Conformal Prediction Under Distribution Shift

Advanced Topics in Statistical Learning, Spring 2023

Ryan Tibshirani

Note: we're following the context, problem setup, notation, etc. from the last lecture on conformal prediction.

In this lecture we cover conformal methods that apply beyond the i.i.d. setting. This is a very active and recent topic of research, and it's possible (or even likely) that what's considered fundamental in this area will change in the next few years. Until then, there will be numerous topics from which we can "pick and choose" for a lecture like this one. We've chosen three such topics—and shamelessly (shamefully?), for two of these, you'll notice an overlap between the authors list and the author of these lecture notes. We should be clear that there is plenty of other interesting work out there that we can't cover in just one lecture.

1 Likelihood-weighted conformal prediction

We first cover a likelihood-weighted conformal prediction method, due to Tibshirani et al. (2019). A primary motivation will be the setting of *covariate shift*, where

$$\begin{aligned} (X_i, Y_i) &\sim P = P_X \times P_{Y|X}, \text{ independently, for } i = 1, \ldots, n, \\ (X_{n+1}, Y_{n+1}) &\sim \tilde{P} = \tilde{P}_X \times P_{Y|X}, \text{ independently.} \end{aligned} \tag{1}$$

Notice that the conditional distribution of $Y|X$ is assumed to be the same for both the training and test data, but the distribution of X is allowed to change, i.e., we allow $\tilde{P}_X \neq P_X$. This is a general framework of great interest, because it encompasses may important problem settings. For example, we could have done some kind of structured covariate sampling for our training set (demographically, geographically, etc.) but then we do prediction "in the wild", where the mix of covariates is different.

The first thing we could ask is: does this even matter for conformal prediction? That is, if we observed data according to (1), and computed the usual conformal prediction intervals, then would we see a problem with coverage? The top row of Figure 1 provides an answer, empirically. This is taken from Tibshirani et al. (2019), and shows the results of an experiment in which, over 5000 repetitions, two test sets are drawn: one without covariate shift (results in red), and one with covariate shift (in blue). The top left panel shows the test coverage of split conformal prediction intervals (drawn as histograms, over the 5000 repetitions). We can see that coverage fails quite noticeably in the covariate shift setting.

To remedy this, we are going to work with a weighed empirical distribution of conformity scores, rather than the usual (unweighted) empirical distribution. And to approach this argument, it helps to build intuition by looking back at the first key idea behind conformal, which recall, used ranks in order to construct adjusted empirical quantiles.

Revisiting the first key idea: rank-based quantiles. The last lecture proved the following fact. If $R_1, \ldots, R_{n+1}$ are exchangeable random variables, then for any $\alpha \in (0, 1)$,

$$\mathbb{P}\bigg\{ R_{n+1} \leq \text{Quantile}\bigg(\frac{\lceil (1-\alpha)(n+1) \rceil}{n}; \frac{1}{n} \sum_{i=1}^{n} \delta_{R_i} \bigg) \bigg\} \geq 1 - \alpha.$$

We can equivalently express this as

$$\mathbb{P}\bigg\{ R_{n+1} \leq \text{Quantile}\bigg(\frac{\lceil (1-\alpha)(n+1) \rceil}{n+1}; \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{R_i} \bigg) \bigg\} \geq 1 - \alpha,$$

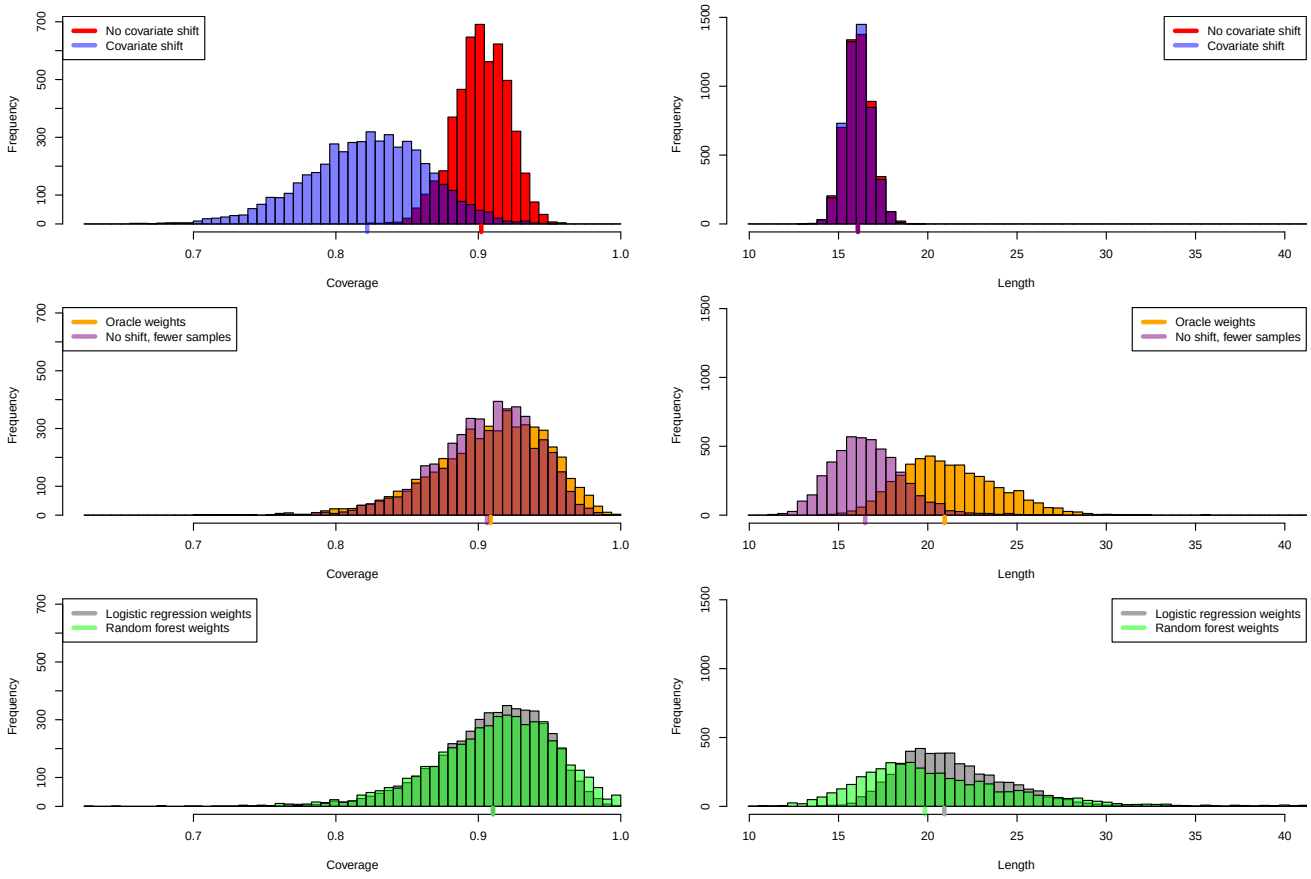

Figure 1: *Experiments for split conformal prediction under covariate shift. All results are aggregated over 5000 repetitions, each of which randomly forms training and test sets. Top row: the usual split conformal prediction with and without covariate shift. Middle row: weighted split conformal using the true (and in general unknown) likelihood ratio between test and training covariate feature distributions, compared to ordinary split conformal without covariate shift but in a problem with a comparable effective sample size. Bottom row: weighted split conformal using an estimated likelihood ratio from running classification with logistic regression or random forests. Credit: Tibshirani et al. (2019).*

because the event in each of the last two displays is equivalent to the statement that R_{n+1} is among the $\lceil (1-\alpha)(n+1) \rceil$ smallest of $R_1, \ldots, R_{n+1}$. The last display is itself equivalent to

$$\mathbb{P}\bigg\{ R_{n+1} \leq \text{Quantile}\bigg(1-\alpha;\ \frac{1}{n+1}\sum_{i=1}^{n+1} \delta_{R_i}\bigg)\bigg\} \geq 1-\alpha,$$

because the quantile function of the empirical distribution of $R_1, \ldots, R_{n+1}$, only changes in increments of $1/(n+1)$ (and will automatically round up to the nearest increment until it captured sufficient probability mass to exceed $1-\alpha$). Finally, it turns out that we can equivalently express the last display as

$$\mathbb{P}\bigg\{ R_{n+1} \leq \text{Quantile}\bigg(1-\alpha;\ \frac{1}{n+1}\sum_{i=1}^{n} \delta_{R_i} + \frac{1}{n+1}\delta_\infty\bigg)\bigg\} \geq 1-\alpha. \tag{2}$$

This can be seen by applying the following fact to the complements of the two events in the previous two displays: for a discrete distribution F with support points $a_1, \ldots, a_k \in \mathbb{R}$, denoting $q = \text{Quantile}(\beta; F)$, if we reassign the points $a_i > q$ to arbitrary values strictly larger than q, yielding a new distribution F', then the level β quantile remains unchanged, $\text{Quantile}(\beta; F) = \text{Quantile}(\beta; F')$.

Alternate proof of the quantile result (2). We will now prove (2) from a new perspective (no longer by reducing it to a statement about the rank of R_{n+1} among $R_1, \ldots, R_{n+1}$) that will enable us to extend this result to a more general setting. The basic idea is to condition on the unlabeled collection of values obtained by our random variables $R_1, \ldots, R_{n+1}$, then inspect the probabilities that the last random variable R_{n+1} attains each one of these values.

Denote by f the probability density function (or mass function, or more generally, Radon-Nikodym derivative with respect to an arbitrary base measure) of the joint sample $R_1, \ldots, R_{n+1}$. Exchangeability means

$$f(r_1, \ldots, r_{n+1}) = f(r_{\sigma(1)}, \ldots, r_{\sigma(n+1)}), \quad \text{for all permutations } \sigma.$$

For simplicity, and without loss of generality, assume that there are almost surely no ties among the scores $R_1, \ldots, R_{n+1}$. Let E_r be the event that $\{R_1, \ldots, R_{n+1}\} = \{r_1, \ldots, r_{n+1}\}$. Then for each i,

$$\begin{aligned}
\mathbb{P}(R_{n+1} = r_i \,|\, E_r) &= \frac{\sum_{\sigma:\sigma(n+1)=i} f(r_{\sigma(1)}, \ldots, r_{\sigma(n+1)})}{\sum_\sigma f(r_{\sigma(1)}, \ldots, r_{\sigma(n+1)})} \\
&= \frac{\sum_{\sigma:\sigma(n+1)=i} f(r_1, \ldots, r_{n+1})}{\sum_\sigma f(r_1, \ldots, r_{n+1})} \\
&= \frac{n!}{(n+1)!} = \frac{1}{n+1}.
\end{aligned}$$

This shows that the distribution of $R_{n+1}|E_r$ is uniform on the set $\{r_1, \ldots, r_{n+1}\}$, that is,

$$R_{n+1}|E_r \sim \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{r_i},$$

and it follows, since $F(Q(t)) \geq t$ for any cumulative distribution function F and corresponding quantile function Q, that

$$\mathbb{P}\left\{ R_{n+1} \leq \text{Quantile}\left(1-\alpha;\, \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{r_i}\right) \,\middle|\, E_r \right\} \geq 1-\alpha,$$

This is the same as

$$\mathbb{P}\left\{ R_{n+1} \leq \text{Quantile}\left(1-\alpha;\, \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{R_i}\right) \,\middle|\, E_r \right\} \geq 1-\alpha,$$

and we can marginalize to obtain

$$\mathbb{P}\left\{ R_{n+1} \leq \text{Quantile}\left(1-\alpha;\, \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{R_i}\right) \right\} \geq 1-\alpha.$$

This is the display right above (2), and by the same argument as given above, it is equivalent to (2).

1.1 Weighted exchangeability: quantile lemma

Though the alternate proof we just gave is a bit longer than the standard reduction to ranks, it is important because it allows us to move past the setting of exchangeable scores $R_1, \ldots, R_{n+1}$. In words, after revealing (conditioning on) the set of values obtained by the scores, we need to be able to answer the following question: *what is the probability with which any given value is that of the test score?*

This question still has a relatively clean answer when $R_1, \ldots, R_{n+1}$ are *weighted exchangeable*, which is a generalization of exchangeability, and specifies that the random variables have a density (or mass function, or more generally, Radon-Nikodym derivative with respect to an arbitrary base measure) of the form

$$f(r_1, \ldots, r_{n+1}) = \prod_{i=1}^{n+1} w_i(r_i) \cdot g(r_1, \ldots, r_{n+1}), \tag{3}$$

where g is any function that is permutation invariant, i.e., $g(r_1, \ldots, r_{n+1}) = g(r_{\sigma(1)}, \ldots, r_{\sigma(n+1)})$, for any permutation σ.

We now have the following extension of (2), stated as a lemma, for concreteness.

Lemma 1. *Let Z_i, $i = 1, \ldots, n+1$ be weighted exchangeable random variables, with respect to weight functions $w_1, \ldots, w_{n+1}$. Assume without loss of generality that these are distinct almost surely. Let*

$$R_i = V(Z_i;\, Z_1, \ldots, Z_{n+1}), \quad i = 1, \ldots, n+1,$$

where V is an arbitrary score function that is symmetric in its last $n+1$ arguments, and define

$$p_i^w(z_1, \ldots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}{\sum_\sigma \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}, \quad i = 1, \ldots, n+1, \tag{4}$$

where the sums are over permutations σ of the numbers $1, \ldots, n+1$. Then for any $\alpha \in (0,1)$,

$$\mathbb{P}\bigg\{R_{n+1} \leq \text{Quantile}\bigg(1-\alpha;\, \sum_{i=1}^n p_i^w(Z_1, \ldots, Z_{n+1})\delta_{R_i} + p_{n+1}^w(Z_1, \ldots, Z_{n+1})\delta_\infty\bigg)\bigg\} \geq 1-\alpha. \tag{5}$$

Proof. We follow the same general strategy from the alternate proof of (2). Let E_z denote the event that $\{Z_1, \ldots, Z_{n+1}\} = \{z_1, \ldots, z_{n+1}\}$, and let $r_i = V(z_i;\, z_1, \ldots, z_{n+1})$, for $i = 1, \ldots, n+1$. Let f be the density function of the joint sample $Z_1, \ldots, Z_{n+1}$. For each i, we have

$$\mathbb{P}(R_{n+1} = r_i \,|\, E_z) = \mathbb{P}(Z_{n+1} = z_i \,|\, E_z) = \frac{\sum_{\sigma:\sigma(n+1)=i} f(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})}{\sum_\sigma f(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})}, \tag{6}$$

and as $Z_1, \ldots, Z_{n+1}$ are weighted exchangeable,

$$\begin{aligned}
\frac{\sum_{\sigma:\sigma(n+1)=i} f(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})}{\sum_\sigma f(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})} &= \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) \cdot g(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})}{\sum_\sigma \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) \cdot g(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})} \\
&= \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) \cdot g(z_1, \ldots, z_{n+1})}{\sum_\sigma \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) \cdot g(z_1, \ldots, z_{n+1})} \\
&= p_i^w(z_1, \ldots, z_{n+1}).
\end{aligned}$$

In other words,

$$R_{n+1} | E_z \sim \sum_{i=1}^{n+1} p_i^w(z_1, \ldots, z_{n+1})\delta_{r_i},$$

which implies that

$$\mathbb{P}\bigg\{R_{n+1} \leq \text{Quantile}\bigg(1-\alpha;\, \sum_{i=1}^{n+1} p_i^w(z_1, \ldots, z_{n+1})\delta_{r_i}\bigg) \,\bigg|\, E_z\bigg\} \geq 1-\alpha.$$

This is equivalent to

$$\mathbb{P}\bigg\{R_{n+1} \leq \text{Quantile}\bigg(1-\alpha;\, \sum_{i=1}^{n+1} p_i^w(Z_1, \ldots, Z_{n+1})\delta_{R_i}\bigg) \,\bigg|\, E_z\bigg\} \geq 1-\alpha,$$

and after marginalizing,

$$\mathbb{P}\bigg\{R_{n+1} \leq \text{Quantile}\bigg(1-\alpha;\, \sum_{i=1}^{n+1} p_i^w(Z_1, \ldots, Z_{n+1})\delta_{R_i}\bigg)\bigg\} \geq 1-\alpha.$$

Finally, by the same arguments as before, we can change the point mass at R_{n+1} to one at ∞, which proves (5) as desired. □

We remark that computation of the probability weights in (4) is very difficult in general, due to the combinatorial form (note that this actually reduces to computing what is known as a *matrix permanent*, which is known to be hard). However, for cetain weighted exchangeable structures, it can be easy, as we will see a bit later for covariate shift.

1.2 Weighted exchangeability: conformal prediction

A weighted version of conformal prediction follows from Lemma 1, which we state next as a theorem, for concreteness.

Theorem 1. *Assume that $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, n+1$ are weighted exchangeable with weight functions $w_1, \ldots, w_{n+1}$. Define a weighted conformal set (based on the first n samples) at a point $x \in \mathcal{X}$, with nominal error level $\alpha \in (0,1)$ as follows. Let*

$$\begin{aligned} R_i^{(x,y)} &= V\Big((X_i, Y_i);\, Z_1, \ldots, Z_n, (x,y)\Big), \quad i = 1, \ldots, n, \\ R_{n+1}^{(x,y)} &= V\Big((x,y);\, Z_1, \ldots, Z_n, (x,y)\Big), \end{aligned} \tag{7}$$

for an arbitrary score function V that is symmetric in its last $n+1$ arguments, and

$$\hat{C}_n^w(x) = \bigg\{ y : R_{n+1}^{(x,y)} \leq \text{Quantile}\Big(1-\alpha;\, \sum_{i=1}^n p_i^w\big(Z_1, \ldots, Z_n, (x,y)\big)\delta_{R_i^{(x,y)}} + p_{n+1}^w\big(Z_1, \ldots, Z_n, (x,y)\big)\delta_\infty\Big)\bigg\}, \tag{8}$$

where p_i^w, $i = 1, \ldots, n+1$ are as in (4). Then $\hat{C}_n^w$ satisfies

$$\mathbb{P}\big(Y_{n+1} \in \hat{C}_n^w(X_{n+1})\big) \geq 1 - \alpha. \tag{9}$$

Proof. Abbreviate $R_i = R_i^{(X_{n+1}, Y_{n+1})}$, $i = 1, \ldots, n+1$. By construction

$$Y_{n+1} \in \hat{C}_n^w(X_{n+1}) \iff R_{n+1} \leq \text{Quantile}\Big(1-\alpha;\, \sum_{i=1}^n p_i^w(Z_1, \ldots, Z_{n+1})\delta_{R_i} + p_{n+1}^w(Z_1, \ldots, Z_{n+1})\delta_\infty\Big),$$

and applying Lemma 1 gives the result. □

Split version. The split conformal version of the above result can be viewed as a special case where the score function relies on a point predictor that has been fit on an external data set. For example, if we take it to be $V(x,y) = |y - \mu_0(x)|$, where μ_0 has been fit on a data set Z_0, then (8) simplifies to

$$\hat{C}_n^w(x) = \mu_0(x) \pm \text{Quantile}\Big(1-\alpha;\, \sum_{i=1}^n p_i^w\big(Z_1, \ldots, Z_n, (x,y)\big)\delta_{|Y_i - \mu_0(X_i)|} + p_{n+1}^w\big(Z_1, \ldots, Z_n, (x,y)\big)\delta_\infty\Big),$$

and by (9), this has coverage at least $1-\alpha$, conditional on Z_0.

CDF form.* The analogous CDF form of the conformal set in (8) is as follows:

$$\hat{C}_n^w(x) = \bigg\{ y : \sum_{i=1}^{n+1} p_i^w\big(Z_1, \ldots, Z_n, (x,y)\big) 1\Big\{R_i^{(x,y)} < R_{n+1}^{(x,y)}\Big\} \leq \lceil 1-\alpha \rceil_w \bigg\}, \tag{10}$$

where $\lceil 1-\alpha \rceil_w = \min\{\tau \in \text{range}(\hat{F}_n^w) : \tau \geq 1-\alpha\}$ and we use $\hat{F}_n^w$ to denote the (random) CDF of the discrete distribution $\sum_{i=1}^n p_i^w(Z_1, \ldots, Z_n, (x,y))\delta_{R_i^{(x,y)}} + p_{n+1}^w(Z_1, \ldots, Z_n, (x,y))\delta_\infty$. Compared to the CDF form of the ordinary unweighted conformal prediction set, from the last lecture, the form in (8) is more complicated—we need to adjust the nominal level of $1-\alpha$ upwards so that it lies in the range of the CDF of the weighted score distribution, and here this distribution is random, so the adjustment is itself random. This is the main reason we worked with the quantile form in (8) in the first place, since we can always use the unadjusted level $1-\alpha$ and completely avoid any such complications.

Auxiliary randomization.* It is worth noting that we can achieve exact coverage by using auxiliary randomization, either in CDF or quantile form. Applying our previous randomization trick (from the last lecture) to the CDF form (10) gives

$$\hat{C}_n^{w,*}(x) = \bigg\{ y : \sum_{i=1}^n p_i^w\big(Z_1,\dots,Z_n,(x,y)\big) 1\Big\{R_i^{(x,y)} < R_{n+1}^{(x,y)}\Big\} + U \sum_{i=1}^{n+1} p_i^w\big(Z_1,\dots,Z_n,(x,y)\big) 1\Big\{R_i^{(x,y)} = R_{n+1}^{(x,y)}\Big\} \leq 1-\alpha \bigg\},$$

where $U \sim \mathrm{Unif}(0,1)$, independent of everything else. This is fairly simple and intuitive: it is free of any level adjustments needed in the unrandomized CDF-based set in (10). Meanwhile, we can also randomize the quantile form (8), as in

$$\begin{aligned}\hat{C}_n^{w,*}(x) = \bigg\{ y : R_{n+1}^{(x,y)} \leq{}& \\ & B^w \cdot \mathrm{Quantile}\bigg(1-\alpha; \sum_{i=1}^n p_i^w\big(Z_1,\dots,Z_n,(x,y)\big)\delta_{R_i^{(x,y)}} + p_{n+1}^w\big(Z_1,\dots,Z_n,(x,y)\big)\delta_\infty\bigg) + \\ & (1-B^w) \cdot \mathrm{Quantile}\bigg(\lfloor 1-\alpha \rfloor_w; \sum_{i=1}^n p_i^w\big(Z_1,\dots,Z_n,(x,y)\big)\delta_{R_i^{(x,y)}} + p_{n+1}^w\big(Z_1,\dots,Z_n,(x,y)\big)\bigg)\bigg\},\end{aligned}$$

where now $\lfloor 1-\alpha \rfloor_w = \max\{\tau \in \mathrm{range}(\hat{F}_n^w) : \tau \leq 1-\alpha\}$ with $\hat{F}_n^w$ denoting the (random) CDF of the discrete distribution $\sum_{i=1}^n p_i^w(Z_1,\dots,Z_n,(x,y))\delta_{R_i^{(x,y)}} + p_{n+1}^w(Z_1,\dots,Z_n,(x,y))\delta_\infty$ as before, and

$$B^w \sim \mathrm{Bernoulli}\bigg(\frac{1-\alpha-\lfloor 1-\alpha \rfloor_w}{\lceil 1-\alpha \rceil_w - \lfloor 1-\alpha \rfloor_w}\bigg),$$

independent of everything else. Arguably, the randomized conformal set from the second-to-last display is actually less simple and intuitive, as we must introduce level adjustments that were not needed in (8). Ultimately, either version of the set $\hat{C}_n^{w,*}$ defined above (it is not clear that the two are equivalent) satisfies

$$\mathbb{P}\big(Y_{n+1} \in \hat{C}_n^{w,*}(X_{n+1})\big) = 1-\alpha.$$

1.3 Conformal prediction for covariate shift

We now show how to apply the above results to get a version of conformal prediction for covariate shift problems, as developed in Tibshirani et al. (2019). However, we note that Theorem 1 can also be used as a basis for developing conformal methods in other non-i.i.d. settings, such as label shift (Podkopaev and Ramdas, 2021), causal inference (Lei and Candès, 2021), experimental design (Fannjiang et al., 2022), and survival analysis (Candès et al., 2023).

Corollary 1. *Assume that $Z_i = (X_i, Y_i)$, $i = 1,\dots,n+1$ obey the model (1). Assume that $\tilde{P}_X$ is absolutely continuous with respect to P_X, and denote $w = d\tilde{P}_X/dP_X$. Define a weighted conformal set (based on the first n samples) at a point $x \in \mathcal{X}$, with nominal error level $\alpha \in (0,1)$, by*

$$\hat{C}_n^w(x) = \bigg\{ y : R_{n+1}^{(x,y)} \leq \mathrm{Quantile}\bigg(1-\alpha; \sum_{i=1}^n \pi_i^w(x)\delta_{R_i^{(x,y)}} + \pi_{n+1}^w(x)\delta_\infty\bigg)\bigg\}, \tag{11}$$

where $R_i^{(x,y)}$, $i = 1,\dots,n+1$ are conformity scores as in (7), for an arbitrary score function V that is symmetric in its last $n+1$ arguments, and

$$\pi_i^w(x) = \frac{w(X_i)}{\sum_{j=1}^n w(X_j) + w(x)},\ i = 1,\dots,n, \quad \text{and} \quad \pi_{n+1}^w(x) = \frac{w(x)}{\sum_{j=1}^n w(X_j) + w(x)}. \tag{12}$$

Then $\hat{C}_n^w$ satisfies

$$\mathbb{P}\big(Y_{n+1} \in \hat{C}_n^w(X_{n+1})\big) \geq 1-\alpha. \tag{13}$$

Proof. It is straightforward to see that the independent draws $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n+1$ are weighted exchangeable (3) with $w_i \equiv 1$ for $i = 1, \ldots, n$, and $w_{n+1}((x, y)) = w(x)$. In this special case, the probabilities in (4) simplify to

$$p_i^w(z_1, \ldots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} w(x_i)}{\sum_\sigma w(x_{\sigma(n+1)})} = \frac{w(x_i)}{\sum_{j=1}^{n+1} w(x_j)}, \quad i = 1, \ldots, n+1,$$

in other words, $p_i^w(Z_1, \ldots, Z_n, (x, y)) = \pi_i^w(x)$, $i = 1, \ldots, n+1$, where the latter are as in (12). Applying Theorem 1 gives the result. □

The same remarks as before apply here: a split conformal version follows as a special case (via a particular score function) and exact coverage in (13) can be achieved by randomizing the quantile in (11).

Looking back at Figure 1, the middle row provides an example of the (split version of the) conformal set in (11), with oracle knowledge (in orange) of the likelihood ratio weight function $w = d\tilde{P}_X/dP_X$. We can see from the middle left panel that its coverage is restored compared to the naive application of conformal in the covariate shift problem, from the top row. However, we also see that the dispersion in the coverage histogram from weighted conformal (over the 5000 repetitions of the experiment) is larger than that of ordinary conformal without covariate shift (in red), from the top row. This is because, with non-uniform weights due to covariate shift, we are effectively operating at a lower sample size. The middle row thus also displays the results (in purple) of usual conformal prediction in a problem without covariate shift but at the same effective sample size, defined as

$$\hat{n} = \left(\frac{\sum_{i=1}^n |w(X_i)|}{\sqrt{\sum_{i=1}^n |w(X_i)|^2}} \right)^2 = \left(\frac{\|w(X_{1:n})\|_1}{\|w(X_{1:n})\|_2} \right)^2,$$

where we abbreviate $w(X_{1:n}) = (w(X_1), \ldots, w(X_n)) \in \mathbb{R}^n$. We see that its coverage dispersion is about the same. Interestingly (and unfortunately for the likelihood-weighted method), even with the effective sample size correction, the usual conformal prediction intervals are shorter than the weighted conformal prediction intervals, as shown in the middle right panel.

1.4 Estimating the likelihood ratio from unlabeled data

Here we describe how to estimate $w = d\tilde{P}_X/dP_X$, the likelihood ratio of interest, when we have access to unlabeled data $X_{n+1}, \ldots, X_{n+m} \in \mathcal{X}$ at prediction time. (This is sometimes called the transductive or semi-supervised setting in machine learning.) We can use any classifier that estimated probabilities of class membership, such as logistic regression or random forests. We proceed as follows: we train the classifier on feature-class pairs (X_i, C_i), $i = 1, \ldots, n+m$, where $C_i = 0$ for $i = 1, \ldots, n$ and $C_i = 1$ for $i = n+1, \ldots, n+m$. Noting that

$$\frac{\mathbb{P}(C = 1|X = x)}{\mathbb{P}(C = 0|X = x)} = \frac{\mathbb{P}(C = 1)}{\mathbb{P}(C = 0)} \frac{d\tilde{P}_X}{dP_X}(x),$$

we can thus view the conditional odds ratio $w(x) = \mathbb{P}(C = 1|X = x)/\mathbb{P}(C = 0|X = x)$ as an equivalent representation for the oracle weight function—since we actually only need to know the likelihood ratio up to a proportionality constant. Therefore, if $\hat{p}(x)$ is an estimate of $\mathbb{P}(C = 1|X = x)$ obtained by fitting a probabilistic classifier to the data (X_i, C_i), $i = 1, \ldots, n+m$, then we can use

$$\hat{w}(x) = \frac{\hat{p}(x)}{1 - \hat{p}(x)} \tag{14}$$

as our estimated weight function for the calculation of probabilities (12), needed for the weighted conformal set (11). The better calibrated the classifier, the better the estimated weighted in (14) will be.

Looking back once again at Figure 1, the bottom row shows the results of using this method to estimate the weights using logistic regression (in gray) and random forests (in green). Both classifiers provide reasonably good prediction sets in the end (logistic regression is actually well-specified in this experiment, so its favorable performance should not be surprising).

1.5 Conformal prediction for structured-X settings

We saw that a particularly simple and computationally efficient application of Theorem 1 was the covariate shift problem. Now we go in the opposite direction: make it even more general and add the same time, even more computationally intractable (at least at face value). The next result essentially already follows from what we proved in Lemma 1 and Theorem 1: we just stop at (6), without simplifying further.

Theorem 2. *Assume that $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n+1$ are distributed according to:*

$$(X_1, \ldots, X_{n+1}) \sim \Lambda,$$
$$Y_i | X_i \sim P_{Y|X}, \text{ independently, for } i = 1, \ldots, n+1.$$

Let λ denote the density (or mass function, or more generally, Radom-Nikodym derivative with respect to an arbitrary base measure) of Λ. Define a weighted conformal set (based on the first n samples) at a point $x \in \mathcal{X}$, with nominal error level $\alpha \in (0,1)$, by

$$\hat{C}_n^{\lambda}(x) = \left\{ y : R_{n+1}^{(x,y)} \leq \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n} p_i^{\lambda}(X_1, \ldots, X_n, x)\delta_{R_i^{(x,y)}} + p_{n+1}^{\lambda}(X_1, \ldots, X_n, x)\delta_{\infty}\right)\right\}, \quad (15)$$

where $R_i^{(x,y)}$, $i = 1, \ldots, n+1$ are conformity scores as in (7), for an arbitrary score function V that is symmetric in its last $n+1$ arguments, and

$$p_i^{\lambda}(x_1, \ldots, x_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} \lambda(x_{\sigma(1)}, \ldots, x_{\sigma(n+1)})}{\sum_{\sigma} \lambda(x_{\sigma(1)}, \ldots, x_{\sigma(n+1)})}, \quad i = 1, \ldots, n+1. \quad (16)$$

Then $\hat{C}_n^{\lambda}$ satisfies

$$\mathbb{P}\big(Y_{n+1} \in \hat{C}_n^{\lambda}(X_{n+1})\big) \geq 1 - \alpha. \quad (17)$$

Computation of the weights in (16) is now even more difficult than (4) in the weighted exchangeable setting (even more difficult than a matrix permanent, since λ could in principle depend in a complicated way on the order of its inputs). That said, the above theorem still produces a conformal set (15) with the very general guarantee (17), which is interesting. This may be useful (and computable) in certain structured-X settings, for example, where the sequence $X_1, \ldots, X_{n+1}$ has some kind of Markov structure.

2 Custom-weighted conformal prediction

We next cover a custom-weight conformal prediction method, due to Barber et al. (2022). In comparison to the likelihood-weighted method in the previous section, the weights considered in the current section will be *fixed* (not a function of the data) but *arbitrary*. The theory, as we'll see, is also quite different; in a sense, it is more general in scope. We'll only cover this at a relatively high level (no proof details).

To state the main theorems, we'll need introduce a few additional pieces of notation. As before, let $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n+1$ be data points (with the last one $Z_{n+1} = (X_{n+1}, Y_{n+1})$ serving as the test point), and V a score function. The additional notation is as follows.

- Denote by $Z = (Z_1, \ldots, Z_{n+1})$ the data vector (an ordered sequence).
- Denote by Z^i the data vector after swapping components i and $n+1$.
- Denote by $R(Z)$ the score vector, with components $R(Z)_j = V(Z_j; Z)$.
- Denote by $R(Z^i)$ the score vector had the data vector been Z^i, with components $R(Z^i)_j = V(Z_j^i; Z^i)$.

2.1 Custom weights, symmetric score function

Note that in the current notation, symmetry of V in its last $n+1$ arguments—which is the typical assumption for the conformity score function—implies that we have $V(Z_j; Z) = V(Z_j; Z_\sigma)$, for any permutation σ, where $Z_\sigma = (Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)})$. Thus under this symmetry condition, we may write $R(Z^i) = R(Z)^i$.

Now we can state the first main result, for custom-weighted conformal prediction.

Theorem 3. *Let $w_i \in [0,1]$, $i = 1,\ldots,n$ be fixed and arbitrary weights, and define*

$$\tilde{w}_i = \frac{w_i}{w_1 + \cdots + w_n + 1}, \; i = 1,\ldots,n, \quad \text{and} \quad \tilde{w}_{n+1} = \frac{1}{w_1 + \cdots + w_n + 1}. \tag{18}$$

Define a weighted conformal set (based on the first n samples) at a point $x \in \mathcal{X}$, with nominal error level $\alpha \in (0,1)$, by

$$\hat{C}_n^w(x) = \left\{ y : R_{n+1}^{(x,y)} \leq \text{Quantile}\Big(1-\alpha; \sum_{i=1}^n \tilde{w}_i \, \delta_{R_i^{(x,y)}} + \tilde{w}_{n+1} \, \delta_\infty\Big) \right\}, \tag{19}$$

where $R_i^{(x,y)}$, $i = 1,\ldots,n+1$ are conformity scores as in (7), for an arbitrary score function V that is symmetric in its last $n+1$ arguments. Then with no *assumptions on the joint distribution of $Z_i = (X_i, Y_i)$, $i = 1,\ldots,n+1$, the set $\hat{C}_n^w$ satisfies*

$$\mathbb{P}\big(Y_{n+1} \in \hat{C}_n^w(X_{n+1})\big) \geq 1 - \alpha - \sum_{i=1}^n \tilde{w}_i \cdot \text{TV}(R(Z), R(Z^i)), \tag{20}$$

where $\text{TV}(A,B)$ is the total variation (TV) distance between the distributions of random variables A, B.

Observe that the normalization step in (18) assigns a unit weight to the test point, and then renormalizes (so that the new weights have unit sum). As the initial weights were all between 0 and 1, this means $\tilde{w}_{n+1} \geq \tilde{w}_i$, for all $i = 1,\ldots,n$. The set in (19) reduces to the (unweighted) conformal prediction method set when $w_1 = \cdots = w_n = 1$.

We can interpret the result in (20) as follows. If the distribution of the test data point Z_{n+1} drifts from that of the training data, but we are able identify a priori which training data points will be most *representative* of the test distribution, then we can upweight these points and downweight the others. This would result in a small coverage gap, where

$$\text{coverage gap} = \sum_{i=1}^n \tilde{w}_i \cdot \text{TV}(R(Z), R(Z^i)) \tag{21}$$

(since we have small weights multiplying large TV distances, and large weights multiplying small TV distances.) Of course, choosing a good weights scheme—identifying a priori which training points are representative of the test data distribution—is an important problem unto itself. In certain structured data settings, such as problems with time series or spatial structure, progress towards general methodology for crafting weights seems more tangible than others. Figure 2 gives an example using an exponentially decaying weight scheme in a time series problem, from Barber et al. (2022).

We now make several further remarks.

Split version. As before, the split conformal version of the above result can be viewed as a special case where the score function relies on a point predictor fit on an external data set. For example, if we take it to be $V(x,y) = |y - \mu_0(x)|$, where μ_0 has been fit on a data set Z_0 then (19) simplifies to

$$\hat{C}_n^w(x) = \mu_0(x) \pm \text{Quantile}\Big(1-\alpha; \sum_{i=1}^n \tilde{w}_i \, \delta_{|Y_i - \mu_0(X_i)|} + \tilde{w}_{n+1} \, \delta_\infty\Big),$$

and (20) becomes

$$\mathbb{P}\big(Y_{n+1} \in \hat{C}_n^w(X_{n+1}) \,\big|\, Z_0\big) \geq 1 - \alpha - \sum_{i=1}^n \tilde{w}_i \cdot \text{TV}\big(R(Z), R(Z^i) \,\big|\, Z_0\big),$$

where $\text{TV}(A, B \,|\, C)$ is the TV distance between the conditional distributions of $A|C$ and $B|C$. If Z_0 is independent of Z, then the coverage gap is just a before, in (21).

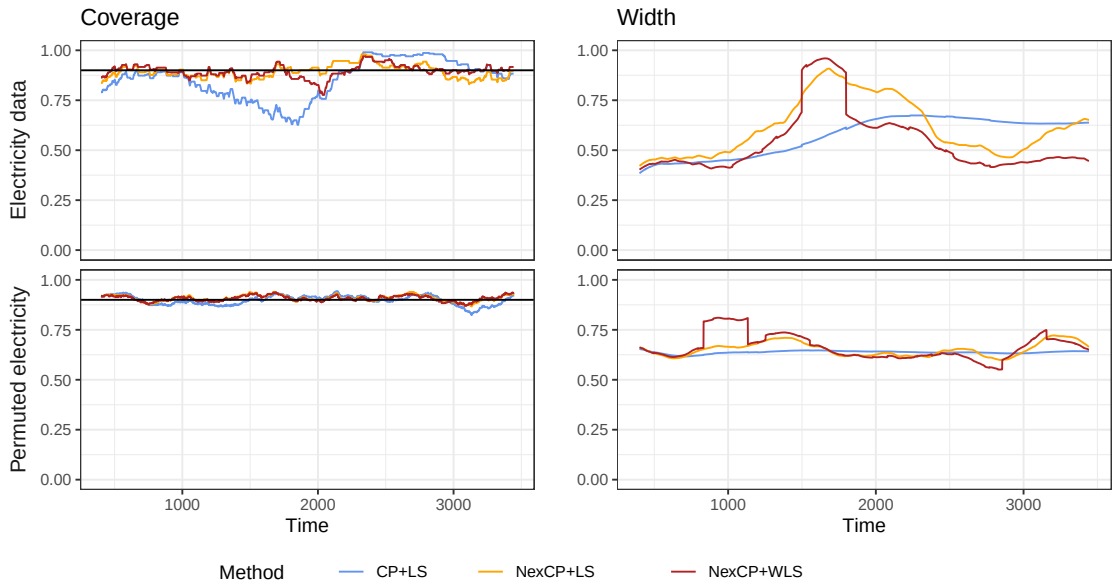

Figure 2: *Experiments for split conformal prediction under distribution drift. Top row: coverage and width in a time series problem, for the usual conformal prediction (CP), weighted conformal with exponentially decaying weights and least squares as the prediction algorithm (NexCP+LS), and the same method but now with weighted least squares as the prediction algorithm (NexCP+WLS). Bottom row: the same metrics and methods on a permuted version of the time series data set. Credit: Barber et al. (2022).*

The i.i.d. setting. When $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n+1$ are i.i.d. (or more generally, exchangeable), we are back to the traditional setting for conformal prediction. In this case, there is no slack in (20), since exchangeability implies $R(Z) \stackrel{d}{=} R(Z^i)$ and thus $\mathrm{TV}(R(Z), R(Z^i)) = 0$ for each i, and hence (20) collapses to an exact $1-\alpha$ coverage guarantee. This not only reproduces the standard result for ordinary conformal, when we take $w_1 = \cdots = w_n = 1$, but also shows us something new: in the i.i.d. (or exchangeable) setting, we can use arbitrary weights and still get exact coverage with weighted conformal prediction.

Coverage gap bounds. It is worth noting a few upper bounds for the coverage gap in (21). First,

$$\text{coverage gap} \leq \sum_{i=1}^{n} \tilde{w}_i \cdot \mathrm{TV}(Z, Z^i),$$

since the TV distance between $f(A), f(B)$ is always less than that between A, B. The bound in the above display is easier to interpret, but can also be much larger than that in (21). (For example, think about the case of a high-dimensional feature space, and a score function that ignores all but a few relevant features for prediction.) Second, if Z_i, $i = 1, \ldots, n+1$ are independent (but not identically distributed), then it can be shown that

$$\text{coverage gap} \leq 2 \sum_{i=1}^{n} \tilde{w}_i \cdot \mathrm{TV}(Z_i, Z_{n+1}).$$

This lends more direct intuition to the idea that we want to assign large weights to training points whose distributions are representative of the test distribution.

CDF form.* The analogous CDF form of the conformal set in (19) is as follows:

$$\hat{C}_n^w(x) = \left\{ y : \sum_{i=1}^{n+1} \tilde{w}_i \, 1\Big\{ R_i^{(x,y)} < R_{n+1}^{(x,y)} \Big\} \leq \lceil 1-\alpha \rceil_w \right\}, \tag{22}$$

where $\lceil 1-\alpha \rceil_w = \min\{\tau \in \text{range}(\hat{F}_n^w) : \tau \geq \beta\}$ and $\hat{F}_n^w$ denotes the (random) CDF of the discrete distribution $\sum_{i=1}^n \tilde{w}_i \, \delta_{R_i^{(x,y)}} + \tilde{w}_{n+1} \, \delta_\infty$. We note that, as before (for the likelihood-weighted case), the adjustment of the probability level needed here is random.

Auxiliary randomization.* We can randomize the CDF form in (22) as follows:

$$\hat{C}_n^{w,*}(x) = \bigg\{ y : \sum_{i=1}^n \tilde{w}_i \, 1\Big\{ R_i^{(x,y)} < R_{n+1}^{(x,y)} \Big\} + U \sum_{i=1}^{n+1} \tilde{w}_i \, 1\Big\{ R_i^{(x,y)} = R_{n+1}^{(x,y)} \Big\} \leq 1-\alpha \bigg\},$$

where $U \sim \text{Unif}(0,1)$, independent of everything else. This is fairly simple and intuitive—it is free of any level adjustments needed in the unrandomized CDF-based set in (22). The set $\hat{C}_n^{w,*}$ satisfies the same guarantee as in (20). It is not clear whether we can say something sharper, but practically, it it gets rid of the discretization error hidden in the inequality (20) by using randomization.

We can also randomize the quantile form in (19) in order to obtain exact coverage (similar to what we did in the likelihood-weighted case) but we omit the details.

2.2 Custom weights, nonsymmetric score function

When the score function is nonsymmetric in its last $n+1$ elements—for example, $V(Z_i; Z)$ when denotes a score assigned to Z_i is based on a model trained on Z that used the *order* of the points in Z, such as an autoregressive model for forecasting in a time series problem—then it is still reasonable to ask what guarantees a weighted conformal prediction set would have. However, this lies outside of the scope of the previous result given in Theorem 3.

It turns out that we can accommodate a completely arbitrary (possibly nonsymmetric) score function V, by injecting a random swap into the construction of the conformal prediction set. This is the second main result, for custom-weighted conformal prediction.

Theorem 4. *Let $w_i \in [0,1]$, $i = 1, \ldots, n$ be fixed and arbitrary weights, and define $\tilde{w}_i$, $i = 1, \ldots, n+1$ as in (18). Define a weighted conformal set (based on the first n samples) at a point $x \in \mathcal{X}$, with nominal error level $\alpha \in (0,1)$, by*

$$\hat{C}_n^w(x) = \bigg\{ y : R_{n+1}^{(x,y),K} \leq \text{Quantile}\Big(1-\alpha; \sum_{i=1}^n \tilde{w}_i \, \delta_{R_i^{(x,y),K}} + \tilde{w}_{n+1} \, \delta_\infty \Big) \bigg\}, \tag{23}$$

where $R_i^{(x,y),K}$, $i = 1, \ldots, n+1$ are conformity scores defined as follows. Let $Z^{(x,y)} = (Z_1, \ldots, Z_n, (x,y))$, and let $(Z^{(x,y)})^K$ denote this vector after swapping components K and $n+1$, where

$$K \sim \sum_{i=1}^{n+1} \tilde{w}_i \, \delta_i,$$

independent of everything else. Then

$$\begin{aligned} R_i^{(x,y),K} &= V\Big((X_i, Y_i); \, (Z^{(x,y)})^K \Big), \quad i = 1, \ldots, n, \\ R_{n+1}^{(x,y),K} &= V\Big((x,y); \, (Z^{(x,y)})^K \Big). \end{aligned} \tag{24}$$

With no *assumptions on V and* no *assumptions on the joint distribution of $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n+1$, the set $\hat{C}_n^w$ satisfies the exact same guarantee as in (20).*

To reiterate, the key difference here is that the scores (24) used in the construction of the conformal set (23) are computed on a data vector in which we swap components K and $n+1$, where K follows a multinomial distribution, with probabilities $\tilde{w}_i$, $i = 1, \ldots, n+1$. To be clear, in the case that the scores are the result of fitting a particular prediction algorithm to the data vector, this swap will alter the fitted model returned by the prediction algorithm. The extent to which this matters undoubtedly depends on the prediction algorithm. In many practical situations, we would not expect this to have a large effect, and we

note generally that applying this swap is much gentler than applying a random permutation to the data vector before computing the scores (which would effectively render V symmetric).

Looking back at Figure 2, we can see this method being applied with weighted least squares as the prediction algorithm (results in red), where the weights themselves decay into the past. This is a nonsymmetric algorithm because the order of the data matters (observation weights get assigned accordingly, for the regression). Compared to unweighted least squares (in yellow), we see comparable coverage and generally shorter prediction intervals.

3 Adaptive conformal inference

Lastly we cover a conformal-like method, for sequential prediction problems, due to Gibbs and Candès (2021). In a way, this is a signifcant departure from the methods we've seen thus far, since the base idea isn't specific to conformal prediction at all. Its core guarantee is quite simple and at the same time quite strong. We'll only cover some of the core properties of this method, skipping more advanced ones.

Assume we are given an infinite sequence of observations $(X_t, Y_t) \in \mathcal{X} \times \mathcal{Y}$, $t = 1, 2, 3, \ldots$, which we can think of as being indexed by time. Assume that at each time t, we have a method for computing a prediction set $C_t^\beta \subseteq \mathcal{Y}$ for Y_t, at any nominal level $\beta \in \mathbb{R}$. The set C_t^β could come from some version of conformal prediction run on the past data (X_s, Y_s), $s < t$, or really, any other method. All that we require is that the method for producing sets *saturates* at any level below 0 or above 1, meaning

$$C_t^\beta = \emptyset \text{ for } \beta \leq 0, \quad \text{and} \quad C_t^\beta = \mathcal{Y} \text{ for } \beta \geq 1. \tag{25}$$

Adaptive conformal inference (ACI)[1] is an algorithm for adjusting the working level $1 - \alpha_t$ of the prediction sets over time $t = 1, 2, 3, \ldots$ so as to try to maintain a realized coverage as close to $1 - \alpha$ as possible, where $\alpha \in (0, 1)$ is some prespecified error tolerance. It is very simple: it initializes $\alpha_0 = \alpha$, and performs updates according to

$$\alpha_{t+1} = \alpha_t - \eta(\mathrm{err}_t - \alpha), \quad t = 0, 1, 2, \ldots, \tag{26}$$

where $\mathrm{err}_t = 1\{Y_t \notin C_t^{1-\alpha_t}\}$ and $\eta > 0$ is a step size. These updates are highly intuitive: if we cover, then we increase the working error level by $\eta\alpha$ (make future prediction sets smaller), and if we miscover, then we decrease the working error level by $\eta(1 - \alpha)$ (make future prediction sets larger).

Boundedness of ACI iterates. The following is a simple but important realization: the ACI iterates in (26) are always uniformly bounded, because, as the proof of the next lemma shows, they have a kind of self-correcting property.

Lemma 2. *The iterates from ACI (26) satisfy $\alpha_t \in [-\eta, 1 + \eta]$, for any $t = 1, 2, 3, \ldots$.*

Proof. Denote $\ell = \inf_{t \geq 1} \alpha_t$ and assume $\ell < -\eta$. Fix any small $\epsilon > 0$ such that $\ell + \epsilon < -\eta$ and $\epsilon < \eta\alpha$. Then there exists some $t \geq 2$ such that $\alpha_t \leq \ell + \epsilon < -\eta$. The update (26) at $t - 1$ tells us that

$$\begin{aligned} \alpha_t &= \alpha_{t-1} - \eta(\mathrm{err}_{t-1} - \alpha) \\ &= \begin{cases} \alpha_{t-1} + \eta\alpha & \text{if } \mathrm{err}_{t-1} = 0 \\ \alpha_{t-1} - \eta(1 - \alpha) & \text{if } \mathrm{err}_{t-1} = 1. \end{cases} \end{aligned}$$

We will be able to derive a contradiction in each of the above cases.

- In the first case α_{t-1} must have been smaller than α_t, but this is impossible, since it would be less than the infimum: $\alpha_{t-1} \leq \ell + \epsilon - \eta\alpha < \ell = \inf_{t \geq 1} \alpha_t$, which is a contradiction.
- In the second case α_{t-1} must have been larger than α_t, but this is impossible, since it would be less than zero, leading to coverage: $\alpha_{t-1} < -\eta + \eta(1 - \alpha) < 0$, so $\mathrm{err}_{t-1} = 0$, which is again a contradiction.

This proves that $\inf_{t \geq 1} \alpha_t \geq -\eta$. The proof that $\sup_{t \geq 1} \alpha_t \leq 1 + \eta$ is similar. □

[1] The method was proposed and named by Gibbs and Candès (2021) in the context of using conformal prediction as the method used to compute prediction sets C_t^β, but as noted, any method for prediction sets will work so long as it saturates in the sense of (25).

Asymptotic coverage of ACI. This boundedness property leads to a simple-to-prove, yet profound fact about ACI: it achieves asymptotic coverage $1-\alpha$, always, meaning, it achieves this no matter the sequence (X_t, Y_t), $t = 1, 2, 3, \ldots$ (even if this sequence were chosen adversarially).

Theorem 5. *For any $t_0 \geq 0$ and $T \geq 1$, the errors from the ACI iterates (26) satisfy*

$$\left| \frac{1}{T} \sum_{t=t_0+1}^{t_0+T} \mathrm{err}_t - \alpha \right| \leq \frac{1+2\eta}{T\eta}. \tag{27}$$

In particular, this implies

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t = \alpha. \tag{28}$$

Proof. Without a loss of generality we prove the result for $t_0 = 0$. Let $s_t = \mathrm{err}_t - \alpha$, and write the update in (26) as $\alpha_t - \eta s_t$. Then

$$\frac{1}{T} \sum_{t=1}^{T} s_t = \frac{\alpha_{T+1} - \alpha_1}{T\eta}.$$

But $\alpha_{T+1} \leq 1+\eta$ and $\alpha_1 \geq -\eta$ by Lemma 2, which proves the desired result. □

To reiterate, Theorem 5 puts no constraints on the data generating distribution (or even assumes that the data is random). It delivers a coverage guarantee over any interval $[t_0+1, t_0+T]$ of $\approx 1-\alpha \pm \frac{1}{T}$, as seen in (27). This of translates into coverage exactly $1-\alpha$ over all time, as seen in (28). Figure 3 gives an example of ACI on financial time series data, from Gibbs and Candès (2021).

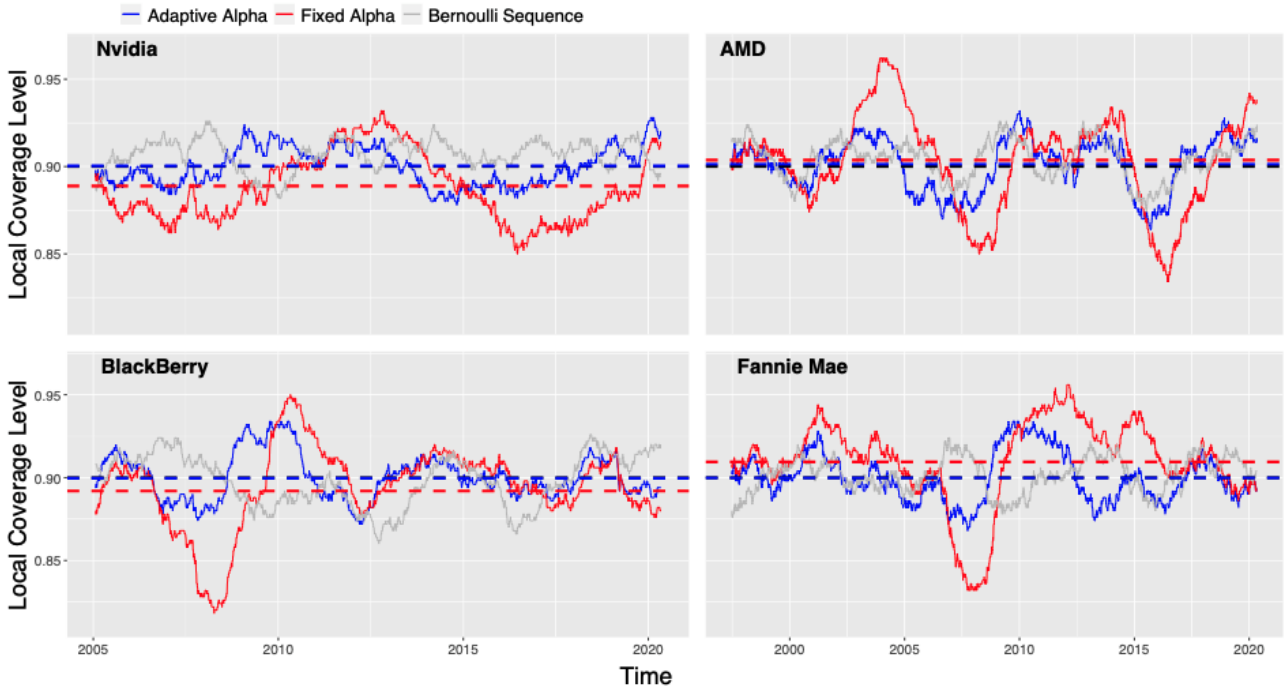

Figure 3: *Experiments for adaptive conformal inference on predicting stock market volatility. The nominal coverage level is 0.9. Shown are the results of local coverage (averaged over a trailing window of 500 days) when using ACI (in blue), versus a fixed level (in red). The coverage fluctuations of ACI are not too dissimilar to the trailing average of i.i.d. Bernoulli coin flips (in gray) with success probability 0.9. Credit: Gibbs and Candès (2021).*

ACI as online gradient descent. Defining $s_t = \mathrm{err}_t - \alpha$, we can show that $s_t \in \partial f_t(\alpha_t)$ for a particular convex loss function f_t, which will make (26) an instance of online gradient descent (technically, the online subgradient method) applied to the convex problem

$$\underset{a}{\text{minimize}} \quad \sum_{t=1}^{T} f_t(a),$$

for some (arbitrarily large) time horizon T. In particular, let

$$\beta_t = \sup\{\beta : Y_t \in C_t^{1-\beta}\},$$

and

$$\begin{aligned} f_t(a) &= \phi_{1-\alpha}(1 - \beta_t - (1-a)) \\ &= \phi_{1-\alpha}(a - \beta_t), \end{aligned}$$

where ϕ_τ denotes the tilted ℓ_1 loss at quantile level τ, that is, $\phi_\tau(x) = \tau|x|$ for $x \geq 0$ and $(1-\tau)|x|$ for $x < 0$. Then a straightforward calculation shows

$$\partial f_t(a) = \begin{cases} \{1-\alpha\} & \text{if } a > \beta_t \\ [-\alpha, 1-\alpha] & \text{if } a = \beta_t \\ \{-\alpha\} & \text{if } a < \beta_t. \end{cases}$$

Furthermore, note that by definition of β_t, we have $a > \beta_t \iff Y_t \notin C_t^{1-a}$. Thus it is clear that

$$s_t = \begin{cases} 1-\alpha & \text{if } \mathrm{err}_t = 1 \\ -\alpha & \text{if } \mathrm{err}_t = 0 \end{cases} \in \partial f_t(\alpha_t).$$

References

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. arXiv: 2202.13415, 2022.

Emmanuel J. Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society: Series B*, 85(1):24–45, 2023.

Clara Fannjiang, Stephen Bates, Anastasios N. Angelopoulos, Jennifer Listgarten, and Michael I. Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, 2021.

Lihua Lei and Emmanuel J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B*, 83(5):911–938, 2021.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2021.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, 2019.