Adaptive Conformal Inference Under Distribution Shift

Anonymous Author(s)
Affiliation
Address
email

Abstract

We develop methods for forming prediction sets in an online setting where the data generating distribution is allowed to vary over time in an unknown fashion. Our framework builds on ideas from conformal inference to provide a general wrapper that can be combined with any black box method that produces point predictions of the unseen label or estimated quantiles of its distribution. While previous conformal inference methods rely on the assumption that the data points are exchangeable, our adaptive approach provably achieves the desired long-term coverage frequency irrespective of the true data generating process. We accomplish this by modelling the distribution shift as a learning problem in a single parameter whose optimal value is varying over time and must be continuously re-estimated. We test our method, *adaptive conformal inference*, on two real world datasets and find that its predictions are robust to visible and significant distribution shifts.

1 Introduction

Machine learning algorithms are increasingly being employed in high stakes decision making processes. For instance, deep neural networks are currently being used in self-driving cars to detect nearby objects [2] and parole decisions are being made with the assistance of complex models that combine over a hundred features [1]. As the popularity of black box methods and the cost of making wrong decisions grow it is crucial that we develop tools to quantify the uncertainty of their predictions.

In this paper we develop methods for constructing prediction sets that are guaranteed to contain the target label with high probability. We focus specifically on an online learning setting in which we observe covariate-response pairs $\{(X_t, Y_t)\}_{t\in\mathbb{N}} \subseteq \mathbb{R}^d \times \mathbb{R}$ in a sequential fashion. At each time step $t \in \mathbb{N}$ we are tasked with using the previously observed data $\{(X_r, Y_r)\}_{1\leq r\leq t-1}$ along with the new covariates, X_t, to form a prediction set $\hat{C}_t$ for Y_t. Then, given a target coverage level $\alpha \in (0, 1)$ our generic goal is to guarantee that Y_t belongs to $\hat{C}_t$ at least $100(1-\alpha)\%$ of the time.

Perhaps the most powerful and flexible tools for solving this problem come from conformal inference [see e.g. 30, 15, 28, 19, 27, 14, 3] . This framework provides a generic methodology for transforming the outputs of any black box prediction algorithm into a prediction set. The generality of this approach has facilitated the development of a large suite of conformal methods, each specialized to a specific prediction problem of interest [e.g. 26, 11, 20, 8, 21, 18]. Common to all of these methods is the guarantee that if the data are exchangeable, then the prediction set has valid marginal coverage $\mathbb{P}(Y_t \in \hat{C}_t) = 1 - \alpha$.

While exchangeability is a common assumption, there are many real-world applications in which we do not expect the marginal distribution of (X_t, Y_t) to be stationary. For example, in finance and economics market behaviour can shift drastically in response to new legislation or major world events. Alternatively, the distribution of (X_t, Y_t) may change as we deploy our prediction method in

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

new environments. This paper develops *adaptive conformal inference* (ACI), a method for forming prediction sets that are robust to changes in the marginal distribution of the data. Our approach is both simple, in that it requires only the tracking of a single parameter that models the shift, and general as it can be combined with any modern machine learning algorithm that produces point predictions or estimated quantiles for the response. We show that over long time intervals ACI achieves the desired target coverage frequency without any assumptions on the data-generating distribution. Moreover, when the distribution shift is small and the prediction algorithm takes a certain simple form we show that ACI will additionally obtain approximate marginal coverage at most time steps.

1.1 Conformal inference

Suppose we are given a fitted regression model for predicting the value of Y from X. Let $y \in \mathbb{R}$ be a candidate value for Y_t. To determine if y is a reasonable estimate of Y_t, we define a conformity score $S : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ that measures how well the value y *conforms* with the predictions of our fitted model. For example, if our regression model produces points predictions $\hat{\mu}(X)$ then we could use a conformity score that measures the distance between $\hat{\mu}(X_t)$ and y. One such example is

$$S(X_t, y) = |\hat{\mu}(X_t) - y|.$$

Alternatively, suppose our regression model produces estimates $\hat{q}(X; p)$ of the pth quantile of the distribution of $Y|X$. Then we could use the method of conformal quantile regression (CQR) [25], which examines the signed distance between y and fitted upper and lower quantiles through the score

$$S(X_t, y) = \max\{\hat{q}(X_t; \alpha/2) - y, y - \hat{q}(X_t; 1 - \alpha/2)\}.$$

Regardless of what conformity score is chosen the key issue is to determine how small $S(X_t, y)$ should be in order to accept y as a reasonable prediction for Y_t. Assume we have a calibration set $\mathcal{D}_{\text{cal}} \subseteq \{(X_r, Y_r)\}_{1 \le r \le t-1}$ that is different from the data that was used to fit the regression model. Using this calibration set we define the fitted quantiles of the conformity scores to be

$$\hat{Q}(p) := \inf \left\{ s : \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{(X_r, Y_r) \in \mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S(X_r, Y_r) \le s\}} \ge p \right\}, \tag{1}$$

and say that y is a reasonable prediction for Y_t if $S(X_t, y) \le \hat{Q}(1 - \alpha)$.

The crucial observation is that if the data $\mathcal{D}_{\text{cal}} \cup \{(X_t, Y_t)\}$ are exchangeable and we break ties uniformly at random then the rank of $S(X_t, Y_t)$ amongst the points $\{S(X_r, Y_r)\}_{(X_r, Y_r) \in \mathcal{D}_{\text{cal}}} \cup \{S(X_t, Y_t)\}$ will be uniform. Therefore,

$$\mathbb{P}(S(X_t, Y_t) \le \hat{Q}(1 - \alpha)) = \frac{\lceil |\mathcal{D}_{\text{cal}}|(1 - \alpha) \rceil}{|\mathcal{D}_{\text{cal}}| + 1}.$$

Thus, defining our prediction set to be $\hat{C}_t := \{y : S(X_t, y) \le \hat{Q}(1 - \alpha)\}$ gives the marginal coverage guarantee

$$\mathbb{P}(Y_t \in \hat{C}_t) = \mathbb{P}(S(X_t, Y_t) \le \hat{Q}(1 - \alpha)) = \frac{\lceil |\mathcal{D}_{\text{cal}}|(1 - \alpha) \rceil}{|\mathcal{D}_{\text{cal}}| + 1}.$$

By introducing additional randomization this generic procedure can be altered slightly to produce a set $\hat{C}_t$ that satisfies the exact marginal coverage guarantee $\mathbb{P}(Y_t \in \hat{C}_t) = 1 - \alpha$ [30]. For the purposes of this paper this adjustment is not critical and so we omit the details here. Additionally, we remark that the method outlined above is often referred to as *split* or *inductive* conformal inference [24, 30, 23]. This refers to the fact that we have split the observed data between a training set used to fit the regression model and a withheld calibration set. The adaptive conformal inference method developed in this paper can also be easily adjusted to work with full conformal inference in which data splitting is avoided at the cost of greater computational resources [30].

2 Adapting conformal inference to distribution shifts

Up until this point we have been working with a single score function $S(\cdot)$ and quantile function $\hat{Q}(\cdot)$. In the general case where the distribution of the data is shifting over time both these functions should

be regularly re-estimated to align with the most recent observations. Therefore, we assume that at each time t we are given a fitted score function $S_t(\cdot)$ and corresponding quantile function $\hat{Q}_t(\cdot)$. We define the realized miscoverage rate of the prediction set $\hat{C}_t(\alpha) := \{y : S_t(X_t, y) \leq \hat{Q}_t(1-\alpha)\}$ as

$$M_t(\alpha) := \mathbb{P}(S_t(X_t, Y_t) > \hat{Q}_t(1-\alpha)),$$

where the probability is over the test point (X_t, Y_t) as well as the data used to fit $S_t(\cdot)$ and $\hat{Q}_t(\cdot)$.

Now, since the distribution generating the data is non-stationary we do not expect $M_t(\alpha)$ to be equal or even close to α. Even so, we may still postulate that if the conformity scores used to fit $\hat{Q}_t(\cdot)$ cover the bulk of the distribution of $S_t(X_t, Y_t)$ then there may be an alternative value $\alpha_t^* \in [0, 1]$ such that $M_t(\alpha_t^*) \cong \alpha$. More rigorously, assume that with probability one, $\hat{Q}_t(\cdot)$ is continuous, non-decreasing and such that $\hat{Q}_t(0) = -\infty$ and $\hat{Q}_t(1) = \infty$. This does not hold for the split conformal quantile functions defined in (1), but in the case where there are no ties amongst the conformity scores we can adjust our definition to guarantee this by smoothing over the jump discontinuities in $\hat{Q}(\cdot)$. Then, $M_t(\cdot)$ will be non-decreasing on $[0, 1]$ with $M_t(0) = 0$ and $M_t(1) = 1$ and so we may define

$$\alpha_t^* := \sup\{\beta \in [0, 1] : M_t(\beta) \leq \alpha\}.$$

Moreover, if we additionally assume that

$$\mathbb{P}(S_t(X_t, Y_t) = \hat{Q}_t(1-\alpha_t^*)) = 0,$$

then we will have that $M_t(\alpha_t^*) = \alpha$. So, in particular we find that by correctly calibrating the argument to $\hat{Q}_t(\cdot)$ we can achieve either approximate or exact marginal coverage.

To perform this calibration we will use a simple online update. This update proceeds by examining the empirical miscoverage frequency of the previous prediction sets and then decreasing (resp. increasing) our estimate of α_t^* if the prediction sets were historically under-covering (resp. over-covering) Y_t. In particular, let α_1 denote our initial estimate (in our experiments we will choose $\alpha_1 = \alpha$). Recursively define the sequence of miscoverage events

$$\text{err}_t := \begin{cases} 1, \text{ if } Y_t \notin \hat{C}_t(\alpha_t), \\ 0, \text{ otherwise,} \end{cases} \qquad \text{where } \hat{C}_t(\alpha_t) := \{y : S_t(X_t, y) \leq \hat{Q}_t(\alpha_t)\}.$$

Then, fixing a step size parameter $\gamma > 0$ we consider the simple online update

$$\alpha_{t+1} := \alpha_t + \gamma(\alpha - \text{err}_t). \tag{2}$$

We refer to this algorithm as *adaptive conformal inference*. Here, err_t plays the role of our estimate of the historical miscoverage frequency. A natural alternative to this is to consider the update

$$\alpha_{t+1} = \alpha_t + \gamma\left(\alpha - \sum_{s=1}^{t} w_s \text{err}_s\right), \tag{3}$$

where $\{w_s\}_{1 \leq s \leq t} \subseteq [0, 1]$ is a sequence of increasing weights with $\sum_{s=1}^{t} w_s = 1$. This update has the appeal of more directly evaluating the recent empirical miscoverage frequency when deciding whether or not to lower or raise α_t. In practice, we find that (2) and (3) produce almost identical results. For example, in Section A.2 in the Appendix we show some sample trajectories for α_t obtained using the update (3) with

$$w_s := \frac{0.95^{t-s}}{\sum_{s'=1}^{t} 0.95^{t-s'}}.$$

We find that these trajectories are very similar to those produced by (2). The main difference is that the trajectories obtained with (3) are smoother with less local variation in α_t. In the remainder of this article we will focus on (2) for simplicity.

2.1 Choosing the step size

The choice of γ gives a tradeoff between adaptability and stability. While raising the value of γ will make the method more adaptive to observed distribution shifts, it will also induce greater volatility in

the value of α_t. In practice, large fluctuations in α_t may be undesirable as it allows the method to oscillate between outputting small conservative and large anti-conservative prediction sets.

In Theorem 4.2 we give an upper bound on $(M_t(\alpha_t)-\alpha)^2$ that is optimized by choosing γ proportional to $\sqrt{|\alpha^*_{t+1} - \alpha^*_t|}$. While not directly applicable in practice, this result supports the intuition that in environments with greater distributional shift the algorithm needs to be more adapatable and thus γ should be chosen to be larger. In our experiments we will work with the value $\gamma = 0.005$. This value was chosen because it was found to give relatively stable trajectories for α_t while still being sufficiently large as to allow α_t to adapt to observed shifts. In agreement with the general principles outlined above we found that larger values of γ also successfully protect against distribution shifts, while taking γ to be too small causes adaptive conformal inference to perform similar to non-adaptive methods that hold $\alpha_t = \alpha$ constant across time.

2.2 Real data example: predicting market volatility

We apply ACI to the prediction of market volatility. Let $\{P_t\}_{1\le t\le T}$ denote a sequence of daily open prices for a stock. For all $t \ge 2$, define the return $R_t := (P_t - P_{t-1})/P_{t-1}$ and realized volatility $V_t = R_t^2$. Our goal is to use the previously observed returns $X_t := \{R_s\}_{1\le s\le t-1}$ to form prediction sets for $Y_t := V_t$. More sophisticated financial models might augment X_t with additional market covariates (available to the analyst at time $t-1$). As the primary purpose of this section is to illustrate adaptive conformal inference we work with only a simple model.

We start off by forming point predictions using a GARCH(1,1) model [4]. This model assumes that $R_t = \sigma_t \epsilon_t$ with $\epsilon_1, \dots, \epsilon_T$ taken to be i.i.d. $\mathcal{N}(0,1)$ and σ_t satisfying the recursive update

$$\sigma_t^2 = \omega + \alpha V_{t-1} + \beta \sigma_{t-1}^2.$$

This is a common model used for forecasting volatility in economics. In practice, shifting market dynamics can cause the predictions of this model to become inaccurate over large time periods. Thus, when forming point predictions we fit the model using only the last 1250 trading days (i.e. approximately 5 years) of market data. More precisely, for all times $t > 1250$ we fit the coefficients $\hat{\omega}_t,\ \hat{\alpha}_t, \hat{\beta}_t$ as well as the sequence of variances $\{\hat{\sigma}_s^t\}_{1\le s\le t-1}$ using only the data $\{R_r\}_{t-1250\le r<t}$. Then, our point prediction for the realized volatility at time t is

$$(\hat{\sigma}_t^t)^2 := \hat{\omega}_t + \hat{\alpha}_t V_{t-1} + \hat{\beta}_t (\hat{\sigma}_{t-1}^t)^2.$$

To form prediction intervals we define the sequence of conformity scores

$$S_t := \frac{|V_t - (\hat{\sigma}_t^t)^2|}{(\hat{\sigma}_t^t)^2}$$

and the corresponding quantile function

$$\hat{Q}_t(p) := \inf\left\{ x : \frac{1}{1250} \sum_{r=t-1250}^{t-1} \mathbb{1}_{S_r \le x} \ge p \right\}.$$

Then, our prediction set at time t is

$$\hat{C}_t(\alpha_t) := \left\{ v : \frac{|v - (\hat{\sigma}_t^t)^2|}{(\hat{\sigma}_t^t)^2} \le \hat{Q}_t(\alpha_t) \right\},$$

where $\{\alpha_t\}$ is initialized with $\alpha_{1250} = \alpha = 0.1$ and then updated recursively as in (2).

We compare this algorithm to a non-adaptive alternative that takes $\alpha_t = \alpha$ fixed. To measure the performance of these methods across time we examine their local coverage frequencies defined as the average coverage rate over the most recent two years, i.e.

$$\text{localCov}_t := 1 - \frac{1}{500} \sum_{r=t-250+1}^{t+250} \text{err}_r. \tag{4}$$

If the methods perform well then we expect the local coverage frequency to stay near the target value $1-\alpha$ across all time points.

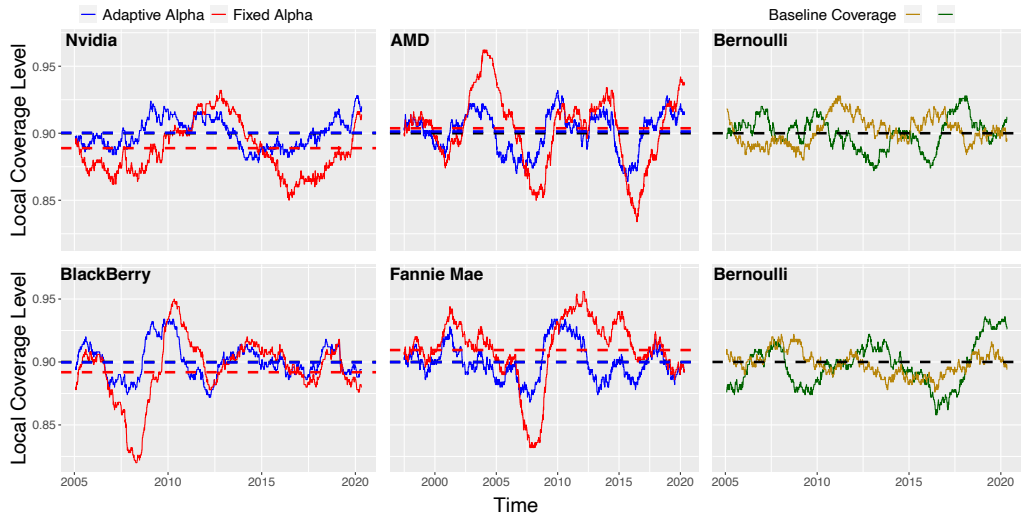

Figure 1: The leftmost four panels show local coverage frequencies for adaptive conformal (blue) and a non-adaptive method that holds $\alpha_t = \alpha$ fixed (red) for the prediction of stock market volatility. The coloured dotted lines mark the average coverage obtained across all time points, while the black line indicates the target level of $1 - \alpha = 0.9$. For a visual comparison the two rightmost panels show the value of the local coverage (4) in cases where $\{\text{err}_t\}_{1 \leq t \leq T}$ is an i.i.d. Bernoulli(0.1) sequence.

Daily open prices were obtained from publicly available datasets published by *The Wall Street Journal*. The realized local coverage frequencies for the non-adaptive and adaptive conformal methods on four different stocks are shown in Figure 1. These stocks were selected out of a total of 13 stocks that we examined because they showed a clear failure of the non-adaptive method. Adaptive conformal inference was found to perform well on all stocks that we tested (data not shown).

As a visual aid, the rightmost panel of Figure 1 shows the moving average $\frac{1}{500}\sum_{r=t-250+1}^{t+250} I_r$ for sequences $\{I_t\}_{1 \leq t \leq T}$ that are i.i.d. Bernoulli(0.1). We see that the local coverage frequencies obtained by adaptive conformal inference (blue lines) always stay within the variation that would be expected from an i.i.d. Bernoulli sequence. On the other hand, the non-adaptive method undergoes large excursions away from the target level of $1 - \alpha = 0.9$ (red lines). For example, in the bottom middle panel we can see that the non-adaptive method fails to cover the realized volatility of Fannie Mae during the 2008 financial crisis, while the adaptive method is robust to this event (see Figure 3 in the Appendix for a plot of the price of Fannie Mae over this time period).

3 Comparison to prior work

Prior work on conformal inference has considered two different types of distribution shift [29, 10]. In both cases the focus was on environments in which the calibration data is drawn i.i.d. from a single distribution P_0, while the test point comes from a second distribution P_1. In this setting Tibshirani et al. [29] showed that valid prediction sets can be obtained by re-weighting the calibration data using the likelihood ratio between P_1 and P_0. However, this requires the conditional distribution of $Y|X$ to be constant between training and testing and the likelihood ratio $P_1(X)/P_0(X)$ to be either known or very accurately estimated. On the other hand, Cauchois et al. [10] develop methods for forming prediction sets that are valid whenever P_1 and P_0 are close in f-divergence. Similar to our work they show that if $D_f(P_1||P_0) \leq \rho$ then there exists a conservative value $\alpha_\rho \in (0, 1)$ such that

$$M(\alpha_\rho) := \mathbb{P}(S(X_t, Y_t) > \hat{Q}(1 - \alpha_\rho)) \leq \alpha.$$

The difference between our work and theirs is twofold. First, while they fix a single conservative value α_ρ our methods aim to estimate the optimal choice α^* satisfying $M(\alpha^*) = \alpha$. This is not possible in the setting of [10] as they do not observe any data from which the size of the distribution shift can be estimated. Second, while they consider only one training and one testing distribution we work in a fully online setting in which the distribution is allowed to shift continuously over time.

4 Coverage guarantees

4.1 General results

In this section we outline the theoretical coverage guarantees of adaptive conformal inference. We will assume throughout that with probability one $\alpha_1 \in [0,1]$ and $\hat{Q}_t$ is non-decreasing with $\hat{Q}_t(x) = -\infty$ for all $x < 0$ and $\hat{Q}_t(x) = \infty$ for all $x > 1$. Our first result shows that over long time intervals adaptive conformal inference obtains the correct coverage frequency irrespective of any assumptions on the data-generating distribution.

Lemma 4.1 *With probability one we have that* $\forall t \in \mathbb{N}$, $\alpha_t \in [-\gamma, 1+\gamma]$.

Proof: Assume by contradiction that with positive probability $\{\alpha_t\}_{t\in\mathbb{N}}$ is such that $\inf_t \alpha_t < -\gamma$ (the case where $\sup_t \alpha_t > 1+\gamma$ is identical). Note that $\sup_t |\alpha_{t+1} - \alpha_t| = \sup_t \gamma|\alpha - \text{err}_t| < \gamma$. Thus, with positive probability we may find $t \in \mathbb{N}$ such that $\alpha_t < 0$ and $\alpha_{t+1} < \alpha_t$. However,

$$\alpha_t < 0 \implies \hat{Q}_t(\alpha_t) = \infty \implies \text{err}_t = 0 \implies \alpha_{t+1} = \alpha_t + \gamma(\alpha - \text{err}_t) \geq \alpha_t$$

and thus $\mathbb{P}(\exists t \text{ such that } \alpha_{t+1} < \alpha_t < 0) = 0$. We have reached a contradiction. □

Proposition 4.1 *With probability one we have that for all* $T \in \mathbb{N}$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} err_t - \alpha \right| \leq \frac{\max\{\alpha_1, 1-\alpha_1\} + \gamma}{T\gamma}. \tag{5}$$

In particular, $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} err_t \overset{a.s.}{=} \alpha$.

Proof: By expanding the recursion defined in (2) and applying Lemma 4.1 we find that

$$[-\gamma, 1+\gamma] \ni \alpha_{T+1} = \alpha_1 + \sum_{t=1}^{T} \gamma(\alpha - \text{err}_t).$$

Rearranging this gives the result. □

Proposition 4.1 makes no assumptions about the data generating distribution. One may immediately ask whether these results can be improved by making mild assumptions on the distribution shifts. We argue that without assumptions on the quality of the initialization the answer to this question is negative. To understand this, consider a very simple example in which $M_t(p) = p$ for all $p \in [0,1]$. Assume for simplicity that $\mathbb{E}[\text{err}_t|\alpha_t] = M_t(\alpha_t)$.[1] Then,

$$\mathbb{E}[\text{err}_t] - \alpha = \mathbb{E}[\alpha_t - \alpha] = \mathbb{E}[\alpha_{t-1} + \gamma(\alpha - \text{err}_{t-1}) - \alpha] = (1-\gamma)\mathbb{E}[\text{err}_{t-1} - \alpha].$$

Repeating this calculation recursively gives that

$$\mathbb{E}[\text{err}_t] = (1-\gamma)^{t-1}\mathbb{E}[\text{err}_1 - \alpha] = (1-\gamma)^{t-1}(\alpha_1 - \alpha),$$

and thus,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\text{err}_t] - \alpha \right| = \frac{1-\gamma^T}{T\gamma} |\alpha_1 - \alpha|.$$

The comparison of this bound to (5) is self-evident. The main difference is that we have replaced $\max\{1-\alpha_1, \alpha_1\}$ with $|\alpha_1 - \alpha|$. This arises from the fact that in this example, choosing $\alpha_t = \alpha$ gives well-calibrated prediction sets. In a more general version of this construction we would need to replace α by the target $\alpha^* \in [0,1]$. As this value is unknown we can view $\max\{\alpha_1, 1-\alpha_1\}$ as an upper bound on $|\alpha_1 - \alpha^*|$. Thus, we view Proposition 4.1 as both an agnostic guarantee that shows that our method gives the correct long-term empirical coverage frequency irrespective of the true data-generating process, and as an approximately tight bound on the worst-case behaviour immediately after initialization.

[1] In general this will not be the case unless $\hat{Q}_t(\cdot)$ and $S_t(\cdot)$ are computed on different subsets of the data.

4.2 Performance in a hidden Markov model

Although we believe Proposition 4.1 to be an approximately tight characterization of the behaviour after initialization, we can still ask whether better bounds can be obtained for large time steps. In this section we answer this question positively by showing that if α_1 is initialized appropriately and the distribution shift is small then tighter coverage guarantees can be given. In order to obtain useful results we will make some simplifying assumptions about the data generating process. While we do not expect these assumptions to hold exactly in any real-world setting, we do consider our results to be representative of the true behaviour of adaptive conformal inference and we expect similar results to hold under alternative models.

4.2.1 Setting

We model the data as coming from a hidden Markov model. In particular, we let $\{A_t\}_{t\in\mathbb{N}} \subseteq \mathcal{A}$ denote the underlying Markov chain for the environment and we assume that conditional on $\{A_t\}_{t\in\mathbb{N}}$, $\{(X_t, Y_t)\}_{t\in\mathbb{N}}$ is an independent sequence with $(X_t, Y_t) \sim P_{A_t}$ for some collection of distributions $\{P_a : a \in \mathcal{A}\}$. In order to simplify our calculations, we assume additionally that the estimated quantile function $\hat{Q}_t(\cdot)$ and score function $S_t(\cdot)$ do not depend t and we denote them by $\hat{Q}(\cdot)$ and $S(\cdot)$. This occurs for example in the split conformal setting with fixed training and calibration sets.

In this setting, $\{(\alpha_t, A_t)\}_{t\in\mathbb{N}}$ forms a Markov chain on $[-\gamma, 1+\gamma] \times \mathcal{A}$. We assume that this chain has a unique stationary distribution π and that $(\alpha_1, A_1) \sim \pi$. This implies that $(\alpha_t, A_t, \text{err}_t)$ is a stationary process and thus will greatly simplify our characterization of the behaviour of err_t. While there is little doubt that the theory can be extended, recall our main goal to get useful and simple results. That said, what we really have in mind here is that $\{A_t\}_{t\in\mathbb{N}}$ is sufficiently well-behaved to guarantee that (α_t, A_t) has a limiting stationary distribution. In Section A.3 we give an example where this is indeed provably the case. Lastly, the assumption that $(\alpha_1, A_1) \sim \pi$ is essentially equivalent to assuming that we have been running the algorithm for long enough to exit the initialization phase described in Section 4.1.

4.2.2 Large deviation bound for the errors

Our first observation is that err_t has the correct average value. More precisely, by Proposition 4.1 we have that $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} \text{err}_t \overset{a.s.}{=} \alpha$ and since err_t is stationary it follows that $\mathbb{E}[\text{err}_t] = \alpha$. Thus, to understand the deviation between $T^{-1}\sum_{t=1}^{T} \text{err}_t$ and α we simply need to characterize the dependence structure of $\{\text{err}_t\}_{t\in\mathbb{N}}$.

We accomplish this in Theorem 4.1, which gives a large deviation bound on $|T^{-1}\sum_{t=1}^{T} \text{err}_t - \alpha|$. The idea behind this result is to decompose the dependence in $\{\text{err}_t\}_{t\in\mathbb{N}}$ into two parts. First, there is dependence due to the fact that α_t is a function of $\{\text{err}_r\}_{1\le r\le t-1}$. In Section A.5 in the Appendix we argue that this dependence induces a negative correlation and thus the errors concentrate around their expectation at a rate no slower than that of an i.i.d. Bernoulli sequence. This gives rise to the first term in (6), which is what would be obtained by applying Hoeffding's inequality to an i.i.d. sequence. Second, there is dependence due to the fact that A_t depends on A_{t-1}. More specifically, consider a setting in which the distribution of $Y|X$ has more variability in some states than others. The goal of adaptive conformal inference is to adapt to the level of variability and thus return larger prediction sets in states where the distribution of $Y|X$ is more spread. However, this algorithm is not perfect and as a result there may be some states $a \in \mathcal{A}$ in which $\mathbb{E}[\text{err}_t|A_t = a]$ is biased away from α. Furthermore, if the environment tends to spend long stretches of time in more variable (or less variable) states this will induce a positive dependence in the errors and cause $T^{-1}\sum_{t=1}^{T} \text{err}_t$ to deviate from α. To control this dependence we use a Bernstein inequality for Markov chains to bound $|T^{-1}\sum_{t=1}^{T} \mathbb{E}[\text{err}_t|A_t] - \alpha|$. This gives rise to the second term in (6).

Theorem 4.1 *Assume that $\{A_t\}_{t\in\mathbb{N}}$ has non-zero absolute spectral gap $1-\eta > 0$. Let*

$$B := \sup_{a\in\mathcal{A}} |\mathbb{E}[\text{err}_t|A_t = a] - \alpha| \quad \textit{and} \quad \sigma_B^2 = \mathbb{E}[(\mathbb{E}[\text{err}_t|A_t] - \alpha)^2].$$

Then,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} \text{err}_t - \alpha\right| \ge \epsilon\right) \le 2\exp\left(-\frac{T\epsilon^2}{8}\right) + 2\exp\left(-\frac{T(1-\eta)\epsilon^2}{8(1+\eta)\sigma_B^2 + 40B\epsilon}\right). \tag{6}$$

A formal proof of this result can be found in Section A.5. The quality of this concentration inequality will depend critically on the size of the bias terms B and σ_B^2. To understand these quantities, let

$$M_a(p) := \mathbb{P}(S(X_t, Y_t) > \hat{Q}(1-p) | A_t = a)$$

denote the realized miscoverage level in state $a \in \mathcal{A}$ obtained by the quantile $\hat{Q}(1-p)$. Assume that $M_a(p)$ is continuous. This will happen for example when $\hat{Q}$ is continuous and $S(X_t, Y_t)|A_t = a$ is continuously distributed. Then, there exists an optimal value α_a^* such that $M_a(\alpha_a^*) = \alpha$. Lemma A.4 in the Appendix shows that if in addition M_a admits a second order Taylor expansion, then

$$B \leq C \left(\gamma + \gamma^{-1} \sup_{a \in \mathcal{A}} \sup_{k \in \mathbb{N}} \mathbb{E}[|\alpha^*_{A_{t+1}} - \alpha^*_{A_t}| \big| A_{t+k} = a] \right) \quad \text{and} \quad \sigma_B^2 \leq B^2.$$

Here, the constant C will depend on how much M_a differs from the ideal case in which $\hat{Q}$ is the true quantile function for $S(X_t, Y_t)|A_t = a$. In this case we would have that M_a is the linear function $M_a(p) = p, \forall p \in [0, 1]$ and $C \leq 2$.

We remark that the term $\mathbb{E}[|\alpha^*_{A_{t+1}} - \alpha^*_{A_t}| \big| A_{t+k} = a]$ can be seen as a quantitative measurement of the size of the distribution shift in terms of a change in the critical value α_a^*. Thus, we interpret these results as showing that if the distribution shift is small and $\forall a \in \mathcal{A}$, $\hat{Q}(\cdot)$ gives reasonable coverage of the distribution of $S(X_t, Y_t)|A_t = a$, then $T^{-1} \sum_{t=1}^T \text{err}_t$ will concentrate well around α.

4.2.3 Achieving approximate marginal coverage

Theorem 4.1 bounds the distance between the average miscoverage rate and the target level over long stretches of time. On the other hand, it provides no information about the marginal coverage frequency at a single time step. The following result shows that if the distribution shift is small, the realized marginal coverage rate $M_{A_t}(\alpha_t)$ will be close to α on average.

Theorem 4.2 *Assume that there exists a constant $L > 0$ such that for all $a \in \mathcal{A}$ and all $\alpha_1, \alpha_2 \in \mathbb{R}$,*

$$|M_a(\alpha_2) - M_a(\alpha_1)| \leq L|\alpha_2 - \alpha_1|.$$

Assume additionally that for all $a \in \mathcal{A}$ there exists $\alpha_a^* \in (0, 1)$ such that $M_a(\alpha_a^*) = \alpha$. Then,

$$\mathbb{E}[(M_{A_t}(\alpha_t) - \alpha)^2] \leq \frac{L(1+\gamma)}{\gamma} \mathbb{E}[|\alpha^*_{A_{t+1}} - \alpha^*_{A_t}|] + \frac{L}{2}\gamma. \tag{7}$$

Proof of this result can be found in Section A.6 of the Appendix. We remark that the right-hand side of (7) is minimized by choosing $\gamma = (2\mathbb{E}[|\alpha^*_{A_{t+1}} - \alpha^*_{A_t}|])^{1/2}$. This choice of γ gives the inequality

$$\mathbb{E}[(M_{A_t}(\alpha_t) - \alpha)^2] \leq L(\sqrt{2} + 1)\sqrt{\mathbb{E}[|\alpha^*_{A_{t+1}} - \alpha^*_{A_t}|]}.$$

As above we have that in the ideal case $\hat{Q}(\cdot)$ is a perfect estimate of the quantiles of $S(X_t, Y_t)|A_t = a$ and thus $M_a(p) = p$ and $L = 1$. Moreover, we once again have the interpretation that $\mathbb{E}[|\alpha^*_{A_{t+1}} - \alpha^*_{A_t}|]$ is a quantitative measurement of the distribution shift. Thus, this result can be interpreted as bounding the average difference between the realized and target marginal coverage in terms of the size of the underlying distribution shift. Finally, note that the choice of $\gamma = (2\mathbb{E}[|\alpha^*_{A_{t+1}} - \alpha^*_{A_t}|])^{1/2}$ formalizes our intuition that γ should be chosen to be larger in domains with greater distribution shift, while not being so large as to cause α_t to be overly volatile.

5 Real data example: election night predictions

During the 2020 presidential election *The Washington Post* used conformalized quantile regression (CQR) (see (1) and Section 1.1) to produce county level predictions of the vote total on election night [12]. Here we replicate the core elements of this method using both fixed and adaptive quantiles.

To make the setting precise, let $\{Y_t\}_{1 \leq t \leq T}$ denote the number of votes cast for Joe Biden in the 2020 presidential election in each of approximately $T = 3000$ counties in the United States. Let X_t denote a set of demographic covariates associated to the tth county. In our experiments X_t will

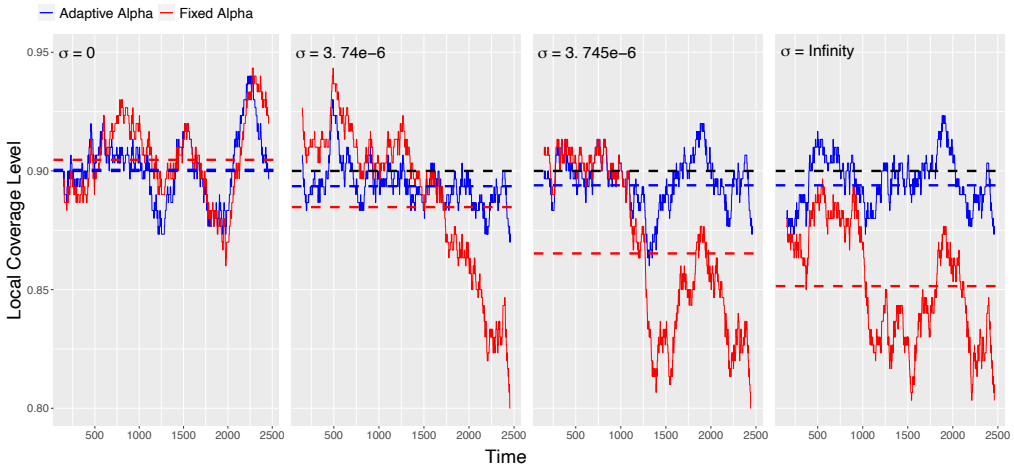

Figure 2: Local coverage frequencies of adaptive conformal (blue) and a non-adaptive method that holds $\alpha_t = \alpha$ fixed (red) on county-level election predictions. Coloured dotted lines show the average coverage across all time points, while the black line indicates the target coverage level of $1 - \alpha = 0.9$. Note that here $\sigma = \infty$ corresponds to ordering the counties exactly by their population size. In our experiments we examined a wider range of values for σ. We have chosen to display results from these four values because they demonstrate a clear transition (in terms of coverage) between the exchangeable case and the perfectly ordered case.

include information on the make-up of the county population by ethnicity, age, sex, median income and education (see Section A.4.1 for details). On election night county vote totals were observed as soon as the vote count was completed. If the order in which vote counts were completed was uniformly random $\{(X_t, Y_t)\}_{1 \leq t \leq T}$ would be an exchangeable sequence on which we could run standard conformal inference methods. In reality, larger urban counties tend to report results later than smaller rural counties and counties on the east coast reported earlier than those on the west coast. Thus, the distribution of (X_t, Y_t) can be viewed as drifting throughout election night.

We apply CQR to predict the county-level vote totals (see Section A.4.2 for details). We consider four different orderings of the counties each obtained by sampling counties without replacement with probability proportional to $\exp(\sigma \cdot \text{county population size})$. Here, $\sigma > 0$ is a parameter whose value we vary. Figure 2 shows the realized local coverage frequency over the most recent 300 counties (see (4)) for the non-adaptive and adaptive conformal methods. When $\sigma = 0$ (leftmost panel) the county data are exchangeable and both the non-adaptive and adaptive method perform well. As σ increases (right panels) the size of the distribution shift increases and the non-adaptive method fails to achieve 90% coverage for later counties. On the other hand, the adaptive method adjusts accordingly and maintains an approximate 90% local coverage rate across all time points. In fact, similar to what was observed in Section 2.2 we find that the maximum deviation of the local coverage rate from 90% is no larger than what is observed in the ideal case where the data are exchangeable.

6 Discussion

There are still many open problems in this area. The methods we develop are specific to cases where Y_t is revealed at each time point. However, there are many settings in which we receive the response in a delayed fashion or in large batches. In addition, our theoretical results in Section 4.2 are limited to a single model for the data generating distribution and the special case where the quantile function $\hat{Q}(\cdot)$ is fixed across time. It would be interesting to determine if similar results can be obtained in settings where $\hat{Q}(\cdot)$ is fit in an online fashion on the most recent data. Another potential area for improvement is in the choice of the step size γ. In Section 2.1 we give some heuristic guidelines for choosing γ based on the size of the distribution shift in the environment. Ideally however we would like to be able to choose γ adaptively without prior knowledge. Finally, our experimental results are limited to just two domains. Additional work is needed to determine if our methods can successfully protect against a wider variety of real-world distribution shifts.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius B. Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago M. Paixão, Filipe Mutz, Lucas de Paula Veronese, Thiago Oliveira-Santos, and Alberto F. De Souza. Self-driving cars: A survey. *Expert Systems with Applications*, 165:113816, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2020.113816. URL https://www.sciencedirect.com/science/article/pii/S095741742030628X.

[3] 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.

[4] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(86)90063-1. URL https://www.sciencedirect.com/science/article/pii/0304407686900631.

[5] United States Census Bureau. County characteristics resident population estimates, 2019. data retrieved from https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html.

[6] United States Census Bureau. 2015-2019 american community survey 5-year average county-level estimates, 2019. data retrieved from https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/.

[7] United States Census Bureau. Small area income and poverty estimates: 2019, 2019. data retrieved from https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/.

[8] Emmanuel J. Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis, 2021.

[9] X. Cao, J. Zhang, and H. V. Poor. On the time-varying distributions of online stochastic optimization. In *2019 American Control Conference (ACC)*, pages 1494–1500, 2019. doi: 10.23919/ACC.2019.8814889.

[10] Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C. Duchi. Robust validation: Confident predictions even when distributions shift, 2020.

[11] Maxime Cauchois, Suyash Gupta, and John C. Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22(81):1–42, 2021. URL http://jmlr.org/papers/v22/20-753.html.

[12] John Cherian and Lenny Bronner. How the washington post estimates outstanding votes for the 2020 presidential election. https://elex-models-prod.s3.amazonaws.com/2020-general/write-up/election_model_writeup.pdf, 2020.

[13] MIT Election Data and Science Lab. County Presidential Election Returns 2000-2016, 2018. URL https://doi.org/10.7910/DVN/VOQCHQ.

[14] 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.

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

[16] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

[17] Bai Jiang, Qiang Sun, and Jianqing Fan. Bernstein's inequality for general markov chains, 2020.

[18] Danijel Kivaranovic, Robin Ristl, Martin Posch, and Hannes Leeb. Conformal prediction intervals for the individual treatment effect, 2020.

[19] 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.

[20] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://doi.org/10.1080/01621459.2017.1307116.

[21] Lihua Lei and Emmanuel J. Candès. Conformal inference of counterfactuals and individual treatment effects, 2021.

[22] David Leip. Dave leip's atlas of u.s. presidential elections., 2020. http://uselectionatlas.org.

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

[24] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Machine Learning: ECML 2002*, pages 345–356, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-36755-0.

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

[26] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/244edd7e85dc81602b7615cd705545f5-Paper.pdf.

[27] 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.

[28] 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.

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

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

[31] Jingyi Zhu and James Spall. Tracking capability of stochastic gradient algorithm with constant gain. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 4522–4527, 12 2016. doi: 10.1109/CDC.2016.7798957.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The primary claims presented in the introduction are that adaptive conformal inference obtains the target long-term coverage frequency under no assumptions on the data-generating distribution and additionally obtains approximate marginal coverage under small distribution shifts. These claims are addressed theoretically in Proposition 4.1 and Theorem 4.2 and empirically in Sections 2.2 and 5.
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [No] This paper develops a wrapper that can be combined with any regression algorithm. Thus, the potential negative societal impacts of our work are the same as the potential negative societal impacts of any generic prediction method. Since these are not specific to our work we do not discuss them here.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Proofs of the results can be found in Sections 4.1, A.5, A.6, A.7, and A.8.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Data for the 2020 election is proprietary. All other datasets are publicly available and citations to the specific locations where they can be found are included in the relevant sections. Code for replicating the main experiments is included in the Supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The key hyperparameter is γ which is specified in Section 2.1. Detailed descriptions of our experimental procedures can be found in Sections 2.2 and A.4.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The experiments in this article do not require significant computational resources. They were all be run in less than a couple of hours on a standard MacBook Pro.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] 2016 election data is under a CC0 licence. 2020 county-level election data is the property of David Leip and we confirmed in private communication with him that our use of the data falls under the accepted terms of use of his custom license agreement. We could not find licensing information for the stock market or Census Bureau data.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We used David Leip's 2020 county-level election data. We confirmed in private communication with him that our use case was acceptable and we included a citation to his dataset in the formated he requested.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]