Journal of Machine Learning Research 25 (2024) 1-36 Submitted 10/22; Revised 5/24; Published 5/24

Conformal Inference for Online Prediction with Arbitrary Distribution Shifts

Isaac Gibbs IGIBBS@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305, USA

Emmanuel Candès CANDES@STANFORD.EDU
Departments of Statistics and Mathematics
Stanford University
Stanford, CA 94305, USA

Editor: Boaz Nadler

Abstract

We consider the problem of forming prediction sets in an online setting where the distribution generating the data is allowed to vary over time. Previous approaches to this problem suffer from over-weighting historical data and thus may fail to quickly react to the underlying dynamics. Here, we correct this issue and develop a novel procedure with provably small regret over all local time intervals of a given width. We achieve this by modifying the adaptive conformal inference (ACI) algorithm of Gibbs and Candès (2021) to contain an additional step in which the step-size parameter of ACI's gradient descent update is tuned over time. Crucially, this means that unlike ACI, which requires knowledge of the rate of change of the data-generating mechanism, our new procedure is adaptive to both the size and type of the distribution shift. Our methods are highly flexible and can be used in combination with any baseline predictive algorithm that produces point estimates or estimated quantiles of the target without the need for distributional assumptions. We test our techniques on two real-world datasets aimed at predicting stock market volatility and COVID-19 case counts and find that they are robust and adaptive to real-world distribution shifts.

Keywords: Conformal inference, online prediction, distribution shift, prediction set, online convex optimization

1. Introduction

We consider a situation in which we observe a data stream $\{(X_t, Y_t)\}_{1\leq t\leq T}$ generated by a dynamic process in which the distribution of (X_t, Y_t) (and more broadly of subsequences $(X_t, Y_t), \dots, (X_{t+s}, Y_{t+s})$) is allowed to vary over time. At each time point t, our goal is to use the previously observed data $\{(X_s, Y_s)\}_{s<t}$, along with the new covariates X_t, to form a prediction set for the target value Y_t. We are motivated by numerous modern applications in which a complex model (e.g. neural network, random forest) is employed to produce a point estimate of Y. While these models have been found to perform well on i.i.d. training and testing data, rigorous guarantees on their accuracy are lacking and their empirical performance has been found to degrade under distribution shift (Koh et al. (2021)). Thus,

©2024 Isaac Gibbs and Emmanuel Candès.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/22-1218.html.

a more robust understanding of the uncertainty underlying these methods' predictions is necessary before they can be deployed in practice.

Prediction sets have become a popular tool for quantifying the accuracy of machine learning models. Formally, we say that $\hat{C}(\cdot) \subseteq \mathbb{R}$ is a $1-\alpha$ prediction set for Y if $\mathbb{P}(Y \in \hat{C}(X)) = 1-\alpha$. Conceptually, by examining the size and scope of $\hat{C}(X)$ the user can gain information above the uncertainty underlying a model's point-prediction.

Many of the most useful tools for computing prediction sets come from the field of conformal inference. This general framework provides a flexible set of methodologies for transforming the point or quantile-estimates output by a black-box machine learning model into valid prediction sets (see e.g. Saunders et al., 1999; Vovk et al., 1999, 2005; Gammerman and Vovk, 2007; Shafer and Vovk, 2008; Lei and Wasserman, 2014; Sadinle et al., 2019; Foygel Barber et al., 2020; Barber et al., 2021). Conformal inference is particularly powerful because it allows users to leverage improvements in the baseline predictor to obtain smaller and more accurate prediction sets (Romano et al. (2019)).

The original conformal inference methods developed by Vovk and colleagues typically require that all the training and testing data be exchangeable (e.g. be i.i.d.), and in particular, require that all points have the same marginal distribution. While the earlier literature does contain some extensions beyond exchangeabiltiy to data sequences that are locally exchangeable or can be transformed to an exchangeable sequence (Vovk et al. (2005)), the applicability of these methods is limited. More recently, many authors have extended conformal inference to account for a wider variety of distribution shifts and dependency structures within the training and testing data. Some examples including methods for stationary time series (Chernozhukov et al. (2018)), cross-sectional time series (Lin et al. (2022)), label shift (Podkopaev and Ramdas (2021)), covariate shift (Tibshirani et al. (2019); Yang et al. (2024)), and generic methods for re-weighting non-identically distributed data (Barber et al. (2023)).

The problem of adjusting a conformal predictor to adapt to arbitrary online distribution shifts was proposed by the present authors in Gibbs and Candès (2021). There, we gave a gradient descent method, called adaptive conformal inference (ACI), that tunes the width of the prediction sets to adapt to the underlying uncertainty in the environment. While that method was found to produce good results both theoretically and empirically, its performance critically relies on a good specification of its step-size parameter. Specifically, it was shown that for optimal performance, the step-size should be set proportional to the underlying rate of change in the environment, which is unknown in practice.

Recently, two alternatives to ACI have been proposed that avoid the need for a user-specified step-size. The most direct approach is that of Zaffran et al. (2022), which gives an expert learning method for adaptively tuning the step-size based off of the historical performance of a set of candidate values. Moving away from gradient descent based methods, Bastani et al. (2022) propose an alternative approach in which the width of the prediction set is chosen directly from a set of candidate thresholds. While these methods can be argued to improve on ACI, both approaches have the shortcoming of heavily weighting older historical data when choosing amongst their candidate values. As our experiments in Section 4.1 show, this can lead to a failure to quickly adapt when abrupt changes occur.

In this article, we propose an alternative expert selection scheme for choosing the step-size in ACI. We show that unlike previous approaches, our method can control the deviation in the coverage probability locally over time and we bound the local coverage of our method in terms of the rate of change of an underlying optimal target parameter. The importance of this result is not purely theoretical, and we provide example settings where alternative methods produce worse adaptivity to the local dynamics than our approach. In addition to the local coverage, we also investigate the average the coverage our method over a long time horizon. We show that for some choices of our method's hyperparameters exact long-term coverage is guaranteed. In practice, we will typically favour alternative hyperparameter settings that are designed to optimize local coverage. For these values, we find empirically that our method can have a bias in the long-term coverage. However, this bias is extremely small, and thus it has a minimal impact on the performance of our method. We evaluate the performance of our method on two real-world prediction tasks aimed towards predicting stock market volatility and COVID-19 case counts, and find that it adapts well to real-world dynamics.

2. Methodology

In this section we outline the main methods of this article. Before developing our new methodology, we begin by reviewing the conformal inference and adaptive conformal inference frameworks.

2.1 Conformal Inference

Let $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ denote a set of observed training data and (X_{n+1}, Y_{n+1}) denote a new test point from which we only observe X_{n+1}. In order to construct a prediction set, conformal inference begins by imputing guesses y for Y_{n+1}. Then, for each candidate value y, a conformity score $S : (\mathbb{R}^d \times \mathbb{R})^n \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is used to measure how well the data point (X_{n+1}, y) conforms with $(X_1, Y_1), \dots, (X_n, Y_n)$. Typically, this is done by first using all $n + 1$ datapoints to fit a regression and then measuring how well y aligns with the prediction of the fitted model at X_{n+1}. For example, we may take $S(\cdot)$ to be the absolute residual

$$S((X_j, Y_j)_{1\leq j\leq n}, (X_{n+1}, y)) := |y - \hat{\mu}(X_{n+1})|, \tag{1}$$

where $\hat{\mu}$ is an estimate of $\mathbb{E}[Y|X]$ fit on $(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, y)$, or the estimated probability

$$S((X_j, Y_j)_{1\leq j\leq n}, (X_{n+1}, y)) := 1 - \hat{\pi}(y|X_{n+1}),$$

where $\hat{\pi}(y|X_{n+1})$ is a fitted estimate of $\mathbb{P}(Y_{n+1} = y|X_{n+1})$. In the final step of conformal inference the value y is added to the prediction set if the test score, $S((X_j, Y_j)_{1\leq j\leq n}, (X_{n+1}, y))$, is small relative to the training scores, $\{S((X_j, Y_j)_{1\leq j\leq n, j\neq i}, (X_{n+1}, y), (X_i, Y_i))\}_{i=1}^n$, e.g. if the residual $|y - \hat{\mu}(X_{n+1})|$ is small relative to $|Y_i - \hat{\mu}(X_i)|$.

In general, the only requirement on $S(\cdot)$ necessary for existing theoretical results to hold is that it is unchanged by permutations of its first n arguments. In the context of this article, the data (X_1, Y_1),…,(X_n, Y_n), (X_{n+1}, Y_{n+1}) will often have a temporal dependence structure. As a result, it may not be sensible to treat all of the arguments to $S(\cdot)$ symmetrically and we will often use conformity scores that are not permutation invariant.

For ease of notation, let $S_i^y := S((X_j, Y_j)_{1\leq j\leq n, j\neq i}, (X_{n+1}, y), (X_i, Y_i))$ and $S_{n+1}^y := S((X_j, Y_j)_{1\leq j\leq n}, (X_{n+1}, y))$. For any $\tau \in \mathbb{R}$ and distribution $\mathcal{D}$, let Quantile $(\tau, \mathcal{D})$ denote the τ_{th} quantile of $\mathcal{D}$ with the convention that $\text{Quantile}\,(\tau, \mathcal{D}) = \infty$ (respectively $-\infty$) for all $\tau \geq 1$ (respectively $\tau \leq 0$). Then, formally, conformal inference outputs the prediction set

$$\hat{C}_{n+1} := \left\{ y : S_{n+1}^{y} \leq \text{Quantile}\left(1-\alpha, \frac{1}{n+1}\sum_{i=1}^{n+1} \delta_{S_i^y}\right)\right\}. \tag{2}$$

As alluded to in the introduction, this set satisfies the following coverage guarantee.

Theorem 1 *If the data $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are exchangeable and $S(\cdot)$ is invariant to permutations of its first n arguments, then*

$$\mathbb{P}(Y_{n+1} \in \hat{C}_{n+1}) \geq 1 - \alpha.$$

Moreover, if in addition the values $\{S_i^{Y_{n+1}}\}_{1 \leq i \leq n+1}$ are distinct with probability one, then

$$\mathbb{P}(Y_{n+1} \in \hat{C}_{n+1}) \leq 1 - \alpha + \frac{1}{n+1}.$$

A full proof of this result can be found in Lemma 1 of Romano et al. (2019) (for an earlier treatment of the first part of the Theorem see also Vovk et al. (2005)).

Unfortunately, in most practical examples, computing $\hat{C}_{n+1}$ requires the user to fit the regression function $\hat{\mu}$ or $\hat{\pi}$ for all possible values of y. As this is not usually computationally feasible, many implementations of conformal inference use a data splitting approach in which the regression estimate is pre-fit in advance using a separate set of training data (Papadopoulos et al. (2002); Vovk et al. (2005); Papadopoulos (2008)). The methods developed in this article do not rely on any formal guarantee of conformal inference and can be used in conjunction with any procedure that produces estimated quantiles of a conformity score. Thus, in our experiments, we avoid extraneous computation by using procedures that do not strictly adhere to the construction given by (2).

2.2 Adaptive Conformal Inference

The methodology developed in this article builds upon the adaptive conformal inference (ACI) algorithm proposed by Gibbs and Candès (2021). This procedure accounts for non-exchangeability by treating the quantile of the conformity scores as a tunable parameter that can be learned in an online fashion. More concretely, let $\mathcal{D}_t^y$ denote our estimate of the conformity score distribution at time-step t with imputed value y. For instance, in our experiments we will often use the empirical distribution of the most recent r conformity scores, $\mathcal{D}_t^y = \frac{1}{r}\sum_{i=t-r+1}^{t} \delta_{S_i^y}$ (recall that standard conformal inference would take $r = t$). Let

$$\hat{C}_t(\beta) := \{y : S_t^y \leq \text{Quantile}\,(1-\beta, \mathcal{D}_t^y)\}, \tag{3}$$

denote the prediction set obtained at timestep t using the $1-\beta$ quantile of $\mathcal{D}_t^y$. Then, without any assumptions on the data generating distribution, we know that $\beta \mapsto \mathbb{P}(Y_t \in \hat{C}_t(\beta))$ is non-increasing with $\mathbb{P}(Y_t \in \hat{C}_t(0)) = 1$ and $\mathbb{P}(Y_t \in \hat{C}_t(1)) = 0$. Now, suppose $\mathcal{D}_t^y$ is a continous distribution (this could be obtained by e.g. smoothing the empirical distribution of the previous conformity scores). Then, it is reasonable to assume that $\beta \mapsto \mathbb{P}(Y_t \in \hat{C}_t(\beta))$ is continuous, and thus, there exists a value α_t^* such that $\mathbb{P}(Y_t \in \hat{C}_t(\alpha_t^*)) = 1 - \alpha$. Since we do not know this optimal value *a priori*, ACI estimates it in an online fashion using a parameter α_t that is updated as

$$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \text{err}_t), \tag{4}$$

where $\gamma > 0$ is a step-size parameter and

$$\text{err}_t := \begin{cases} 0, & \text{if } Y_t \in \hat{C}_t(\alpha_t), \\ 1, & \text{if } Y_t \notin \hat{C}_t(\alpha_t). \end{cases}$$

In simple terms, the update (4) can be seen as increasing/decreasing the size of the prediction set in response to the historical under/over coverage of the algorithm.

A natural criticism of this approach, originally raised in Bastani et al. (2022), is that ACI can obtain good coverage not because it successfully learns α_t^*, but rather simply due to the fact that it reactively corrects its past mistakes. In particular, it may be the case that α_t oscillates between being well below and well above α_t^* and thus good coverage is obtained only through a cancellation of positive and negative errors. In Section 4.2.2 we give empirical evidence indicating that the new methods developed in this article do not exhibit such pathological behaviour.

Returning to (4), the critical difficulty in implementing ACI is the choice of γ. Gibbs and Candès (2021) gives theoretical results suggesting that γ should be chosen proportional to the size of the variation in α_t^* across time. However, this value is unknown and no procedure for estimating it is given. Additionally, much of the theory given in Gibbs and Candès (2021) is only valid under two additional assumptions.

1. The conformity score function $S((X_s, Y_s)_{s<t}, (X_t, y)) = S(X_t, y)$ is a fixed function that depends only on the new data point (X_t, y) and does not use the most recent data $(X_s, Y_s)_{s<t}$ to recalibrate its predictions. For example, under this assumption, the conformity score (1) would use a regression function $\hat{\mu}(\cdot)$ that is fixed in advance and not updated as time progresses.

2. Instead of using an adaptive distribution, $\mathcal{D}_t^y$ to generate quantiles for the prediction set in (3) we instead have some fixed reference distribution, $\mathcal{D}$ that the conformity scores are compared to, i.e. the prediction set can be written as

$$\hat{C}_t(\alpha_t) = \{y : S(X_t, y) \leq \text{Quantile}(1 - \alpha_t, \mathcal{D})\}.$$

These two assumptions are clearly problematic since under distribution shift the most recent data should be used to recalibrate both the regression function and the estimated distribution of the scores. The most obvious consequence of using fixed models is an increase in the size of the prediction sets over time as the true model drifts and the errors in the point predictions made by the regression model grow. More subtly, holding $\mathcal{D}$ fixed can lead to large oscillations in α_t^* as the past conformity scores no longer reflect the current situation. This will increase the difficulty of the online learning problem and may cause the coverage proability, $\mathbb{P}(Y_t \in \hat{C}_t(\alpha_t))$ to sharply deviate from $1 - \alpha$. In the following sections, we develop a new method and novel theoretical results that make no assumptions on $S(\cdot)$ and $\mathcal{D}$ and thus allow these quantities to be updated over time.

2.3 Dynamically-Tuned Adaptive Conformal Inference

In order to describe our new method, it is useful to first observe that the ACI update (4) can be viewed as a gradient descent step with respect to the pinball loss. To see this, let

$$\beta_t := \sup\{\beta : Y_t \in \hat{C}_t(\beta)\},$$

be the value of β such that $\hat{C}_t(\beta_t)$ is the smallest prediction set containing Y_t. Recall the definition of the pinball loss,

$$\ell(\beta_t, \theta) := \alpha(\beta_t - \theta) - \min\{0, \beta_t - \theta\}.$$

Then, one can verify that (4) is equivalent to the update

$$\alpha_{t+1} := \alpha_t - \gamma \nabla_\theta \ell(\beta_t, \alpha_t).^*$$

Through this lens, ACI can be viewed as a gradient descent procedure with respect to the sequence of convex losses $\{\ell(\beta_t, \cdot)\}$. Thus, in order to learn γ we can utilise popular methods from the online convex optimization literature. In particular, we will employ an exponential re-weighting scheme that chooses a value for γ based off of the historical performance of a set of candidate values. Methods of this type have a long history dating back to the original work of Vovk (1990). Our specific procedure is a small modification of an algorithm proposed by Gradu et al. (2023). In that article, Gradu et al. (2023) consider a control setting in which present actions affect the future states of the system. This leads them to consider a surrogate loss function that accounts for the long-term dependence structure induced by the actions. Because we have no such dependence, we consider a simplified version of their method here.

We refer to the resulting procedure as dynamically-tuned adaptive conformal inference (DtACI, Algorithm 1). This algorithm takes as input a candidate set of values for γ and constructs a corresponding candidate set of values for α_t by running multiple versions of ACI in parallel. In the convex optimization literature these parallel sequences are typically referred to as experts. The final value of α_t output at time t is then chosen from among these experts by evaluating their historical performance. In effect, we learn the optimal value of γ in an online fashion, enabling dynamic calibration of the prediction set to the size of the distribution shift in the environment.

Before moving on, we note that Algorithm 1 is not completely parameter free. In fact, while we have removed the need for an unknown step-size parameter, this has come at the cost of adding two unknown weight parameters, η and σ. While this may initially appear to be problematic, in Section 3, we will outline a simple procedure for choosing σ and η that does not involve any unknown quantities. This contrasts sharply with the situation for γ in which an optimal choice requires an in-depth knowledge of the distribution shift. Moreover, in some environments the size of the distribution shift can vary over time and a single constant value for γ can perform poorly. For example, in Section 4.1 we demonstrate a setting in which adaptively tuning γ allows us to quickly respond to an abrupt change in the environment, while a more stationary choice of γ lags behind. This issue does not

[*] Here we have ignored the edge case $\beta_t = \alpha_t$. In this case to match the original ACI update one should take the smallest subgradient of $\ell(\beta_t, \alpha_t)$, i.e. the value 0.

Algorithm 1: DtACI, modified version of Algorithm 1 in Gradu et al. (2023).

```
Data: Observed values $\{\beta_t\}_{1\leq t\leq T}$, set of candidate $\gamma$ values $\{\gamma_i\}_{1\leq i\leq k}$, starting
        points $\{\alpha_1^i\}_{1\leq i\leq k}$, and parameters $\sigma$ and $\eta$.
$w_1^i \leftarrow 1,\ 1\leq i\leq k$;
for $t = 1, 2, \dots, T$ do
    Define the probabilities $p_t^i := w_t^i / \sum_{1\leq j\leq k} w_t^j$, $\forall 1\leq i\leq k$;
    Output $\alpha_t = \alpha_t^i$ with probability $p_t^i$;
    $\bar{w}_t^i \leftarrow w_t^i \exp(-\eta\ell(\beta_t, \alpha_t^i)),\ \forall 1\leq i\leq k$;
    $\bar{W}_t \leftarrow \sum_{1\leq i\leq k} \bar{w}_t^i$;
    $w_{t+1}^i \leftarrow (1-\sigma)\bar{w}_t^i + \bar{W}_t\sigma/k$;
    $\mathrm{err}_t^i := \mathbb{1}\{Y_t \notin \hat{C}_t(\alpha_t^i)\}$, $\forall 1\leq i\leq k$;
    $\mathrm{err}_t := \mathbb{1}\{Y_t \notin \hat{C}_t(\alpha_t)\}$;
    $\alpha_{t+1}^i = \alpha_t^i + \gamma_i(\alpha - \mathrm{err}_t^i)$, $\forall 1\leq i\leq k$;
```

exist for σ and η and we provide extensive empirical evidence demonstrating that a single choice of these parameters performs well across a large variety of environments. A more theoretical discussion on the optimal settings for σ and η that justifies a specific fixed choice for these parameters can be found at the end of Section 3.1.

2.4 Comparison to Existing Methods

As mentioned in the introduction, two other alternatives to ACI have been proposed in the literature. Most closely related to the present paper is the AgACI method of Zaffran et al. (2022), which aims to learn the value of γ in ACI using the adaptive Bernstein online expert aggregation scheme of Wintenberger (2017). To describe this method, let $\{L_t^i\}$ and $\{\eta_t^i\}$ denote the cumulative loss and learning rate for expert $i \in \{1, \dots, k\}$ given by the recursive updates, $L_t^i = 0$, $\eta_t^i = 0$, and

$$\begin{aligned}
&\ell_t^i := (\mathrm{err}_{t-1} - \alpha)(\alpha_t^i - \alpha_t),\\
&L_t^i := L_{t-1}^i + \frac{1}{2}(\ell_t^i(1 + \eta_{t-1}^i\ell_t^i) + 2^{\lceil \log_2(\max_{1\leq s\leq t}|\ell_s^i|)\rceil + 1}\mathbb{1}\{\eta_{t-1}^i\ell_t^i > 1/2\}),\\
&\eta_t^i := \max\left\{2^{-\lceil \log_2(\max_{1\leq s\leq t}|\ell_s^i|)\rceil - 1}, \sqrt{\frac{\log(1/k)}{\sum_{s=1}^t(\ell_s^i)^2}}\right\}.
\end{aligned}$$

The details of these definitions are not critical. The most important thing to note is that ℓ_t^i is exactly equal to the first order linearization of the difference $\ell(\beta_{t-1}, \alpha_{t-1}^i) - \ell(\beta_{t-1}, \alpha_t)$ and thus L_t^i can be interpreted as a cumulative loss over time. With these definitions in hand AgACI defines the probabilities

$$\widetilde{p}_t^i := \frac{\eta_t^i \exp(-\eta_t^i L_t^i)}{\sum_{j=1}^k \eta_t^j \exp(-\eta_t^j L_t^j)},$$

and outputs the estimate $\alpha_t := \sum_{i=1}^k \widetilde{p}_t^i \alpha_t^i$.

The primary difference between AgACI and our method is the relative weight given to the historical performance of the experts. To see this, we first observe that by unravelling the DtACI updates, the DtACI weights can be re-written as a mixture distribution where element s considers the most recent s losses. More precisely, we have

$$w_{t+1}^i = \sum_{s=0}^{t} (1-\sigma)^{t-s} \bar{W}_s \left(\frac{\sigma}{k}\right)^{\mathbb{1}\{s\neq 0\}} \exp\left(-\eta \sum_{j=s+1}^{t} \ell(\beta_j, \alpha_j^i)\right),$$

where for ease of notation we have set $\bar{W}_0 = 1$. Without any formal analysis, it can be immediately seen that the more recent datapoints appear more often in this mixture and thus contribute more to our choice of weights. On the other hand, the AgACI weights are based off a cumulative sum of all previous losses and thus assign a similar degree of importance to all historical data-points. The upside of this choice is that in environments where the rate of distribution shift is constant, AgACI can effectively converge on a single optimal step-size. However, this comes at the cost of reduced adaptivity over time. For instance, if the environment starts in a state of slow distribution drift, but then undergoes an abrupt shift, AgACI can fail to increase the step-size quickly and thus be slow to react to the change. Empirical examples demonstrating these properties are given in Section 4.1.

The second alternative to ACI that we consider is the multivalid conformal prediction (MVP) method of Bastani et al. (2022). Instead of targeting the optimal parameter α_t^*, this algorithm chooses α_t in order to explicitly obtain the desired long-term miscoverage frequency $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} \text{err}_t = \alpha$. In addition, MVP is also designed to satisfy threshold-calibrated coverage, i.e. for every τ, $\lim_{T\to\infty}(\sum_{t=1}^{T}\mathbb{1}\{\alpha_t = \tau\})^{-1}\sum_{t=1}^{T}\text{err}_t\mathbb{1}\{\alpha_t=\tau\} = \alpha$. At a high level, this is accomplished by setting a grid of possible choices for α_t, and then at each time step outputting the value in the grid that has produced the best historical coverage.

While MVP can perform well in stationary environments where there exists a single optimal choice for the threshold, it does not give significant adaptivity to local changes. This is demonstrated by our experiments in Section 4.1 where MVP fails to adjust to the local variation in α_t^*. For a more complete description of the MVP algorithm see Section A of the Appendix.

Finally, we emphasize that the good local coverage properties of DtACI are not solely an empirical phenomena. Indeed, in the next section we give bounds that control the difference between the estimates α_t produced by DtACI and the optimal values, α_t^* over any local time interval. This theory is new to our methods and no similar results exist for AgACI or MVP.

3. Coverage Properties of DtACI

In this section we outline the main coverage guarantees of DtACI. We begin by drawing from known results in the online convex optimization literature that bound a quantity known as the dynamic regret of DtACI. We then draw a connection between this regret and the coverage. Finally, we evaluate the long-term coverage in a specialized case where the hyperparameters η and σ decay to 0 over time. All proofs are deferred to the Appendix.

3.1 Dynamic Regret of DtACI

Our first result quantifies the error in the expert aggregation scheme by bounding the difference between the loss we obtain and that of the best expert.

Lemma 2 (Modified version of Lemma A.2 in Gradu et al. (2023)) *Assume that* $\sigma \leq 1/2$*. Then, for any interval* $I = [r,s] \subseteq [T]$ *and any* $1 \leq i \leq k$*,*

$$\sum_{t=r}^{s} \mathbb{E}[\ell(\beta_t, \alpha_t)] \leq \sum_{t=r}^{s} \ell(\beta_t, \alpha_t^i) + \eta \sum_{t=r}^{s} \mathbb{E}[\ell(\beta_t, \alpha_t)^2] + \frac{1}{\eta}\left(\log(k/\sigma) + |I| 2\sigma\right), \tag{5}$$

where the expectation is over the randomness in Algorithm 1 and the data $\beta_1, \dots, \beta_T$ *can be viewed as fixed.*

With this lemma in hand, we now turn to our true target, namely the values α_t^* defined in Section 2.3. Our first step is to recall the following regret bound for gradient descent with a dynamic target.

Lemma 3 (Application of Theorem 10.1 of Hazan (2019)) *For any fixed interval* $I = [r,s]$*, sequence* $\alpha_r^*, \dots, \alpha_s^*$ *, and* $1 \leq i \leq k$*,*

$$\sum_{t=r}^{s} \ell(\beta_t, \alpha_t^i) - \sum_{t=r}^{s} \ell(\beta_t, \alpha_t^*) \leq \frac{3}{2\gamma_i}(1+\gamma_i)^2 \left(\sum_{t=r+1}^{s} |\alpha_t^* - \alpha_{t-1}^*| + 1 \right) + \frac{1}{2}\gamma_i |I|.$$

By combining the previous two lemmas we obtain the main result of this section.

Theorem 4 *Let* $\gamma_{\max} := \max_{1 \leq i \leq k} \gamma_i$ *and assume that* $\gamma_1 < \gamma_2 < \cdots < \gamma_k$ *with* $\gamma_{i+1}/\gamma_i \leq 2$ *for all* $1 < i \leq k$*. Assume additionally that* $\gamma_k \geq \sqrt{1 + 1/|I|}$ *and* $\sigma \leq 1/2$*. Then, for any interval* $I = [r,s] \subseteq [T]$ *and any sequence* $\alpha_r^*, \dots, \alpha_s^* \in [0,1]$*,*

$$\frac{1}{|I|} \sum_{t=r}^{s} \mathbb{E}[\ell(\beta_t, \alpha_t)] - \frac{1}{|I|} \sum_{t=r}^{s} \ell(\beta_t, \alpha_t^*) \leq \frac{\log(k/\sigma) + 2\sigma|I|}{\eta |I|} + \frac{\eta}{|I|} \sum_{t=r}^{s} \mathbb{E}[\ell(\beta_t, \alpha_t)^2]$$
$$+ 4(1+\gamma_{\max})^2 \max\left\{ \sqrt{\frac{\sum_{t=r+1}^{s} |\alpha_t^* - \alpha_{t-1}^*| + 1}{|I|}}, \gamma_1 \right\},$$

where the expectation is over the randomness in Algorithm 1 and the data $\beta_1, \dots, \beta_T$ *can be viewed as fixed.*

If we assume that $\gamma_1 \leq \sqrt{\frac{\sum_{t=r+1}^s |\alpha_t^* - \alpha_{t-1}^*| + 1}{|I|}}$ and take the optimal choices $\sigma = 1/(2|I|)$ and $\eta = \sqrt{\frac{\log(2k|I|)+1}{\sum_{t=r}^s \mathbb{E}[\ell(\beta_t, \alpha_t)^2]}}$, we obtain the much simpler bound,

$$\begin{aligned}\frac{1}{|I|}\sum_{t=r}^s \mathbb{E}[\ell(\beta_t,\alpha_t)] - \frac{1}{|I|}\sum_{t=r}^s \ell(\beta_t,\alpha_t^*) &\leq 2\sqrt{\frac{\log(2k|I|)+1}{|I|}}\sqrt{\frac{1}{|I|}\sum_{t=r}^s \mathbb{E}[\ell(\beta_t,\alpha_t)^2]} \\ &\quad + 4(1+\gamma_{\max})^2\sqrt{\frac{\sum_{t=r+1}^s |\alpha_t^* - \alpha_{t-1}^*| + 1}{|I|}} \\ &= O\left(\sqrt{\frac{\log(|I|)}{|I|}}\right) + O\left(\sqrt{\frac{\sum_{t=r+1}^s |\alpha_t^* - \alpha_{t-1}^*|}{|I|}}\right).\end{aligned}$$

The quantity $\frac{\sum_{t=r+1}^s |\alpha_t^* - \alpha_{t-1}^*|}{|I|}$ can be viewed as a one-dimensional quantification of the size of the distribution shift in the environment. Thus, Theorem 4 gives a direct control on the average peformance of DtACI in terms of the distribution shift. We emphasize that this result holds over *any* interval $|I|$ of a fixed length, justifying our earlier claim that DtACI is able to adapt to the distribution shift locally over all time steps.

Unfortunately, the values for σ and η specified above are not usable in practice since they depend on both the size of the time interval $|I|$ and the non-constant value $\sum_{t=r}^s \mathbb{E}[\ell(\beta_t, \alpha_t)]^2$. For this first issue, the user can pick any interval size of interest, with the consideration that choosing larger intervals gives a tighter bound at the cost of weaker local guarantees. In our experiments, we will set η and σ using the choice $|I| = 500$.

For the second issue, we give two options. The first is to note that in the idealized setting, where there is no distribution shift, we would have $\beta_t \sim \text{Unif}(0,1)$ and $\alpha_t^i \cong \alpha_t^* = \alpha$. Plugging in these approximations we obtain

$$\frac{1}{|I|}\sum_{t=r}^s \mathbb{E}[\ell(\beta_t,\alpha_t)^2] \cong \mathbb{E}_{\beta\sim\text{Unif}(0,1)}[\ell(\beta,\alpha)^2] = \frac{(1-\alpha)^2\alpha^2}{3},$$

and substituting this value into the expression for η above gives the choice $\eta = \sqrt{\frac{3}{500}}\sqrt{\frac{\log(2k\cdot 500)+1}{(1-\alpha)^2\alpha^2}}$.

Our second option, is to simply update η in an online fashion through the equation

$$\eta = \eta_t := \sqrt{\frac{\log(2k\cdot 500)+1}{\sum_{s=t-501}^t \mathbb{E}[\ell(\beta_s,\alpha_s)^2]}}.$$

This choice would allow us to adaptively track any changes in $\frac{1}{500}\sum_{s=t-501}^t \mathbb{E}[\ell(\beta_s,\alpha_s)^2]$ across time. In the Appendix, we prove a generalization of Theorem 4 that allows $\eta = \eta_t$ to vary across time. We find that a dynamic choice of η_t offers the same regret guarantees as a fixed choice so long as the variability in η_t is not too large. Hence, adaptive values for η_t can be used to minimize the regret bound of Theorem 4. On the other hand, empirically, we find that on real data the approximation $\frac{1}{|I|}\sum_{t=r}^s \mathbb{E}[\ell(\beta_t,\alpha_t)^2] \cong \frac{(1-\alpha)^2\alpha^2}{3}$ is highly accurate. Thus, the two different choices for η give nearly identical results in practice. For ease of presentation in the sections that follow, we will only display results using the first fixed, heuristic choice of η. Results for the variable choice are given in the Appendix.

3.2 Bounds on the Short-Term Coverage

The previous section gives bounds on the performance of α_t in terms of the pinball loss $\ell(\beta_t, \alpha_t)$. However, the pinball loss is not our true objective, and our primary goal is to obtain a value of α_t that is close to α_t^*. Our next result provides a direct connection between bounds on $\ell(\beta_t, \alpha_t)$ and bounds on $(\alpha_t - \alpha_t^*)^2$.

Proposition 5 *Let β be a random variable and assume that there exists a value α^* such that $\mathbb{P}(\beta < \alpha^*) = \alpha$. Then, for any τ,*

$$\mathbb{E}[\ell(\beta, \tau)] - \mathbb{E}[\ell(\beta, \alpha^*)] = \begin{cases} \mathbb{E}[(\tau - \beta)\mathbb{1}_{\alpha^* < \beta \leq \tau}], & \text{if } \tau \geq \alpha^*, \\ \mathbb{E}[(\beta - \tau)\mathbb{1}_{\tau < \beta \leq \alpha^*}], & \text{if } \tau < \alpha^*. \end{cases}$$

So, in particular, if β has a density $p(\cdot)$ on $[0,1]$ with $p(x) \geq p > 0$ for all $x \in [0,1]$, then

$$\mathbb{E}[\ell(\beta, \tau)] - \mathbb{E}[\ell(\beta, \alpha^*)] \geq \frac{p(\tau - \alpha^*)^2}{2}.$$

Now, let α_t^* be any value satisfying $\mathbb{P}(Y_t \in \hat{C}_t(\alpha_t^*) | \{\beta_s\}_{s<t}) = 1 - \alpha$. Then, combining Proposition 5 with the results from Section 3.1 we obtain the desired bound on $(\alpha_t - \alpha_t^*)^2$,

$$\frac{1}{|I|} \sum_{t=r}^{s} \frac{p\mathbb{E}[(\alpha_t - \alpha_t^*)^2]}{2} \leq O\left(\sqrt{\frac{\log(|I|)}{|I|}}\right) + O\left(\sqrt{\frac{\sum_{t=r+1}^{s} \mathbb{E}[|\alpha_t^* - \alpha_{t-1}^*|]}{|I|}}\right), \tag{6}$$

where the expectation is now over the randomness in both Algorithm 1 and $\{\beta_t\}_{t \leq s}$, and p is any lower bound on the density of $\beta_t, |\{\beta_r\}_{r<s}, \forall t \leq s$. Similarly, if we additionally assume that $\beta \mapsto \mathbb{P}(Y_t \in \hat{C}_t(\beta) | \{\beta_s\}_{s<t})$ is L-Lipschitz, then $|\mathbb{P}(Y_t \in \hat{C}_t(\beta) | \{\beta_s\}_{s<t}) - (1 - \alpha)| \leq L|\alpha_t - \alpha_t^*|$ and thus (6) can also be read as a bound on the local coverage of DtACI. Such a Lipschitz assumption may be reasonable if the distribution of the conformity scores and our estimates of its quantiles are sufficiently smooth.

In Gibbs and Candès (2021), it was shown that adaptive conformal inference satisfies a similar bound to (6). However, that result required three major assumptions: 1) the size of the distribution shift $\frac{\sum_{t=r+1}^{s} |\alpha_t^* - \alpha_{t-1}^*|}{|I|}$ is known, 2) (X_t, Y_t) is generated by a hidden Markov model, and 3) the regression model is not re-fit across time. In stark contrast, here we make no such assumptions and, in addition, give a result that is adaptive to changes in the size of the distribution shift across time. In practical settings, none of these three assumptions can be reasonably expected to hold and thus our new results constitute a significant generalization of those in Gibbs and Candès (2021).

3.3 Bounds on the Long-Term Coverage

The results of the previous section show that DtACI obtains a coverage rate close to $1 - \alpha$ over any local time interval. It is natural to ask if elongating this interval leads to an average coverage of exactly $1 - \alpha$. Here, we show that this is indeed the case if the parameters $\eta = \eta_t \to 0$ and $\sigma = \sigma_t \to 0$. At a high-level, sending η and σ to 0 causes the method to put more weight on older historical data and thus gives a version of DtACI that is closer to

AgACI and MVP. Prior work has shown that MVP also obtains exact long-term coverage (Bastani et al. (2022)), while for AgACI, this property has only been observed empirically (Zaffran et al. (2022)).

Theorem 6 *Consider a modified version of Algorithm 1 in which on iteration t the parameters η and σ are replaced by values η_t and σ_t. Let $\gamma_{\min} := \min_i \gamma_i$ and $\gamma_{\max} := \max_i \gamma_i$. Then,*

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\text{err}_t] - \alpha \right| \leq \frac{1+2\gamma_{\max}}{T\gamma_{\min}} + \frac{(1+2\gamma_{\max})^2}{\gamma_{\min}} \frac{1}{T} \sum_{t=1}^{T} \eta_t e^{\eta_t(1+2\gamma_{\max})} + 2\frac{1+\gamma_{\max}}{\gamma_{\min}} \frac{1}{T} \sum_{t=1}^{T} \sigma_t,$$

where the expectation is over the randomness in Algorithm 1 and the data $\beta_1, \dots, \beta_T$ can be viewed as fixed. So, in particular, if $\lim_{t\to\infty} \eta_t = \lim_{t\to\infty} \sigma_t = 0$, then $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \text{err}_t \overset{a.s.}{=} \alpha$.

Following the discussion of the previous sections, decaying values of η_t and σ_t should only be used if the size of the distribution shift in the environment is known to be stationary. Since we do not consider this to be a realistic assumption in most situations, we advocate for using constant or slowly varying values for η and σ. For these choices, we empirically find that DtACI can produce intervals that are biased in the sense that $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \text{err}_t \neq \alpha$. However, in all the examples we have investigated, this bias is sufficiently small to be of little practical consequence.

3.4 Removing Randomness in the Choice of α_t

In practical settings, the randomness in the choice of α_t may be undesirable. To rectify this, we provide an alternative approach that replaces the choice $\alpha_t \sim \sum_{i=1}^{k} p_t^i \delta_{\alpha_t^i}$ with $\bar{\alpha}_t = \sum_{i=1}^{k} p_t^i \alpha_t^i$. The full version of this procedure is stated in Algorithm 2. Importantly, this new method admits the same regret bound as our original procedure.

Corollary 7 *Under the same conditions, the conclusion of Theorem 4 holds with α_t replaced by $\bar{\alpha}_t$ on the left-hand side of the display.*

Proof This is an immediate consequence of Jensen's inequality. ■

Unsurprisingly, we find that in practice Algorithms 1 and 2 produce nearly identical results. Thus, we will abuse terminology and also refer to Algorithm 2 as DtACI. For simplicity, the following sections show results only for Algorithm 2.

4. Empirical Results

In this section we investigate the performance of DtACI as well as the previously proposed AgACI and MVP methods. In all experiments the set of candidate γ values is taken to be $\{0.001, 0.002, 0.004, 0.008, 0.0160, 0.032, 0.064, 0.128\}$. Code for reproducing these results can be found at `https://github.com/isgibbs/DtACI`.

Algorithm 2:

```
Data: Observed values $\{\beta_t\}_{1\leq t\leq T}$, set of candidate $\gamma$ values $\{\gamma_i\}_{1\leq i\leq k}$, starting
      points $\{\alpha_1^i\}_{1\leq i\leq k}$, and parameters $\sigma$ and $\eta$.
$w_1^i \leftarrow 1,\ 1 \leq i \leq k$;
for $t = 1, 2, \ldots, T$ do
    Define the probabilities $p_t^i := w_t^i / \sum_{1\leq j\leq k} w_t^j$, $\forall 1 \leq i \leq k$;
    Output $\bar{\alpha}_t = \sum_{1\leq i\leq k} p_t^i \alpha_t^i$;
    $\bar{w}_t^i \leftarrow w_t^i \exp(-\eta \ell(\beta_t, \alpha_t^i)),\ \forall 1 \leq i \leq k$;
    $\bar{W}_t \leftarrow \sum_{1\leq i\leq k} \bar{w}_t^i$;
    $w_{t+1}^i \leftarrow (1-\sigma)\bar{w}_t^i + \bar{W}_t \sigma / k$;
    $\mathrm{err}_t^i := \mathbb{1}\{Y_t \notin \hat{C}_t(\alpha_t^i)\}$, $\forall 1 \leq i \leq k$;
    $\mathrm{err}_t := \mathbb{1}\{Y_t \notin \hat{C}_t(\bar{\alpha}_t)\}$;
    $\alpha_{t+1}^i = \alpha_t^i + \gamma_i(\alpha - \mathrm{err}_t^i)$, $\forall 1 \leq i \leq k$;
```

4.1 Simulated Examples

We begin by considering a set of simulated examples in which we can exactly measure the local coverage properties of the methods. To make the results interpretable, we focus on a simple setting in which we observe a sequence of independent random variables $\{Y_t\}_{t=1}^T$, where $Y_t \sim \mathcal{N}(\mu_t, 1)$, and form prediction sets using the standard normal distribution as our quantile estimator, i.e. we set

$$\hat{C}_t(\alpha_t) := \{y : y \leq \text{Quantile}(1-\alpha_t, \mathcal{N}(0,1))\}.$$

We consider three different choices for the sequence of means $\{\mu_t\}_{t=1}^T$:

- A *stationary* setting in which $\mu_t = 0$ is held constant and thus the data are i.i.d..
- A *smooth* shift setting in which μ_t drifts continuously across time. More precisely, we set $\mu_0 = 0$ and
$$\mu_{t+1} = \mu_t + \frac{1}{2}(\mu_t - \mu_{t-1}) + \frac{1}{2}\epsilon_t,$$
where $\{\epsilon_t\} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.006)$.
- A setting with *jump* shifts in which μ_t undergoes jump discontinuities of various sizes. In particular, we divide the time period into three equally sized intervals. In the first and third interval μ_t oscillates between -0.075 and 0.075, while in the second interval the distribution shifts are larger and μ_t oscillates between -1.5 and 1.5.

The final trajectories of μ_t generated in all three settings are shown in Figure 4.1. Figure 4.2 shows the corresponding trajectories of α_t^* as well as the estimates, α_t, produced by DtACI, AgACI, and MVP. We can immediately see that DtACI and AgACI accurately adapt to the local changes in α_t^* across all three settings, while MVP only performs well in the stationary environment.

To get a more fine-grained comparison of the relative performance of DtACI, AgACI, and MVP we additionally compute the time-instantaneous coverages of these methods. Namely,

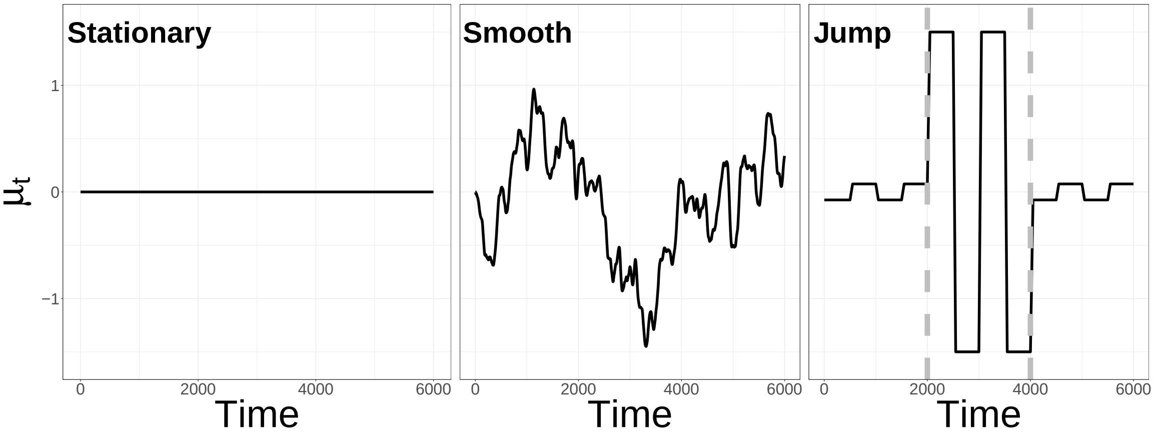

Figure 4.1: Trajectories for μ_t in the *stationary*, *smooth*, and *jump* settings. To aid readability, the trajectory of μ_t in the center panel has been locally averaged over a moving time interval of width 50. Vertical grey dotted lines in the jump shift plot denote the regime switches where the size of the distribution shift changes.

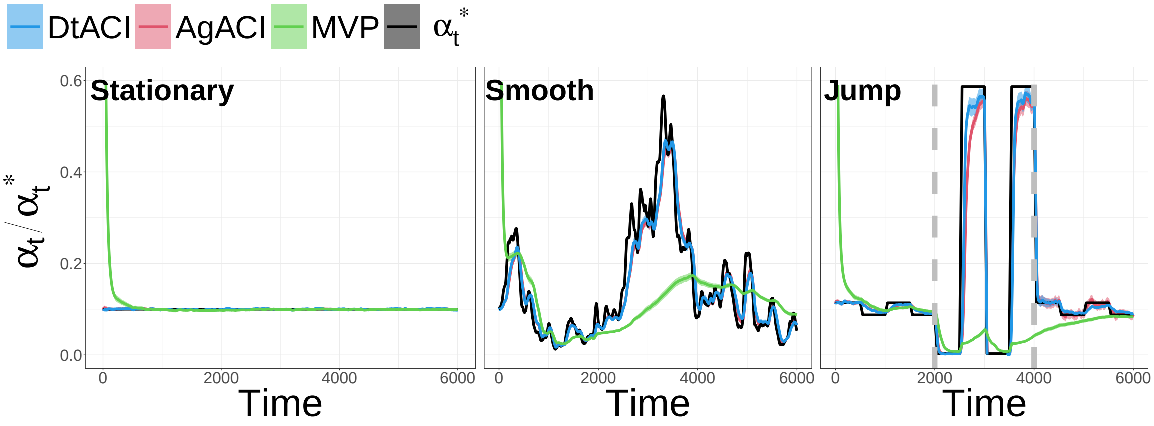

Figure 4.2: Comparison of the simulated trajectories of α_t^* (black) against the estimated values, α_t output by DtACI (blue), AgACI (red), and MVP (green). Solid lines display averages across 100 trials with μ_t (and thus α_t^*) held fixed and α_t regenerated using independent draws of $\{Y_t\}$. Shaded regions show confidence intervals for the corresponding means. To aid readability, means and confidence intervals have been locally averaged over a moving time interval of width 50. Finally, the vertical grey dotted lines in the jump plot denote the regime switches where the size of the distribution shift changes.

let α_t^D, α_t^A, and α_t^M denote the values output by DtACI, AgACI, and MVP at time step t. Then, we compute the instantaneous coverage gaps $\text{CG}_x = |\mathbb{P}(Y_t \in \hat{C}(\alpha_t^x)|\alpha_t^x) - (1-\alpha)|$ for $x \in \{D, A, M\}$ and we plot the difference between the values obtained by AgACI and MVP and those obtained by DtACI (i.e. the values $\text{CG}_A - \text{CG}_D$ and $\text{CG}_M - \text{CG}_D$). Thus, in the results that follow a value of 0 indicates identical performance, while positive/negative values indicate that DtACI performs better/worse than the competitors.

The resulting coverage performances are shown in Figure 4.3. Overall, we find that DtACI offers greater adaptivity and more precise coverage than AgACI and MVP in the non-stationary settings, while suffering only a slight degradation in performance under

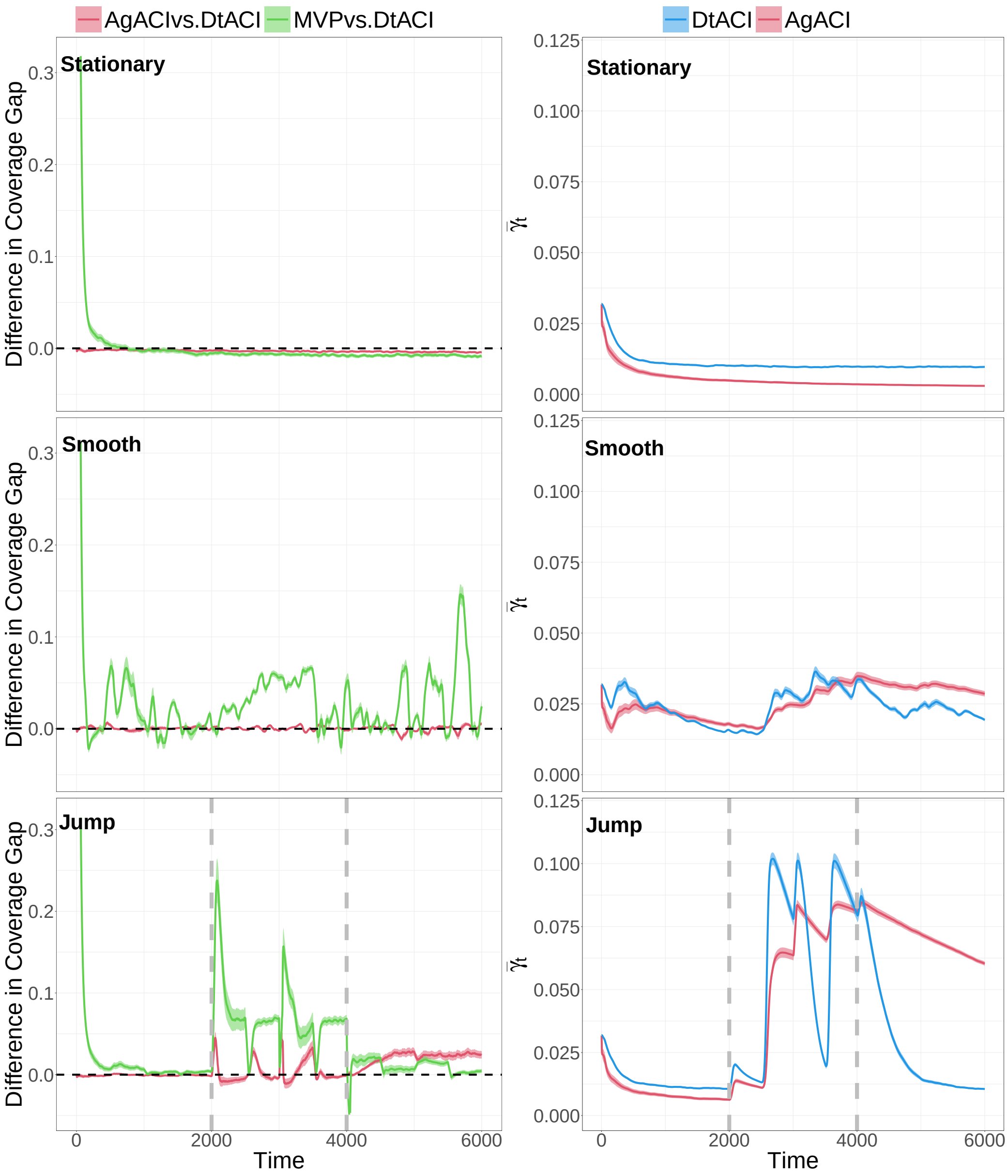

Figure 4.3: Comparison of the coverage gaps obtained by AgACI and MVP against the baseline of DtACI (left-panels) and the mean step-size trajectories output by DtACI and AgACI (right-panels). Solid lines display averages across 100 trials, while shaded regions show confidence intervals for the corresponding means. To aid readability, means and confidence intervals have been locally averaged over a moving time interval of width 50. Finally, vertical grey lines in the jump shift plots denote the regime switches where the size of the distribution shift changes, while horizontal black lines in the coverage gap plots denote the value 0 at which the performance of DtACI exactly matches the competing methods.

stationarity. More specifically, our results for each of the three settings can be summarized as follows.

- *Stationary setting:* In the stationary setting MVP and AgACI slightly outperform DtACI. This is due to the fact that MVP and AgACI are able to more precisely converge to the single optimal value for α_t^*, while DtACI can never set its step-size to exactly 0 and thus maintains some minor fluctionations around α_t^* across all time steps.

- *Smooth shift setting:* For non-stationary data MVP now gives significantly worse performance and shows little adaptivity to the distribution shifts. On the other hand, both AgACI and DtACI perform well and are able to approximately track the changes in α_t^* over time. Overall, the results for AgACI and DtACI are nearly identical, which is expected since a single choice of the step-size is sufficient to give good performance in this environment.

- *Jump shift setting:* Once again MVP fails to adapt to the distribution shifts. Additionally, while AgACI does show reasonable adaptivity, it fails to adjust its step-size to track the changes in the environment. We visualize this behaviour in the right panels of Figure 4.3, which show the average step-sizes, $\bar{\gamma}_t = \sum_{i=1}^k p_t^i \gamma_i$ produced by AgACI and DtACI. We see that AgACI correctly gives a small step-size in the first phase when the distribution shifts are small. However, once when we enter the second stage and the distribution shifts jump in magnitude, AgACI is slow to react and its step-size lags behind that of DtACI. This behaviour is amplified in the third stage during which AgACI never decreases its step-size to match the smaller distribution shifts and thus suffers large coverage errors throughout.

Overall, we find that while AgACI and MVP perform slightly better than DtACI in the stationary setting, this comes at the cost of a greater loss of adaptivity in the non-stationary settings.

4.2 Real Data Examples

We now compare AgACI, MVP, and DtACI on two real-world datasets.

4.2.1 Online Prediction in the Stock Market

For our first real-world data example, we return to a stock market prediction task that was originally used to evaluate ACI (Gibbs and Candès (2021)). In this problem, the goal is to use the previously observed values of a stock price, $\{P_s\}_{s=0,\dots,t-1}$, to predict its volatilty at the next step, defined as

$$V_t := \left(\frac{P_t - P_{t-1}}{P_{t-1}}\right)^2.$$

To predict the volatility, we model the stock returns $R_t := \frac{P_t - P_{t-1}}{P_{t-1}}$ as coming from a GARCH(1,1) design. This is a classical financial model for market dynamics in which it is assumed that $R_t = \sigma_t \epsilon_t$ with $\epsilon_t \sim \mathcal{N}(0,1)$ and

$$\sigma_t^2 = \omega + \tau V_{t-1} + \lambda \sigma_{t-1}^2,$$

for some unknown parameters ω, τ, λ. At each time step, t, we use the most recent 1250 days of returns $\{R_s\}_{t-1250\leq s<t}$ to produce estimates $\hat{\omega}_t$, $\hat{\tau}_t$, $\hat{\lambda}_t$, $\{\hat{\sigma}_t^s\}_{s<t}$, and a one-step ahead prediction $(\hat{\sigma}_t^t)^2 = \hat{\omega}_t + \hat{\tau}_t V_{t-1} + \hat{\lambda}_t(\hat{\sigma}_t^{t-1})^2$. We then construct prediction sets using the equation

$$\hat{C}_t(\alpha_t) := \left\{ v : S_t(v) \leq \text{Quantile}\left(1-\alpha_t, \frac{1}{1250}\sum_{s=t-1250}^{t-1} \delta_{S_s(V_s)}\right)\right\},$$

where $S_t(v)$ is taken to be either the normalized conformity score $S_t(v) = |v - (\hat{\sigma}_t^t)^2|/(\hat{\sigma}_t^t)^2$ or the unnormalized score $S_t(v) := |v - (\hat{\sigma}_t^t)^2|$.

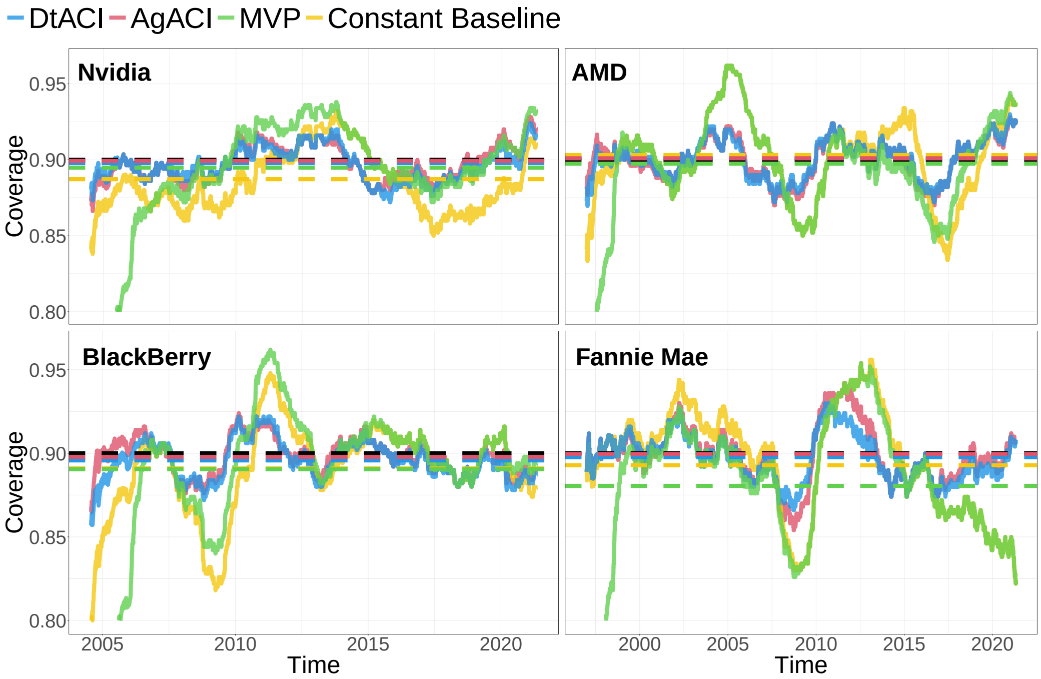

Figure 4.4: Estimation of stock market volatility using conformity score $S_t(v) = |v - (\hat{\sigma}_t^t)^2|/(\hat{\sigma}_t^t)^2$. Solid lines show the local coverage level for DtACI (blue), AgACI (red), MVP (green), and a naive baseline that holds $\alpha_t = \alpha$ fixed (yellow). Dashed lines indicate the global coverage frequency over all time steps for the same methods. Finally, the black dashed lines indicates the target level of $1-\alpha = 0.9$. Note that in some of the panels (e.g. the top-right AMD panel) the yellow and green lines exactly overlap for many time steps leaving only the green line clearly visible. Some overlap is also observed in the red and blue lines.

In Gibbs and Candès (2021) it was found that the normalized and unnormalized conformity scores lead to distribution shifts of drastically different sizes. To gain some intuition as to why this is the case, let $W \sim \chi_1^2$ and observe that if the GARCH(1,1) model is true, and moreover $(\hat{\sigma}_t^t)^2 = \sigma_t^2$ is an exactly accurate prediction, then the normalized score is distributed as $S_t(V_t) \overset{D}{=} |W-1|$, while the unnormalized score follows $S_t(V_t) \overset{D}{=} \sigma_t^2|W-1|$. Thus, in this setting, α_t^* will be invariant across time for the normalized score and highly variable for the unnormalized score. Consistent with this intuition, Gibbs and Candès (2021) found that the distribution shift is much larger for the unnormalized score than the normalized score and, thus, in order to obtain good local coverage, ACI requires different (and *a priori* unknown) step sizes for the two scores. In contrast, as we will show shortly, DtACI obtains good performance in both conditions without any prior knowledge of the distribution shift.

Similar to the previous section, we measure the performance of the prediction sets by their local coverage. Since the true time-instantaneous coverage is no longer an observable quantity, we instead compute empirical local average coverage rates over a moving 500-day window, i.e. we compute the moving average

$$\text{LocalCov}_t := 1 - \frac{1}{500}\sum_{t-250+1}^{t+250} \text{err}_t. \tag{7}$$

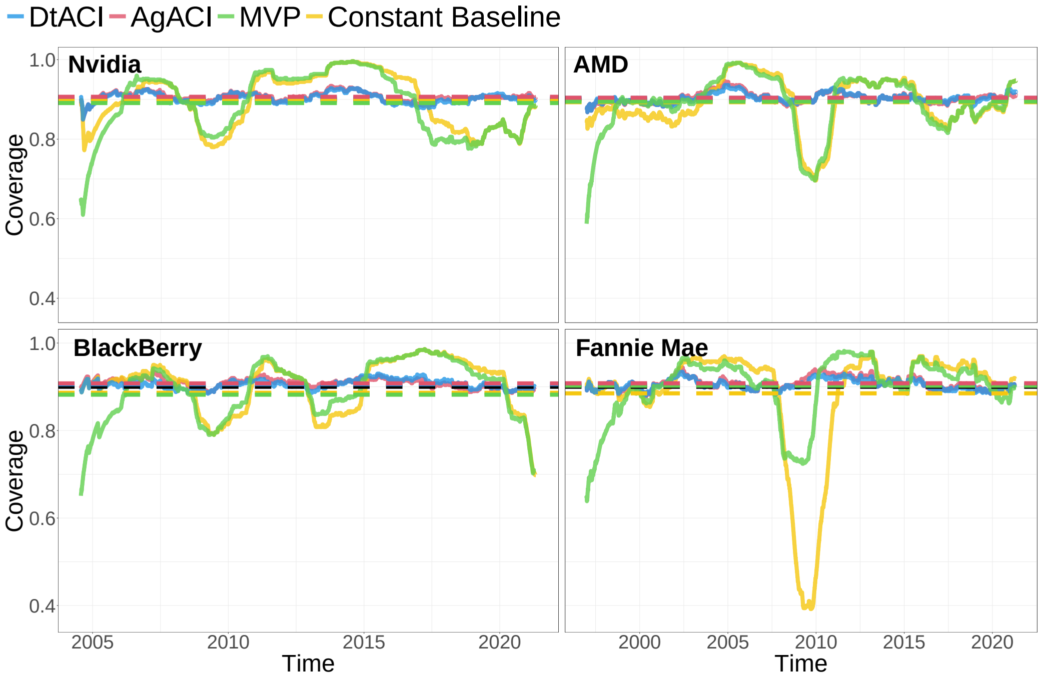

Figure 4.5: Estimation of stock market volatility using conformity score $S_t(v) = |v - (\hat{\sigma}_t^t)^2|$. Solid lines show the local coverage level for DtACI (blue), AgACI (red), MVP (green), and a naive baseline that holds $\alpha_t = \alpha$ fixed (yellow). Dashed lines indicate the global coverage frequency over all time steps for the same methods. Finally, the black dashed lines indicates the target level of $1 - \alpha = 0.9$.

Figures 4.4 and 4.5 show the local average coverage rates obtained by DtACI, AgACI, and MVP, using the normalized and unnormalized conformity scores on four different stocks. In addition, we have also plotted the coverage values for a naive baseline method that holds $\alpha_t = \alpha$ fixed. At a high level, this baseline essentially measures the size of the underlying shifts in the environment. Finally, as an additional reference, Figure D.6 in the Appendix shows the prices of the stocks over the same period.

All four stocks demonstrate obvious price swings leading to clear distribution shifts in the data. Similar to our simulated examples, we find that DtACI is able adapt to both the presence and size of these shifts and obtain a local coverage rate near the target level of $1 - \alpha = 0.9$ over all time steps and conditions. Overall, the size of the distribution shifts in these environments appears relatively constant, and thus AgACI produces nearly identical results to DtACI. Perhaps the only exception to this is the Fannie Mae data in Figure 4.4 for which there is a small time window following the 2009 financial crisis where AgACI is slower to react to the sharp rise in volatility. Finally, we find that MVP shows minimal adaptivity to the underlying shifts throughout and often performs nearly identically to the fixed baseline method.

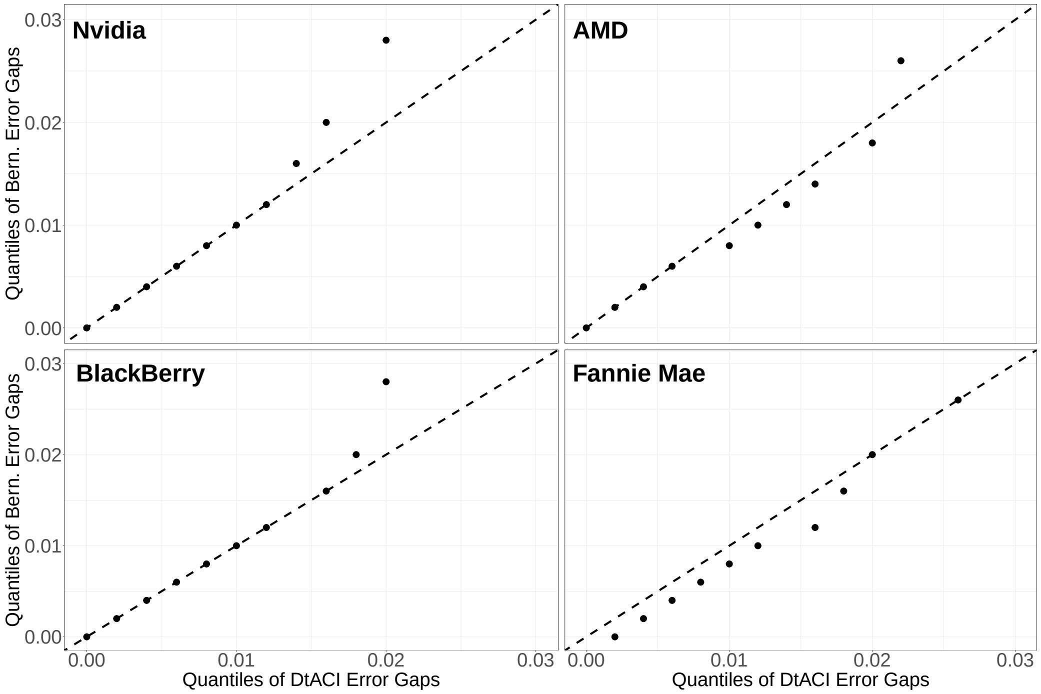

Figure 4.6: Q-Q plots comparing the distribution of local coverage gaps obtained by an i.i.d. Bernoulli(α) sequence against those realized by DtACI. The dashed line indicates the ideal situation of exact equality.

While DtACI seems to perform reasonably well across all settings, one may still wonder if additional improvements can be made. Namely, is it possible for a sensible method to produce local coverage errors that are tighter to the $1 - \alpha$ line? To answer this question, we compare the coverage properties of DtACI against the ideal situation in which the coverage errors are an i.i.d. Bernoulli(α) sequence. Figure 4.6 shows Q-Q plots comparing the empirical distributions of $|\text{LocalCov}_t - (1-\alpha)|$ realized by DtACI against the distribution of the same quantity for an i.i.d. Bernoulli(α) sequence. We find that the two distributions tightly align across all four stocks, indicating that in some sense the local coverage properties of DtACI are difficult to improve. A small exception occurs in the right tail on the Nvidia and Blackberry data (left two panels), where the local coverage error produced by our method is smaller than what would be expected from Bernoulli(α) random variables. As a final remark, note that while Figure 4.6 only shows results for the normalized conformity scores, we also obtained similar results for the unnormalized scores (see Figure D.4 in the Appendix).

4.2.2 Evaluating the Reactivity of DtACI

In Bastani et al. (2022), adaptive conformal inference was criticized for failing to provide a coverage guarantee conditional on the value of α_t. Indeed, one may be concerned that since ACI acts reactively by widening/shrinking its prediction sets in response to past mistakes, good local coverage is obtained not due to successfully learning α_t^*, but rather as a result of the simple tendency of the algorithm to correct its prior under/over-coverage. Our simulations in Section 4.1 already partially show that DtACI can successfully learn α_t^*. Here, we provide further evidence demonstrating that DtACI is not simply acting reactively on real data.

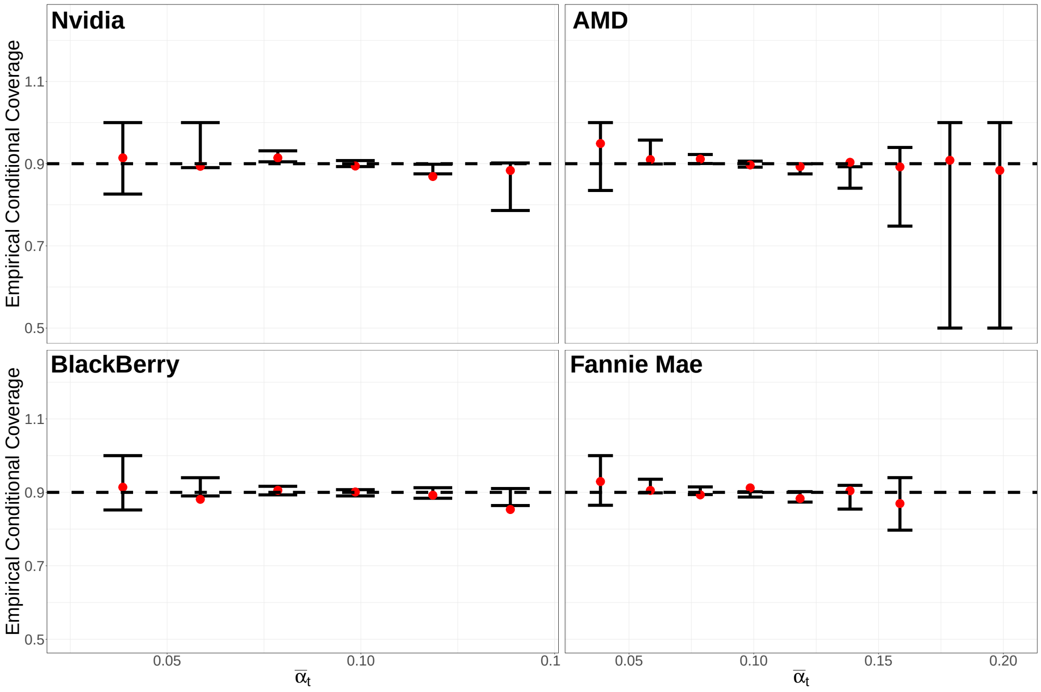

Figure 4.7: Empirical conditional coverage of $\bar{\alpha}_t$ for the estimation of stock market volatility with conformity score $S_t(v) = |v - (\hat{\sigma}_t^t)^2|/(\hat{\sigma}_t^t)^2$. Red points show the empirical conditional coverage given $\bar{\alpha}_t \in B_i$ with error bars indicating the corresponding 0.025 and 0.975 quantiles across 100 block bootstrap resamples of the data $\{(X_t, Y_t)\}$ with block-size 100. For visual clarity all error bars are truncated to the range $[0.5, 1]$. Black dashed lines shows the target level of $1 - \alpha = 0.9$.

To do this, we divide the interval $[0, 1]$ into m evenly sized sub-intervals $B_1, \dots, B_m$ and evaluate the conditional coverage levels

$$\text{CondCoverage}_i := \frac{1}{|\{t : \bar{\alpha}_t \in B_i\}|} \sum_{t:\bar{\alpha}_t \in B_i} \text{err}_t. \tag{8}$$

We apply this to the volatility dataset outlined in the previous section using the normalized conformity score and display our results in Figure 4.7. As a visual aid, this figure contains error bars indicating the 0.025 and 0.975 quantiles of CondCoverage_i across block bootstrap re-samples of the data (see Algorithm 4 in the Appendix for details). These error bars would be expected to give a valid confidence interval for the conditional coverage if, for instance,

$\{(X_t, Y_t)\}$ was a stationary time-series. However, since there is distribution shift in these examples, the error bars are *not* accompanied by any coverage guarantee. Thus, we present them simply as a visual aid to help the reader judge the distance between CondCoverage_i and $1-\alpha$ relative to the sample size $|\{t : \bar{\alpha}_t \in B_i\}|$.

Overall, we find that almost all of the error bars cover the target level of $1-\alpha = 0.9$ and for a large majority of the bins, CondCoverage_i is nearly exactly equal to 0.9. If DtACI was simply acting reactively to its past mistakes we would expect to observe over-coverage at small values of α_t and under-coverage at large values of α_t. Thus, Figure 4.7 provides strong evidence that DtACI is not enacting any such pathological behaviour. Similar results were also obtained using the unnormalized conformity scores (see Figure D.5 in the Appendix).

4.2.3 Predicting Covid-19 Case Counts

Our final example considers the problem of predicting future COVID-19 case counts. We base our methods on the work of Tibshirani (2020) and work with a simple model for generating one-week ahead forecasts of the seven day moving average of the number of confirmed cases of COVID-19 in each county in the United States. In this model, future forecasts are generated based off of the historical prevalence of COVID-19 across the US and Facebook survey data that provides us with a moving seven day average of the proportion of people who report knowing someone in their local community with COVID-19. All data is obtained from a public repository made available by the DELPHI group at Carnegie Mellon (Reinhart et al. (2021)).

Let $\{CO_{t,i}\}_{t,i}$ and $\{F_{t,i}\}_{t,i}$ denote the time series of COVID-19 case counts and Facebook survey responses, respectively, where t indexes time and i indexes one of the 3243 counties in the US. At each time step t we predict $\{CO_{t+7,i}\}_i$ by using least-squares regression to fit the model

$$CO_{s,i} \sim \beta_0^t + \sum_{j=1}^{3} \lambda_j^t CO_{s-7j,i} + \sum_{j=1}^{3} \kappa_j^t F_{s-7j,i},\ \ s = t-14\dots,t,\ \ i = 1,\dots,3243,$$

and setting

$$\widehat{CO}_{t+7,i} = \hat{\beta}_0^t + \sum_{j=1}^{3} \hat{\lambda}_j^t CO_{t-7j,i} + \sum_{j=1}^{3} \hat{\kappa}_j^t F_{t-7j,i},\ \ i = 1,\dots,3243.$$

Because Facebook survey data is not available for all counties at all time steps, we restrict our analysis to those counties with no missing values in the above expressions.

To compute prediction sets for county i, we define the conformity scores $S_{t,i} := |\widehat{CO}_{t,i} - CO_{t,i}|/|CO_{t-7,i} - CO_{t,i}|$ and counts $n_t := |\{i : \text{County } i \text{ has available data at time } t-1\}|$, and set

$$\hat{C}_{t,i}(\alpha_{t,i}) := \left\{ c : \frac{|\widehat{CO}_{t,i} - c|}{|CO_{t-7,i} - c|} \leq \text{Quantile}\left(1 - \alpha_{t,i}, \frac{1}{n_t}\sum_i \delta_{S_{t-1,i}}\right)\right\}.$$

Since different counties will undergo different dynamics at different stages of the pandemic, we run DtACI, AgACI, and MVP separately for each county to obtain a set of county-specific trajectories, $\{\alpha_{t,i}\}_{t,i}$ for α_t.

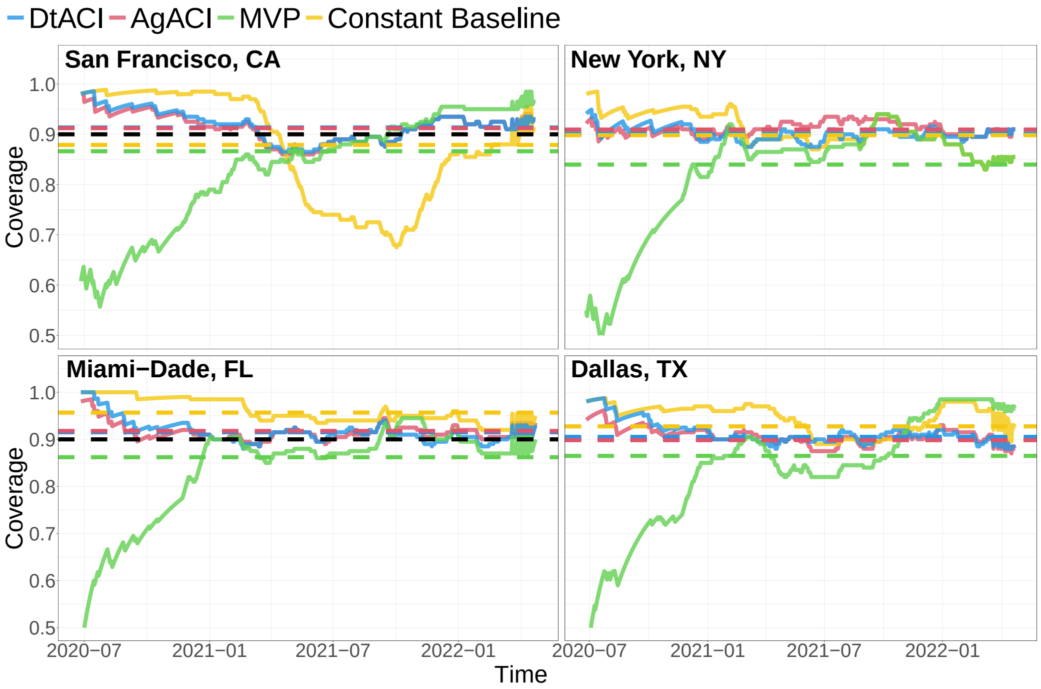

Figure 4.8: Coverage results for the prediction of county-level COVID-19 case counts. Solid lines show the local coverage level for DtACI (blue), AgACI (red), MVP (green), and a naive baseline that holds $\alpha_t = \alpha$ fixed (yellow). Dashed lines indicate the global coverage frequency over all time steps for the same methods. The black dashed lines indicates the target level of $1 - \alpha = 0.9$.

Figure 4.8 shows the empirical local coverage rates over the nearest 200 time steps (i.e equation (7), with 250 replaced by 100) for four US counties. These four counties were chosen because they are large urban centres for which data was available over the entire time window we considered. All four of these counties have undergone multiple waves of COVID-19, each of which caused a large swing in the observed case count (see Figure D.7) and thus induced a clear distribution shift into the data. Much like the previous example, we find that DtACI and AgACI successfully correct for these shifts, while MVP provides inconsistent coverage across the four examples. This contrasts sharply with the baseline method that holds $\alpha_t = \alpha$ fixed, which, depending on the example, undergoes large swings (e.g. top-left panel) or displays a systematic bias (e.g. bottom-left panel) away from the target coverage frequency.

Acknowledgments

E.C. was supported by the Office of Naval Research grant N00014-20-1-2157, the National Science Foundation grant DMS-2032014, the Simons Foundation under award 814641, and the ARO grant 2003514594. I.G. was supported under the Office of Naval Research grant N00014-20-1-2157 and the Simons Foundation under award 814641 as well as by the Overdeck Fellowship Fund. The authors thank Elad Hazan for useful pointers to the online learning literature and John Cherian for helpful comments on an earlier version of this work.

Appendix A. Detailed Description of Multivalid Conformal Prediction

The generic version of multivalid conformal prediction (MVP) proposed by Bastani et al. (2022) is designed to give simultaneous coverage over a collection of subsets of the covariate space. This is accomplished by constructing prediction sets of the form $\{y : S_t(X_t, y) \leq q\}$, where q is chosen from a set of candidate values based off of their performance on historical data. To make this method comparable to DtACI and AgACI we implement a modified version that does not consider any subsets of the covariate space and treats $1 - \beta_t$ as the conformity score. More specifically, our implementation outputs prediction sets of the form $\{y : 1 - \beta_t(y) \leq q\}$, where q is chosen using the MVP algorithm and

$$\beta_t(y) := \max\{0 \leq \beta \leq 1 : S_t^y \leq \text{Quantile}(1 - \beta, \mathcal{D}_t^y)\}.$$

With this construction the value of $1 - q$ output by MVP is exactly anologous to the values α_t output by DtACI and AgACI.

The full details of our implementation are given in Algorithm 3 below. Following the original implementation of MVP in Bastani et al. (2022) we take our hyperparameters to be $m = 40$, $\eta = \sqrt{\frac{\log(m)}{4.2m}}$, and $r = 800000$. Finally, in what follows $f(\cdot)$ refers to the function $f(n) := \sqrt{(n+1)\log(n+2)^2}$

Algorithm 3: Modified version of the MVP algorithm of Bastani et al. (2022).

```
Data: Observed values $\{\beta_t\}_{1 \leq t \leq T}$, target coverage level $\alpha$, number of candidate
      thresholds $m$, hyperparameters $\eta$, $r$.
for $t = 1, 2, \dots, T$ do
    for $i = 0, 1, \dots, m - 1$ do
        $n_t^i \leftarrow \sum_{s<t} \mathbb{1}\{i/m \leq q_s < (i+1)/m \text{ or } i = m-1, q_s = 1\}$;
        $V_t^i \leftarrow \sum_{s<t} \mathbb{1}\{i/m \leq q_s < (i+1)/m \text{ or } i = m-1, q_s = 1\}(\alpha - \mathbb{1}\{\beta_s < 1 - q_s\})$;
        $C_t^i = (\exp(\eta V_t^i / f(n_t^i)) - \exp(-\eta V_t^i / f(n_t^i)))/f(n_t^i)$;
    if $C_t^i > 0$ for all $i \in [m]$ then
        Output $q_t = 0$;
    else if $C_t^i < 0$ for all $i \in [m]$ then
        Output $q_t = 1$;
    else
        Set $i^* \in [m-1]$ to be the minimum index such that $C_t^{i^*} C_t^{i^*+1} \leq 0$;
        $p_t \leftarrow |C_t^{i^*+1}|/(|C_t^{i^*+1}| + |C_t^{i^*}|)$, with the convention $0/0 = 1$;
        Output $q_t = i^*/m - 1/(rm)$ with probability $p_t$ and $q_t = i^*/m$ with
          probability $1 - p_t$;
    Output prediction set $\{y : 1 - \beta_t(y) \leq q_t\}$;
```

Appendix B. Details of the Block Bootstrap for Section 4.2.2

Algorithm 4 below outlines the block bootstrap procedure used to generate the error bars in Figures 4.7 and D.5. Both figures were generated using the choices $M = 100$ and $b = 100$.

```
Algorithm 4: Block bootstrap
  Data: Sequence of stock returns $\{R_t\}_{1\leq t\leq T}$, number of bootstrap samples $M$,
        block size $b$, partition $B_1,\dots,B_k$ of $[0,1]$.
  Define the blocks of data $D_i := \{R_{ib+1},\dots,R_{(i+1)b}\}$, $0 \leq i \leq T/b-1$;
  for $j = 1,2,\dots,M$ do
      Sample $i_1^j,\dots,i_{T/b}^j \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{0,\dots,T/b-1\})$;
      Run the procedure outlined in Section 4.2.1 on dataset $\{D_{i_1^j},\dots,D_{i_{T/b}^j}\}$ to
        obtain sequences $\{\bar{\alpha}_t^j\}_{1\leq t\leq T}$ and $\{\text{err}_t^j\}_{1\leq t\leq T}$;
      For $1 \leq \ell \leq k$ compute
          $$\text{CondCoverage}_\ell^j := \frac{1}{|\{t : \bar{\alpha}_t^j \in B_\ell\}|}\sum_{t:\bar{\alpha}_t^j \in B_\ell} \text{err}_t^j.$$
  For all $1 \leq \ell \leq k$ output the 0.025 and 0.975 empirical quantiles of
    $\{\text{CondCoverage}_\ell^j\}_{1\leq j\leq M}$.
```

Appendix C. Proofs for Section 3

This section contains the proofs of Lemma 2, Lemma 3, Theorem 4, Proposition 5, and Theorem 6. In addition we prove a modified version of Theorem 4 that allows η to vary over time.

C.1 Proof of Lemma 2

Proof We follow the calculations of Gradu et al. (2023). Let $\ell(\beta_t) := (\ell(\beta_t,\alpha_t^1),\dots,\ell(\beta_t,\alpha_t^k))$, $\ell(\beta_t)^2 := (\ell(\beta_t,\alpha_t^1)^2,\dots,\ell(\beta_t,\alpha_t^k)^2)$ and $p_t := (p_t^1,\dots,p_t^k)$. By construction notice that $W_{t+1} := \sum_{i=1}^k w_{t+1}^i = \bar{W}_t$. Thus, using the inequalities $\exp(-x) \leq 1-x+x^2$ and $1-y \leq \exp(-y)$ we find that

$$\frac{W_{t+1}}{W_t} = \sum_{i=1}^k p_t^i \exp(-\eta\ell(\beta_t,\alpha_t^i)) \leq \exp(-\eta p_t^T\ell(\beta_t) + \eta^2 p_t^T \ell(\beta_t)^2),$$

and thus inductively

$$W_{s+1}/W_r \leq \exp\left(-\sum_{t=r}^s \eta p_t^T \ell(\beta_t) + \eta^2 p_t^T\ell(\beta_t)^2\right).$$

On the other hand, for any i, $w_{t+1}^i \geq w_t^i(1-\sigma)\exp(-\eta\ell(\beta_t,\alpha_t^i))$ which gives

$$\frac{W_{s+1}}{W_r} \geq \frac{w_{s+1}^i}{W_r} \geq (1-\sigma)^{|I|}\exp\left(-\sum_{t=r}^s \eta\ell(\beta_t,\alpha_t^i)\right)p_r^i \geq (1-\sigma)^{|I|}\exp\left(-\sum_{t=r}^s \eta\ell(\beta_t,\alpha_t^i)\right)\frac{\sigma}{k}.$$

Combining these two inequalities and taking a logarithm yields

$$\log(\sigma/k) + |I|\log(1-\sigma) - \sum_{t=r}^s \eta\ell(\beta_t,\alpha_t^i) \leq -\sum_{t=r}^s \eta p_t^T\ell(\beta_t) + \eta^2 p_t^T\ell(\beta_t)^2.$$

Finally, since $\sigma \leq 1/2$ we may use the inequality $\log(1-\sigma) \geq -2\sigma$ to get the final result

$$\sum_{t=r}^{s} \mathbb{E}[\ell(\beta_t, \alpha_t)] \leq \sum_{t=r}^{s} \ell(\beta_t, \alpha_t^i) + \eta \sum_{t=r}^{s} \mathbb{E}[\ell(\beta_t, \alpha_t)^2] + \frac{1}{\eta}\left(\log(k/\sigma) + |I|2\sigma\right).$$

■

C.2 Proof of Lemma 3

Proof Lemma 4.1 in Gibbs and Candès (2021) shows that for all values t, $\alpha_t^i \in [-\gamma_i, 1+\gamma_i]$. Since $\beta_t \in [0,1]$ we then also have that $\ell_t(\beta_t, \alpha_t^i) \leq \max\{\alpha, 1-\alpha\}|\beta_t - \alpha_t^i| \leq 1 + \gamma_i$. Plugging this fact into Theorem 10.1 of Hazan (2019) gives the result. ■

C.3 Proof of Theorem 4

Proof Fix any $i \in [k]$ and write

$$\begin{aligned}&\sum_{t=r}^{s} \mathbb{E}[\ell(\beta_t, \alpha_t)] - \sum_{t=r}^{s} \ell(\beta_t, \alpha_t^*)\\ &= \left(\sum_{t=r}^{s} \mathbb{E}[\ell(\beta_t, \alpha_t)] - \sum_{t=r}^{s} \ell(\beta_t, \alpha_t^i)\right) + \left(\sum_{t=r}^{s} \ell(\beta_t, \alpha_t^i) - \sum_{t=r}^{s} \ell(\beta_t, \alpha_t^*)\right).\end{aligned}$$

Applying Lemma 2 to the first term and Lemma 3 to the second term gives

$$\begin{aligned}\sum_{t=r}^{s} \mathbb{E}[\ell(\beta_t, \alpha_t)] - \sum_{t=r}^{s} \ell(\beta_t, \alpha_t^*) \leq & \eta \sum_{t=r}^{s} \mathbb{E}[\ell(\beta_t, \alpha_t)^2] + \frac{1}{\eta}\left(\log(|E|/\sigma) + |I|2\sigma\right)\\ &+ \frac{3}{2\gamma_i}(1+\gamma_i)^2 \left(\sum_{t=r+1}^{s} |\alpha_t^* - \alpha_{t-1}^*| + 1\right) + \frac{1}{2}\gamma_i |I|.\end{aligned}$$

Now, there are two cases. If

$$\sqrt{\frac{\sum_{t=r+1}^{s} |\alpha_t^* - \alpha_{t-1}^*| + 1}{|I|}} \geq \gamma_1 \tag{9}$$

then since $\sqrt{\frac{\sum_{t=r+1}^{s} |\alpha_t^* - \alpha_{t-1}^*| + 1}{|I|}} \leq \sqrt{1 + 1/|I|} \leq \gamma_k$ we may find a value γ_i such that

$$\sqrt{\frac{\sum_{t=r+1}^{s} |\alpha_t^* - \alpha_{t-1}^*| + 1}{|I|}} \leq \gamma_i \leq 2\sqrt{\frac{\sum_{t=r+1}^{s} |\alpha_t^* - \alpha_{t-1}^*| + 1}{|I|}}.$$

Plugging this value into the previous expression gives the desired result. Otherwise if (9) does not hold, then we may simply plug in γ_1 for γ_i. ■

C.4 Results for Variable η

We conclude this section by stating and proving a modified version of Theorem 4 that allows η to vary over time. In particular, we consider a modified version of DtACI in which the update for $\bar{w}_t^i$ is replaced by $\bar{w}_t^i \leftarrow w_t^i \exp(-\eta_t \ell_t(\beta_t, \alpha_t^i))$. Now, recall that our regret bounds consider a target interval length. Call this target length L. Then, Theorem 8 below shows that if the normalized variation in η_t is of order $\sqrt{1/L}$, DtACI will have dynamic regret of size $O(\sqrt{\log(L)/L} + \max\{\gamma_1, \sqrt{\frac{1}{L}\sum_{t=r+1}^s |\alpha_t^* - \alpha_{t-1}^*|}\})$.

The primary case of interest is that in which $\eta_t = \sqrt{\frac{\log(2Lk)+1}{\sum_{s=t-L+1}^t \mathbb{E}[\ell(\beta_s,\alpha_s)^2]}}$. Here, the assumption that η_t has small variability is motivated by a time-series model in which $(\beta_t, (\alpha_t^i, w_t^i)_{i\in[k]}))$ has a stationary distribution. For example, the original work of Gibbs and Candès (2021) considers a setting in which $\{(X_t, Y_t)\}$ follows a hidden Markov model and $S_t^y = S(X_t, y)$ and $\mathcal{D}_t^y = \mathcal{D}$ are fixed quantities that do not depend on t. In this set-up, it follows that $(\beta_t, (\alpha_t^i, w_t^i)_{i\in[k]})$ forms a Markov chain and thus under reasonable mixing assumptions on (X_t, Y_t), one can expect $\frac{1}{L}\sum_{s=t-L+1}^t \mathbb{E}[\ell(\beta_s, \alpha_s)^2]$ to have variations of order $1/\sqrt{L}$.

Theorem 8 *Let $L \in \mathbb{N}$ denote a fixed target interval length and $I = [r, s]$ be any interval of length L with $r > L$. Set $\sigma = 1/(2L)$ and for $t > L$ set $\eta_t := \sqrt{\frac{\log(2Lk)+1}{\sum_{s=t-L+1}^t \mathbb{E}[\ell(\beta_s,\alpha_s)^2]}}$. Let $\{\alpha_t^i, p_t^i\}_{t\in[T],i\in[k]}$ be generated by a modified version of Algorithm 1 in which the update for $\bar{w}_t^i$ is replaced by $\bar{w}_t^i \leftarrow w_t^i \exp(-\eta_t \ell_t(\beta_t, \alpha_t^i))$. Assume that*

$$\frac{1}{L\eta_s}\sum_{t=r}^s |\eta_t - \eta_s|, \frac{1}{L\eta_s}\sum_{t=r}^s |\eta_t^2 - \eta_s^2| \leq O\left(\frac{1}{\sqrt{L}}\right).$$

Then, under the conditions of Theorem 4,

$$\begin{aligned}\frac{1}{L}\sum_{t=r}^s \mathbb{E}[\ell(\beta_t, \alpha_t)] - \frac{1}{L}\sum_{r=1}^s \ell(\beta_t, \alpha_t^*) \leq 2\sqrt{\frac{\log(2Lk)+1}{L}}&\sqrt{\frac{1}{L}\sum_{t=r}^s \mathbb{E}[\ell(\beta_t,\alpha_t)^2]} \\ &+ 4(1+\gamma_{\max})^2 \max\left\{\sqrt{\frac{\sum_{t=r+1}^s |\alpha_t^* - \alpha_{t-1}^*| + 1}{L}}, \gamma_1\right\} \\ &+ O\left(\frac{1}{\sqrt{L}}\right) \\ &= O\left(\sqrt{\frac{\log(L)}{L}}\right) + O\left(\max\{\gamma_1, \sqrt{\frac{1}{L}\sum_{t=r+1}^s |\alpha_t^* - \alpha_{t-1}^*|}\}\right),\end{aligned}$$

where the expectation is over the randomness in DtACI and the data $\beta_1, \ldots, \beta_T$ can be viewed as fixed.

Proof Proceeding identically to the proof of Lemma 2 we have that for any $i \in [k]$,

$$\sum_{t=r}^s \eta_t \mathbb{E}[\ell(\beta_t, \alpha_t)] \leq \sum_{t=r}^s \eta_t \ell(\beta_t, \alpha_t^i) + \sum_{t=r}^s \eta_t^2 \mathbb{E}[\ell(\beta_t, \alpha_t)^2] + \log(k/\sigma) + |I|2\sigma$$

Now, by Lemma 4.1 of Gibbs and Candès (2021) we know that $\alpha_t^i \in [-\gamma_i, 1+\gamma_i]$ and thus that $\ell(\beta_t, \alpha_t^i), \ell(\beta_t, \alpha_t) \leq 1+\gamma_{\max}$. Hence,

$$\begin{aligned}\eta_s \sum_{t=r}^s \mathbb{E}[\ell(\beta_t, \alpha_t)] \leq \eta_s \sum_{t=r}^s \ell(\beta_t, \alpha_t^i) + \eta_s^2 \sum_{t=r}^s \mathbb{E}[\ell(\beta_t, \alpha_t)^2] + \log(k/\sigma) + |I|2\sigma \\ + 2(1+\gamma_{\max})\sum_{t=r}^s |\eta_t - \eta_s| + (1+\gamma_{\max})^2 \sum_{t=r}^s |\eta_t^2 - \eta_s^2|.\end{aligned}$$

The remainder of the proof is identical to that of Theorem 4.

■

C.5 Proof of Proposition 5

Proof For simplicity we will only prove the case where $\tau > \alpha^*$. The case where $\tau \leq \alpha^*$ is identical. By a direct computation we have that

$$\begin{aligned}
&\mathbb{E}[\ell(\beta, \tau)] - \mathbb{E}[\ell(\beta, \alpha^*)] \\
&= \mathbb{E}[\alpha(\beta - \tau)\mathbb{1}_{\beta \geq \tau}] + \mathbb{E}[(1-\alpha)(\tau - \beta)\mathbb{1}_{\beta < \tau}] - \mathbb{E}[\alpha(\beta - \alpha^*)\mathbb{1}_{\beta \geq \alpha^*}] - \mathbb{E}[(1-\alpha)(\alpha^* - \beta)\mathbb{1}_{\beta < \alpha^*}] \\
&= -\mathbb{E}[\alpha(\tau - \alpha^*)\mathbb{1}_{\beta \geq \tau}] + \mathbb{E}[(1-\alpha)(\tau - \alpha^*)\mathbb{1}_{\beta < \alpha^*}] + \mathbb{E}[((1-\alpha)(\tau - \beta) - \alpha(\beta - \alpha^*))\mathbb{1}_{\alpha^* \leq \beta < \tau}] \\
&= -\mathbb{E}[\alpha(\tau - \alpha^*)\mathbb{1}_{\beta \geq \alpha^*}] + \mathbb{E}[(1-\alpha)(\tau - \alpha^*)\mathbb{1}_{\beta < \alpha^*}] + \mathbb{E}[\alpha(\tau - \alpha^*)\mathbb{1}_{\alpha^* \leq \beta < \tau}] \\
&\quad + \mathbb{E}[(1-\alpha)(\tau - \alpha^*)\mathbb{1}_{\alpha^* \leq \beta < \tau}] - \mathbb{E}[(\beta - \alpha^*)\mathbb{1}_{\alpha^* \leq \beta < \tau}] \\
&= -\alpha(1-\alpha)(\tau - \alpha^*) + \alpha(1-\alpha)(\tau - \alpha^*) + \mathbb{E}[(\tau - \alpha^*)\mathbb{1}_{\alpha^* \leq \beta < \tau}] - \mathbb{E}[(\beta - \alpha^*)\mathbb{1}_{\alpha^* \leq \beta < \tau}] \\
&= \mathbb{E}[(\tau - \beta)\mathbb{1}_{\alpha^* \leq \beta < \tau}].
\end{aligned}$$

This proves the first part of Proposition 5. For the second part, note that if β has a density on $[0,1]$ that is lower bounded by p, then

$$\mathbb{E}[(\tau - \beta)\mathbb{1}_{\alpha^* \leq \beta < \tau}],\ \mathbb{E}[(\beta - \tau)\mathbb{1}_{\tau \leq \beta < \alpha^*}] \geq \int_0^{|\tau - \alpha^*|} xpdx = p\frac{(\tau - \alpha^*)^2}{2}.$$

■

C.6 Proof of Theorem 6

Proof Let

$$\tilde{\alpha}_t := \sum_i \frac{p_t^i \alpha_t^i}{\gamma_i}$$

and observe that

$$\begin{aligned}\tilde{\alpha}_t = \sum_i \frac{p_t^i(\alpha_{t+1}^i - \gamma_i(\alpha - \text{err}_t^i))}{\gamma_i} = \sum_i \frac{p_t^i \alpha_{t+1}^i}{\gamma_i} + \sum_i p_t^i(\text{err}_t^i - \alpha) \\ = \tilde{\alpha}_{t+1} + \sum_i \frac{(p_t^i - p_{t+1}^i)\alpha_{t+1}^i}{\gamma_i} + \sum_i p_t^i(\text{err}_t^i - \alpha).\end{aligned}$$

Thus,

$$\mathbb{E}[\mathrm{err}_t] - \alpha = \tilde{\alpha}_t - \tilde{\alpha}_{t+1} + \sum_i \frac{(p^i_{t+1} - p^i_t)\alpha^i_{t+1}}{\gamma_i}. \tag{10}$$

Now, for ease of notation let $W_t := \sum_i w^i_t$ and $\tilde{p}^i_{t+1} := \frac{p^i_t \exp(-\eta_t \ell(\beta_t, \alpha^i_t))}{\sum_{i'} p^{i'}_t \exp(-\eta_t \ell(\beta_t, \alpha^{i'}_t))}$. Recall that by definition,

$$p^i_{t+1} = \frac{w^i_{t+1}}{\sum_{i'} w^{i'}_{t+1}} = (1-\sigma_t)\tilde{p}^i_{t+1} + \frac{\sigma_t}{k}.$$

Moreover, by a direct computation

$$\begin{aligned}
\tilde{p}^i_{t+1} - p^i_t &= \frac{p^i_t \exp(-\eta_t \ell(\beta_t, \alpha^i_t))}{\sum_{i'} p^{i'}_t \exp(-\eta_t \ell(\beta_t, \alpha^{i'}_t))} - p^i_t \\
&= p^i_t \frac{\exp(-\eta_t \ell(\beta_t, \alpha^i_t)) - \sum_{i'} p^{i'}_t \exp(-\eta_t \ell(\beta_t, \alpha^{i'}_t))}{\sum_{i'} p^{i'}_t \exp(-\eta_t \ell(\beta_t, \alpha^{i'}_t))} \\
&= p^i_t \frac{\sum_{i'} p^{i'}_t (\exp(-\eta_t \ell(\beta_t, \alpha^i_t)) - \exp(-\eta_t \ell(\beta_t, \alpha^{i'}_t)))}{\sum_{i'} p^{i'}_t \exp(-\eta_t \ell(\beta_t, \alpha^{i'}_t))} \\
&= p^i_t \frac{\sum_{i'} p^{i'}_t \exp(-\eta_t \ell(\beta_t, \alpha^{i'}_t))(\exp(\eta_t \ell(\beta_t, \alpha^{i'}_t) - \eta_t \ell(\beta_t, \alpha^i_t)) - 1)}{\sum_{i'} p^{i'}_t \exp(-\eta_t \ell(\beta_t, \alpha^{i'}_t))} \\
&= p^i_t \sum_{i'} \tilde{p}^{i'}_{t+1} (\exp(\eta_t \ell(\beta_t, \alpha^{i'}_t) - \eta_t \ell(\beta_t, \alpha^i_t)) - 1).
\end{aligned}$$

By Lemma 4.1 of Gibbs and Candès (2021) we know that $\alpha^i_t \in [-\gamma_i, 1+\gamma_i]$ and thus that $|\ell(\beta_t, \alpha^{i'}_t) - \ell(\beta_t, \alpha^i_t)| \leq \max\{\alpha, 1-\alpha\}|\alpha^{i'}_t - \alpha^i_t| \leq 1 + 2\gamma_{\max}$. Hence, by the mean value theorem,

$$|\exp(\eta_t \ell(\beta_t, \alpha^{i'}_t) - \eta_t \ell(\beta_t, \alpha^i_t)) - 1| \leq \eta_t(1+2\gamma_{\max})\exp(\eta_t(1+2\gamma_{\max})),$$

and thus also,

$$|\tilde{p}^i_{t+1} - p^i_t| \leq p^i_t \eta_t (1+2\gamma_{\max})\exp(\eta_t(1+2\gamma_{\max})).$$

Applying Lemma 4.1 of Gibbs and Candès (2021) again we conclude that

$$\begin{aligned}
\left|\sum_i \frac{(p^i_{t+1} - p^i_t)\alpha^i_{t+1}}{\gamma_i}\right| &\leq (1-\sigma_t)\sum_i \left|\frac{(\tilde{p}^i_{t+1} - p^i_t)\alpha^i_{t+1}}{\gamma_i}\right| + \sigma_t \sum_i \left|\frac{(1/k - p^i_t)\alpha^i_{t+1}}{\gamma_i}\right| \\
&\leq \frac{\eta_t(1+2\gamma_{\max})^2}{\gamma_{\min}}\exp(\eta_t(1+2\gamma_{\max})) + 2\sigma_t \frac{1+\gamma_{\max}}{\gamma_{\min}}.
\end{aligned}$$

So, taking a sum over t in equation 10 we arrive at the inequality,

$$\left|\frac{1}{T}\sum_{t=1}^T \mathbb{E}[\mathrm{err}_t] - \alpha\right| \leq \frac{|\tilde{\alpha}_1 - \tilde{\alpha}_{T+1}|}{T} + \frac{(1+2\gamma_{\max})^2}{\gamma_{\min}} \frac{1}{T}\sum_{t=1}^T \eta_t e^{\eta_t(1+2\gamma_{\max})} + 2\frac{1+\gamma_{\max}}{\gamma_{\min}} \frac{1}{T}\sum_{t=1}^T \sigma_t.$$

Applying Lemma 4.1 of Gibbs and Candès (2021) one final time gives $\gamma_{\min}\tilde{\alpha}_t \in [-\gamma_{\max}, 1 + \gamma_{\max}]$ and thus $|\tilde{\alpha}_1 - \tilde{\alpha}_{T+1}| \leq (1+2\gamma_{\max})/\gamma_{\min}$. Plugging this into the previous expression gives the desired upperbound.

Now, recall that if $\{a_t\}_{t=1}^{\infty}$ is a sequence such that $a_t \to 0$, then $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} a_t = 0$ (see Lemma 9 below). So, applying this to the above expression we find that if $\eta_t, \sigma_t \to 0$, then

$$\lim_{T\to\infty} \left| \frac{1}{T}\sum_{t} \mathbb{E}[\text{err}_t] - \alpha \right| = 0.$$

A standard application of the law of large numbers then gives that,

$$\lim_{T\to\infty} \left| \frac{1}{T}\sum_{t} \text{err}_t - \alpha \right| = \lim_{T\to\infty} \left| \frac{1}{T}\sum_{t} \mathbb{E}[\text{err}_t] - \alpha \right| = 0,$$

as desired.

■

We conclude this section by stating a standard analysis fact that was used in the proof of Theorem 6.

Lemma 9 *Let $\{a_t\}_{t=1}^{\infty}$ be a sequence such that $a_t \to 0$. Then, $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} a_t = 0$.*

Proof For any fixed $S \in \mathbb{N}$ we have the bound,

$$\left| \frac{1}{T}\sum_{t=1}^{T} a_t \right| \leq \frac{S \max_{1\leq t\leq S} |a_s|}{T} + \max_{t>S} |a_t|.$$

So, sending $T \to \infty$ gives,

$$\limsup_{T\to\infty} \left| \frac{1}{T}\sum_{t=1}^{T} a_t \right| \leq \max_{t>S} a_t,$$

and sending $S \to \infty$ gives the desired result. ■

Appendix D. Additional Figures

We begin by giving three figures showing the results for Section 4.2 in the case where we use a variable value for η, given by $\eta = \eta_t = \sqrt{\frac{\log(2k\cdot 500)+1}{\sum_{s=t-501}^{t} \mathbb{E}[\ell(\beta_s,\alpha_s)^2]}}$. We see that the results are nearly exactly identical to those obtained with the fixed heurisitic choice of η.

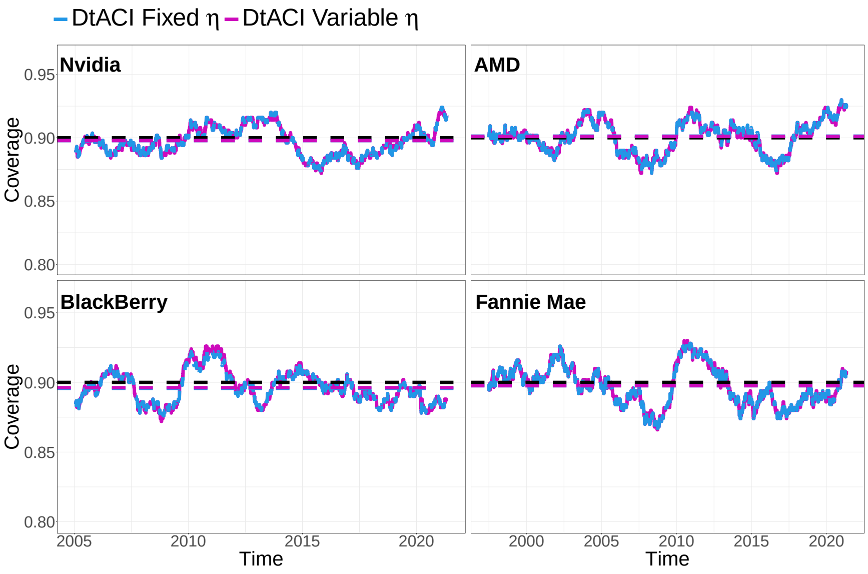

Figure D.1: Estimation of stock market volatility using conformity score $S_t(v) = |v - (\hat{\sigma}_t^t)^2|/(\hat{\sigma}_t^t)^2$ and either η fixed at 2.72 (blue) or a variable value (purple) of $\eta_t = \sqrt{\frac{\log(2k\cdot 500)+1}{\sum_{s=t-501}^{t}\mathbb{E}[\ell(\beta_s,\alpha_s)^2]}}$. Solid lines show the local coverage level for $\bar{\alpha}_t$, while dashed lines indicate the global coverage frequency over all time steps. The black dashed lines indicates the target level of $1-\alpha = 0.9$.

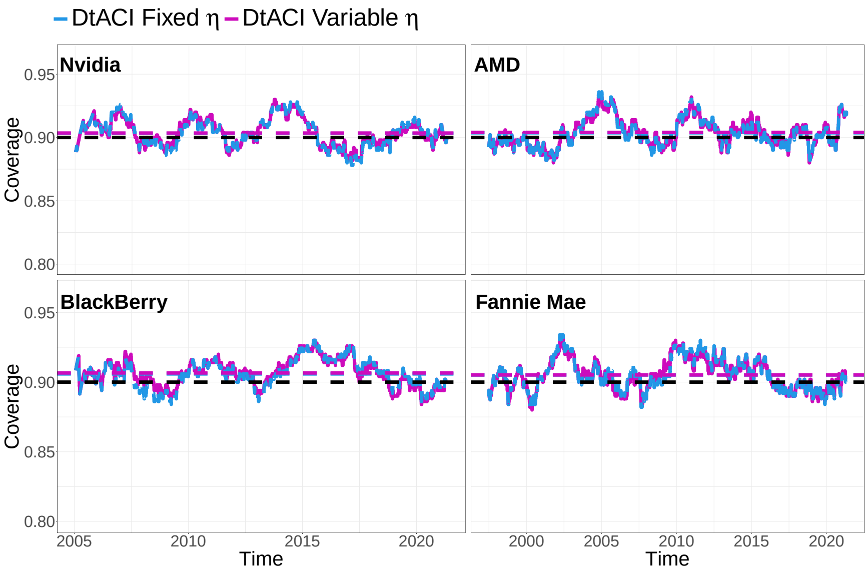

Figure D.2: Estimation of stock market volatility using conformity score $S_t(v) = |v - (\hat{\sigma}_t^t)^2|$ and either η fixed at 2.72 (blue) or a variable value (purple) of $\eta_t = \sqrt{\frac{\log(2k\cdot 500)+1}{\sum_{s=t-501}^{t}\mathbb{E}[\ell(\beta_s,\alpha_s)^2]}}$. Solid lines show the local coverage level for $\bar{\alpha}_t$, while dashed lines indicate the global coverage frequency over all time steps. The black dashed lines indicates the target level of $1-\alpha = 0.9$.

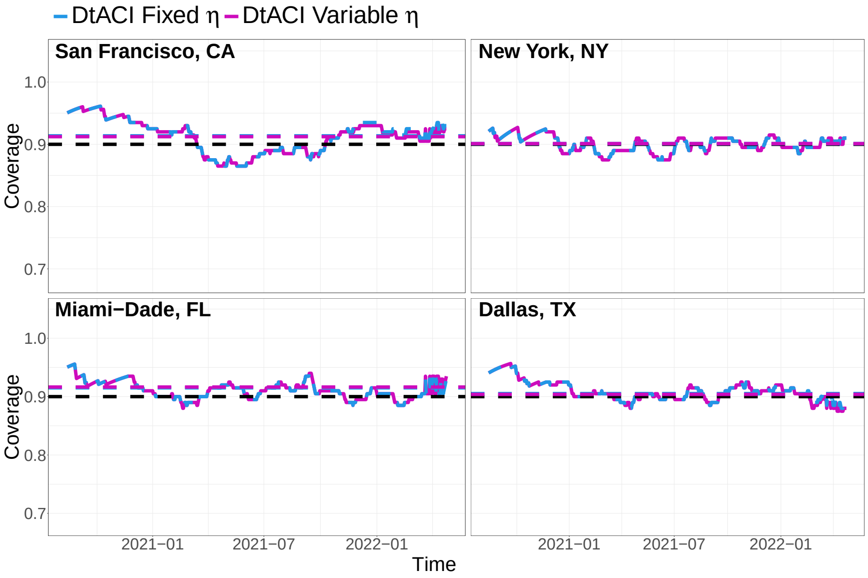

Figure D.3: Coverage results for the prediction of county-level COVID-19 case counts using either η fixed at 2.72 (blue) or a variable value (purple) of $\eta_t = \sqrt{\frac{\log(2k\cdot 500)+1}{\sum_{s=t-500}^{t-1}\mathbb{E}[\ell(\beta_s,\alpha_s)^2]}}$. Solid lines show the local coverage level for $\bar{\alpha}_t$, while dashed lines indicate the global coverage frequency over all time steps. The black dashed lines indicates the target level of $1-\alpha = 0.9$.

Our next figure shows Q-Q plots comparing the quantiles of the local coverage gaps realized by DtACI against those for an i.i.d. Bernoulli(0.9) sequence for the volatility dataset of Section 4.2.1 and the unnormalized conformity score $S_t(v) = |v - (\hat{\sigma}_t^t)^2|$.

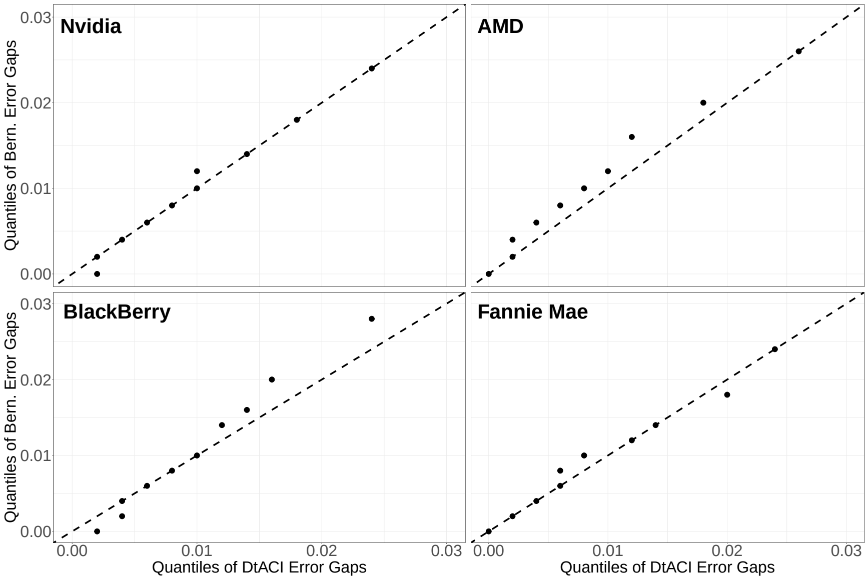

Figure D.4: Q-Q plots comparing the distribution of local coverage gaps obtained by an i.i.d. Bernoulli(α) sequence against those realized by DtACI. Results for DtACI were generated with conformity score $S_t(v) = |v - (\hat{\sigma}_t^t)^2|$. The dashed line indicates the ideal situation of exact equality.

The next figure shows the empirical conditional coverage (8) for the estimation of stock market volatility with the unnormalized conformity score. As for the normalized conformity score, we observe conditional coverages close to the target level across all values of $\bar{\alpha}_t$.

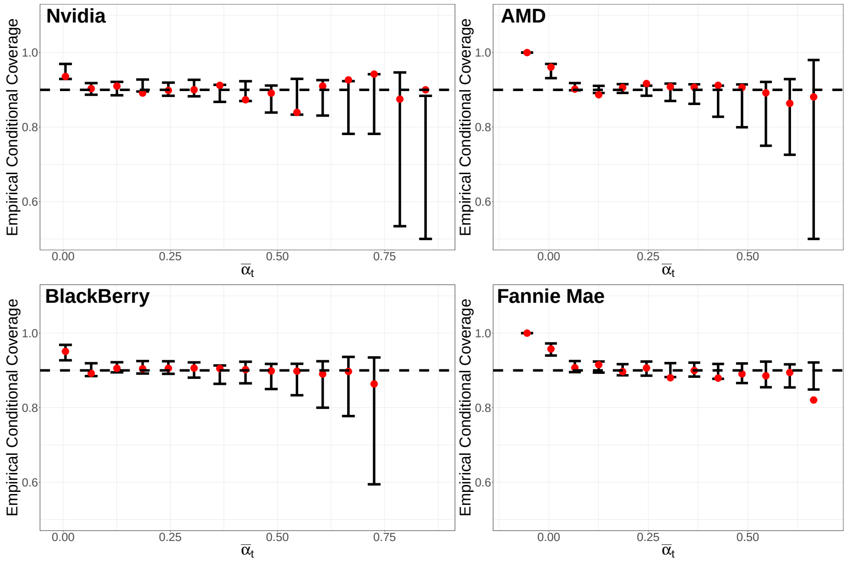

Figure D.5: Empirical conditional coverage of $\bar{\alpha}_t$ for the estimation of stock market volatility with conformity score $S_t(v) = |v - (\hat{\sigma}_t^t)^2|$. Red points show the empirical conditional coverage given $\bar{\alpha}_t \in B_i$ with error bars indicating the corresponding 0.025 and 0.975 quantiles across 100 block bootstrap resamples of the data $\{(X_t, Y_t)\}$ with block-size set to 100. For visual clarity all error bars are truncated to the range $[0.5, 1]$. Black dashed lines shows the target level of $1 - \alpha = 0.9$.

Finally, Figures D.6 and D.7 show the stock prices and the COVID-19 case counts for the datasets considered in Section 4.2.

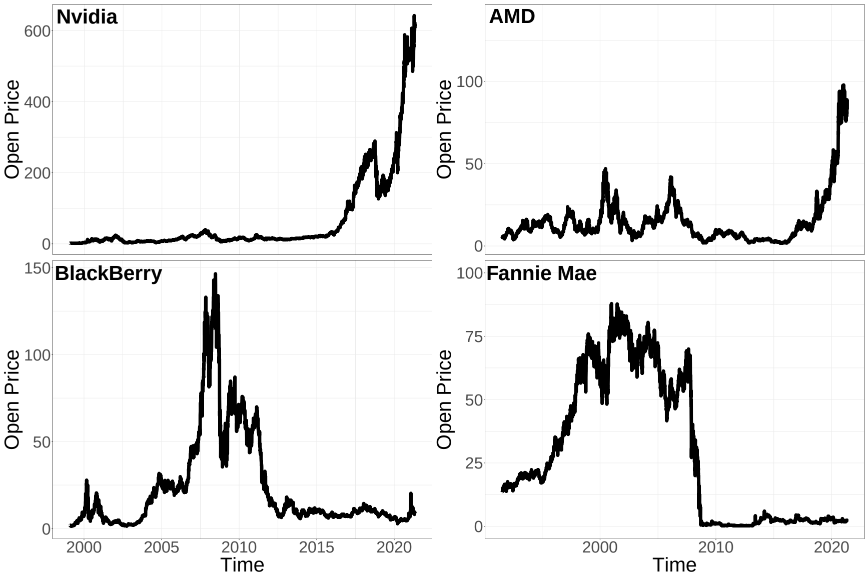

Figure D.6: Daily open prices for the four stocks considered in Section 4.2.1.

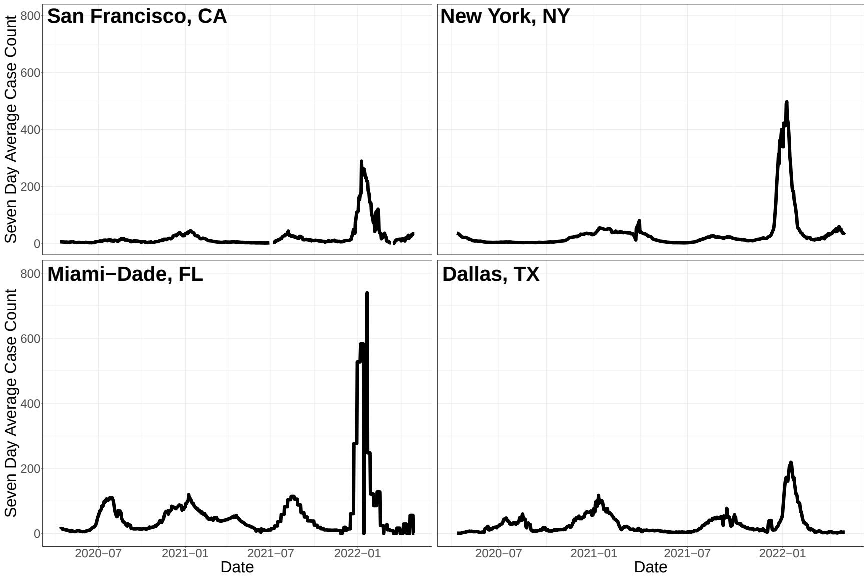

Figure D.7: Moving seven day averages of the number of new confirmed COVID-19 cases per 100,000 people in the four counties considered in Section 4.2.3.

References

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. 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.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816 – 845, 2023. doi: 10.1214/23-AOS2276. URL https://doi.org/10.1214/23-AOS2276.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=QNjyrDBx6tz.

Victor Chernozhukov, Kaspar Wüthrich, and Zhu Yinchu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 732–749. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/chernozhukov18a.html.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 08 2020. ISSN 2049-8772. doi: 10.1093/imaiai/iaaa017. URL https://doi.org/10.1093/imaiai/iaaa017. iaaa017.

Alexander Gammerman and Vladimir Vovk. Hedging predictions in machine learning. *The Computer Journal*, 50(2):151–163, 2007. doi: 10.1093/comjnl/bxl065.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/0d441de75945e5acbc865406fc9a2559-Paper.pdf.

Paula Gradu, Elad Hazan, and Edgar Minasyan. Adaptive regret for control of time-varying dynamics. In *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, volume 211, pages 560–572. PMLR, 15–16 Jun 2023. URL https://proceedings.mlr.press/v211/gradu23a.html.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint*, 2019. arXiv:1909.05207.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 01 2014. doi: 10.1111/rssb.12021.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Conformal prediction with temporal quantile adjustments. In *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=PM5gVmG2Jj`.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in Artificial Intelligence,*, pages 315–330, 2008.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, pages 345–356, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-36755-0.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. *arXiv preprint*, 2021. arXiv:2103.03323.

Alex Reinhart, Logan Brooks, Maria Jahja, Aaron Rumack, Jingjing Tang, Sumit Agrawal, Wael Al Saeed, Taylor Arnold, Amartya Basu, Jacob Bien, Ángel A. Cabrera, Andrew Chin, Eu Jing Chua, Brian Clark, Sarah Colquhoun, Nat DeFries, David C. Farrow, Jodi Forlizzi, Jed Grabman, Samuel Gratzl, Alden Green, George Haff, Robin Han, Kate Harwood, Addison J. Hu, Raphael Hyde, Sangwon Hyun, Ananya Joshi, Jimi Kim, Andrew Kuznetsov, Wichada La Motte-Kerr, Yeon Jin Lee, Kenneth Lee, Zachary C. Lipton, Michael X. Liu, Lester Mackey, Kathryn Mazaitis, Daniel J. McDonald, Phillip McGuinness, Balasubramanian Narasimhan, Michael P. O'Brien, Natalia L. Oliveira, Pratik Patil, Adam Perer, Collin A. Politsch, Samyak Rajanala, Dawn Rucker, Chris Scott, Nigam H. Shah, Vishnu Shankar, James Sharpnack, Dmitry Shemetov, Noah Simon, Benjamin Y. Smith, Vishakha Srivastava, Shuyi Tan, Robert Tibshirani, Elena Tuzhilina, Ana Karina Van Nortwick, Valérie Ventura, Larry Wasserman, Benjamin Weaver, Jeremy C. Weiss, Spencer Whitman, Kristin Williams, Roni Rosenfeld, and Ryan J. Tibshirani. An open repository of real-time covid-19 indicators. *Proceedings of the National Academy of Sciences*, 118(51):e2111452118, 2021. doi: 10.1073/pnas.2111452118. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2111452118`.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf`.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234, 2019. doi: 10.1080/01621459.2017.1395341. URL `https://doi.org/10.1080/01621459.2017.1395341`.

C. Saunders, A. Gammerman, and V. Vovk. Transduction with confidence and credibility. In *Sixteenth International Joint Conference on Artificial Intelligence (IJCAI '99) (01/01/99)*, pages 722–726, 1999. URL `https://eprints.soton.ac.uk/258961/`.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. URL `http://jmlr.org/papers/v9/shafer08a.html`.

R. J. Tibshirani. Can symptoms surveys improve covid-19 forecasts? `https://delphi.cmu.edu/blog/2020/09/21/can-symptoms-surveys-improve-covid-19-forecasts/`, 2020. Accessed: 2022-06-17.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf`.

V. Vovk, A. Gammerman, and C. Saunders. Machine-learning applications of algorithmic randomness. In *Sixteenth International Conference on Machine Learning (ICML-1999) (01/01/99)*, pages 444–453, 1999. URL `https://eprints.soton.ac.uk/258960/`.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

Volodimir G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, COLT '90, page 371–386, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. ISBN 1558601465.

Olivier Wintenberger. Optimal learning with bernstein online aggregation. *Mach. Learn.*, 106(1):119–141, jan 2017. ISSN 0885-6125. doi: 10.1007/s10994-016-5592-6. URL `https://doi.org/10.1007/s10994-016-5592-6`.

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae009, 03 2024. ISSN 1369-7412. doi: 10.1093/jrsssb/qkae009. URL `https://doi.org/10.1093/jrsssb/qkae009`.

Margaux Zaffran, Olivier Feron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 25834–25866. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/zaffran22a.html`.