$\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t - \sigma_{t+1|t}^{-2}\boldsymbol{\Sigma}_t\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}_{t+1})\boldsymbol{\phi}_{\mathbf{v}}^\top(\mathbf{x}_{t+1})\boldsymbol{\Sigma}_t\,.$$

Here, the per-slot complexity is only $\mathcal{O}((2D)^2)$, significantly more scalable than the vanilla GP with cubic complexity (3). Further, the resulting online scalable (OS) GP approach is *memoryless* that requires no storage of past data.

B. OS-GP with CP

In the Bayes credible set (9), the value of c_β is constant over time. In the case of model mis-specification, the set $\mathcal{K}_\beta(\mathbf{x}_{t+1})$ might have poor coverage, that is, $\mathbb{P}(y_{t+1}\in\mathcal{K}_\beta(\mathbf{x}_{t+1})) \neq \beta$. Moreover, when data exchangeability is broken, the prediction set constructed by standard CP no longer has theoretical guarantees. To mitigate such inconsistency and enable provably convergent coverage, we will wed OS-GP with CP, abbreviated as "OS-GP-CP" hereafter.

Considering the *Bayesian* nature of the GP predictor, the score function is chosen to be the negative predictive log-likelihood in (8), namely, $s_t(\mathbf{x},y) := -\log p(y|\mathcal{D}_t,\mathbf{x})$. For a new test input $\mathbf{x}_{t+1}$ at slot $t+1$, one can obtain the conformal Bayes prediction set as

$$\begin{aligned}\mathcal{C}_t(\mathbf{x}_{t+1}) &= \{y\in\mathcal{Y} : s_t(\mathbf{x}_{t+1},y)\le q_t\}\\ &= \left[\hat{y}_{t+1|t} - c_{t+1}\sigma_{t+1|t}, \hat{y}_{t+1|t} + c_{t+1}\sigma_{t+1|t}\right]\end{aligned} \tag{12}$$

where, unlike c_β in (9), $c_{t+1} = \sqrt{2q_t - \log(2\pi\sigma_{t+1|t}^2)}$ changes over time, adapting to new data samples.

Upon receiving the true label y_t, we will not only update the posterior of $\boldsymbol{\theta}$ in (10), but also the value of q_{t+1} based on the feedback that whether y_{t+1} is covered by the prediction set $\mathcal{C}_t(\mathbf{x}_{t+1})$. Specifically, q_{t+1} is updated with the rule [11]

$$q_{t+1} = q_t + \eta_t\left(\mathbb{I}(y_{t+1}\notin\mathcal{C}_t(\mathbf{x}_{t+1})) - \alpha\right) \tag{13}$$

where $\mathbb{I}(\cdot)$ is an indicator function, which takes the value of 1 (0) if the statement inside is true (false). Intuitively, if the label y_{t+1} is not covered by $\mathcal{C}_t(\mathbf{x}_{t+1})$, the value of q_t will be increased. This update rule (13) can also be interpreted as the online (sub)gradient descent on the quantile loss $\rho_{1-\alpha}(u) = (1-\alpha)\max\{u,0\} + \alpha\max\{-u,0\}$ as [11]

$$q_{t+1} = q_t - \eta_t\nabla\rho_{1-\alpha}\left(s_t(\mathbf{x}_{t+1},y_{t+1}) - q_t\right). \tag{14}$$

The learning rate η_t plays a pivotal role in the long-term coverage performance. Although a judiciously chosen constant learning rate ($\eta_t = \eta,\ \forall t$) can achieve guaranteed coverage asymptotically [11], it will yield high variability in the instantaneous coverage even when the data samples are i.i.d. [27]. To address this issue, a decaying learning rate $\eta_t \propto t^{-a}$ for some $a\in(0,1)$ can be chosen as long as the data are stationary or has slow-varying dynamics within the time horizon [27]. In the case when the data exhibit *sudden* distributional shift, the learning rate has to be reset to recover coverage more quickly. How to detect these change points is critical. As the model fit deteriorates after the change points, the size of the prediction set will become larger to reflect the higher uncertainty. Thus, we can rely on the size of the prediction set, namely, $|\mathcal{C}_t(\mathbf{x}_{t+1})| = 2c_{t+1}\sigma_{t+1|t}$ (cf. (12)), to detect change points. Specifically, we calculate the average of $|\mathcal{C}_t(\mathbf{x}_{t+1})|$ over a window with W slots. If the average prediction set size increases over r consecutive steps, we will declare a distributional shift and then reset the learning rate.

The long-run coverage of OS-GP-CP with the time-varying learning rate η_t can be guaranteed via the following theorem [27].

Theorem. *For an arbitrary sequence of data points* $(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\ldots$, *and an arbitrary positive sequence of learning rates* $\{\eta_t\}$, *if the score function* $s_t(\mathbf{x},y):\mathcal{X}\times\mathcal{Y}\to[0,B]$ *and the initial threshold* $q_0\in[0,B]$, *the long-run coverage rate of OS-GP-CP is given by*

$$\left|\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{I}(y_{t+1}\in\mathcal{C}_t(\mathbf{x}_{t+1})) - (1-\alpha)\right| \le \frac{CN_T}{T} \tag{15}$$

where $C = 2(B + \max_{0\le t\le T-1}\eta_t)/(\min_{0\le t\le T-1}\eta_t)$ *and* $N_T = \sum_{\tau=1}^{T}\mathbb{I}(\eta_\tau > \eta_{\tau-1})$ *is the number of times the learning rate is increased.*

Thus, as long as the step size does not decay too quickly, and the number of resets N_T is sublinear in T, OS-GP-CP can achieve long-run coverage convergence.

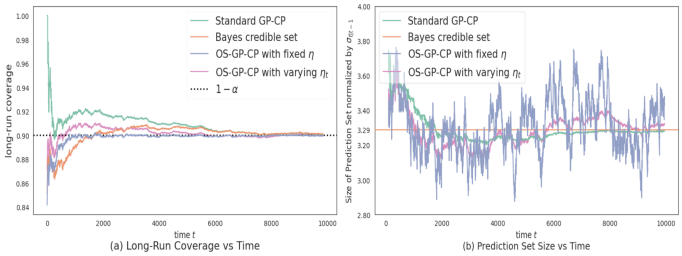

Fig. 1. Synthetic dataset without distributional shift.

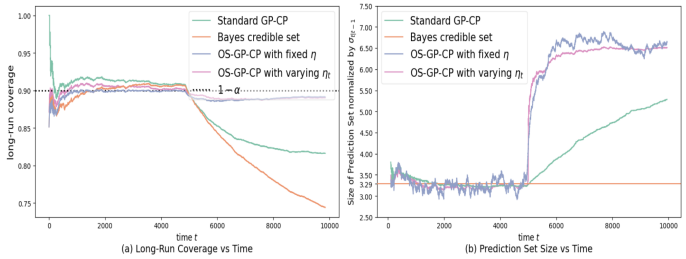

Fig. 2. Synthetic dataset with distributional shift.

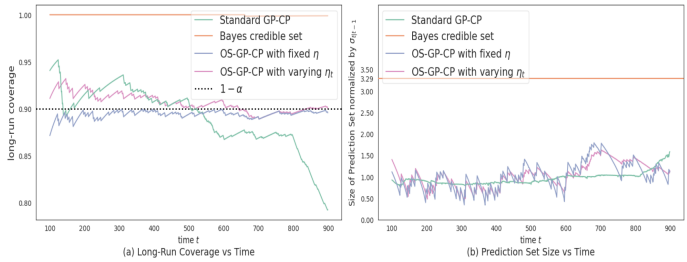

Fig. 3. Real stock price dataset.

IV. NUMERICAL EXPERIMENTS

This section assesses the coverage performance of the proposed OS-GP-CP approach with a constant $\eta = 0.05$ and time-varying η_t through numerical experiments. For the latter, we leverage on the approach in Sec. III-B with window size $W = 15$ and $r = 100$ to detect the change points, at which the learning rate is reset. During any two change points, a decaying learning rate $\eta_t = t^{-3/5}$ is adopted.

For comparison, the competing alternatives consist of the Bayes credible set (9) with $\beta = 1-\alpha$ and standard CP set (4) with q_t estimated as $1-\alpha$ quantile of past scores. For the GP model, the kernel is given by the radial basis function (RBF) $\kappa(\mathbf{x},\mathbf{x}') = \sigma_\theta^2 \exp(-\|\mathbf{x}-\mathbf{x}'\|_2^2/\sigma_l^2)$, where the kernel magnitude σ_θ^2 and characteristic lengthscale σ_l^2 are optimized together with the noise variance σ_n^2 by maximizing the marginal likelihood using the first 100 data points. In the RF approximation, the number D of spectral features is 200. The competing methods are tested on three datasets, namely, a synthetic one with i.i.d. (exchangeable) samples, a synthetic one with distributional shift, and real-world [28]. The performance metrics are the long-run average coverage up to slot t, given by $1/t\sum_{\tau=0}^{t-1}\mathbb{I}(y_{\tau+1}\in\mathcal{C}_\tau(\mathbf{x}_{\tau+1}))$, and the size of the prediction set. The value of α is set to be 0.1 in all the tests.

Synthetic data with i.i.d. samples. $10,000$ samples are generated based on $y_t = \sin(x_t) + n_t$, where the input feature $x_t \sim \mathcal{U}(0,10)$ and $n_t \sim \mathcal{N}(0,0.1^2)$. In this idea setting with i.i.d. (exchangeable) samples, the coverage of all the prediction sets converges to $1-\alpha = 0.9$ as observed in Fig. 1. While the size of standard CP set converges to that of the Bayes credible set, the size of OS-GP-CP with constant η has noticeable oscillation, corroborating the high variation in the instantaneous coverage [27]. On the other hand, OS-GP-CP with varying η_t produces prediction sets with more stable size.

Synthetic data with distributional shift. The second dataset, also consisting of $10,000$ data points, was generated using $y_t = \sin(\mathbf{x}_t) + n_t\mathbb{I}(t\le 5000) + \epsilon_t\mathbb{I}(t>5000)$, where $n_t \sim \mathcal{N}(0,0.1^2)$ and $\epsilon_t \sim \mathcal{N}(0,0.2^2)$. Apparently, there is a distributional shift at slot 5000. As shown in Fig. 2, the Bayes credible set shows a marked coverage drop after 5000 points due to model mis-specification. Similarly, the coverage for standard CP also falls below 0.9 as it struggles with non-exchangeable data. OS-GP-CP with both fixed and varying step sizes quickly recover coverage around 0.9. Note that in the latter, a change point is detected and the learning rate is reset at slot 5003, validating the efficacy of the change point detection mechanism in Sec.III-B.

Real stock price time-series data. This dataset tracks the closing stock price (y_t) of Apple Inc. (AAPL) using its opening, high, and low prices, collected in $\mathbf{x}_t$, from January 4, 2016, to July 31, 2019 [28]. Here, continuous distribution changes over time can be observed in the data samples. While the traditional Bayes credible set achieves full coverage as shown in Fig. 3, it did so by excessively enlarging the prediction set. Standard CP also struggles with the evolving data distribution, leading to a drop in coverage below 0.8. By contrast, our proposed OS-GP-CP with both constant and time-varying learning rates yield a more modest prediction set size with average coverage at 0.9.

V. CONCLUSIONS

This work relied on the CP framework to allow for prediction sets with valid coverage guarantees to be constructed by the GP predictor. To effect scalability with the online arrival of data, the RF-based parametric OS-GP model was adopted that has constant model update per slot. To further combat against distributional shift, we leveraged the feedback whether the true label resides in the prediction set to adaptively adjust the key threshold parameter when constructing the prediction set. Numerical results showcased that the resulting OS-GP-CP approach has improved coverage performance relative to the competing baselines under the distributional shift.

REFERENCES

[1] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

[2] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Springer, 2005, vol. 29.

[3] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018.

[4] A. G. Wilson and P. Izmailov, "Bayesian deep learning and a probabilistic perspective of generalization," *Advances in neural information processing systems*, vol. 33, pp. 4697–4708, 2020.

[5] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, "Deep kernel learning," *Proc. Int. Conf. Artif. Intel. and Stats.*, pp. 370–378, 2016.

[6] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Foundations and Trends® in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023.

[7] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, pp. 371–421, jun 2008.

[8] E. Fong and C. C. Holmes, "Conformal Bayesian computation," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 18 268–18 279, 2021.

[9] S. Stanton, W. Maddox, and A. G. Wilson, "Bayesian optimization with conformal prediction sets," *Proc. Int. Conf. Artif. Intel. and Stats.*, pp. 959–986, 2023.

[10] H. Papadopoulos, "Guaranteed coverage prediction intervals with Gaussian process regression," *IEEE Trans. Pattern Anal. Mach. Intel.*, 2024.

[11] I. Gibbs and E. Candes, "Adaptive conformal inference under distribution shift," *Advances in Neural Information Processing Systems*, 2021.

[12] I. Gibbs and E. Candès, "Conformal inference for online prediction with arbitrary distribution shifts," *Journal of Machine Learning Research*, vol. 25, pp. 1–36, 2024, submitted 10/22; Revised 5/24; Published 5/24.

[13] A. Bhatnagar, H. Wang, C. Xiong, and Y. Bai, "Improved online conformal prediction via strongly adaptive online learning," *International Conference on Machine Learning*, vol. 2023, pp. 2337–2363, 2023.

[14] M. Zaffran, O. Féron, Y. Goude, J. Josse, and A. Dieuleveut, "Adaptive conformal predictions for time series," *International Conference on Machine Learning*, vol. 2022, pp. 25 834–25 866, 2022.

[15] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *The Annals of Statistics*, vol. 51, no. 2, pp. 816–845, 2023.

[16] W. Chen, K.-J. Chun, and R. F. Barber, "Discretized conformal prediction for efficient distribution-free inference," *Stat*, vol. 7, no. 1, p. e173, 2018.

[17] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," *The Twelfth International Conference on Learning Representations*, 2024.

[18] A. N. Angelopoulos, S. Bates, M. Jordan, and J. Malik, "Uncertainty sets for image classifiers using conformal prediction," *International Conference on Learning Representations*, 2021.

[19] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," *The Annals of Statistics*, vol. 49, no. 1, 2021.

[20] H. Papadopoulos, A. Gammerman, and V. Vovk, "Normalized nonconformity measures for regression conformal prediction," *Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications*, pp. 64–69, 2008.

[21] H. Liu, Y.-S. Ong, X. Shen, and J. Cai, "When Gaussian process meets big data: A review of scalable GPs," *IEEE Trans. Neural Net. and Learn. Syst.*, vol. 31, no. 11, pp. 4405–4423, 2020.

[22] J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.

[23] M. Lázaro-Gredilla, J. Quiñonero Candela, C. E. Rasmussen, and A. Figueiras-Vidal, "Sparse spectrum Gaussian process regression," *J. Mach. Learn. Res.*, vol. 11, no. Jun, pp. 1865–1881, 2010.

[24] Q. Lu, G. V. Karanikolas, and G. B. Giannakis, "Incremental ensemble Gaussian processes," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 45, no. 2, pp. 1876–1893, 2022.

[25] Q. Lu, K. D. Polyzos, B. Li, and G. B. Giannakis, "Surrogate modeling for Bayesian optimization beyond a single Gaussian process," *IEEE Trans. Pattern Anal. Mach. Intel.*, 2023.

[26] Q. Lu, G. Karanikolas, Y. Shen, and G. B. Giannakis, "Ensemble Gaussian processes with spectral features for online interactive learning with scalability," *Proc. Int. Conf. Artif. Intel. and Stats.*, pp. 1910–1920, 2020.

[27] A. N. Angelopoulos, R. Barber, and S. Bates, "Online conformal prediction with decaying step sizes," *Forty-first International Conference on Machine Learning*, 2024.

[28] Ran Aroussi, "yfinance: Yahoo! Finance market data downloader," 2024, python library used for data retrieval. [Online]. Available: https://github.com/ranaroussi/yfinance