Improved Online Conformal Prediction via Strongly Adaptive Online Learning

Aadyot Bhatnagar [1] **Huan Wang** [1] **Caiming Xiong** [1] **Yu Bai** [1]

Abstract

We study the problem of uncertainty quantification via prediction sets, in an online setting where the data distribution may vary arbitrarily over time. Recent work develops *online conformal prediction* techniques that leverage regret minimization algorithms from the online learning literature to learn prediction sets with approximately valid coverage and small regret. However, standard regret minimization could be insufficient for handling changing environments, where performance guarantees may be desired not only over the full time horizon but also in all (sub-)intervals of time. We develop new online conformal prediction methods that minimize the *strongly adaptive regret*, which measures the worst-case regret over all intervals of a fixed length. We prove that our methods achieve near-optimal strongly adaptive regret for all interval lengths simultaneously, and approximately valid coverage. Experiments show that our methods consistently obtain better coverage and smaller prediction sets than existing methods on real-world tasks, such as time series forecasting and image classification under distribution shift.

1. Introduction

Modern machine learning models make highly accurate predictions in many settings. In high stakes decision-making tasks, it is just as important to estimate the model's uncertainty by quantifying how much the true label may deviate from the model's prediction. A common approach for uncertainty quantification is to learn *prediction sets* that associate each input with a set of candidate labels, such as prediction intervals for regression, and label sets for classification. The most important requirement for learned prediction sets is to achieve valid *coverage*, i.e. they should cover the true label with at least $1-\alpha$ (such as 90%) probability.

Conformal prediction (Vovk et al., 2005) is a powerful framework for augmenting any *base predictor* (such as a pretrained model) into prediction sets with valid coverage guarantees (Angelopoulos & Bates, 2021). These guarantees require almost no assumptions on the data distribution, except *exchangeability* (i.i.d. data is a sufficient condition). However, exchangeability fails to hold in many real-world settings such as time series data (Chernozhukov et al., 2018) or data corruption (Hendrycks et al., 2018), where the data may exhibit *distribution shift*. Various approaches have been proposed to handle such distribution shift, such as reweighting (Tibshirani et al., 2019; Barber et al., 2022) or distributionally robust optimization (Cauchois et al., 2022).

A recent line of work develops *online conformal prediction* methods for the setting where the data arrives in a sequential order (Gibbs & Candès, 2021; 2022; Zaffran et al., 2022; Feldman et al., 2022). At each step, their methods output a prediction set parameterized by a single *radius* parameter that controls the size of the set. After receiving the true label, they adjust this parameter adaptively via *regret minimization* techniques—such as Online Gradient Descent (OGD) (Zinkevich, 2003)—on a certain quantile loss over the radius. These methods are shown to achieve empirical coverage frequency close to $1-\alpha$, regardless of the data distribution (Gibbs & Candès, 2021). In addition to coverage, importantly, these methods achieve sublinear regret with respect to the quantile loss (Gibbs & Candès, 2022). Such regret guarantees ensure that the size of the prediction set is reasonable, and rule out "trivial" algorithms that achieve valid coverage by alternating between predicting the empty set and full set (cf. Section 2 for a discussion).

While regret minimization techniques achieve coverage and regret guarantees, they may fall short in more dynamic environments where we desire a strong performance not just over the entire time horizon (as captured by the regret), but also within every *sub-interval* of time. For example, if the data distribution shifts abruptly for a few times, we rather desire strong performance within each contiguous interval between two consecutive shifts, in addition to the entire

[1]Salesforce AI Research, Palo Alto, CA, USA. Correspondence to: Aadyot Bhatnagar <aadyotb@gmail.com>, Yu Bai <yu.bai@salesforce.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

The code for our experiments can be found at https://github.com/salesforce/online_conformal.

horizon. Gibbs & Candès (2022) address this issue partially by proposing the Fully Adaptive Conformal Inference (FACI) algorithm, a meta-algorithm that aggregates multiple *experts* (base learners) that are OGD instances with different learning rates. However, their algorithm may not be best suited for achieving such interval-based guarantees, as each expert still runs over the full time horizon and is not really localized. This is also reflected in the fact that FACI achieves a near-optimal $\widetilde{\mathcal{O}}(\sqrt{k})$ regret within intervals of a fixed length k, but is unable to achieve this over all lengths $k \in [T]$ simultaneously.

In this paper, we design improved online conformal prediction algorithms by leveraging *strongly adaptive regret minimization*, a technique for attaining strong performance on all sub-intervals simultaneously in online learning (Daniely et al., 2015; Jun et al., 2017). Our proposed algorithm, Strongly Adaptive Online Conformal Prediction (SAOCP), is a new meta-algorithm that manages multiple experts, with the key difference that each expert now only operates on its own *active interval*. We summarize our contributions:

- We propose SAOCP, a new algorithm for online conformal prediction. SAOCP is a meta-algorithm that maintains multiple experts each with its own active interval, building on strongly adaptive regret minimization techniques (Section 3). We instantiate the experts as Scale-Free OGD (SF-OGD), an anytime variant of the OGD, which we also study as an independent algorithm.
- We prove that SAOCP achieves a near-optimal strongly adaptive regret of $\widetilde{\mathcal{O}}(\sqrt{k})$ regret over all intervals of length k simultaneously, and that both SAOCP and SF-OGD achieve approximately valid coverage (Section 4).
- We show experimentally that SAOCP and SF-OGD attain better coverage in localized windows and smaller prediction sets than existing methods, on two real-world tasks: time series forecasting and image classification under distribution shift (Section 5).

1.1. Related work

Conformal prediction The original idea of conformal prediction (utilizing exchangeable data) is developed in the early work of Vovk et al. (1999; 2005); Shafer & Vovk (2008). Learning prediction sets via conformal prediction has since been adopted as a major approach for uncertainty quantification in regression (Papadopoulos, 2008; Vovk, 2012; Lei & Wasserman, 2014; Vovk et al., 2018; Romano et al., 2019; Gupta et al., 2019; Barber et al., 2021; 2022) and classification (Lei et al., 2013; Romano et al., 2020; Cauchois et al., 2020; 2022; Angelopoulos et al., 2021b), with further applications in general risk control (Bates et al., 2021; Angelopoulos et al., 2021a; 2022a), biological imaging (Angelopoulos et al., 2022b), and protein design (Fannjiang et al., 2022), to name a few.

Recent work also proposes to optimize the prediction sets' *efficiency* (e.g. width or cardinality) in addition to coverage (Pearce et al., 2018; Park et al., 2020; Yang & Kuchibhotla, 2021; Stutz et al., 2022; Angelopoulos et al., 2021a;b; Bai et al., 2022). The regret that we consider can be viewed as a (surrogate) measure for efficiency in the online setting.

Conformal prediction under distribution shift For the more challenging case where data may exhibit distribution shift (and thus are no longer exchangeable), several approaches are proposed to achieve approximately valid coverage, such as reweighting (using prior knowledge about the data's dependency structure) (Tibshirani et al., 2019; Podkopaev & Ramdas, 2021; Candès et al., 2021; Barber et al., 2022), distributionally robust optimization (Cauchois et al., 2020), or doubly robust calibration (Yang et al., 2022).

Our work makes an addition to the online conformal prediction line of work (Gibbs & Candès, 2021; 2022; Zaffran et al., 2022; Feldman et al., 2022), which uses regret minimization techniques from the online learning literature (Zinkevich, 2003; Hazan, 2022) to adaptively adjust the size of the prediction set based on recent observations. Closely related to our work is the FACI algorithm of Gibbs & Candès (2022), which is a meta-algorithm that uses multiple experts for handling changing environments. Our meta-algorithm SAOCP differs in style from theirs, in that our experts only operate on their own active intervals, and it achieves a better guarantee on the strongly adaptive regret.

A related line of work studies conformal prediction for time series data. Chernozhukov et al. (2018); Xu & Xie (2021); Sousa et al. (2022) use randomization and ensembles to produce valid prediction sets for time series that are ergodic in a certain sense. Some other works directly apply vanilla conformal prediction to time series either without theoretical guarantees or requiring weaker notions of exchangeability (Dashevskiy & Luo, 2008; Wisniewski et al., 2020; Kath & Ziel, 2021; Stankeviciute et al., 2021; Sun & Yu, 2022).

Strongly adaptive online learning Our algorithms adapt techniques from the online learning literature, notably strongly adaptive regret minimization (Daniely et al., 2015; Jun et al., 2017; Zhang et al., 2018) and scale-free algorithms for achieving other kinds of adaptive (e.g. anytime) regret guarantees (Orabona & Pál, 2018).

2. Preliminaries

We consider standard learning problems in which we observe examples $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and wish to predict a label y from input x. A *prediction set* $C : \mathcal{X} \to 2^{\mathcal{Y}}$ is a set-valued function that maps any input x to a *set* of predicted labels $C(x) \subset \mathcal{Y}$. Two prevalent examples are *prediction intervals*

for regression in which $\mathcal{Y} = \mathbb{R}$ and $C(x)$ is an interval, and *label prediction sets* for (m-class) classification in which $\mathcal{Y} = [m]$ and $C(x)$ is a subset of $[m]$. Prediction sets are a popular approach to quantify the uncertainty associated with the point prediction $\widehat{y} = f(x)$ of a black box model.

We study the problem of learning prediction sets in the *online* setting, in which the data $(X_1, Y_1), \dots, (X_T, Y_T)$ arrive sequentially. At each time step $t \in [T]$, we output a prediction set $\widehat{C}_t = \widehat{C}_t(X_t)$ based on the current input X_t and past observations $\{(X_i, Y_i)\}_{i \le t-1}$, *before* observing the true label Y_t. The primary goal of the prediction set is to achieve *valid coverage*: $\mathbb{P}[Y_t \in \widehat{C}_t(X_t)] = 1 - \alpha$, where $1-\alpha \in (0,1)$ is the target coverage level pre-determined by the user. Standard choices for α include $\{0.1, 0.05\}$, which correspond to $\{90\%, 95\%\}$ target coverage respectively.

Throughout this paper, we use the standard notation $\mathcal{O}(\cdot)$ to suppress absolute constants, and $\widetilde{\mathcal{O}}(\cdot)$ to suppress absolute constants and polylogarithmic factors, e.g. $\log(T)$.

Online conformal prediction We now review the idea of *online conformal prediction*, initiated by Gibbs & Candès (2021; 2022). This framework for learning prediction sets in the online setting achieves coverage guarantees even under distribution shift.

At each time $t \in [T]$, online conformal prediction assumes that we have a family of prediction sets $\mathcal{C}_t = \{\widehat{C}_t(x,s)\}_{x \in \mathcal{X}, s \in \mathbb{R}}$ specified by a *radius* parameter $s \in \mathbb{R}$, and we need to predict $\widehat{s}_t \in \mathbb{R}$ and output prediction set $\widehat{C}_t = \widehat{C}_t(X_t, \widehat{s}_t) \subset \mathcal{Y}$. The family $(\mathcal{C}_t)_{t \in [T]}$ is typically defined through *base predictors* $\widehat{f}_t$ (for example, $\widehat{f}_t \equiv f$ can be a fixed pretrained model). A standard example in regression is that we have a base predictor $\widehat{f}_t : \mathcal{X} \to \mathbb{R}$, and we can choose $\widehat{C}_t(X_t, s) := [\widehat{f}_t(X_t) - s, \widehat{f}_t(X_t) + s]$ to be a prediction interval around $\widehat{f}_t(X_t)$, in which case the radius s is the (half) width of the interval. In general, we allow any $\mathcal{C}_t$ that are *nested sets* (Gupta et al., 2019) in the sense that $\widehat{C}_t(x,s) \subseteq \widehat{C}_t(x,s')$ for all $x \in \mathcal{X}$ and $s \le s'$, so that a larger radius always yields a larger set.

Online conformal prediction adopts online learning techniques to learn $\widehat{s}_t$ based on past observations. Defining the "true radius" $S_t := \inf\{s \in \mathbb{R} : Y_t \in \widehat{C}_t(X_t, s)\}$ (i.e. the smallest radius s such that $\widehat{C}_t$ covers Y_t), we consider the $(1-\alpha)$-*quantile loss* (aka *pinball loss* (Koenker & Bassett Jr, 1978)) between S_t and any predicted radius $\widehat{s}$:

$$\begin{aligned} \ell^{(t)}(\widehat{s}) &= \ell_{1-\alpha}(S_t, \widehat{s}) \\ &:= \max\{(1-\alpha)(S_t - \widehat{s}), \alpha(\widehat{s} - S_t)\}. \end{aligned} \tag{1}$$

Throughout the rest of this paper, we assume that all true radii are bounded: $S_t \in [0, D]$ almost surely for all $t \in [T]$.

After observing X_t, predicting the radius $\widehat{s}_t$, and observing the label Y_t (and hence S_t), the gradient[1] $\nabla \ell^{(t)}(\widehat{s}_t)$ can be evaluated and has the following simple form:

$$\nabla \ell^{(t)}(\widehat{s}_t) = \alpha - \mathbb{1}[\widehat{s}_t < S_t] = \alpha - \underbrace{\mathbb{1}[Y_t \notin \widehat{C}_t]}_{:= \mathrm{err}_t}, \tag{2}$$

where err_t is the indicator of miscoverage at time t ($\mathrm{err}_t = 1$ if $\widehat{C}_t$ did not cover Y_t). Gibbs & Candès (2021) perform an Online Gradient Descent (OGD) step to obtain $\widehat{s}_{t+1}$:

$$\widehat{s}_{t+1} = \widehat{s}_t - \eta \nabla \ell^{(t)}(\widehat{s}_t) = \widehat{s}_t + \eta(\mathrm{err}_t - \alpha), \tag{3}$$

where $\eta > 0$ is a learning rate, and the algorithm is initialized at some $\widehat{s}_1 \in \mathbb{R}$. Update (3) increases the predicted radius if $\widehat{C}_t$ did not cover Y_t ($\mathrm{err}_t = 1$), and decreases the radius otherwise. This makes intuitive sense as an approach for adapting the radius to recent observations.

Adaptive Conformal Inference (ACI) The ACI algorithm of Gibbs & Candès (2021) uses update (3) in conjunction with a specific choice of $\mathcal{C}_t = \mathcal{C}_t^{\mathsf{ACI}}$, where

$$\widehat{C}_t^{\mathsf{ACI}}(X_t, \widehat{s}_t) = \left\{ y : \widetilde{S}_t(X_t, y) \le Q_{\widehat{s}_t}\left(\{\widetilde{S}_\tau\}_{\tau=1}^{t-1} \right) \right\}, \tag{4}$$

where $\widetilde{S}_t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is any function (termed the *score* function), $\widetilde{S}_\tau := \widetilde{S}_\tau(X_\tau, y_\tau)$ denotes score of the τ-th observation, and $Q_s(\cdot)$ denotes the $s \in (0,1)$-th empirical quantile of a set (cf. (32)). In words, ACI's confidence set contains all y whose score $\widetilde{S}_t(X_t, y)$ is below the $\widehat{s}_t$-th quantile of the past scores, and they use (3) to learn this quantile level. The framework presented here generalizes the ACI algorithm where we allow any choice of $\mathcal{C}_t$ that are nested sets, including but not limited to (4). For convenience of discussions, unless explicitly specified, we will also refer to this general version of (3) as ACI throughout this paper.

Empirically, we show in Section 5 that FACI (Gibbs & Candès, 2022) (an extension of ACI) performs better when trained to predict S_t directly under our more general parameterization, i.e. $\widehat{C}(X_t, \widehat{s}_t) = \{y : \widetilde{S}_t(X_t, y) \le \widehat{s}_t\}$.

Coverage and regret guarantees Gibbs & Candès (2021) show that ACI[2] achieves approximately valid (empirical) coverage in the sense that the empirical miscoverage frequency is close to the target level α:

$$\mathrm{CovErr}(T) := \left| \tfrac{1}{T} \textstyle\sum_{t=1}^T \mathrm{err}_t - \alpha \right| \le \tfrac{D+\eta}{\eta T}. \tag{5}$$

In addition to coverage, by standard online learning analyses, ACI achieves a *regret* bound on the quantile losses $\{\ell^{(t)}\}_{t \in [T]}$: we have $\mathrm{Reg}(T) \le \mathcal{O}(D^2/\eta + \eta T) \le \mathcal{O}(D\sqrt{T})$ (with optimally chosen η), where

$$\mathrm{Reg}(T) := \textstyle\sum_{t=1}^T \ell^{(t)}(\widehat{s}_t) - \inf_{s^\star \in \mathbb{R}} \sum_{t=1}^T \ell^{(t)}(s^\star). \tag{6}$$

[1] More precisely, $\nabla \ell^{(t)}(\widehat{s}_t)$ is a subgradient.

[2] Their results are established on the specific choice of $\mathcal{C}_t$ in (4), but can be extended directly to any $\mathcal{C}_t$ that are nested sets.

Algorithm 1 Strongly Adaptive Online Conformal Prediction (SAOCP), adapted from Jun et al. (2017).

Input : Target coverage $1-\alpha \in (0,1)$; maximum possible radius $D > 0$

for $t = 1, \dots, T$ **do**

  `// Obtain prediction interval by aggregating active experts`

  Initialize new expert $\mathcal{A}_t = \texttt{SF-OGD}\big(\alpha \leftarrow \alpha;\ \eta \leftarrow D/\sqrt{3};\ \widehat{s}_1 \leftarrow \widehat{s}_{t-1}\big)$ (Algorithm 2), and set weight $w_{t,t} = 0$

  Compute active set $\mathrm{Active}(t) = \{i \in [T] : t - L(i) < i \le t\}$, where $L(i)$ is defined in (8)

  Compute prior probability $\pi_i \propto i^{-2}(1 + \lfloor \log_2 i \rfloor)^{-1}\mathbb{1}[i \in \mathrm{Active}(t)]$

  Compute un-normalized probability $\widehat{p}_i = \pi_i [w_{t,i}]_+$ for all $i \in [t]$

  Normalize $p = \widehat{p}/\left\|\widehat{p}\right\|_1 \in \Delta^t$ if $\left\|\widehat{p}\right\|_1 > 0$, else $p = \pi$

  Compute predicted radius $\widehat{s}_t = \sum_{i \in \mathrm{Active}(t)} p_i \widehat{s}_{i,t}$ (for $t \ge 2$), and $\widehat{s}_t = 0$ for $t = 1$

  Observe input $X_t \in \mathcal{X}$ and **return** prediction set $\widehat{C}_t(X_t, \widehat{s}_t)$

  `// Use meta loss and per-expert losses to update experts`

  Observe true label $Y_t \in \mathcal{Y}$, compute true radius $S_t = \inf\{s \in \mathbb{R} : Y_t \in \widehat{C}_t(X_t, s)\}$ and loss function $\ell^{(t)}(\cdot) = \ell_{1-\alpha}(S_t, \cdot)$

  for $i \in \mathrm{Active}(t)$ **do**

    Update expert $\mathcal{A}_i$ with (X_t, Y_t) and obtain next predicted radius $\widehat{s}_{i,t+1}$

    Compute $g_{i,t} = \begin{cases} \frac{1}{D}\left(\ell^{(t)}(\widehat{s}_t) - \ell^{(t)}(\widehat{s}_{i,t})\right) & w_{i,t} > 0 \\ \frac{1}{D}\left[\ell^{(t)}(\widehat{s}_t) - \ell^{(t)}(\widehat{s}_{i,t})\right]_+ & w_{i,t} \le 0 \end{cases}$

    Update expert weight $w_{i,t+1} = \frac{1}{t-i+1}\left(\sum_{j=i}^{t} g_{i,j}\right)\left(1 + \sum_{j=i}^{t} w_{i,j} g_{i,j}\right)$

One advantage of the regret as an additional performance measure alongside coverage is that it rules out certain algorithms that achieve good coverage in a "trivial" fashion and are not useful in practice — for example, $\widehat{C}_t$ may simply alternate between the empty set and the full set $\{\alpha, 1-\alpha\}$ proportion of the time, which satisfies the coverage bound (5) on arbitrary data distributions, but suffers from linear regret even on certain simple data distributions (cf. Appendix A.2).

We remark that the regret has a connection to the coverage error in that $\mathrm{CovErr}(T) = |\frac{1}{T}\sum_{t=1}^{T} \nabla \ell^{(t)}(\widehat{s}_t)|$, i.e. the coverage error is equal to the average gradient (derivative) of the losses. However, without further distributional assumptions, regret bounds and coverage bounds do not imply each other in a black-box fashion (see e.g. Gibbs & Candès (2021, Appendix A)) and need to be established separately for each algorithm.

3. Strongly Adaptive Online Conformal Prediction

Our approach is motivated from the observation that regret minimization is in a sense limited, as the regret measures performance over the *entire time horizon* $[T]$, which may be insufficient when the algorithm encounters changing environments. For example, if $S_t = 1$ for $1 \le t \le T/2$ and $S_t = 100$ for $T/2 < t \le T$, then achieving small regret on all *(sub)-intervals* of size $T/2$ is a much stronger guarantee than achieving small regret over $[T]$. For this reason, we seek *localized* guarantees over all intervals simultaneously, to prevent worst-case scenarios such as significant miscoverage or high radius within a specific interval.

The *Strongly Adaptive Regret* (SARegret) (Daniely et al., 2015; Zhang et al., 2018) has been proposed in the online learning literature as a generalization of the regret that captures the performance of online learning algorithms over all intervals simultaneously. Concretely, for any $k \in [T]$, the SARegret of length k of any algorithm is defined as

$$\mathrm{SAReg}(T, k) := \max_{[\tau, \tau+k-1] \subseteq [T]} \left(\sum_{t=\tau}^{\tau+k-1} \ell^{(t)}(\widehat{s}_t) - \inf_{s^\star} \sum_{t=\tau}^{\tau+k-1} \ell^{(t)}(s^\star) \right) \quad (7)$$

$\mathrm{SAReg}(T, k)$ measures the maximum regret over all intervals of length k, which reduces to the usual regret at $k = T$, but may in addition be smaller for smaller k.

Algorithm: SAOCP We leverage techniques for minimizing the strongly adaptive regret to perform online conformal prediction. Our main algorithm, Strongly Adaptive Online Conformal Prediction (SAOCP, described in Algorithm 1), adapts the work of Jun et al. (2017) to the online conformal prediction setting. At a high level, SAOCP is a meta-algorithm that manages multiple *experts*, where each expert is itself an arbitrary online learning algorithm taking charge of its own *active interval* that has a finite *lifetime*. At each $t \in [T]$, Algorithm 1 instantiates a new expert $\mathcal{A}_t$ with active interval $[t, t + L(t) - 1]$, where $L(t)$ is its lifetime:

$$L(t) := g \cdot \max_{n \in \mathbb{Z}} \{2^n : t \equiv 0 \bmod 2^n\}, \quad (8)$$

and $g \in \mathbb{Z}_{\ge 1}$ is a multiplier for the lifetime of each expert. It is straightforward to see that at most $g\lfloor \log_2 t \rfloor$ experts are active at any time t under choice (8), granting Algorithm 1 a total runtime of $\mathcal{O}(T \log T)$ for any $g = \Theta(1)$. Then, at any time t, the predicted radius $\widehat{s}_t$ is obtained by aggregating

Algorithm 2 Scale-Free Online Gradient Descent (SF-OGD), adapted from Orabona & Pál (2018).

Input : $\alpha \in (0,1)$, learning rate $\eta > 0$, init. $\widehat{s}_1 \in \mathbb{R}$
for $t \geq 1$ **do**
  Observe input $X_t \in \mathcal{X}$
  Return prediction set $\widehat{C}_t(X_t, \widehat{s}_t)$
  Observe true label $Y_t \in \mathcal{Y}$ and compute true radius $S_t = \inf\{s \in \mathbb{R} : Y_t \in \widehat{C}_t(X_t, s)\}$.
  Compute loss $\ell^{(t)}(\cdot) = \ell_{1-\alpha}(S_t, \cdot)$
  Update predicted radius

$$\widehat{s}_{t+1} = \widehat{s}_t - \eta \frac{\nabla \ell^{(t)}(\widehat{s}_t)}{\sqrt{\sum_{i=1}^{t} \|\nabla \ell^{(i)}(\widehat{s}_i)\|_2^2}} \tag{10}$$

the predictions of active experts (Line 7):

$$\widehat{s}_t = \textstyle\sum_{i \in \text{Active}(t)} p_{i,t} \widehat{s}_{i,t},$$

where the weights $\{p_{i,t}\}_{i \in [t]}$ (Lines 4-6) rely on the $\{w_{i,t}\}_{i \in [t]}$ computed by the *coin betting* scheme (Orabona & Pál, 2016; Jun et al., 2017) in Lines 12-13.

Choice of expert In principle, SAOCP allows any choice of the expert that is a good regret minimization algorithm over its own active interval satisfying *anytime* regret guarantees. We choose the experts to be Scale-Free OGD (SF-OGD; Algorithm 2) (Orabona & Pál, 2018), a variant of OGD that decays its effective learning rate based on cumulative past gradient norms (cf. (10)).

On the quantile loss (1) (executed over the full horizon $[T]$ with learning rate $\eta = \Theta(D)$; $\eta = D/\sqrt{3}$ is optimal), SF-OGD enjoys an anytime regret guarantee (Proposition A.2)

$$\text{Reg}(t) \leq \mathcal{O}(D\sqrt{t}) \quad \text{for all } t \in [T], \tag{9}$$

which follows directly by applying Orabona & Pál (2018, Theorem 2). Plugging SF-OGD into Line 2 of Algorithm 1 gives our full SAOCP algorithm.

SF-OGD as an independent algorithm As a strong regret minimization algorithm itself, SF-OGD can also be run independently (over $[T]$) as an algorithm for online conformal prediction (described in Algorithm 2). We find empirically that SF-OGD itself already achieves strong performances in many scenarios (Section 5).

4. Theory

4.1. Strongly Adaptive Regret

We begin by showing the SARegret guarantee of SAOCP. As we instantiate SAOCP with SF-OGD as the experts, the proof follows directly by plugging the regret bound for SF-OGD (9) into the SARegret guarantee for SAOCP (Jun et al., 2017), and can be found in Appendix B.1.

Proposition 4.1 (SARegret bound for SAOCP)**.** *Algorithm 1 achieves the following SARegret bound* simultaneously for all lengths $k \in [T]$*:*

$$\text{SAReg}(T,k) \leq 15D\sqrt{k(\log T + 1)} \leq \widetilde{\mathcal{O}}(D\sqrt{k}). \tag{11}$$

The $\widetilde{\mathcal{O}}(D\sqrt{k})$ rate achieved by SAOCP is near-optimal for general online convex optimization problems, due to the standard regret lower bound $\Omega(D\sqrt{k})$ over any fixed interval of length k (Orabona, 2019, Theorem 5.1).

Comparison with FACI The SARegret guarantee of SAOCP in (11) improves substantially over the FACI (Fully Adaptive Conformal Inference) algorithm (Gibbs & Candès, 2022), an extension of ACI. Concretely, (11) holds *simultaneously for all lengths* k. By contrast, FACI achieves $\text{SAReg}(T,k) \leq \widetilde{\mathcal{O}}(D^2/\eta + \eta k)$ in our setting (cf. their Theorem 3.2), where $\eta > 0$ is their meta-algorithm learning rate. This can imply the same rate $\widetilde{\mathcal{O}}(D\sqrt{k})$ for a *single* k by optimizing η, but not multiple values of k simultaneously. As an example, for small interval sizes of order $k = \mathcal{O}(T^c) = o(T)$ with $c < 1$, FACI with the optimized $\eta \asymp T^{-c/2}$ can achieve the optimal rate $\text{SAReg}(T,k) \leq \widetilde{\mathcal{O}}(DT^{c/2})$ on *this particular* c, but will perform suboptimally on other values of c. By contrast, SAOCP achieves $\text{SAReg}(T,k) \leq \widetilde{\mathcal{O}}(DT^{c/2})$ simultaneously for all $c \in (0,1)$.

Also, in terms of algorithm styles, while both SAOCP and FACI are meta-algorithms that maintain multiple experts (base algorithms), a main difference between them is that all experts in FACI differ in their learning rates and are all active over $[T]$, whereas experts in SAOCP differ in their active intervals (cf. (8)).

Dynamic regret The (worst-case) dynamic regret—which measures the performance of an online learning algorithm against the worst-case (strongest) comparator sequence—is another generalization of the regret for capturing the performance in changing environments (Zinkevich, 2003; Besbes et al., 2015). Building on the reduction from dynamic regret to strongly adaptive regret (Zhang et al., 2018), we show that SAOCP also achieves a worst-case dynamic regret bound on any interval within $[T]$, with rate depending on a certain *path length* of the true radii $\{S_t\}_{t \geq 1}$; see Proposition B.1 and the discussions thereafter.[3]

[3]We remark that a more general version of dynamic regret which measures the regret against an *arbitrary* comparator sequence has also been considered in the literature. However, the relationship between the SARegret and the general dynamic regret remains unclear; see e.g. Zhao et al. (2022) for a discussion.

4.2. Coverage

Recall the empirical coverage error defined in (5):

$$\mathrm{CovErr}(T) = \left|\tfrac{1}{T}\sum_{t=1}^{T} \mathrm{err}_t - \alpha\right|.$$

Without any distributional assumptions, we show that SF-OGD achieves $\mathrm{CovErr}(t) \le \mathcal{O}(t^{-1/4}\log t)$ for any $t \in [T]$. So its empirical coverage converges to the target $1-\alpha$ as $T \to \infty$, similar to ACI (though with a slightly slower rate). The proof (Appendix B.3) builds on a grouping argument and the fact that the effective learning rate $\eta/\sqrt{\sum_{\tau=1}^{t} \left\|\nabla \ell^{(\tau)}(\widehat{s}_\tau)\right\|_2^2}$ of SF-OGD changes slowly in t.

Theorem 4.2 (Coverage bound for SF-OGD)**.** *Algorithm 2 with any learning rate $\eta = \Theta(D)$ and any initialization $\widehat{s}_1 \in [0, D]$ achieves $\mathrm{CovErr}(T) \le \mathcal{O}(\alpha^{-2}T^{-1/4}\log T)$ for any $T \ge 1$.*

We now provide a distribution-free coverage bound for SAOCP, building on a similar grouping argument as in Theorem 4.2.

Theorem 4.3 (Coverage bound for SAOCP; Informal version of Theorem B.3)**.** *For any $T \ge 1$, a randomized variant of Algorithm 1 where Line 7 is replaced by sampling an expert $i \sim p_t$ and predicting $\widehat{s}_t := \widehat{s}_{t,i}$ achieves*

$$\mathrm{CovErr}(T) \le \mathcal{O}\big(\inf_\beta (T^{1/2-\beta} + T^{\beta-1}S_\beta(T))\big). \quad (12)$$

Theorem 4.3 considers a randomized variant of SAOCP, and its coverage bound depends on a quantity $S_\beta(T)$ (full definition in Theorem B.3) that measures the smoothness of the expert weights and the cumulative gradient norms for each individual expert. Both are expected for technical reasons and also appear in coverage bounds for other expert-style meta-algorithms such as FACI (Gibbs & Candès, 2022). For instance, if there exists $\beta \in (1/2, 1)$ so that $S_\beta(T) \le \widetilde{\mathcal{O}}(T^\gamma)$ for some $\gamma < 1-\beta$, then (12) implies a coverage bound $\mathrm{CovErr}(T) \le \widetilde{\mathcal{O}}(T^{-\min\{1/2-\beta, \beta-1+\gamma\}}) = o_T(1)$.

We remark that Theorem 4.3 also holds more generically for other choices of the expert weights $\{p_t\}_{t\in[T]}$ (Line 5-6) and active intervals, not just those specified in Algorithm 1. In particular, SF-OGD is the special case where there is only a single active expert over $[T]$. In this case, we can recover the $\widetilde{\mathcal{O}}(\alpha^{-2}T^{-1/4})$ bound of Theorem 4.2 (see Appendix B.4.1 for a detailed discussion).

Additional coverage guarantee under distributional assumptions Under some mild regularity assumptions on the distributions of $S_1, \dots, S_T$, we show in Theorem C.3 that SAOCP achieves approximately valid coverage on *every* sub-interval of time. Its coverage error on any interval $I = [\tau, \tau+k-1] \subseteq [1, T]$ is $\widetilde{\mathcal{O}}(k^{-1/(2q)} + (\mathrm{Var}_I/k)^{1/q})$ for a certain $q \ge 2$ that quantifies the regularity of the distribution, and Var_I is a certain notion of variation between the true quantiles of S_t over $t \in I$ (cf. (29)). In particular, we obtain an approximately valid coverage on any interval I for which $\mathrm{Var}_I = o(k)$.

5. Experiments

We test SF-OGD (Algorithm 2) and SAOCP (Algorithm 1) empirically on two representative real-world online uncertainty quantification tasks: time series forecasting (Section 5.1) and image classification under distribution shift (Section 5.2). Choices of the prediction sets $\{\widehat{C}_t(x,s)\}_{x,s}$ will be described within each experiment. In both experiments, we compare against the following methods:

1. SCP: standard Split Conformal Prediction (Vovk et al., 2005) adapted to the online setting, which simply predicts the $(1-\alpha)$-quantile of the past radii. SCP does not admit a valid coverage guarantee in our settings as the data may not be exchangeable in general;
2. NExCP: Non-Exchangeable SCP (Barber et al., 2022), a variant of SCP that handles non-exchangeable data by reweighting. We follow their recommendations and use an exponential weighting scheme that upweights more recent observations;
3. FACI (Gibbs & Candès, 2022) with their specific "quantile parametrization" (4), and score function $\widetilde{S}_t$ corresponding to our choice of $\widehat{C}_t$;
4. FACI-S: Generalized version of FACI applied to predicting the radius $\widehat{s}_t$'s on our choice of $\widehat{C}_t$ directly.

Additional details about all methods can be found in Appendix E. Throughout this section we choose the target coverage level to be the standard $1-\alpha = 90\%$.

5.1. Time Series Forecasting

Setup We consider multi-horizon time series forecasting problems with real-valued observations $\{y_t\}_{t\ge1} \in \mathbb{R}$, where the base predictor $\widehat{f}$ uses the history $X_t := y_{1:t}$ to predict H steps into the future, i.e. $\widehat{f}(X_t) = \{\widehat{f}^{(h)}(X_t)\}_{h\in[H]} = \{\widehat{y}^{(h)}_{t+h}\}_{h\in[H]}$, where $\widehat{y}^{(h)}_{t+h}$ is a prediction for y_{t+h}. Using $\widehat{f}(X_t)$, we produce fixed-width prediction intervals

$$\widehat{C}^{(h)}_t(X_t, \widehat{s}^{(h)}_t) := [\widehat{y}^{(h)}_{t+h} - \widehat{s}^{(h)}_t, \widehat{y}^{(h)}_{t+h} + \widehat{s}^{(h)}_t], \quad (13)$$

where $\widehat{s}^{(h)}_t$ is predicted by an independent copy of the online conformal prediction algorithm for each $h \in [H]$ (so that there are H such algorithms in parallel). We form our online setting using a standard rolling window evaluation loop, wherein each *batch* consists of predicting all H intervals $\{\widehat{C}^{(h)}_t\}_{h\in[H]}$, observing all H true values $\{y_{t+h}\}_{h\in[H]}$, and moving to the next batch by setting $t \to t+H$. For each $h \in [H]$, we only evaluate y_{t+h} against one interval $\widehat{C}^{(h)}_t(X_t, \widehat{s}^{(h)}_t)$. After the evaluation is done, we use all pairs $\{(y_{t+k}, \widehat{y}^{(h)}_{t+k})\}_{k\in[H]}$ to update $\widehat{s}^{(h)}_t \to \widehat{s}^{(h)}_{t+H}$. To set

Method	LGBM (MAE = 0.06) Coverage	Width	LCE_k	SAReg_k	ARIMA (MAE = 0.18) Coverage	Width	LCE_k	SAReg_k	Prophet (MAE = 0.12) Coverage	Width	LCE_k	SAReg_k
SCP	$.844_{.004}$	$.127_{.004}$	$.252_{.007}$	$.017_{.001}$	$.871_{.004}$	$.245_{.029}$	$.237_{.008}$	$.039_{.021}$	$.783_{.008}$	$.178_{.007}$	$.355_{.012}$	$.019_{.001}$
NExCP	$.875_{.002}$	$.134_{.004}$	$.197_{.006}$	$.013_{.001}$	$.871_{.004}$	$.245_{.034}$	$.227_{.007}$	$.040_{.024}$	$.856_{.003}$	$.187_{.007}$	$.231_{.007}$	$.010_{.001}$
FACI	$.866_{.002}$	$\mathbf{.113_{.003}}$	$.180_{.005}$	$\mathbf{.009_{.001}}$	$.866_{.004}$	$\underline{.232}_{.034}$	$.214_{.006}$	$\underline{.034}_{.022}$	$.867_{.002}$	$\underline{.175}_{.007}$	$.184_{.005}$	$\underline{.006}_{.000}$
SF-OGD	$.889_{.002}$	$.138_{.004}$	$\underline{.154}_{.004}$	$.011_{.001}$	$.877_{.003}$	$.250_{.036}$	$\underline{.195}_{.006}$	$.037_{.022}$	$.888_{.001}$	$.186_{.007}$	$\mathbf{.138_{.003}}$	$.007_{.001}$
FACI-S	$.883_{.002}$	$.128_{.003}$	$.163_{.004}$	$.010_{.001}$	$.872_{.004}$	$.238_{.034}$	$.201_{.006}$	$.035_{.021}$	$.885_{.001}$	$.180_{.007}$	$.144_{.003}$	$\underline{.006}_{.000}$
SAOCP	$.882_{.002}$	$\underline{.121}_{.003}$	$\mathbf{.143_{.004}}$	$\mathbf{.009_{.000}}$	$.864_{.003}$	$\mathbf{.221_{.031}}$	$\mathbf{.190_{.005}}$	$\mathbf{.033_{.022}}$	$.872_{.001}$	$\mathbf{.173_{.007}}$	$\underline{.143}_{.003}$	$\mathbf{.005_{.000}}$

Table 1. Results on M4 Hourly (414 time series) with target coverage $1-\alpha = 0.9$ and interval size $k = 20$. Results are formatted as mean_{std}. Best results are **bold**, while second best are underlined, as long as the method's global coverage is in $(0.85, 0.95)$ (green). For all base predictors, SAOCP achieves the best or second-best width, local coverage error, and strongly adaptive regret.

Method	LGBM (MAE = 0.13) Coverage	Width	LCE_k	SAReg_k	ARIMA (MAE = 0.06) Coverage	Width	LCE_k	SAReg_k	Prophet (MAE = 0.32) Coverage	Width	LCE_k	SAReg_k
SCP	$.769_{.004}$	$.184_{.002}$	$.466_{.005}$	$.031_{.001}$	$.896_{.002}$	$.122_{.002}$	$.290_{.004}$	$.018_{.001}$	$.599_{.005}$	$.349_{.004}$	$.614_{.005}$	$.051_{.001}$
NExCP	$.818_{.002}$	$.183_{.002}$	$.420_{.004}$	$.015_{.000}$	$.891_{.001}$	$.116_{.002}$	$.296_{.004}$	$.012_{.001}$	$.715_{.003}$	$.356_{.004}$	$.559_{.004}$	$.019_{.000}$
FACI	$.846_{.002}$	$.169_{.002}$	$.308_{.003}$	$.008_{.000}$	$.886_{.001}$	$\underline{.101}_{.002}$	$.259_{.003}$	$\mathbf{.008_{.000}}$	$.767_{.003}$	$.344_{.004}$	$.397_{.004}$	$.014_{.001}$
SF-OGD	$.873_{.001}$	$.173_{.002}$	$.246_{.003}$	$.011_{.001}$	$.892_{.001}$	$.106_{.002}$	$.245_{.003}$	$.011_{.001}$	$.862_{.001}$	$.354_{.004}$	$.220_{.002}$	$.008_{.001}$
FACI-S	$.875_{.001}$	$\underline{.169}_{.002}$	$\underline{.240}_{.003}$	$\underline{.010}_{.001}$	$.891_{.001}$	$.103_{.002}$	$\underline{.243}_{.003}$	$.010_{.001}$	$.866_{.002}$	$\mathbf{.349_{.004}}$	$\underline{.210}_{.002}$	$\underline{.007}_{.001}$
SAOCP	$.869_{.001}$	$\mathbf{.162_{.002}}$	$\mathbf{.213_{.002}}$	$\mathbf{.007_{.000}}$	$.875_{.001}$	$\mathbf{.093_{.002}}$	$\mathbf{.238_{.002}}$	$\mathbf{.008_{.001}}$	$.867_{.001}$	$\underline{.352}_{.004}$	$\mathbf{.172_{.001}}$	$\mathbf{.005_{.000}}$

Table 2. Results on M4 Daily (4227 time series) with target coverage $1-\alpha = 0.9$ and interval size $k = 20$. Results are formatted as mean_{std}. Best results are **bold**, while second best are underlined, as long as the method's global coverage is in $(0.85, 0.95)$ (green). SAOCP achieves the best width, local coverage error, and strongly adaptive regret for all base predictors. The only methods which achieve global coverage in $(0.85, 0.95)$ for LGBM and Prophet are the ones that predict $\widehat{s}_{t+1}$ directly, not as a quantile of $S_1, \ldots, S_t$.

the maximum radius for SF-OGD and SAOCP, we choose $D/\sqrt{3}$ for each horizon h to be the largest h-step residual observed on the calibration split of the training data.

Base predictors We consider three diverse types of base predictors (models), and we use their implementations in Merlion v2.0.0 (Bhatnagar et al., 2021):

1. LGBM: A model which uses gradient boosted trees to predict $\widehat{y}^{(h)}_{t+h} = \widehat{f}^{(h)}(y_{t-L+1}, \ldots, y_t)$. This approach attains strong performance on many time series benchmarks (Elsayed et al., 2021; Bhatnagar et al., 2021).
2. $\text{ARIMA}(10, d^\star, 10)$: The classical AutoRegressive Integrated Moving Average stochastic process model for a time series, where the difference order $d^\star$ is chosen by KPSS stationarity test (Kwiatkowski et al., 1992).
3. Prophet (Taylor & Letham, 2017): A popular Bayesian model which directly predicts the value y as a function of time, i.e. $\widehat{y}_t = \widehat{f}(t)$.

Datasets We evaluate on four datasets totaling 5111 time series: the hourly (414 time series), daily (4227 time series), and weekly (359 time series) subsets of the M4 Competition, a dataset of time series from many domains including industries, demographics, environment, finance, and transportation (Makridakis et al., 2018); and NN5, a dataset of 111 time series of daily banking data (Ben Taieb et al., 2012). We normalize each time series to lie in $[0, 1]$.

We use horizons H of 24, 30, and 26 for hourly, daily, and weekly data, respectively. Each time series of length L is split into a training set of length $L - 120$ with 80% for training the base predictor and 20% for initializing the UQ methods, and a test set of length 120 to test the UQ methods.

Metrics For each experiment, we average the following statistics across all time series: global coverage, median width, worst-case local coverage error

$$\text{LCE}_k := \max_{[\tau, \tau+k-1] \subseteq [1,T]} \left| \alpha - \frac{1}{k} \sum_{t=\tau}^{\tau+k-1} \text{err}_t \right|, \quad (14)$$

and strongly adaptive regret $\text{SAReg}(T, k)$ (7), which we abbreviate as SAReg_k. In all cases, we use an interval length of $k = 20$. We also report the average mean absolute error (MAE) of each base predictor.

Results We report results on M4 Hourly and M4 Daily in Tables 1, 2, and on M4 Weekly and NN5 Daily in Tables 4, and 5 (Appendix D).

SAOCP consistently achieves global coverage in $(0.85, 0.95)$, and it obtains the best or second-best interval width, local coverage error, and strongly adaptive regret for all base predictors on all 3 M4 datasets. FACI-S generally achieves better LCE_k and SAReg_k than FACI, showing the benefits of predicting $\widehat{s}_{t+1}$ directly, rather than as a quantile of $S_1, \ldots, S_t$. The relative performance of FACI-S and SF-OGD varies, though FACI-S is usually a bit better.

Method	LCE_{10}	LCE_{20}	LCE_{30}	LCE_{40}	LCE_{50}	LCE_{60}
SCP	.585	.466	.389	.336	.295	.263
NExCP	.592	.420	.323	.262	.221	.191
FACI	.467	.308	.232	.187	.157	.134
SF-OGD	.394	.246	.183	.146	.123	<u>.104</u>
FACI-S	<u>.385</u>	<u>.240</u>	<u>.179</u>	<u>.144</u>	<u>.122</u>	<u>.104</u>
SAOCP	**.337**	**.213**	**.159**	**.128**	**.107**	**.091**

Table 3. Local coverage error of the LGBM model evaluated on the M4 Daily dataset (4227 time series) with target coverage $1-\alpha = 0.9$ and multiple interval sizes k. Best results are **bold**, while second-best are <u>underlined</u>. SAOCP achieves the best LCE_k for all k. Standard errors for all reported values are below 0.005.

However, SAOCP consistently achieves better LCE_k and SAReg_k than both FACI-S and SF-OGD.

There are multiple instances where all of SCP/NExCP/FACI fail to attain global coverage in $(0.85, 0.95)$ (Tables 2 and 4). The base predictor's MAE is at least 0.13 in all these cases, suggesting an advantage of predicting $\widehat{s}_{t+1}$ directly as in SF-OGD/SAOCP when the underlying base predictor is inaccurate.

Evaluations with different interval lengths Tables 1, 2, 4, and 5 evaluate LCE_k of various online conformal prediction methods with an interval size of $k = 20$ across a wide range of experimental settings. To determine whether these results hold for different interval lengths, we report LCE_k for the LGBM model on the M4 Daily dataset for $k \in \{10, 20, 30, 40, 50, 60\}$ (Table 3). Out of the box, SAOCP achieves the best LCE for all k, further exhibiting its robustness.

Additional experiments with ensemble models In Appendix F, we use EnbPI (Xu & Xie, 2021) to train a bootstrapped ensemble, and we compare EnbPI's results with those obtained by applying NExCP, FACI, SF-OGD, and SAOCP to the residuals produced by that ensemble. The results largely mirror those in the main paper.

5.2. Image Classification Under Distribution Shift

Datasets and Setup We evaluate the ability of conformal prediction methods to maintain coverage when the underlying distribution shifts in a systematic manner. We use a ResNet-50 classifier (He et al., 2016) pre-trained on ImageNet and implemented in PyTorch (Paszke et al., 2019). Here, $x \in \mathcal{X}$ is an image, and $y \in \mathcal{Y} = [m]$ is its class. To construct structured distribution shifts away from the training distribution, we use TinyImageNet-C and ImageNet-C (Hendrycks & Dietterich, 2019), which are corrupted versions of the TinyImageNet ($m = 200$ classes) (Le & Yang, 2015) and ImageNet ($m = 1000$ classes) (Deng et al., 2009) test sets designed to evaluate model robustness. These corrupted datasets apply 15 visual corruptions at 5 different severity levels to each image in the original test set.

We consider two regimes: **sudden shifts** where the corruption level alternates between 0 (the base test set) and 5, and **gradual shifts** where the corruption level increases in the order of $\{0, 1, \ldots, 5\}$. We randomly sample 500 data points for each corruption level before changing to the next level.

Prediction Sets We follow Angelopoulos et al. (2021b) to construct our prediction sets. Let $\widehat{f} : \mathbb{R}^d \to \Delta^m$ be a classifier that outputs a probability distribution on the m-simplex. At each t, we sample $U_t \sim \text{Unif}[0, 1]$ and let

$$S_t(x, y) = \lambda\sqrt{[k_y - k_{reg}]_+} + U_t\widehat{f}_y(x) + \textstyle\sum_{i=1}^{k_y-1} \widehat{f}_{\pi(i)}(x)$$
$$\widehat{C}_t(X_t) = \{y : S_t(X_t, y) \le \widehat{s}_t\}, \tag{15}$$

where π is the permutation that ranks $\widehat{f}(x)$ in decreasing order, $\pi(k_y) = y$, and λ and k_{reg} are regularization parameters designed to reduce the size of the prediction set. For TinyImageNet, we use $\lambda = 0.01$ and $k_{reg} = 20$. For ImageNet, we use $\lambda = 0.01$ and $k_{reg} = 10$. Accordingly, we set the maximum radius $D = \lambda\sqrt{m - k_{reg}}$ for SF-OGD and SAOCP, which we note is the maximum value of $S_t(\cdot, \cdot)$.

Metrics When evaluating the UQ methods, we plot the local coverage and prediction set size (PSS) of each method using an interval length of $k = 100$,

$$\text{LocalCov}(t) = \tfrac{1}{100}\textstyle\sum_{i=t}^{t+99} \mathbb{1}[Y_i \in \widehat{C}_i(X_i)]$$
$$\text{LocalPSS}(t) = \tfrac{1}{100}\textstyle\sum_{i=t}^{t+99} \left|\widehat{C}_i(X_i)\right|.$$

We compare the local coverage to a target of $1-\alpha$, while we compare the local PSS to the $1-\alpha$ empirical quantile of the oracle set sizes $\text{PSS}_t^\star = |\{y : S_t(X_t, y) \le S_t(X_t, Y_t)\}|$. These targets are the "best fixed" values in each window. We also report the worst-case local coverage error LCE_{100} (14).

Results We evaluate the UQ methods on TinyImageNet and TinyImageNet-C in Figure 1, and on ImageNet and ImageNet-C in Figure 2 (Appendix G). In both sudden and gradual distribution shift, the local coverage of SAOCP and SF-OGD remains the closest to the target of 0.9. The difference is more notable when the distribution shifts suddenly. When the distribution shifts more gradually, NExCP, FACI, and FACI-S have worse coverage than SAOCP and SF-OGD at the first change point, which is where the largest change in the best set size occurs.

All methods besides SCP predict sets of similar sizes, though FACI's, FACI-S's, and NExCP's prediction set sizes adapt more slowly to changes in the best fixed size (e.g. $t \in [500, 700]$ for gradual shift in Figure 2). On TinyImageNet, SAOCP obtains slightly better local coverage than

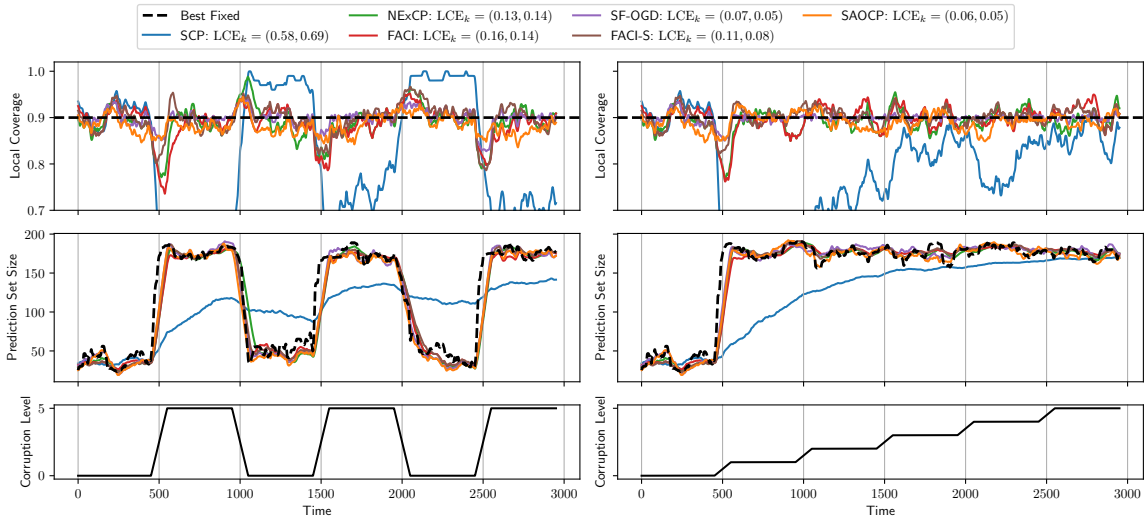

Figure 1. Local coverage (top row) and prediction set size (second row) achieved by various UQ methods when the distribution shifts between TinyImageNet and TinyImageNet-C every 500 steps. We plot moving averages with a window size of 100. Left: sudden shifts between corruption level 0 and 5. Right: gradual shift from level 0 to 5. SAOCP and SF-OGD's local coverage remain the closest to the target of 0.9, especially at the change points. While their prediction sets have similar size, LCE_k is lower for SAOCP than SF-OGD.

SF-OGD, and they both have similar prediction set sizes (Figure 1). On ImageNet, SAOCP and SF-OGD attain similar local coverages, but SAOCP tends to attain that coverage with a smaller prediction set (Figure 2).

6. Conclusion

This paper develops new algorithms for online conformal prediction under arbitrary distribution shifts. Our algorithms achieve approximately valid coverage and better strongly adaptive regret than existing work. On real-world experiments, our proposed algorithms achieve coverage closer to the target within local windows, and they produce smaller prediction sets than existing methods. Our work opens up many questions for future work, such as obtaining stronger coverage guarantees, or characterizing the optimality of the learned radii under various settings with distribution shift.

References

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

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021a.

Angelopoulos, A. N., Bates, S., Jordan, M., and Malik, J. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=eNdiU_DbM9.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022a.

Angelopoulos, A. N., Kohli, A. P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., and Romano, Y. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pp. 717–730. PMLR, 2022b.

Bai, Y., Mei, S., Wang, H., Zhou, Y., and Xiong, C. Efficient and differentiable conformal prediction with general function classes. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Ht85_jyihxp.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486 – 507, 2021. doi: 10.1214/20-AOS1965. URL https://doi.org/10.1214/20-AOS1965.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability, 2022. URL https://arxiv.org/abs/2202.13415.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Ben Taieb, S., Bontempi, G., Atiya, A. F., and Sorjamaa, A. A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. *Expert Systems with Applications*, 39(8):7067–7083, 2012. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2012.01.039. URL https://www.sciencedirect.com/science/article/pii/S0957417412000528.

Besbes, O., Gur, Y., and Zeevi, A. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Bhatnagar, A., Kassianik, P., Liu, C., Lan, T., Yang, W., Cassius, R., Sahoo, D., Arpit, D., Subramanian, S., Woo, G., Saha, A., Jagota, A. K., Gopalakrishnan, G., Singh, M., Krithika, K. C., Maddineni, S., Cho, D., Zong, B., Zhou, Y., Xiong, C., Savarese, S., Hoi, S., and Wang, H. Merlion: A machine learning library for time series. 2021. URL https://arxiv.org/abs/2109.09265.

Candès, E. J., Lei, L., and Ren, Z. Conformalized survival analysis, 2021. URL https://arxiv.org/abs/2103.09763.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. Robust validation: Confident predictions even when distributions shift, 2020. URL https://arxiv.org/abs/2008.04267.

Cauchois, M., Gupta, S., and Duchi, J. C. Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. *J. Mach. Learn. Res.*, 22 (1), jul 2022. ISSN 1532-4435.

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. Exact and robust conformal inference methods for predictive machine learning with dependent data. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 732–749. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/chernozhukov18a.html.

Daniely, A., Gonen, A., and Shalev-Shwartz, S. Strongly adaptive online learning. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1405–1411, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/daniely15.html.

Dashevskiy, M. and Luo, Z. Network traffic demand prediction with confidence. In *IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference*, pp. 1–5, 2008. doi: 10.1109/GLOCOM.2008.ECP.284.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Elsayed, S., Thyssens, D., Rashed, A., Jomaa, H. S., and Schmidt-Thieme, L. Do we really need deep learning models for time series forecasting?, 2021. URL https://arxiv.org/abs/2101.02118.

Fannjiang, C., Bates, S., Angelopoulos, A., Listgarten, J., and Jordan, M. I. Conformal prediction for the design problem. *arXiv preprint arXiv:2202.03613*, 2022.

Feldman, S., Ringel, L., Bates, S., and Romano, Y. Risk control for online learning models, 2022. URL https://arxiv.org/abs/2205.09095.

Gibbs, I. and Candès, E. Adaptive conformal inference under distribution shift. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=6vaActvpcp3.

Gibbs, I. and Candès, E. Conformal inference for online prediction with arbitrary distribution shifts, 2022. URL https://arxiv.org/abs/2208.08401.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. K. Nested conformal prediction and quantile out-of-bag ensemble methods. *arXiv preprint arXiv:1910.10562*, 2019.

Hazan, E. *Introduction to Online Convex Optimization*. MIT Press, Cambridge, MA, USA, 2022. ISBN 9780262046985.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Hendrycks, D., Mazeika, M., Wilson, D., and Gimpel, K. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in neural information processing systems*, 31, 2018.

Jun, K.-S., Orabona, F., Wright, S., and Willett, R. Improved Strongly Adaptive Online Learning using Coin Betting. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 943–951. PMLR, 20–22 Apr 2017.

URL https://proceedings.mlr.press/v54/jun17a.html.

Kath, C. and Ziel, F. Conformal prediction interval estimation and applications to day-ahead and intraday power markets. *International Journal of Forecasting*, 37(2):777–799, 2021. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2020.09.006. URL https://www.sciencedirect.com/science/article/pii/S0169207020301473.

Koenker, R. and Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159–178, 1992. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(92)90104-Y. URL https://www.sciencedirect.com/science/article/pii/030440769290104Y.

Le, Y. and Yang, X. S. Tiny imagenet visual recognition challenge. 2015.

Lei, J. and Wasserman, L. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014. doi: https://doi.org/10.1111/rssb.12021. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12021.

Lei, J., Robins, J., and Wasserman, L. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013. doi: 10.1080/01621459.2012.751873. URL https://doi.org/10.1080/01621459.2012.751873. PMID: 25237208.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808, 2018.

Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Orabona, F. and Pál, D. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Orabona, F. and Pál, D. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018. ISSN 0304-3975. doi: https://doi.org/10.1016/j.tcs.2017.11.021. URL https://www.sciencedirect.com/science/article/pii/S0304397517308514. Special Issue on ALT 2015.

Papadopoulos, H. Inductive conformal prediction: Theory and application to neural networks. In Fritzsche, P. (ed.), *Tools in Artificial Intelligence*, chapter 18. IntechOpen, Rijeka, 2008. doi: 10.5772/6078. URL https://doi.org/10.5772/6078.

Park, S., Bastani, O., Matni, N., and Lee, I. Pac confidence sets for deep neural networks via calibrated prediction. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJxVI04YvB.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Pearce, T., Brintrup, A., Zaki, M., and Neely, A. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4075–4084. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/pearce18a.html.

Podkopaev, A. and Ramdas, A. Distribution-free uncertainty quantification for classification under label shift. In de Campos, C. and Maathuis, M. H. (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 844–853. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/podkopaev21a.html.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf.

Romano, Y., Sesia, M., and Candès, E. J. Classification with valid and adaptive coverage. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Sousa, M., Tomé, A. M., and Moreira, J. A general framework for multi-step ahead adaptive conformal heteroscedastic time series forecasting, 2022. URL https://arxiv.org/abs/2207.14219.

Stankeviciute, K., M. Alaa, A., and van der Schaar, M. Conformal time-series forecasting. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6216–6228. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/312f1ba2a72318edaaa995a67835fad5-Paper.pdf.

Steinwart, I. and Christmann, A. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17 (1):211 – 225, 2011. doi: 10.3150/10-BEJ267. URL https://doi.org/10.3150/10-BEJ267.

Stutz, D., Dvijotham, K. D., Cemgil, A. T., and Doucet, A. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=t8O-4LKFVx.

Sun, S. and Yu, R. Copula conformal prediction for multi-step time series forecasting, 2022. URL https://arxiv.org/abs/2212.03281.

Taylor, S. J. and Letham, B. Forecasting at scale. *PeerJ Preprints*, 5(e3190v2), Sept 2017. doi: 10.7287/peerj.preprints.3190v2.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf.

Vovk, V. Conditional validity of inductive conformal predictors. In Hoi, S. C. H. and Buntine, W. (eds.), *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR. URL https://proceedings.mlr.press/v25/vovk12.html.

Vovk, V., Gammerman, A., and Saunders, C. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pp. 444–453, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

Vovk, V., Nouretdinov, I., Manokhin, V., and Gammerman, A. Cross-conformal predictive distributions. In Gammerman, A., Vovk, V., Luo, Z., Smirnov, E., and Peeters, R. (eds.), *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*, volume 91 of *Proceedings of Machine Learning Research*, pp. 37–51. PMLR, 11–13 Jun 2018. URL https://proceedings.mlr.press/v91/vovk18a.html.

Wisniewski, W., Lindsay, D., and Lindsay, S. Application of conformal prediction interval estimations to market makers' net positions. In Gammerman, A., Vovk, V., Luo, Z., Smirnov, E., and Cherubin, G. (eds.), *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pp. 285–301. PMLR, 09–11 Sep 2020. URL https://proceedings.mlr.press/v128/wisniewski20a.html.

Xu, C. and Xie, Y. Conformal prediction interval for dynamic time-series. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11559–11569. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/xu21h.html.

Yang, Y. and Kuchibhotla, A. K. Finite-sample efficient conformal prediction, 2021. URL https://arxiv.org/abs/2104.13871.

Yang, Y., Kuchibhotla, A. K., and Tchetgen, E. T. Doubly robust calibration of prediction sets under covariate shift, 2022. URL https://arxiv.org/abs/2203.01761.

Zaffran, M., Feron, O., Goude, Y., Josse, J., and Dieuleveut, A. Adaptive conformal predictions for time series. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25834–25866. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zaffran22a.html.

Zhang, L., Yang, T., rong jin, and Zhou, Z.-H. Dynamic regret of strongly adaptive methods. In Dy, J. and Krause,

A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5882–5891. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/zhang18o.html.

Zhao, P. and Zhang, L. Improved analysis for dynamic regret of strongly convex and smooth functions. In *Learning for Dynamics and Control*, pp. 48–59. PMLR, 2021.

Zhao, P., Xie, Y.-F., Zhang, L., and Zhou, Z.-H. Efficient methods for non-stationary online learning. *Advances in Neural Information Processing Systems*, 35:11573–11585, 2022.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pp. 928–935. AAAI Press, 2003. ISBN 1577351894.

A. Basic Properties of Online Conformal Prediction Algorithms

A.1. Properties of SF-OGD

We consider the SF-OGD algorithm (Algorithm 2). We first show that the iterates of SF-OGD are bounded within a range slightly larger than the range of the true radii. The proof is similar to Gibbs & Candès (2021, Lemma 4.1).

Lemma A.1 (Bounded iterates for SF-OGD)**.** *Suppose the true radii are bounded: $S_t \in [0, D]$ for all $t \in [T]$. Then Algorithm 2 with any initialization $\widehat{s}_1 \in [-\eta, D+\eta]$ and learning rate $\eta > 0$ admits bounded iterates:*

$$\widehat{s}_t \in [-\eta, D+\eta] \quad \textit{for all } t \in [T].$$

Proof. Recall by (2) that

$$\nabla \ell^{(t)}(\widehat{s}_t) = \alpha - \mathbb{1}\left[\widehat{s}_t < S_t\right] = \alpha - \mathrm{err}_t \in \{-(1-\alpha), \alpha\} \subset [-1, 1]. \tag{16}$$

for all $t \in [T]$. Therefore, Algorithm 2 satisfies for any $t \geq 1$ that

$$|\widehat{s}_{t+1} - \widehat{s}_t| = \eta \left| \frac{\alpha - \mathrm{err}_t}{\sqrt{\sum_{\tau=1}^{t} (\alpha - \mathrm{err}_\tau)^2}} \right| \leq \eta. \tag{17}$$

We prove the lemma by contradiction. Suppose there exists some t such that $\widehat{s}_t \notin [-\eta, D+\eta]$. Let $t \geq 2$ be the smallest such time index (abusing notation slightly). Suppose $\widehat{s}_t > D + \eta$, then by (17) we must have $\widehat{s}_{t-1} > D$ but $\widehat{s}_{t-1} \leq D + \eta$. Note that $\widehat{s}_{t-1} > D \geq S_{t-1}$ by our precondition, so that the $(t-1)$-th prediction set must cover and thus $\mathrm{err}_{t-1} = 0$. Therefore by the algorithm update (10) we have

$$\widehat{s}_t = \widehat{s}_{t-1} - \eta \frac{\alpha - \mathrm{err}_{t-1}}{\sqrt{\sum_{\tau=1}^{t-1} (\alpha - \mathrm{err}_\tau)^2}} < \widehat{s}_{t-1} \leq D + \eta,$$

contradicting with our assumption that $\widehat{s}_t > D + \eta$. A similar contradiction can be derived for the other case where $\widehat{s}_t < -\eta$. This proves the desired result. □

The following regret bound follows directly by applying the generic regret bound of Scale-Free OGD (Orabona & Pál, 2018, Theorem 2) to the quantile loss (1).

Proposition A.2 (Anytime regret bound for SF-OGD)**.** *Suppose the true radii are bounded: $S_t \in [0, D]$ for all $t \in [T]$. Then Algorithm 2 with any initialization $\widehat{s}_1 \in [0, D]$ and learning rate $\eta = D/\sqrt{3}$ achieves the following regret bound for any $t \in [T]$:*

$$\mathrm{Reg}(t) \leq (\sqrt{3}+1) D \sqrt{\sum_{\tau=1}^{t} \left\| \nabla \ell^{(\tau)}(\widehat{s}_\tau) \right\|_2^2} \leq \mathcal{O}(D\sqrt{t}).$$

Proof. The second inequality follows directly by (16).

To prove the first inequality (the regret bound), we note that Algorithm 2 is a special case of the Scale-Free Mirror Descent algorithm of Orabona & Pál (2018, Section 4) with convex loss $\ell^{(t)}(\cdot) = \ell_{1-\alpha}(S_t, \cdot)$, and regularizer $R(s) := s^2/(2\eta)$ (in their notation) which is $\lambda = 1/\eta$-strongly convex with respect to the ℓ_2 norm on $\mathbb{R}$. Further, by Lemma A.1 we have $\widehat{s}_t \in [-\eta, D+\eta]$ for all $t \in [T]$. Therefore, applying Orabona & Pál (2018, Theorem 2) gives that for any $t \in [T]$,

$$\begin{aligned}
\sum_{\tau=1}^{t} \ell^{(\tau)}(\widehat{s}_\tau) - \inf_{s^\star \in [0,D]} \sum_{\tau=1}^{t} \ell^{(\tau)}(s^\star) &\leq \left(\frac{1}{\lambda} + \sup_{\tau \geq 1} B_R(s^\star, \widehat{s}_\tau) \right) \cdot \sqrt{\sum_{\tau=1}^{t} \left\| \nabla \ell^{(\tau)}(\widehat{s}_\tau) \right\|_2^2} \\
&= \left(\eta + \sup_{s^\star \in [0,D], s' \in [-\eta, D+\eta]} \frac{1}{2\eta} (s^\star - s')^2 \right) \cdot \sqrt{\sum_{\tau=1}^{t} \left\| \nabla \ell^{(\tau)}(\widehat{s}_\tau) \right\|_2^2}
\end{aligned}$$

$$= \left(\eta + \frac{(D+\eta)^2}{2\eta}\right) \cdot \sqrt{\sum_{\tau=1}^{t} \left\|\nabla \ell^{(\tau)}(\widehat{s}_\tau)\right\|_2^2},$$

where $B_R(\cdot,\cdot)$ denotes the Bregman divergence associated with R. Choosing $\eta = D/\sqrt{3}$, the leading coefficient is $(\sqrt{3}+1)D$. The desired result follows by noting that

$$\mathrm{Reg}(t) = \sum_{\tau=1}^{t} \ell^{(\tau)}(\widehat{s}_\tau) - \inf_{s^\star \in \mathbb{R}} \sum_{\tau=1}^{t} \ell^{(\tau)}(s^\star) = \sum_{\tau=1}^{t} \ell^{(\tau)}(\widehat{s}_\tau) - \inf_{s^\star \in [0,D]} \sum_{\tau=1}^{t} \ell^{(\tau)}(s^\star)$$

by our assumption that $S_t \in [0, D]$ and basic properties of the quantile losses $\left\{\ell^{(\tau)}(\cdot)\right\}_{\tau \geq 1}$. □

A.2. Example of "Trivial" Algorithm with Coverage Guarantee

We consider the following "trivial" online conformal prediction algorithm that does not utilize the data at all: Simply predict the maximum radius D for $(1-\alpha)$ proportion of the time steps, then predict the minimum radius 0 for α proportion of the time steps:

$$\begin{cases} \widehat{s}_t := D \quad \text{for } t \in \{1, \ldots, \lfloor (1-\alpha)T \rfloor\} =: T_{\text{full}}, \\ \widehat{s}_t := 0 \quad \text{for } t \in \{\lfloor (1-\alpha)T \rfloor + 1, \ldots, T\} =: T_{\text{empty}}. \end{cases} \tag{18}$$

By our assumption that $S_t \in [0, D]$ almost surely and the nested set structure of $\widehat{C}_t(X_t, \cdot)$, we have $\mathrm{err}_t = \mathbb{1}[Y_t \in \widehat{C}_t] = \mathbb{1}[\widehat{s}_t \geq S_t] = 0$ for all $t \in T_{\text{full}}$, and similarly $\mathrm{err}_t = 1$ for all $t \in T_{\text{empty}}$. Therefore, algorithm (18) directly satisfies

$$\mathrm{CovErr}(T) = \left| \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t - \alpha \right| = \left| \frac{|T_{\text{empty}}|}{T} - \alpha \right| = \left| \frac{T - \lfloor (1-\alpha)T \rfloor}{T} - \alpha \right| \leq \frac{1}{T}, \tag{19}$$

i.e. the algorithm achieves approximately $(1-\alpha)$ empirical coverage, with error $O(1/T)$. It is also straightforward to see that, by slightly modifying the definition of $T_{\text{full}}, T_{\text{empty}}$ (making the two index sets alternate), we can make the above coverage bound hold in an anytime sense (for $t \in [T]$).

However, it is straightforward to construct examples of data distributions for which the trivial algorithm (18) suffers linear regret on the quantile loss $\ell^{(t)}$ defined in (1), and such data distributions can be chosen to be fairly simple. For example, suppose all data points admit the same true radius $D/2$, i.e.

$$S_t \equiv D/2 \quad \text{for all } t \in [T].$$

Then for $s^\star = D/2$ we have $\ell^{(t)}(s^\star) = \ell_{1-\alpha}(D/2, D/2) = 0$ for all $t \in [T]$, which achieves total loss $\sum_{t=1}^{T} \ell^{(t)}(s^\star) = 0$ (the smallest possible, since $\ell^{(t)}(\cdot) \geq 0$). On the other hand, algorithm (18) achieves loss

$$\ell^{(t)}(\widehat{s}_t) = \ell_{1-\alpha}(S_t, \widehat{s}_t) = \begin{cases} \ell_{1-\alpha}(D/2, D) = \alpha(D/2) \quad \text{for } t \in T_{\text{full}}, \\ \ell_{1-\alpha}(D/2, 0) = (1-\alpha)(D/2) \quad \text{for } t \in T_{\text{empty}}. \end{cases}$$

Therefore, we have

$$\begin{aligned} \mathrm{Reg}(T) &= \sum_{t=1}^{T} \ell^{(t)}(\widehat{s}_t) - \inf_{s^\star} \ell^{(t)}(s_\star) = \sum_{t=1}^{T} \ell^{(t)}(\widehat{s}_t) = \alpha D/2 \cdot |T_{\text{full}}| + (1-\alpha)D/2 \cdot |T_{\text{empty}}| \\ &= \alpha D/2 \cdot \lfloor (1-\alpha)T \rfloor + (1-\alpha)D/2 \cdot (T - \lfloor (1-\alpha)T \rfloor) \geq \alpha(1-\alpha)DT = \Omega(T), \end{aligned}$$

i.e. algorithm (18) suffers from linear regret. This demonstrates sublinear regret as a sensible criterion for ruling out trivial algorithms like (18).

B. Proofs for Section 4

B.1. Proof of Proposition 4.1

The proof follows by plugging in the regret bound for SF-OGD (Proposition A.2) into Jun et al. (2017, Theorem 2). Define $u(t) := \max_n\{2^n : t \equiv 0 \bmod 2^n\}$. Fix any $k \in [T]$ and $\tau \in [T-k+1]$. Their proof starts by splitting the interval $[\tau, \tau+k-1]$ into consecutive sub-intervals $\bar{J}^{(1)}, \ldots, \bar{J}^{(n)}$, where $\bar{J}^{(i)} = [\tau_i, \max\{\tau+k, \tau_i + u(\tau_i)\} - 1]$ is a prefix of expert $\mathcal{A}_{\tau_i}$'s active interval.

We have for any fixed $s^\star \in \mathbb{R}$ that

$$\begin{aligned}
\mathrm{Regret}_\tau^k(s^\star) &:= \sum_{t=\tau}^{\tau+k-1} \ell^{(t)}(\widehat{s}_t) - \sum_{t=\tau}^{\tau+k-1} \ell^{(t)}(s^\star) \\
&= \sum_{i=1}^{n} \sum_{t\in\bar{J}^{(i)}} \Big(\ell^{(t)}(\widehat{s}_t) - \ell^{(t)}(\widehat{s}_{t,\tau_i})\Big) + \sum_{i=1}^{n} \sum_{t\in\bar{J}^{(i)}} \Big(\ell^{(t)}(\widehat{s}_{t,\tau_i}) - \ell^{(t)}(s^\star)\Big) \\
&\le D \sum_{i=1}^{n} \underbrace{\sqrt{\big|\bar{J}^{(i)}\big|(7\log T + 5)}}_{\text{Jun et al. (2017, Lemma 2)}} + D \sum_{i=1}^{n} \underbrace{\sqrt{\big|\bar{J}^{(i)}\big|(1+\sqrt{3})}}_{\text{Proposition A.2}} \\
&\le D\Big(\sqrt{7\log T + 5} + \sqrt{1+\sqrt{3}}\Big) \underbrace{\sum_{j=0}^{\infty} \sqrt{k 2^{-j}}}_{\text{Jun et al. (2017, Lemma 3)}} \\
&= \frac{D\sqrt{2}}{\sqrt{2}-1}\sqrt{k}\Big(\sqrt{7\log T + 5} + \sqrt{1+\sqrt{3}}\Big) \le 15D\sqrt{k(\log T + 1)}
\end{aligned}$$

Taking supremum over all $s^\star \in \mathbb{R}$ and all intervals $[\tau, \tau+k-1] \subset [T]$, we obtain the desired bound on $\mathrm{SAReg}(T, k)$. □

B.2. Dynamic Regret for SAOCP

Proposition B.1 (Dynamic regret bound for SAOCP)**.** *Algorithm 1 achieves the following (worst-case) dynamic regret bound: For any interval $[\tau, \tau+k-1] \subset [T]$ of length $k \in [T]$, we have*

$$\sum_{t=\tau}^{\tau+k-1} \ell^{(t)}(\widehat{s}_t) - \min_{s^\star_{\tau:\tau+k-1}} \sum_{t=\tau}^{\tau+k-1} \ell^{(t)}(s_t^\star) = \sum_{t=\tau}^{\tau+k-1} \Big[\ell^{(t)}(\widehat{s}_t) - \ell^{(t)}(S_t)\Big] \le \widetilde{\mathcal{O}}\Big(D\Big[V_{[\tau,\tau+k-1]}^{1/3} k^{2/3} + \sqrt{k}\Big]\Big), \tag{20}$$

where

$$V_{[\tau,\tau+k-1]} := \sum_{t=\tau+1}^{\tau+k-1} |S_t - S_{t-1}|$$

is the path length *of the true radii within* $[\tau, \tau+k-1]$.

The dynamic regret[4] $\widetilde{\mathcal{O}}(V_{[\tau,\tau+k-1]}^{1/3} k^{2/3} + \sqrt{k})$ obtained in Proposition B.1 matches minimax optimal rate (Besbes et al., 2015) for general online convex optimization problems under gradient feedback; though we remark that faster rates are achievable under full-information feedback of the entire loss function (Zhao & Zhang, 2021, Theorem 3).

Comparison of dynamic regret with FACI Dividing (20) by k, we obtain the following average dynamic regret bound for SAOCP on $[\tau, \tau+k-1]$:

$$\widetilde{\mathcal{O}}\Big(D\Big[(V_{[\tau,\tau+k-1]}/k)^{1/3} + 1/\sqrt{k}\Big]\Big),$$

simultaneously for all lengths k and $\tau \in [T-k+1]$.

[4] More precisely, the intermediate result with $V_{[\tau,\tau+k-1]}$ replaced by the standard total variation of losses $\widetilde{V}_{[\tau,\tau+k-1]}$ in (21).

In comparison, the FACI algorithm (adapted to our setting) with learning rate η acheives average dynamic regret bound (Gibbs & Candès, 2022, Theorem 3.2)

$$\widetilde{\mathcal{O}}\Big(D\Big[(V_{[\tau,\tau+k-1]}/k)^{1/2}+\eta/D+D/(\eta k)\Big]\Big).$$

When the path length $V_{[\tau,\tau+k-1]}=o(k)$, FACI achieves a better dependence on the average path length ($V_{[\tau,\tau+k-1]}/k=o(1)$, yet a worse dependence on k itself due to the inability to choose the optimal η simultaneously for all k, similar as the comparison of their SARegret bounds (Section 4.1).

Proof of Proposition B.1 We apply the dynamic regret bound of Zhang et al. (2018, Corollary 5) for the SAOCP algorithm on the interval $[\tau,\tau+k-1]$, and note that our iterates $\widehat{s}_t\in[-\eta,D+\eta]\subset[-D,2D]$ by Lemma A.1 and our choice $\eta=D/\sqrt{3}$ in Algorithm 1. Therefore we obtain

$$\sum_{t=\tau}^{\tau+k-1}\ell^{(t)}(\widehat{s}_t)-\min_{s^\star_{\tau:\tau+k-1}}\sum_{t=\tau}^{\tau+k-1}\ell^{(t)}(s_t^\star)\le\widetilde{\mathcal{O}}\Big(D\Big[\widetilde{V}_{[\tau,\tau+k-1]}^{1/3}k^{2/3}+\sqrt{k}\Big]\Big),$$

where

$$\widetilde{V}_{[\tau,\tau+k-1]}=\sum_{t=\tau+1}^{\tau+k-1}\sup_{s'\in[0,D]}\Big|\ell^{(t)}(s')-\ell^{(t-1)}(s')\Big|\overset{(i)}{\le}\sum_{t=\tau+1}^{\tau+k-1}|S_t-S_{t-1}|=V_{[\tau,\tau+k-1]},\tag{21}$$

where (i) follows by the fact that $\big|\ell^{(t)}(s')-\ell^{(t-1)}(s')\big|=|\ell_{1-\alpha}(s',S_t)-\ell_{1-\alpha}(s',S_{t-1})|\le|S_t-S_{t-1}|$ by the 1-Lipschitzness of the quantile loss (1) with respect to the second argument. This proves the desired result. □

B.3. Proof of Theorem 4.2

We first note that, by (2) and (10), Algorithm 2 simplifies to the update

$$\widehat{s}_{t+1}=\widehat{s}_t+\eta\frac{\mathrm{err}_t-\alpha}{\sqrt{\sum_{s=1}^t(\mathrm{err}_s-\alpha)^2}}=\widehat{s}_1+\eta\sum_{s=1}^t\frac{\mathrm{err}_s-\alpha}{\sqrt{\sum_{i=1}^s(\mathrm{err}_i-\alpha)^2}}.\tag{22}$$

Note that we have $\widehat{s}_{t+1}\in[-\eta,D+\eta]$ for all $t\ge0$ (Lemma A.1), which implies that

$$\left|\sum_{t=t_0+1}^{t_f}\frac{\mathrm{err}_t-\alpha}{\sqrt{\sum_{s=1}^t(\mathrm{err}_s-\alpha)^2}}\right|=\frac{1}{\eta}\big|\widehat{s}_{t_f+1}-\widehat{s}_{t_0+1}\big|\le\frac{D+2\eta}{\eta}\quad\text{for any }0\le t_0<t_f.$$

Note that $|\mathrm{err}_t-\alpha|\in[\alpha,1]$ for all t. Therefore, we can invoke Lemma B.2 below with $a_t=\mathrm{err}_t-\alpha$ and $M=(D+2\eta)/\eta$ to obtain that for any $T\ge1$,

$$\left|\frac{1}{T}\sum_{t=1}^T\mathrm{err}_t-\alpha\right|\le2\left(\frac{D+3\eta}{\eta}+\alpha^{-2}\log T\right)T^{-1/4}\le\mathcal{O}(\alpha^{-2}T^{-1/4}\log T),$$

where the later bound holds for any $\eta=\Theta(D)$. This proves Theorem 4.2. □

Lemma B.2. *Suppose the sequence $\{a_t\}_{t\in[T]}\in\mathbb{R}$ satisfies $\alpha\le|a_t|\le1$ for some $\alpha>0$, and*

$$\left|\sum_{t=t_0+1}^{t_f}\frac{a_t}{\sqrt{\sum_{s=1}^t a_s^2}}\right|\le M\quad\textit{for any }\ 0\le t_0<t_f\le T.$$

Then we have

$$\left|\frac{1}{T}\sum_{t=1}^T a_t\right|\le2\big(M+1+\alpha^{-2}\log T\big)T^{-1/4}.$$

Proof. The proof builds on a grouping argument. Define integers

$$L = \lceil T^{\beta} \rceil, \quad K = \lceil T/L \rceil \le T^{1-\beta} + 1,$$

where $\beta \in (0,1)$ is a parameter to be chosen. For any $k \in [K]$, define the k-th group to be

$$G_k = \{t_{k-1}+1, \dots, t_k\} := \{(k-1)L+1, \dots, \min\{kL, T\}\}, \tag{23}$$

so that we have $\bigcup_{k=1}^{K} G_k = [T]$, $|G_k| = L$ for all $k \in [K-1]$, and $|G_K| \le L$.

Next, for any fixed $k \ge 2$, define sums

$$S_k := \sum_{t \in G_k} \frac{a_t}{\sqrt{\sum_{s=1}^{t} a_s^2}}, \quad \widetilde{S}_k := \sum_{t \in G_k} \frac{a_t}{\sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}}.$$

By our precondition, we have $|S_k| \le M$ for all $k \in [K]$. Further, we have

$$\begin{aligned}
\left|S_k - \widetilde{S}_k\right| &\le \sum_{t \in G_k} |a_t| \cdot \left(\frac{1}{\sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}} - \frac{1}{\sqrt{\sum_{s=1}^{t} a_s^2}} \right) \le |G_k| \cdot \left(\frac{1}{\sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}} - \frac{1}{\sqrt{\sum_{s=1}^{t_k} a_s^2}} \right) \\
&\overset{(i)}{\le} L \cdot \frac{\sum_{s=t_{k-1}+1}^{t_k} a_s^2}{2\left(\sum_{s=1}^{t_{k-1}} a_s^2\right)^{3/2}} \overset{(ii)}{\le} L \cdot \frac{L}{2(\alpha^2(k-1)L) \cdot \left(\sum_{s=1}^{t_{k-1}} a_s^2\right)^{1/2}} = \frac{L}{2\alpha^2(k-1) \cdot \sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}},
\end{aligned}$$

where (i) uses the inequality $\frac{1}{\sqrt{x}} - \frac{1}{\sqrt{x+y}} \le \frac{y}{2x^{3/2}}$ for $x, y \ge 0$, and (ii) uses the bounds $\sum_{s=t_{k-1}+1}^{t_k} a_s^2 \le (t_k - t_{k-1}) \le L$ and $\sum_{s=1}^{t_{k-1}} a_s^2 \ge \alpha^2 t_{k-1} = \alpha^2(k-1)L$. By the triangle inequality, this implies that

$$\left|\widetilde{S}_k\right| \le |S_k| + \left|\widetilde{S}_k - S_k\right| \le M + \frac{L}{2\alpha^2(k-1) \cdot \sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}},$$

and thus for any $k \ge 2$ that

$$\begin{aligned}
\left|\sum_{t \in G_k} a_t\right| &= \underbrace{\left|\sum_{t \in G_k} \frac{a_t}{\sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}}\right|}_{|\widetilde{S}_k|} \cdot \sqrt{\sum_{s=1}^{t_{k-1}} a_s^2} \le \left(M + \frac{L}{2\alpha^2(k-1) \cdot \sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}} \right) \cdot \sqrt{\sum_{s=1}^{t_{k-1}} a_s^2} \\
&\le M \sqrt{\sum_{s=1}^{t_{k-1}} a_s^2} + \frac{L}{2\alpha^2(k-1)} \le M\sqrt{(k-1)L} + \frac{L}{2\alpha^2(k-1)}.
\end{aligned}$$

For $k = 1$, we have trivially $\left|\sum_{t \in G_1} a_t\right| \le |G_1| \le L$. Summing the bounds over $k \in [K]$ yields

$$\begin{aligned}
\left|\sum_{t=1}^{T} a_t\right| &\le L + \sum_{k=1}^{K} \left|\sum_{t \in G_k} a_t\right| \le L + M\sqrt{L} \cdot \sum_{k=2}^{K} \sqrt{k-1} + \frac{L}{2\alpha^2} \sum_{k=2}^{K} \frac{1}{k-1} \\
&\le L + \frac{2}{3} M\sqrt{L} K^{3/2} + \frac{L}{2\alpha^2} \log_2 K \\
&\le \lceil T^{\beta} \rceil + \frac{2}{3} M \sqrt{\lceil T^{\beta} \rceil} \cdot T^{3(1-\beta)/2} + \frac{1}{2\alpha^2} \lceil T^{\beta} \rceil \log_2(T^{1-\beta}) \\
&\le 2T^{\beta} + 2MT^{3/2-\beta} + \frac{2}{\alpha^2} T^{\beta} \log T.
\end{aligned}$$

Choosing $\beta = 3/4$, we obtain

$$\left|\sum_{t=t_0+1}^{t_f} a_t\right| \le 2\big(M + 1 + \log T/\alpha^2\big) T^{3/4}.$$

Dividing by T on both sides yields the desired result. □

B.4. Coverage of SAOCP

Theorem B.3 (Coverage bound for SAOCP)**.** *Consider a randomized version of Algorithm 1 where Line 7 is changed to sampling an expert $i \sim p_{t,\cdot} \in \Delta([t])$ and outputting radius $\widehat{s}_{t,i}$. Consider the corresponding expected miscoverage error*

$$\widetilde{\mathrm{err}}_t := \sum_{i=1}^{t} p_{t,i} \underbrace{\mathbb{1}\left[\widehat{s}_{t,i} < S_t\right]}_{:=\mathrm{err}_{t,i}}. \tag{24}$$

Then we have for any $T \geq 1$ that

$$\left|\frac{1}{T}\sum_{t=1}^{T}\widetilde{\mathrm{err}}_t - \alpha\right| \leq \mathcal{O}\left(\inf_{\beta\in(1/2,1)}\left\{T^{1/2-\beta} + T^{\beta-1} \times \underbrace{\left(1 + \sum_{j=2}^{\lceil T^{1-\beta}\rceil} \max_{t\in G_j}\sum_{i=1}^{t}\left|p_{t,i} - p_{t_{j-1},i}\frac{G^i_{i:t_{j-1}}}{G^i_{i:t}}\right|\right)}_{S_\beta(\{p_t\}_{t\in[T]},\{\sum_{\tau=i}^{t}\|\nabla\ell^{(\tau)}(\widehat{s}_{\tau,i})\|_2^2\}_{i\leq t}\}) =: S_\beta(T)}\right\}\right)$$

(understanding $p_{t_{j-1},i} := 0$ for any $i > t_{j-1}$), where for each $\beta \in (1/2,1)$, $\{G_j\}_{j=1}^{\lceil T^{1-\beta}\rceil}$ with $|G_j| \leq \lceil T^\beta\rceil$, $G_j = \{t_{j-1}+1,\ldots,\min\{t_j,T\}\}$ is the even grouping of $[T]$ as in (23), and

$$G^i_{i:t} := \sqrt{\sum_{\tau=i}^{t}\left\|\nabla\ell^{(\tau)}(\widehat{s}_{\tau,i})\right\|_2^2} = \sqrt{\sum_{\tau=i}^{t}(\mathrm{err}_{\tau,i}-\alpha)^2}$$

is the cumulative squared gradients received by expert $\mathcal{A}_i$ for any $t > i$ (understanding experts as running until time T even after they become inactive).

Proof. Fix any $i \in [T]$. As Algorithm 1 chooses each expert $\mathcal{A}_i$ to be SF-OGD (Algorithm 2), we have by (10) that for all $t \geq i$,

$$\widehat{s}_{t+1,i} - \widehat{s}_{t,i} = \frac{\eta}{G^i_{i:t}} \cdot (\mathrm{err}_{t,i} - \alpha). \tag{25}$$

Now fix any $\beta \in (1/2,1)$. For any group $2 \leq j \leq \lceil T^{1-\beta}\rceil$ and $t \in G_j$, plugging the above into definition (24) gives that

$$\begin{aligned}
&\widetilde{\mathrm{err}}_t - \alpha = \sum_{i=1}^{t} p_{t,i}(\mathrm{err}_{t,i}-\alpha) = \frac{1}{\eta}\sum_{i=1}^{t} p_{t,i}G^i_{i:t}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i}) \\
&= \frac{1}{\eta}\sum_{i=1}^{t_{j-1}} p_{t_{j-1},i}G^i_{i:t_{j-1}}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i}) + \frac{1}{\eta}\sum_{i=1}^{t}\left(p_{t,i}G^i_{i:t} - p_{t_{j-1},i}G^i_{i:t_{j-1}}\right)(\widehat{s}_{t+1,i} - \widehat{s}_{t,i}) \\
&= \frac{1}{\eta}\sum_{i=1}^{t_{j-1}} p_{t_{j-1},i}G^i_{i:t_{j-1}}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i}) + \frac{1}{\eta}\sum_{i=1}^{t}\left(p_{t,i} - p_{t_{j-1},i}\frac{G^i_{i:t_{j-1}}}{G^i_{i:t}}\right)\cdot G^i_{i:t}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i}).
\end{aligned}$$

Summing this over $t \in G_j$ and noting that the coefficients $p_{t_{j-1},i}G^i_{i:t_{j-1}}$ in the first sum does not depend on t, we get

$$\begin{aligned}
&\left|\sum_{t\in G_j}(\widetilde{\mathrm{err}}_t - \alpha)\right| \\
&\leq \left|\frac{1}{\eta}\sum_{i=1}^{t_{j-1}} p_{t_{j-1},i}G^i_{i:t_{j-1}}(\widehat{s}_{t_j+1,i} - \widehat{s}_{t_{j-1}+1,i})\right| + |G_j|\cdot\max_{t\in G_j}\left|\frac{1}{\eta}\sum_{i=1}^{t}\left(p_{t,i} - p_{t_{j-1},i}\frac{G^i_{i:t_{j-1}}}{G^i_{i:t}}\right)\cdot G^i_{i:t}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i})\right| \\
&\leq \frac{1}{\eta}\max_{i\in[t_{j-1}]} G^i_{i:t_{j-1}}\left|\widehat{s}_{t_j+1,i} - \widehat{s}_{t_{j-1}+1,i}\right| + |G_j|\cdot\max_{t\in G_j}\sum_{i=1}^{t}\left|\frac{1}{\eta}\left(p_{t,i} - p_{t_{j-1},i}\frac{G^i_{i:t_{j-1}}}{G^i_{i:t}}\right)\cdot G^i_{i:t}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i})\right|
\end{aligned}$$

$$\overset{(i)}{\leq} \frac{D+2\eta}{\eta}\sqrt{T} + |G_j| \cdot \max_{t\in G_j} \sum_{i=1}^{t} \left| p_{t,i} - p_{t_{j-1},i} \frac{G^i_{i:t_{j-1}}}{G^i_{i:t}} \right|$$
$$\overset{(ii)}{\leq} C\sqrt{T} + |G_j| \cdot \max_{t\in G_j} \sum_{i=1}^{t} \left| p_{t,i} - p_{t_{j-1},i} \frac{G^i_{i:t_{j-1}}}{G^i_{i:t}} \right|$$

Above, (i) used $G^i_{i:t_{j-1}} \leq \sqrt{t_{j-1}-i+1} \leq \sqrt{T}$ by the definition of $G^i_{i:t_{j-1}}$, the bound $\left|\widehat{s}_{t_j+1,i} - \widehat{s}_{t_{j-1}+1,i}\right| \leq (D+2\eta)$ which follows by the fact that each expert is initialized within $[-\eta, D+\eta]$ and applying Lemma A.1, and the bound $\left|G^i_{i:t}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i})\right| \leq \eta$ by (25); (ii) used the fact that $\eta = D/\sqrt{3}$ in Algorithm 1, so that $(D+2\eta)/\eta = 2+\sqrt{3} =: C$ is an absolute constant. Also, note that for group $j=1$, we directly have

$$\left| \sum_{t\in G_1} (\widetilde{\mathrm{err}}_t - \alpha) \right| \leq |G_1|.$$

Summing all the above bounds over $j \in \left[\lceil T^{1-\beta} \rceil\right]$ gives

$$\left| \sum_{t=1}^{T} (\widetilde{\mathrm{err}}_t - \alpha) \right| \leq \sum_{j=1}^{\lceil T^{1-\beta}\rceil} \left| \sum_{t\in G_j} (\widetilde{\mathrm{err}}_t - \alpha) \right|$$
$$\leq \mathcal{O}\left(T^{3/2-\beta} + |G_1| + \sum_{j=2}^{\lceil T^{1-\beta}\rceil} |G_j| \times \max_{t\in G_j} \sum_{i=1}^{t} \left| p_{t,i} - p_{t_{j-1},i} \frac{G^i_{i:t_{j-1}}}{G^i_{i:t}} \right| \right)$$
$$\leq \mathcal{O}\left(T^{3/2-\beta} + T^{\beta} \left(1 + \sum_{j=2}^{\lceil T^{1-\beta}\rceil} \max_{t\in G_j} \sum_{i=1}^{t} \left| p_{t,i} - p_{t_{j-1},i} \frac{G^i_{i:t_{j-1}}}{G^i_{i:t}} \right| \right) \right)$$

Dividing both sides by T proves the desired bound for this fixed β. Further taking supremum over $\beta \in (1/2, 1)$ gives the desired result.

□

B.4.1. Discussions & SF-OGD as a Special Case

We first note that, the proof of Theorem B.3 does not rely on the specific structure of either the expert weights $\{p_{t,i}\}_{i<t}$ or the active intervals. Therefore, the result of Theorem B.3 holds generically for any other aggregation scheme over experts with arbitrary active intervals, in addition to that specified in Algorithm 1.

In particular, by setting $p_{t,1} = 1$ and $p_{t,i} = 0$ for $i \geq 2$, and defining the first expert $\mathcal{A}_1$ to be active over $[T]$, Algorithm 1 (either with or without the randomization, since there is only one active expert) recovers Algorithm 2. In this case, we show that $S_\beta(T) \leq \widetilde{\mathcal{O}}(\alpha^{-2})$ for any $\beta \in (1/2, 1)$, so that Theorem B.3 (and its informal version in Theorem 4.3) indeed subsumes Theorem 4.2 as a special case by choosing $\beta = 3/4$, as claimed in Section 4.2.

We have

$$S_\beta(T) = 1 + \sum_{j=2}^{\lceil T^{1-\beta}\rceil} \max_{t\in G_j} \sum_{i=1}^{t} \left| p_{t,i} - p_{t_{j-1},i} \frac{G^i_{i:t_{j-1}}}{G^i_{i:t}} \right| \overset{(i)}{=} 1 + \sum_{j=2}^{\lceil T^{1-\beta}\rceil} \max_{t\in G_j} \left| 1 - \frac{G^1_{1:t_{j-1}}}{G^1_{1:t}} \right|, \tag{26}$$

where (i) used the fact that $p_{t,1} = 1$ and $p_{t,i} = 0$ for $i \geq 2$. For any $t \in G_j$, we have

$$\left| 1 - \frac{G^1_{1:t_{j-1}}}{G^1_{1:t}} \right| = 1 - \frac{\sqrt{\sum_{s=1}^{t_{j-1}} (\mathrm{err}_s - \alpha)^2}}{\sqrt{\sum_{s=1}^{t} (\mathrm{err}_s - \alpha)^2}} \overset{(i)}{\leq} \frac{\sum_{s=t_{j-1}+1}^{t} (\mathrm{err}_s - \alpha)^2}{2\sum_{s=1}^{t_{j-1}} (\mathrm{err}_s - \alpha)^2} \overset{(ii)}{\leq} \frac{t - t_{j-1}}{2\alpha^2 t_{j-1}} \overset{(iii)}{\leq} \frac{\lceil T^\beta \rceil}{2\alpha^2 \cdot (j-1)\lceil T^\beta \rceil} = \frac{1}{2\alpha^2(j-1)},$$

where (i) follows from the inequality $1 - \frac{\sqrt{x}}{\sqrt{x+y}} = \frac{\sqrt{x+y}-\sqrt{x}}{\sqrt{x+y}} \leq \frac{\sqrt{x+y}-\sqrt{x}}{\sqrt{x}} = \sqrt{1+\frac{y}{x}} - 1 \leq \frac{y}{2x}$ for any $x, y \geq 0$; (ii) follows by the bound $|\mathrm{err}_s - \alpha| \in [\alpha, 1]$ for any s; (iii) follows by definition of the grouping (23). Plugging the above bound

into (26), we obtain

$$S_\beta(T) \le 1 + \sum_{j=2}^{\lceil T^{1-\beta} \rceil} \frac{1}{2\alpha^2(j-1)} \le \mathcal{O}\big(\alpha^{-2} \log T\big) = \widetilde{\mathcal{O}}(\alpha^{-2}),$$

proving the claim.

C. Distribution-Aware Coverage Guarantees for SAOCP

In this section, we show that under mild density lower bound assumptions on the true radii, a probabilistic variant of the coverage error of SAOCP (Algorithm 1) is bounded by $\widetilde{\mathcal{O}}(k^{-1/(2q)}) + \widetilde{\mathcal{O}}((\mathrm{Var}_k/k)^{1/q})$ for every interval of length k, where $q \ge 2$ is a parameter of the density lower bound assumption, and Var_k measures a certain variance (over intervals of length k) in the $1-\alpha$ conditional quantiles of the true radii. The proof builds on the strongly adaptive regret guarantee (in the quantile loss) for SAOCP (Proposition 4.1), and bounding parameter estimation errors by excess quantile losses using a *self-calibration inequality* type argument (Steinwart & Christmann, 2011).

Setting We consider the online conformal prediction setting described in Section 2. For any $t \ge 1$, let $\mathcal{F}_t := \sigma(\{X_i, \widehat{s}_i, S_i\}_{i\in[t-1]}, X_t)$ be the σ-algebra by all observed data $\{(X_i, \widehat{s}_i, S_i)\}_{i\le t-1}$ as well as X_t. Note that by definition of the online conformal prediction setting, the predicted radius $\widehat{s}_t$ can only depend on information within $\mathcal{F}_t$ as well as (possibly) external randomness. Consequently, we have $S_t \perp\!\!\!\perp \widehat{s}_t \mid \mathcal{F}_t$, i.e. S_t and $\widehat{s}_t$ are conditionally independent given $\mathcal{F}_t$.

We now state our assumptions on the distributions of the true radii.

Assumption C.1 (Density upper bounds)**.** For all $t \in [T]$, there exists a constant $L > 0$ such that $S_t \mid \mathcal{F}_t$ is a continuous random variable that is bounded within $[0, D]$ and has a density $f_t : [0, D] \to \mathbb{R}_{\ge 0}$ with $f_t(s) \le L/D$ for all $s \in [0, D]$.

Assumption C.2 (Density lower bounds)**.** For all $t \in [T]$, $S_t \mid \mathcal{F}_t$ is a continuous random variable that is bounded within $[0, D]$ and has a density $f_t : [0, D] \to \mathbb{R}_{\ge 0}$. With probability one, there exist constants $b > 0, q \ge 2, \Delta_t > 0$ such that

$$f_t(s) \ge \frac{2b}{D}\left|\frac{2(s - s_t^\star)}{D}\right|^{q-2} \tag{27}$$

for all $s \in [s_t^\star - \Delta_t, s_t^\star + \Delta_t]$, where

$$s_t^\star := Q_{1-\alpha}(S_t \mid \mathcal{F}_t) \tag{28}$$

is the $1-\alpha$ conditional quantile of S_t.

As examples for Assumption C.2, the case where $q = 2$ corresponds to a constant lower bound on the conditional density $f_t(\cdot)$ *locally* around $s_t^\star$, which holds e.g. if each $f_t(\cdot)$ itself has a constant lower bound over $[0, D]$ (this is the assumption made by Gibbs & Candès (2022)). A larger q makes the density lower bound (27) easier to satisfy and thus specifies a more relaxed assumption. We also note that $s_t^\star$ is itself a random variable which is measurable on $\mathcal{F}_t$.

For any interval $I = [\tau, \tau + k - 1] \subseteq [T]$, our coverage result depends on a certain variance between $s_\tau^\star, \ldots, s_{\tau+k-1}^\star$. Concretely, define the interval quantile variation

$$\mathrm{Var}_I := \sum_{t=\tau}^{\tau+k-1} \mathbb{E}\left[\left(\frac{s_t^\star}{D} - \frac{1}{kD}\sum_{i=\tau}^{\tau+k-1} \mathbb{E}[s_i^\star \mid \mathcal{F}_\tau]\right)^2\right]. \tag{29}$$

Then, the expected absolute difference between SAOCP's predictions $\widehat{s}_\tau, \ldots, \widehat{s}_{\tau+k-1}$ and the true conditional quantiles $s_\tau^\star, \ldots, s_{\tau+k-1}^\star$ is $\widetilde{\mathcal{O}}(k^{-1/(2q)}) + \widetilde{\mathcal{O}}(k^{-1/q}\mathrm{Var}_I^{1/q})$. Due to the Lipschitzness of the CDFs (by Assumption C.1), the coverage error has a similar form. So SAOCP achieves better coverage when the interval quantile variation is lower, and it achieves approximately valid coverage as long as $\mathrm{Var}_I \le o(|I|)$. More formally, we have:

Theorem C.3. *Let Assumptions C.1 & C.2 hold. Fix any interval $I = [\tau, \tau + k - 1] \subseteq [T]$. Then, letting $\Delta_I = \min\{\Delta_t : t \in I\}$, Algorithm 1 achieves quantile estimation error*

$$\frac{1}{|I|}\sum_{t\in I} \mathbb{E}\left[\left|\frac{\widehat{s}_t - s_t^\star}{D}\right|\right] \le \mathcal{O}\left(\frac{1}{b^{1/q}}\frac{D}{\Delta_I}\left(\frac{\log T}{|I|}\right)^{1/(2q)}\right) + \mathcal{O}\left(\frac{L^{1/q}}{b^{1/q}}\frac{D}{\Delta_I}\left(\frac{\mathrm{Var}_I}{|I|}\right)^{1/q}\right)$$

and interval miscoverage error

$$\frac{1}{|I|}\sum_{t\in I}\left|\mathbb{P}[Y_t \in \widehat{C}_t(X_t)] - (1-\alpha)\right| \le \mathcal{O}\left(\frac{L}{b^{1/q}}\frac{D}{\Delta_I}\left(\frac{\log T}{|I|}\right)^{1/(2q)}\right) + \mathcal{O}\left(\frac{L^{1+1/q}}{b^{1/q}}\frac{D}{\Delta_I}\left(\frac{\mathrm{Var}_I}{|I|}\right)^{1/q}\right).$$

In Theorem C.3, q is a parameter quantifying the difficulty of lower bounding the distribution of $S_t \mid \mathcal{F}_t$ away from its $1-\alpha$ conditional quantile $s_t^\star$. If q is higher, then closeness to $s_t^\star$ is less correlated with the expected regret on the quantile loss (1). Meanwhile, the term D/Δ_I grows larger as α grows smaller. The inclusion of this term mirrors the inclusion of α^{-2} in Theorem 4.2, and it indicates that more extreme quantiles are harder to learn.

Proof of Theorem C.3 The key ingredient is the technical Lemma C.4, which uses the expected dynamic regret of a sequence $\widehat{s}_\tau, \dots, \widehat{s}_{\tau+k-1}$ to upper bound the expected distance between that sequence and the true $1-\alpha$ conditional quantiles of $S_\tau, \dots, S_{\tau+k-1}$. We decompose the expected dynamic regret into the interval regret of Algorithm 1 (which we can upper bound by Proposition 4.1) and a term which we can use Var_I to upper bound. The desired coverage bound follows by the Lipschitzness of the CDFs of the S_t's (implied by the density upper bound Assumption C.1). In this proof, we use $\mathbb{E}_{\mathcal{F}_t}[X]$ as short-hand for the conditional expectation $\mathbb{E}[X \mid \mathcal{F}_t]$.

Lemma C.4 (Bounding quantile estimation error by dynamic regret)**.** *Fix any interval $I = [\tau, \tau+k-1] \subseteq [T]$, and let $\Delta_I = \min\{\Delta_t : t \in I\}$. Under Assumption C.2, we have*

$$\sum_{t\in I}\mathbb{E}\left[\left|\frac{\widehat{s}_t - s_t^\star}{D}\right|\right]^q \le \frac{2q(q-1)}{bD}\left(\frac{D}{2\Delta_I}\right)^q \sum_{t\in I}\mathbb{E}[\ell_{1-\alpha}(S_t, \widehat{s}_t) - \ell_{1-\alpha}(S_t, s_t^\star)].$$

To prove Theorem C.3, we follow a similar technique to Zhang et al. (2018) and decompose the expected dynamic regret

$$\begin{aligned}\mathbb{E}\left[\sum_{t\in I}\ell_{1-\alpha}(S_t,\widehat{s}_t) - \ell_{1-\alpha}(S_t, s_t^\star)\right] &= \underbrace{\mathbb{E}\left[\sum_{t\in I}\ell_{1-\alpha}(S_t,\widehat{s}_t) - \inf_s\sum_{t\in I}\ell_{1-\alpha}(S_t,s)\right]}_{A} \\ &\quad + \underbrace{\mathbb{E}\left[\inf_s\sum_{t\in I}\ell_{1-\alpha}(S_t,s) - \ell_{1-\alpha}(S_t,s_t^\star)\right]}_{B}.\end{aligned}$$

We first observe that term A is simply the expected interval regret on I. Since Proposition 4.1 bounds the strongly adaptive regret with probability one, we can bound $A \le 15D\sqrt{|I|(\log T + 1)}$. Now, we analyze term B, and note that

$$\mathbb{E}\left[\inf_s\sum_{t\in I}\ell_{1-\alpha}(S_t,s) - \ell_{1-\alpha}(S_t,s_t^\star)\right] \le \inf_s \mathbb{E}\left[\sum_{t\in I}\mathbb{E}_{\mathcal{F}_t}[\ell_{1-\alpha}(S_t,s) - \ell_{1-\alpha}(S_t,s_t^\star))]\right]$$

by Jensen's inequality and the tower property of conditional expectation. Now, for any $t \in I$ and $s \in [0, D]$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{F}_t}[\ell_{1-\alpha}(S_t,s) - \ell_{1-\alpha}(S_t,s_t^\star)] \\
&= \mathbb{E}_{\mathcal{F}_t}[(1-\alpha)(S_t-s)\mathbb{1}[S_t > s] + \alpha(s-S_t)\mathbb{1}[S_t \le s] - (1-\alpha)(S_t - s_t^\star)\mathbb{1}[S_t > s_t^\star] - \alpha(s_t^\star - S_t)\mathbb{1}[S_t \le s_t^\star]] \\
&= \begin{cases}\mathbb{E}_{\mathcal{F}_t}[(S_t-s)\mathbb{1}[S_t > s] - \alpha(S_t - s) - \alpha(s_t^\star - S_t)] & S_t \le s_t^\star \\ \mathbb{E}_{\mathcal{F}_t}[(s-S_t)\mathbb{1}[S_t \le s] - (1-\alpha)(s-S_t) - (1-\alpha)(s_t^\star - S_t)] & S_t > s_t^\star\end{cases} \\
&= \begin{cases}\mathbb{E}_{\mathcal{F}_t}[(S_t-s)\mathbb{1}[s < S_t \le s_t^\star] + \alpha(s - s_t^\star)] & S_t \le s_t^\star \\ \mathbb{E}_{\mathcal{F}_t}[(s-S_t)\mathbb{1}[s_t^\star \le S_t \le s] - (1-\alpha)(s-s_t^\star)] & S_t > s_t^\star\end{cases} \\
&\overset{(i)}{=} \mathbb{E}_{\mathcal{F}_t}[(S_t-s)\mathbb{1}[s \le S_t \le s_t^\star] + (s-S_t)\mathbb{1}[s_t^\star \le S_t \le s]] + \alpha(s-s_t^\star)\mathbb{P}_{\mathcal{F}_t}[S_t \le s_t^\star] - (1-\alpha)(s-s_t^\star)\mathbb{P}_{\mathcal{F}_t}[S_t \le s_t^\star] \\
&\overset{(ii)}{=} \mathbb{E}_{\mathcal{F}_t}[(S_t-s)\mathbb{1}[s \le S_t \le s_t^\star] + (s-S_t)\mathbb{1}[s_t^\star \le S_t \le s]] \\
&= \left|\int_s^{s_t^\star}(x-s)f_t(x)\mathrm{d}x\right| \overset{(iii)}{\le} \frac{L(s_t^\star - s)^2}{2D}
\end{aligned}$$

Above, (i) uses the fact that $s_t^\star$ is $\mathcal{F}_t$-measurable, (ii) uses the fact that $\mathbb{P}_{\mathcal{F}_t}[S_t \le s_t^\star] = 1-\alpha$ by definition, and (iii) uses the density upper bound $f_t(x) \le \frac{L}{D}$ (Assumption C.1). Therefore, taking the infimum over $s \in \mathbb{R}$ and by definition of Var_I (29), we have $B \le \frac{LD}{2}\mathrm{Var}_I$, and

$$\mathbb{E}\left[\sum_{t\in I} \ell_{1-\alpha}(S_t, \widehat{s}_t) - \ell_{1-\alpha}(S_t, s_t^\star)\right] \le 15D\sqrt{|I|(\log T + 1)} + \frac{LD}{2}\mathrm{Var}_I.$$

We combine this result with the power-mean inequality, Lemma C.4, and the facts that $(q(q-1))^{1/q} = \mathcal{O}(1)$ and $(x+y)^{1/q} \le x^{1/q} + y^{1/q}$ to prove the first part of Theorem C.3,

$$\frac{1}{|I|}\sum_{t\in I}\mathbb{E}\left[\left|\frac{\widehat{s}_t - s_t^\star}{D}\right|\right] \le \left(\frac{1}{|I|}\sum_{t\in I}\mathbb{E}\left[\left|\frac{\widehat{s}_t - s_t^\star}{D}\right|\right]^q\right)^{1/q} \le \mathcal{O}\left(\frac{D}{b^{1/q}\Delta_I}\left(\frac{\log T}{|I|}\right)^{1/(2q)}\right) + \mathcal{O}\left(\frac{D}{\Delta_I}\left(\frac{L\mathrm{Var}_I}{b|I|}\right)^{1/q}\right). \quad (30)$$

To prove the desired coverage bound, we note that

$$\frac{1}{|I|}\sum_{t\in I}\left|\mathbb{P}[Y_t \in \widehat{C}_t(X_t)] - (1-\alpha)\right| = \frac{1}{|I|}\sum_{t\in I}|\mathbb{E}[\mathbb{E}_{\mathcal{F}_t}[\mathbb{1}[S_t \le \widehat{s}_t] - \mathbb{1}[S_t \le s_t^\star]]]| \le \frac{1}{|I|}\sum_{t\in I}\mathbb{E}[|F_t(\widehat{s}_t) - F_t(s_t^\star)|],$$

where the final inequality uses Jensen's inequality and the fact that $\mathbb{E}_{\mathcal{F}_t}[\mathbb{1}[S_t \le s]] = F_t(s)$ is the CDF of $S_t \mid \mathcal{F}_t$. The result follows by combining (30) with Assumption C.1, which implies that the CDF F_t is L/D-Lipschitz. □

Proof of Lemma C.4 We first consider the following general situation, where S is any continuous random variable bounded in $[0, D]$ with a density f. Define $s^\star = Q_{1-\alpha}(S)$. Let g be the density of the normalized random variable $X = \frac{2S-D}{D} \in [-1, 1]$, and let $x^\star = Q_{1-\alpha}(X) = \frac{2s^\star - D}{D}$. As in Steinwart & Christmann (2011, Example 2.3), assume that there exist constants $b > 0, q \ge 2, \Delta > 0$ such that

$$g(x) \ge b|x - x^\star|^{q-2} \iff f(s) \ge \frac{2b}{D}\left|\frac{2(s-s^\star)}{D}\right|^{q-2} \quad (31)$$

for all $s \in [s^\star - \Delta, s^\star + \Delta]$. Let $\beta = \frac{b}{q-1}$ and $\gamma = \beta(\frac{2\Delta}{D})^{q-1}$. By Steinwart & Christmann (2011, Theorem 2.7),

$$\begin{aligned}
|x - x^\star| &\le 2^{1-1/q} q^{1/q} \gamma^{-1/q} (\mathbb{E}[\ell_{1-\alpha}(X, x) - \ell_{1-\alpha}(X, x^\star)])^{1/q} \\
&= 2\left(\frac{q}{2\gamma}\mathbb{E}[\ell_{1-\alpha}(X, x) - \ell_{1-\alpha}(X, x^\star)]\right)^{1/q} \\
&= 2\left(\frac{q(q-1)}{b}\left(\frac{D}{2\Delta}\right)^{q-1}\mathbb{E}[\ell_{1-\alpha}(X, x) - \ell_{1-\alpha}(X, x^\star)]\right)^{1/q}
\end{aligned}$$

Since $|x - x^\star| = \frac{2}{D}|s - s^\star|$, $\ell_{1-\alpha}(X, x) = \frac{2}{D}\ell_{1-\alpha}(S, s)$, and we can obtain

$$\left|\frac{s - s^\star}{D}\right|^q \le \frac{2q(q-1)}{bD}\left(\frac{D}{2\Delta}\right)^q \mathbb{E}[\ell_{1-\alpha}(S, s) - \ell_{1-\alpha}(S, s^\star)].$$

Lemma C.4 now follows by fixing any $t \in I$, defining $\mathcal{G}_t = \sigma(\mathcal{F}_t, \widehat{s}_t)$ and noticing that $S_t \mid \mathcal{G}_t \overset{\text{dist}}{=} S_t \mid \mathcal{F}_t$ (as $\mathcal{G}_t$ only involves possibly an additional "external" randomness of the online prediction algorithm over $\mathcal{F}_t$), so $s_t^\star = Q_{1-\alpha}(S_t \mid \mathcal{F}_t) = Q_{1-\alpha}(S_t \mid \mathcal{G}_t)$. Since $\widehat{s}_t$ and $s_t^\star$ are $\mathcal{G}_t$-measurable, we can bound

$$\mathbb{E}_{\mathcal{G}_t}\left[\left|\frac{\widehat{s}_t - s_t^\star}{D}\right|^q\right] \le \frac{2q(q-1)}{bD}\left(\frac{D}{2\Delta_I}\right)^q \mathbb{E}_{\mathcal{G}_t}[\ell_{1-\alpha}(S_t, \widehat{s}_t) - \ell_{1-\alpha}(S_t, s_t^\star)].$$

The result follows by taking unconditional expectations of both sides, observing that $\mathbb{E}[|s_t - s_t^\star|]^q \le \mathbb{E}[|s_t - s_t^\star|^q]$ by Jensen's inequality, and summing over all $t \in I$. □

Method	LGBM (MAE = 0.19)				ARIMA (MAE = 0.09)				Prophet (MAE = 0.41)			
	Coverage	Width	LCE_k	SAReg_k	Coverage	Width	LCE_k	SAReg_k	Coverage	Width	LCE_k	SAReg_k
SCP	$.706_{.018}$	$.288_{.008}$	$.443_{.020}$	$.040_{.003}$	$.911_{.005}$	$.203_{.012}$	$.257_{.013}$	$.017_{.002}$	$.555_{.019}$	$.453_{.016}$	$.571_{.019}$	$.058_{.004}$
NExCP	$.764_{.013}$	$.268_{.009}$	$.411_{.017}$	$.014_{.001}$	$.904_{.004}$	$.185_{.012}$	$.254_{.011}$	$.009_{.001}$	$.681_{.013}$	$.462_{.018}$	$.508_{.017}$	$.017_{.001}$
FACI	$.822_{.008}$	$.254_{.009}$	$.282_{.012}$	$.008_{.001}$	$.899_{.003}$	$.169_{.011}$	$.205_{.008}$	$\mathbf{.007}_{\mathbf{.001}}$	$.776_{.009}$	$.458_{.018}$	$.338_{.013}$	$.007_{.001}$
SF-OGD	$.872_{.004}$	$.262_{.009}$	$.208_{.008}$	$.011_{.002}$	$.901_{.003}$	$.170_{.011}$	$\underline{.191}_{.008}$	$.010_{.002}$	$.871_{.003}$	$.475_{.019}$	$.209_{.008}$	$.009_{.002}$
FACI-S	$.863_{.005}$	$\mathbf{.240}_{\mathbf{.009}}$	$\underline{.207}_{.008}$	$\underline{.010}_{.002}$	$.891_{.004}$	$\mathbf{.152}_{\mathbf{.009}}$	$.197_{.008}$	$.010_{.002}$	$.856_{.004}$	$\mathbf{.459}_{\mathbf{.018}}$	$\underline{.197}_{.006}$	$\mathbf{.007}_{\mathbf{.002}}$
SAOCP	$.872_{.004}$	$\underline{.248}_{.009}$	$\mathbf{.170}_{\mathbf{.006}}$	$\mathbf{.008}_{\mathbf{.001}}$	$.886_{.003}$	$\underline{.155}_{.010}$	$\mathbf{.178}_{\mathbf{.007}}$	$\underline{.008}_{.001}$	$.868_{.002}$	$\underline{.473}_{.019}$	$\mathbf{.158}_{\mathbf{.004}}$	$\mathbf{.007}_{\mathbf{.001}}$

Table 4. Results on M4 Weekly (359 time series) with target coverage $1-\alpha = 0.9$ and interval size $k = 20$. Results are formatted as mean_{std}. Best results are **bold**, while second best are underlined, as long as the method's global coverage is in $(0.85, 0.95)$ (green). For all base predictors, SAOCP achieves the best local coverage error, best strongly adaptive regret, and second-best width. The only methods which achieve global coverage in $(0.85, 0.95)$ for LGBM and Prophet are the ones that predict $\widehat{s}_{t+1}$ directly, not as a quantile of $S_1, \ldots, S_t$.

Method	LGBM (MAE = 0.08)				ARIMA (MAE = 0.07)				Prophet (MAE = 0.11)			
	Coverage	Width	LCE_k	SAReg_k	Coverage	Width	LCE_k	SAReg_k	Coverage	Width	LCE_k	SAReg_k
SCP	$.932_{.003}$	$.205_{.004}$	$\mathbf{.123}_{\mathbf{.006}}$	$.012_{.001}$	$.938_{.003}$	$.199_{.004}$	$\underline{.126}_{.007}$	$.012_{.001}$	$.912_{.006}$	$.221_{.009}$	$.154_{.009}$	$.010_{.000}$
NExCP	$.922_{.002}$	$.187_{.003}$	$.133_{.006}$	$.011_{.000}$	$.922_{.002}$	$.175_{.003}$	$.136_{.006}$	$.012_{.001}$	$.908_{.003}$	$.208_{.010}$	$.146_{.006}$	$.010_{.000}$
FACI	$.910_{.002}$	$\mathbf{.179}_{\mathbf{.004}}$	$.130_{.005}$	$\mathbf{.010}_{\mathbf{.000}}$	$.906_{.002}$	$\mathbf{.162}_{\mathbf{.004}}$	$.132_{.005}$	$\mathbf{.011}_{\mathbf{.001}}$	$.900_{.002}$	$\mathbf{.200}_{\mathbf{.010}}$	$.131_{.005}$	$\mathbf{.009}_{\mathbf{.000}}$
SF-OGD	$.904_{.002}$	$.190_{.004}$	$\mathbf{.123}_{\mathbf{.004}}$	$.011_{.000}$	$.901_{.002}$	$.176_{.004}$	$.130_{.005}$	$.012_{.000}$	$.898_{.002}$	$.216_{.010}$	$\underline{.128}_{.004}$	$.011_{.000}$
FACI-S	$.909_{.002}$	$\mathbf{.179}_{\mathbf{.003}}$	$.127_{.005}$	$\mathbf{.010}_{\mathbf{.000}}$	$.910_{.002}$	$\underline{.166}_{.003}$	$\mathbf{.123}_{\mathbf{.004}}$	$\mathbf{.011}_{\mathbf{.000}}$	$.904_{.002}$	$\underline{.203}_{.010}$	$\mathbf{.125}_{\mathbf{.004}}$	$\mathbf{.009}_{\mathbf{.000}}$
SAOCP	$.892_{.002}$	$\mathbf{.179}_{\mathbf{.003}}$	$.132_{.004}$	$.012_{.000}$	$.895_{.002}$	$\underline{.166}_{.003}$	$.127_{.005}$	$.012_{.000}$	$.885_{.002}$	$.207_{.010}$	$.134_{.004}$	$.013_{.000}$

Table 5. Results on NN5 Daily (111 time series) with target coverage $1-\alpha = 0.9$ and interval size $k = 20$. Results are formatted as mean_{std}. Best results are **bold**, while second best are underlined, as long as the method's global coverage is in $(0.85, 0.95)$ (green). All methods perform similarly well.

D. Additional Time Series Experiments

Here, we report the results of our time series experiments (as described in Section 5.1) on M4 Weekly and NN5 Daily (the two smaller datasets) in Tables 4 and 5, respectively. The results on M4 Weekly (Table 4) are quite similar to those on M4 Daily (Table 2). All methods do reasonably well on NN5. Considering that even split conformal attains strong worst-case local coverage error, the residuals likely have a near-exchangeable distribution on NN5 (Barber et al., 2022, Theorems 2a, 3).

E. Additional Experimental Details

We provide specific implementation details of all methods here. We use $Q_{1-\alpha}(\cdot)$ to denote the (empirical) $(1-\alpha)$-th quantile of a set of scalars, defined as

$$Q_{1-\alpha}\left(\{S_\tau\}_{\tau=1}^t\right) := \inf\left\{s \in \mathbb{R} : \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{1}\left[S_\tau \le s\right] \ge 1-\alpha\right\}. \tag{32}$$

1. SCP: Split conformal prediction (Vovk et al., 2005) predicts $\widehat{s}_{t+1} = Q_{1-\alpha}(\frac{1}{t}\sum_{\tau=1}^{t}\delta_{S_t})$.
2. NExCP: Non-exchangeable conformal prediction extends SCP by using a *weighted* quantile function $\widehat{s}_{t+1} = Q_{1-\alpha}(\frac{1}{t}\sum_{\tau=1}^{t} w_\tau \delta_{S_t})$. Barber et al. (2022) suggest using geometrically decaying weights to adapt NExCP to situations with distribution shift, so we use $w_t = (1-3\alpha/4)^{1-t}w_1$.
3. FACI: Fully Adaptive Conformal Inference has 4 hyperparameters: the individual expert learning rates $\gamma_1, \ldots, \gamma_N$; a target interval length k; and the meta-algorithm learning rate η; and a smoothing parameter σ. We set $k = 100$ and follow Gibbs & Candès (2022) to set $N = 8$, $\sigma = \frac{1}{2k}$, $\gamma = \{0.001, 0.002, 0.004, 0.008, 0.016, 0.032, 0.064, 0.128\}$, and
$$\eta_t = \sqrt{\frac{\log(Nk)+2}{\sum_{\tau=t-k}^{t-1}\mathbb{E}[\ell_\alpha(\beta_t,\alpha_t)^2]}},$$
where the expectation is over α_t. We also tried $k = 20$ for the time series experiments (to match our evaluation metrics), but the results were worse.

Method	LGBM (MAE = 0.05) Coverage	Width	LCE_k	SAReg_k	ARIMA (MAE = 0.14) Coverage	Width	LCE_k	SAReg_k	Prophet (MAE = 0.08) Coverage	Width	LCE_k	SAReg_k
EnbPI	.803$_{.006}$	.088$_{.003}$	.331$_{.009}$	.017$_{.001}$	.916$_{.003}$	.220$_{.021}$	.186$_{.007}$	.032$_{.014}$	.834$_{.007}$	.141$_{.005}$	.299$_{.011}$	.018$_{.001}$
EnbNEx	.864$_{.002}$	.097$_{.003}$	.218$_{.006}$	.010$_{.001}$	.907$_{.003}$	.198$_{.018}$	.193$_{.007}$	.027$_{.012}$	.892$_{.003}$	.151$_{.005}$	.195$_{.006}$	.010$_{.001}$
EnbFACI	.856$_{.002}$	**.081$_{.003}$**	.200$_{.005}$	.007$_{.001}$	.900$_{.003}$	**.181$_{.019}$**	.161$_{.005}$	**.021$_{.007}$**	.884$_{.002}$	**.130$_{.005}$**	.164$_{.005}$	.007$_{.001}$
EnbSF-OGD	.871$_{.002}$	.098$_{.004}$	.173$_{.004}$	.008$_{.001}$	.906$_{.003}$	.201$_{.020}$	.165$_{.005}$	.027$_{.009}$	.898$_{.001}$	.145$_{.005}$	.144$_{.004}$	.008$_{.001}$
EnbSAOCP	.884$_{.002}$	.091$_{.003}$	**.134$_{.002}$**	**.005$_{.000}$**	.893$_{.003}$	.192$_{.024}$	**.150$_{.004}$**	.053$_{.028}$	.888$_{.002}$	**.130$_{.005}$**	**.130$_{.002}$**	**.005$_{.000}$**

Table 6. Ensemble results on M4 Hourly with target coverage $1-\alpha = 0.9$ and interval size $k = 20$. Results are formatted as mean_{std}. Best results are **bold**, while second best are underlined, as long as the method's global coverage is in $(0.85, 0.95)$ (green). EnbSAOCP achieves the best local coverage error and strongly adaptive regret for all models, except ARIMA where its strongly adaptive regret is somewhat high.

Method	LGBM (MAE = 0.11) Coverage	Width	LCE_k	SAReg_k	ARIMA (MAE = 0.10) Coverage	Width	LCE_k	SAReg_k	Prophet (MAE = 0.18) Coverage	Width	LCE_k	SAReg_k
EnbPI	.512$_{.005}$	.093$_{.001}$	.716$_{.004}$	.077$_{.001}$	.894$_{.002}$	.121$_{.002}$	.289$_{.004}$	.056$_{.034}$	.791$_{.005}$	.288$_{.002}$	.374$_{.005}$	.062$_{.001}$
EnbNEx	.749$_{.002}$	.142$_{.002}$	.520$_{.004}$	.020$_{.000}$	.894$_{.001}$	.116$_{.003}$	.293$_{.004}$	.035$_{.022}$	.847$_{.003}$	.343$_{.003}$	.367$_{.005}$	.029$_{.000}$
EnbFACI	.776$_{.002}$	.130$_{.002}$	.392$_{.003}$	.015$_{.000}$	.887$_{.001}$	.100$_{.007}$	.248$_{.003}$	**.015$_{.005}$**	.853$_{.003}$	**.232$_{.003}$**	.230$_{.003}$	.025$_{.001}$
EnbSF-OGD	.798$_{.002}$	.143$_{.002}$	.396$_{.004}$	.021$_{.001}$	.898$_{.001}$	.106$_{.002}$	.241$_{.003}$	.047$_{.034}$	.900$_{.002}$	.292$_{.003}$	.195$_{.002}$	.023$_{.001}$
EnbSAOCP	.875$_{.001}$	**.138$_{.002}$**	**.203$_{.002}$**	**.007$_{.000}$**	.908$_{.001}$	**.096$_{.002}$**	**.187$_{.002}$**	.044$_{.033}$	.917$_{.001}$	.233$_{.002}$	**.139$_{.001}$**	**.011$_{.000}$**

Table 7. Ensemble on M4 Daily with target coverage $1-\alpha = 0.9$ and interval size $k = 20$. Results are formatted as mean_{std}. Best results are **bold**, while second best are underlined, as long as the method's global coverage is in $(0.85, 0.95)$ (green). EnbSAOCP achieves the best or second best width, local coverage error, and strongly adaptive regret for all models. It is also the only method which achieves valid coverage for LGBM.

4. SF-OGD: Scale-Free Online Gradient Descent. The only hyperparameter is the maximum radius D. For the time series experiments (Section 5.1, Appendix F), we set $D/\sqrt{3}$ for each horizon h equal to the largest h-step residual observed on the calibration split of the training data. For the m-way image classification experiments (Section 5.2, Appendix G), we set $D = 1 + \lambda\sqrt{m - k_{reg}}$, where λ and k_{reg} are the width regularization parameters in (15).
5. FACI-S: FACI applied to S_t rather than α_t. The hyperparameters are the same, except the losses used to compute η_t are $\ell_{1-\alpha}(S_t, \widehat{s}_t)$, and the learning rates are multiplied by D. We set D in the same way as SF-OGD.
6. SAOCP: There are 2 hyperparameters: the maximum radius D and the lifetime multiplier g in (8). We set D in the same way as SF-OGD. We set $g = 8$ for the time series experiments and $g = 32$ for the image classification experiments.

F. Time Series Experiments with Ensemble Models

In this section, we replicate the experiments of Section 5.1 using ensemble models trained with the method of EnbPI (Xu & Xie, 2021). Specifically, we train base learner $\widehat{f}^{(b)}$ on $(X_t, Y_T)_{t \in I_b}$, where I_b is sampled randomly from $[T]$. Then, we obtain the residual $S_t^y = |y - \phi(\widehat{f}^{(b)}(X_t) : I_b \not\ni t)|$ by aggregating all models not trained on (X_t, Y_t). Finally, the residual of a new observation (X_{T+1}, Y_{T+1}) is $|Y_{T+1} - \phi(\widehat{f}^{(b)}(X_{T+1}) : b \in [B])|$. We use $B = 5$ models in the ensemble.

EnbPI predicts the radius $\widehat{s}_{t+1}$ as the $1-\alpha$ empirical quantile of the previously observed residuals, as in split conformal prediction. However, these prediction sets can be obtained via an arbitrary function of the scores, including NExCP, FACI, SF-OGD, or SAOCP. Besides EnbPI, we call these hybrid methods EnbNEx, EnbFACI, EnbSF-OGD, and EnbSAOCP respectively. We use the same hyperparameters as described in Appendix E.

This approach puts the other methods on even footing with EnbPI, because EnbPI removes the need for a train/calibration split and changes the underlying model from a single learner to a more accurate ensemble. We consider the contributions of EnbPI orthogonal to our own, and this section shows that their method can successfully be combined with ours.

The results mirror those of Section 5.1. EnbSAOCP generally obtains the best or second-best interval width, worst-case local coverage error, and strongly adaptive regret on all M4 datasets (Tables 6, 7, 8). On NN5 (Table 9), all methods obtain similar strongly adaptive regret. However, EnbSF-OGD and EnbSAOCP obtain the best and second-best worst-case local coverage error at the cost of having slightly wider intervals. Across the board, EnbSAOCP has narrower intervals than EnbSF-OGD.

Method	LGBM (MAE = 0.12) Coverage	Width	LCE_k	SAReg_k	ARIMA (MAE = 0.07) Coverage	Width	LCE_k	SAReg_k	Prophet (MAE = 0.16) Coverage	Width	LCE_k	SAReg_k
EnbPI	$.540_{.019}$	$.118_{.006}$	$.596_{.019}$	$.079_{.005}$	$.893_{.008}$	$.174_{.008}$	$.266_{.015}$	$.021_{.003}$	$.785_{.016}$	$.288_{.010}$	$.369_{.019}$	$.048_{.004}$
EnbNEx	$.720_{.012}$	$.168_{.007}$	$.483_{.016}$	$.015_{.001}$	$.910_{.004}$	$.161_{.009}$	$.252_{.011}$	$.012_{.001}$	$.875_{.008}$	$.375_{.013}$	$.305_{.015}$	$.027_{.001}$
EnbFACI	$.784_{.009}$	$.159_{.006}$	$.336_{.013}$	$.010_{.001}$	$.905_{.004}$	$\underline{.142}_{.008}$	$.193_{.008}$	$\underline{.009}_{.001}$	$.881_{.006}$	$\underline{.235}_{.007}$	$.201_{.010}$	$\underline{.017}_{.001}$
EnbSF-OGD	$.795_{.011}$	$.172_{.007}$	$.329_{.015}$	$.016_{.002}$	$.908_{.003}$	$.152_{.009}$	$\underline{.184}_{.008}$	$.012_{.002}$	$.915_{.004}$	$.304_{.009}$	$\underline{.164}_{.008}$	$.023_{.002}$
EnbSAOCP	$.874_{.004}$	$\mathbf{.165}_{.006}$	$\mathbf{.173}_{.006}$	$\mathbf{.005}_{.000}$	$.909_{.003}$	$\mathbf{.131}_{.006}$	$\mathbf{.151}_{.006}$	$\mathbf{.008}_{.001}$	$.907_{.003}$	$\mathbf{.232}_{.006}$	$\mathbf{.133}_{.004}$	$\mathbf{.012}_{.001}$

Table 8. Ensemble results on M4 Weekly with target coverage $1-\alpha = 0.9$ and interval size $k = 20$. Results are formatted as mean_{std}. Best results are **bold**, while second best are underlined, as long as the method's global coverage is in $(0.85, 0.95)$ (green). EnbSAOCP achieves the best width, local coverage error, and strongly adaptive regret for all models. It is also the only method which achieves valid coverage for LGBM.

Method	LGBM (MAE = 0.09) Coverage	Width	LCE_k	SAReg_k	ARIMA (MAE = 0.07) Coverage	Width	LCE_k	SAReg_k	Prophet (MAE = 0.07) Coverage	Width	LCE_k	SAReg_k
EnbPI	$.859_{.005}$	$\mathbf{.164}_{.004}$	$.225_{.010}$	$.012_{.001}$	$.910_{.003}$	$\underline{.163}_{.004}$	$.159_{.008}$	$\underline{.010}_{.001}$	$.904_{.004}$	$\underline{.163}_{.005}$	$.160_{.007}$	$\mathbf{.010}_{.001}$
EnbNEx	$.882_{.003}$	$\underline{.177}_{.004}$	$.177_{.007}$	$\underline{.010}_{.000}$	$.920_{.002}$	$.174_{.003}$	$.138_{.006}$	$.012_{.001}$	$.923_{.002}$	$.178_{.006}$	$.126_{.004}$	$.012_{.001}$
EnbFACI	$.883_{.002}$	$\underline{.177}_{.004}$	$.166_{.006}$	$\mathbf{.009}_{.000}$	$.905_{.002}$	$\mathbf{.156}_{.003}$	$.131_{.005}$	$\mathbf{.009}_{.001}$	$.907_{.002}$	$\mathbf{.156}_{.003}$	$.123_{.004}$	$\mathbf{.010}_{.000}$
EnbSF-OGD	$.895_{.002}$	$.194_{.004}$	$\underline{.132}_{.004}$	$\underline{.010}_{.000}$	$.912_{.002}$	$.176_{.004}$	$\mathbf{.120}_{.004}$	$.012_{.001}$	$.912_{.002}$	$.178_{.004}$	$\underline{.120}_{.004}$	$.013_{.001}$
EnbSAOCP	$.886_{.002}$	$.188_{.004}$	$\mathbf{.131}_{.005}$	$.012_{.000}$	$.906_{.002}$	$.168_{.003}$	$\underline{.122}_{.004}$	$.012_{.000}$	$.905_{.002}$	$.170_{.003}$	$\mathbf{.118}_{.003}$	$.013_{.000}$

Table 9. Ensemble results on NN5 Daily with target coverage $1-\alpha = 0.9$ and interval size $k = 20$. Results are formatted as mean_{std}. Best results are **bold**, while second best are underlined, as long as the method's global coverage is in $(0.85, 0.95)$ (green). EnbSAOCP and EnbSF-OGD achieve the best and second-best worst-case local coverage error error at the cost of having slightly wider intervals.

G. Image Classification on ImageNet/ImageNet-C

We replicate the experiments of Section 5.2 using a ResNet-50 classifier on the ImageNet (Deng et al., 2009) base dataset and its corrupted version ImageNet-C (Hendrycks & Dietterich, 2019). We train the model using SGD with learning rate 0.1 (annealed by a factor of 10 every 7 epochs), momentum 0.9, batch size 256, and early stopping if validation accuracy stops improving for 10 epochs. The model achieved a final test accuracy of 52.8%. Figure 2 shows the results.

When performing uncertainty quantification, we use the conformal score (15) with width regularization parameters $\lambda = 0.01$ and $k_{reg} = 20$. As in Section 5.2, SAOCP and SF-OGD's local coverages remain the closest to the target of 0.9. The differences are most apparent when the distribution shift is sudden, suggesting that they are able to adapt to these distribution shifts more quickly than other methods. While all methods attain similar prediction set sizes, NExCP and FACI adapt more slowly to the best fixed prediction set size than SAOCP and SF-OGD. SAOCP also has better coverage than SF-OGD.

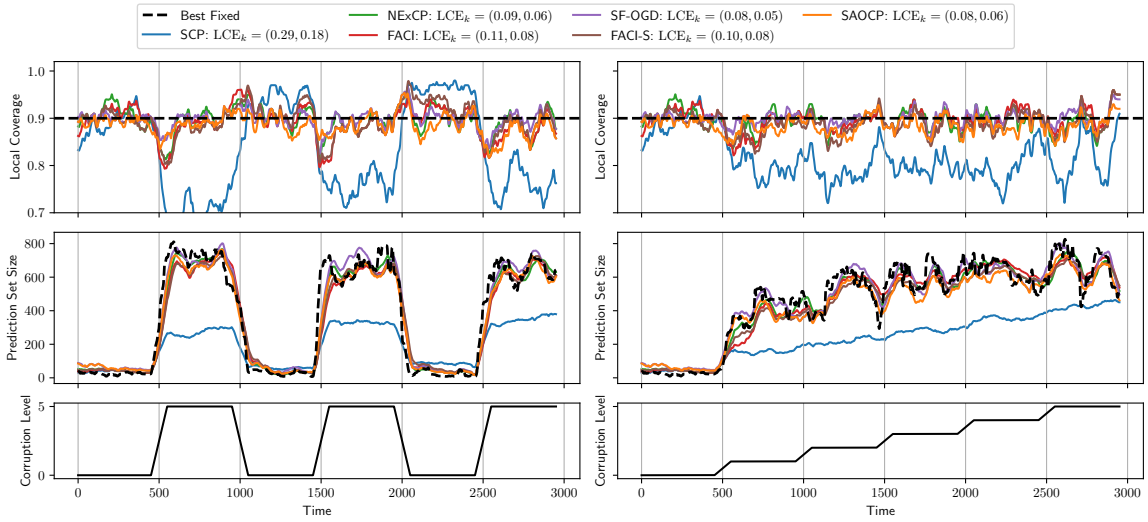

Figure 2. Local coverage (top row) and prediction set size (second row) achieved by various UQ methods when the distribution shifts between ImageNet and ImageNet-C every 500 steps. We plot moving averages with window size $k = 100$. Left: sudden shifts between corruption level 0 and 5. Right: gradual shift from level 0 to 5. SAOCP and SF-OGD's local coverage remain the closest to the target of 0.9, especially at the change points. While the two methods attain similar local coverage, SAOCP returns smaller prediction sets than SF-OGD.