Online conformal prediction with decaying step sizes

Anastasios N. Angelopoulos [*1] **Rina Foygel Barber** [*2] **Stephen Bates** [*3]

Abstract

We introduce a method for online conformal prediction with decaying step sizes. Like previous methods, ours possesses a retrospective guarantee of coverage for arbitrary sequences. However, unlike previous methods, we can simultaneously estimate a population quantile when it exists. Our theory and experiments indicate substantially improved practical properties: in particular, when the distribution is stable, the coverage is close to the desired level *for every time point*, not just on average over the observed sequence.

1. Introduction

We study the problem of online uncertainty quantification, such as that encountered in time-series forecasting. Our goal is to produce a *prediction set* at each time, based on all previous information, that contain the true label with a specified coverage probability. Such prediction sets are useful to the point of being requirements in many sequential problems, including medicine (Robinson, 1978), robotics (Lindemann et al., 2023), finance (Mykland, 2003), and epidemiology (Cramer et al., 2022). Given this broad utility, it comes as no surprise that prediction sets have been studied for approximately one hundred years (and possibly more; see Section 1.1 of Tian et al. (2022)).

Formally, consider a sequence of data points $(X_t, Y_t) \in \mathcal{X} \times \mathcal{Y}$, for $t = 1, 2, \ldots$. At each time t, we observe X_t and seek to cover Y_t with a set $\mathcal{C}_t(X_t)$, which depends on a *base model* trained on all past data (as well as the current feature X_t). After predicting, we observe Y_t, and the next time-step ensues. Note that we have not made any assumptions yet about the data points and their dependencies.

This paper introduces a method for constructing the prediction sets $\mathcal{C}_t$ that has **simultaneous best-case and worst-case guarantees**—that is, a "best of both worlds" property. We will describe the method shortly in Section 1.1. Broadly speaking, the method can gracefully handle both arbitrary adversarial sequences data points and also independent and identically distributed (I.I.D.) sequences. In the former case, our method will remain robust, ensuring that the historical fraction of miscovered labels converges to the desired error rate, $\alpha \in (0, 1)$. In the latter case, our method will converge, eventually producing the optimal prediction sets. We summarize our results below:

1. **Worst-case guarantee (Theorem 1):** When the data points are arbitrary, our algorithm achieves

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{Y_t \in \mathcal{C}_t(X_t)} \in \left(1 - \alpha \pm \frac{C}{T^{1/2-\epsilon}}\right), \qquad (1)$$

   for a constant C and any fixed $\epsilon > 0$. We call this a *long-run coverage* guarantee.

2. **Best-case guarantee (Theorem 3):** When the data points are I.I.D., our algorithm achieves

$$\lim_{T\to\infty} \mathbb{P}\left(Y_T \in \mathcal{C}_T(X_T)\right) \to 1 - \alpha. \qquad (2)$$

   We call this a *convergence guarantee*.

Our algorithm is the first to satisfy both guarantees simultaneously. Moreover, the decaying step size yields more stable behavior than prior methods, as we will see in experiments. See Section 1.2 for a discussion of the relationship with other methods, such as those of Gibbs & Candes (2021), Angelopoulos et al. (2023), and Xu & Xie (2021).

1.1. Method and Setup

We now describe our prediction set construction. Borrowing from conformal prediction, consider a bounded conformal score function $s_t : \mathcal{X} \times \mathcal{Y} \to [0, B]$, at each time t. This score $s_t = s_t(X_t, Y_t)$ is large when the predictions of the

[*]Equal contribution [1]Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley CA USA [2]Department of Statistics, University of Chicago, Chicago IL USA [3]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge MA USA. Correspondence to: Anastasios N. Angelopoulos <angelopoulos@berkeley.edu>, Rina Foygel Barber <rina@uchicago.edu>, Stephen Bates <stephenbates@mit.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

base model disagree greatly with the observed label; an example would be the residual score, $s_t(x, y) = |y - \hat{f}_t(x)|$, for a model $\hat{f}_t : \mathcal{X} \to \mathbb{R}$ trained online. This concept is standard in conformal prediction (Vovk et al., 2005), and we refer the reader to Angelopoulos & Bates (2023) for a recent overview. Given this score function, define

$$\mathcal{C}_t(x) = \{y \in \mathcal{Y} : s_t(x, y) \leq q_t\}, \tag{3}$$

where the threshold q_t is updated with the rule

$$q_{t+1} = q_t + \eta_t(\mathbb{1}_{Y_t \notin \mathcal{C}_t(X_t)} - \alpha). \tag{4}$$

In particular, if we fail to cover Y_t at time t, then the threshold increases to make the procedure slightly more conservative at the next time step (and vice versa).

Familiar readers will notice the similarity of the update step (4) to that of Gibbs & Candes (2021); Bhatnagar et al. (2023); Feldman et al. (2023); Angelopoulos et al. (2023), the main difference being that here, η_t can change over time—later on we will see that $\eta_t \propto t^{-1/2-\epsilon}$, for some small $\epsilon \in (0, 1/2)$, leads to guarantees (1) and (2) as described above. We remark also that the update step for q_t can be interpreted as an online (sub)gradient descent algorithm on the quantile loss $\rho_{1-\alpha}(t) = (1-\alpha)\max\{t, 0\} + \alpha \max\{-t, 0\}$ (Koenker & Bassett Jr, 1978), i.e., we can equivalently write the update step (4) as

$$q_{t+1} = q_t - \eta_t \nabla \rho_{1-\alpha}(s_t - q_t).$$

In this work, we will consider two different settings:

Setting 1 (Adversarial setting)**.** *We say that we are in the adversarial setting if we allow* $(X_1, Y_1), (X_2, Y_2), \ldots$ *to be an arbitrary sequence of elements in* $\mathcal{X} \times \mathcal{Y}$*, and* $s_1, s_2, \ldots$ *to be an arbitrary sequence of functions from* $\mathcal{X} \times \mathcal{Y}$ *to* $[0, B]$*.*

Setting 2 (I.I.D. setting)**.** *We say that we are in the I.I.D. setting if we require that* $(X_t, Y_t) \overset{\text{iid}}{\sim} P$ *for some distribution* P*, and require that the choice of the function* $s_t : \mathcal{X} \times \mathcal{Y} \to [0, B]$ *depends only on* $\{(X_r, Y_r)\}_{r<t}$*, for each* t *(i.e., the model is trained online).*

Of course, any result proved for Setting 1 will hold for Setting 2 as well. We remark that Setting 2 can be relaxed to allow for randomness in the choice of the score functions s_t—our results for the I.I.D. setting will hold as long as the function s_t is chosen independently of $\{(X_r, Y_r)\}_{r \geq t}$.

Our method, like all conformal methods, has coverage guarantees that hold for *any* underlying model and data stream. Still, the quality of the output (e.g., the size of the prediction sets) does critically depend on the quality of the underlying model. This general interplay between conformal methods and models is discussed throughout the conformal literature (e.g., Vovk et al., 2005; Angelopoulos & Bates, 2023).

1.2. Related work

We begin by reviewing the most closely related literature. Set constructions of the form in (3), which "invert" the score function, are commonplace in conformal prediction (Vovk et al., 2005), with q_t chosen as a sample quantile of the previous conformal scores. However, the exchangeability-based arguments of the standard conformal framework cannot give any guarantees in Setting 1. The idea to set q_t via online gradient descent with a *fixed* step size appears first in Gibbs & Candes (2021), which introduced online conformal prediction in the adversarial setting. The version we present here builds also on the work of Bhatnagar et al. (2023), Feldman et al. (2023), and Angelopoulos et al. (2023); in particular, Angelopoulos et al. (2023) call the update in (4) the "quantile tracker". These papers all have long-run coverage guarantees in Setting 1, but do not have convergence guarantees in Setting 2.

Subsequent work to these has explored time-varying step sizes that respond to distribution shifts, primarily for the purpose of giving other notions of validity, such as regret analyses (Gibbs & Candès, 2022; Zaffran et al., 2022; Bastani et al., 2022; Noarov et al., 2023; Bhatnagar et al., 2023). From an algorithmic perspective, these methods depart significantly from the update in (4), generally by incorporating techniques from online learning—such as strongly adaptive online learning (Daniely et al., 2015), adaptive regret (Gradu et al., 2023), and adaptive aggregation of experts (Cesa-Bianchi & Lugosi, 2006). To summarize, the long-run coverage and regret bounds in these papers apply to substantially different, usually more complicated algorithms than the simple expression we have in (4). We remark that "best of both worlds" guarantees appear in the online learning literature (e.g., Bubeck & Slivkins, 2012; Koolen et al., 2016; Zimmert & Seldin, 2021; Jin et al., 2021; Chen et al., 2023; Dann et al., 2023), where the aim is to find a single algorithm whose *regret* is optimal both in a stochastic setting (i.e., data sampled from a distribution) and in an adversarial setting. A crucial difference, however, is that our paper's guarantees are concerned with inference and predictive coverage, rather than with estimation or regret.

Farther afield from our work, there have been several other explorations of conformal prediction in time-series, but these are quite different. For example, the works of Barber et al. (2022) and Chernozhukov et al. (2018) provide conformal-type procedures with coverage guarantees under certain relaxations of exchangeability; both can provide marginal coverage in Setting 2, but cannot give any guarantees in Setting 1. Xu & Xie (2021; 2023) study the behavior of conformal methods under classical nonparametric assumptions such as model consistency and distributional smoothness for its validity, and thus cannot give

distribution-free guarantees in Settings 1 or 2. Lin et al. (2022) studies the problem of cross-sectional coverage for multiple exchangeable time-series. The online conformal prediction setup was also considered early on by Vovk (2002) for exchangeable sequences. These works are not directly comparable to ours, the primary point of difference being the adversarial guarantee we can provide in Setting 1.

Finally, traditional solutions to the prediction set problem have historically relied on Bayesian modeling (e.g., Foreman-Mackey et al., 2017) or distributional assumptions such as autoregression, smoothness, or ergodicity (e.g., Biau & Patra, 2011). A parallel literature on calibration exists in the adversarial sequence model (e.g., Foster & Vohra, 1998). Our work, like that of Gibbs & Candes (2021), is clearly related to the literatures on both calibration and online convex optimization (Zinkevich, 2003), and we hope these connections will continue to reveal themselves; our work takes online conformal prediction one step closer to online learning by allowing the use of decaying step sizes, which is typical for online gradient descent.

1.3. Our contribution

We provide the first analysis of the online conformal prediction update in (4) with an arbitrary step size. Our analysis gives strong long-run coverage bounds for appropriately decaying step sizes, even in the adversarial setting (Setting 1). We also give a simultaneous convergence guarantee in the I.I.D. setting (Setting 2), showing that the parameter q_t converges to the optimal value q^*. Importantly, this type of convergence does *not* hold with a fixed step size (the case previously analyzed in the online conformal prediction literature). In fact, we show that with a fixed step size, online conformal prediction returns meaningless prediction sets (i.e., either $\emptyset$ or $\mathcal{Y}$) infinitely often. From the theoretical point of view, therefore, our method is the first to provide this type of "best-of-both-worlds" guarantee.

While these theoretical results show an improvement (relative to the fixed-step-size method) in an I.I.D. setting, from the practical perspective we will see that a decaying step size also enables substantially better results and more stable behavior on real time series data, which lies somewhere between the I.I.D. and the adversarial regime.

2. Main results in the adversarial setting

We now present our main results for the adversarial setting, Setting 1, which establish long-run coverage guarantees with no assumptions on the data or the score functions.

2.1. Decreasing step sizes

Our first main result shows that, for a nonincreasing step size sequence, the long-run coverage rate

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{Y_t \in \mathcal{C}_t(X_t)} \tag{5}$$

will converge to the nominal level $1-\alpha$.

Theorem 1. *Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be an arbitrary sequence of data points, and let $s_t : \mathcal{X} \times \mathcal{Y} \to [0, B]$ be arbitrary functions. Let η_t be a positive and nonincreasing sequence of step sizes, and fix an initial threshold $q_1 \in [0, B]$.*

Then online conformal prediction satisfies

$$\left| \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{Y_t \in \mathcal{C}_t(X_t)} - (1-\alpha) \right| \leq \frac{B + \eta_1}{\eta_T T}$$

for all $T \geq 1$.

As a special case, if we choose a constant step size $\eta_t \equiv \eta$ then this result is analogous to Gibbs & Candes (2021, Proposition 4.1). On the other hand, if we choose $\eta_t \propto t^{-a}$ for some $a \in (0, 1)$, then the long-run coverage at time T has error bounded as $\mathcal{O}(\frac{1}{T^{1-a}})$.

2.2. Arbitrary step sizes

As discussed above, if the data appears to be coming from the same distribution then a decaying step size can be advantageous, to stabilize the behavior of the prediction sets over time. However, if we detect a sudden distribution shift and start to lose coverage, we might want to increase the step size η_t to recover coverage more quickly. To accommodate this, the above theorem can be generalized to an arbitrary step size sequence, as follows.

Theorem 2. *Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be an arbitrary sequence of data points, and let $s_t : \mathcal{X} \times \mathcal{Y} \to [0, B]$ be arbitrary functions. Let η_t be an arbitrary positive sequence, and fix an initial threshold $q_1 \in [0, B]$.*

Then online conformal prediction satisfies

$$\left| \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{Y_t \in \mathcal{C}_t(X_t)} - (1-\alpha) \right| \leq \frac{B + \max_{1 \leq t \leq T} \eta_t}{T} \cdot \|\Delta_{1:T}\|_1$$

for all $T \geq 1$, where the sequence Δ is defined with values

$$\Delta_1 = \eta_1^{-1}, \text{ and } \Delta_t = \eta_t^{-1} - \eta_{t-1}^{-1} \text{ for all } t \geq 2.$$

We can see that Theorem 1 is indeed a special case of this more general result, because in the case of a nonincreasing sequence η_t, we have $\max_{1 \leq t \leq T} \eta_t = \eta_1$, and

$$\|\Delta_{1:T}\|_1 = |\eta_1^{-1}| + \sum_{t=2}^{T} |\eta_t^{-1} - \eta_{t-1}^{-1}| = \eta_1^{-1} + \sum_{t=2}^{T} (\eta_t^{-1} - \eta_{t-1}^{-1}) = \eta_T^{-1}.$$

But Theorem 2 can be applied much more broadly. For example, we might allow the step size to decay during long stretches of time when the distribution seems stationary, but then reset to a larger step size whenever we believe the distribution may have shifted. In this case, we can obtain an interpretable error bound from the result of Theorem 2 by observing that $\|\Delta_{1:T}\|_1 \leq \frac{2N_T}{\min_{1\leq t\leq T}\eta_t}$, where $N_T = \sum_{t=2}^{T} \mathbb{1}_{\eta_t > \eta_{t-1}}$ is the number of times we increase the step size. Thus, as long as the step size does not decay too quickly, and the number of "resets" N_T is $o(T)$, the upper bound of Theorem 2 will still be vanishing.

3. Results for I.I.D. data

We now turn to studying the setting of I.I.D. data, Setting 2, where $(X_1, Y_1), (X_2, Y_2), \ldots$ are sampled I.I.D. from some distribution P on $\mathcal{X} \times \mathcal{Y}$. While Theorems 1 and 2 show that the coverage of the procedure converges in a weak sense, as in (5), for *any* realization of the data (or even with a nonrandom sequence of data points), we would also like to understand whether the procedure might satisfy stronger notions of convergence with "nice" data. Will the sequence of prediction intervals converge in a suitable sense? We will see that decaying step size does indeed lead to convergence, whereas a constant step size leads to oscillating behavior.

In order to make our questions precise, we need to introduce one more piece of notation to capture the notion of coverage at a particular time t—the "instantaneous" coverage. Let

$$\mathsf{Coverage}_t(q) = \mathbb{P}_P\left(s_t(X, Y) \leq q \mid s_t\right),$$

where the probability is calculated with respect to a data point $(X, Y) \sim P$ drawn independently of s_t. Then, at time t, the prediction set $\mathcal{C}_t(X_t)$ has coverage level $\mathsf{Coverage}_t(q_t)$, by construction. We will see in our results below that for an appropriately chosen decaying step size, $\mathsf{Coverage}_t(q_t)$ will concentrate around $1-\alpha$ over time, while if we choose a constant step size, then $\mathsf{Coverage}_t(q_t)$ will be highly variable.

3.1. Results with a pretrained score function

To begin, we assume that the score function is pretrained, i.e., that $s_1 = s_2 = \ldots$ are all equal to some fixed function $s : \mathcal{X} \times \mathcal{Y} \to [0, B]$. The reader should interpret this as the case where the underlying model is not updated online (e.g., $s(x, y) = |y - \hat{f}(x)|$ for a pretrained model $\hat{f}$ that is no longer being updated). This simple case is intended only as an illustration of the trends we might see more generally; in Section 3.2 below we will study a more realistic setting, where model training is carried out online as the data is collected.

In this setting, since the score function does not vary with t, we have $\mathsf{Coverage}_t(\cdot) \equiv \mathsf{Coverage}(\cdot)$ where

$$\mathsf{Coverage}(q) = \mathbb{P}_P\left(s(X, Y) \leq q\right),$$

i.e., instantaneous coverage at time t is $\mathsf{Coverage}(q_t)$.

First, we will see that choosing a *constant* step size leads to undesirable behavior: while coverage will hold on average over time (as recorded in Theorem 1 and in the earlier work of Gibbs & Candes (2021)), there will be high variability in $\mathsf{Coverage}(q_t)$ over time—for instance, we may have $\mathcal{C}_t(X_t) = \emptyset$ infinitely often.

Proposition 1. *Let $(X_t, Y_t) \overset{\text{iid}}{\sim} P$ for some distribution P. Suppose also that $s_t \equiv s$ for some fixed function $s : \mathcal{X} \times \mathcal{Y} \to [0, B]$, and that $\eta_t \equiv \eta$ for a positive constant step size $\eta > 0$. Assume also that α is a rational number.*

Then online conformal prediction satisfies

$$\mathsf{Coverage}(q_t) = 0 \text{ for infinitely many } t,$$

and

$$\mathsf{Coverage}(q_t) = 1 \text{ for infinitely many } t,$$

almost surely.

In other words, even in the simplest possible setting of I.I.D. data and a fixed model, we cannot expect convergence of the method if we use a constant step size.

On the other hand, if we choose a sequence of step sizes η_t that decays at an appropriate rate (such as $\eta_t \propto t^{-1/2-\epsilon}$, for some $\epsilon \in (0, 1/2)$, as mentioned earlier) then over time, this highly variable behavior can be avoided. Instead, we will typically see coverage converging to $1-\alpha$ for *each* constructed prediction set $\mathcal{C}_t(X_t)$, i.e., $\mathsf{Coverage}(q_t) \to 1-\alpha$. We will need one more assumption: defining q^* as the $(1-\alpha)$-quantile of $s(X, Y)$, we assume that q^* is unique:

$$\begin{aligned} &\mathsf{Coverage}(q) < 1-\alpha \text{ for all } q < q^*, \\ &\mathsf{Coverage}(q) > 1-\alpha \text{ for all } q > q^*. \end{aligned} \tag{6}$$

Theorem 3. *Let $(X_t, Y_t) \overset{\text{iid}}{\sim} P$ for some distribution P. Suppose also that $s_t \equiv s$ for some fixed function $s : \mathcal{X} \times \mathcal{Y} \to [0, B]$. Assume that η_t is a fixed nonnegative step size sequence satisfying*

$$\sum_{t=1}^{\infty} \eta_t = \infty, \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty. \tag{7}$$

Assume also that q^* is unique as in (6).

Then online conformal prediction satisfies

$$q_t \to q^* \text{ almost surely.}$$

With an additional assumption, this immediately implies convergence of the coverage, $\mathsf{Coverage}(q_t)$:

Corollary 1. *Under the setting and assumptions of Theorem 3, assume also that $s(X,Y)$ has a continuous distribution (under $(X,Y) \sim P$). Then*

$$\mathsf{Coverage}(q_t) \to 1-\alpha \text{ almost surely.}$$

That is, instead of the high variance in coverage incurred by a constant step size (as in Proposition 1), here the coverage converges to the nominal level $1-\alpha$. Finally, with additional assumptions, we can also characterize the *rate* at which the threshold q_t converges to q^*:

Proposition 2. *Under the setting and assumptions of Corollary 1, assume also that the distribution of $s(X,Y)$ (under $(X,Y) \sim P$) has density lower-bounded by γ in the range $[q^*-\delta, q^*+\delta]$, for some $\gamma, \delta > 0$. Take the step size sequence $\eta_t = ct^{-1/2-\epsilon}$, for some $c > 0$ and $\epsilon \in (0, 1/2)$. Then it holds for all $t \geq 1$ that*

$$\mathbb{E}\left[(q_t - q^*)^2\right] \leq bt^{-1/2-\epsilon},$$

where b is a constant that depends only on $B, \gamma, \delta, c, \epsilon$.

3.2. Results with online training of the score function

The result of Theorem 3 above is restricted to a very simple setting, where the score functions are given by $s_t \equiv s$ for some fixed s, i.e., we are using a pretrained model. We now consider the more interesting setting where the model is trained online. Formally, we consider Setting 2 where we allow the score function s_t to depend arbitrarily on the data observed *before time t*, i.e., on $\{(X_r, Y_r)\}_{r<t}$.

First, we will consider a constant step size $\eta_t \equiv \eta$.

Proposition 3. *Let $(X_t, Y_t) \overset{\text{iid}}{\sim} P$ for some distribution P, and assume the score functions $s_t : \mathcal{X} \times \mathcal{Y} \to [0, B]$ are trained online. Let $\eta_t \equiv \eta$ for a positive constant step size $\eta > 0$.*

Then online conformal prediction satisfies

$$\liminf_{t\to\infty} \mathsf{Coverage}_t(q_t) = 0, \quad \limsup_{t\to\infty} \mathsf{Coverage}_t(q_t) = 1$$

almost surely.

This result is analogous to Proposition 1 for the case of a pretrained score function (but with a slightly weaker conclusion due to the more general setting). As before, the conclusion we draw is that a constant step size inevitably leads to high variability in $\mathsf{Coverage}_t(q_t)$.

On the other hand, if we take a decaying step size, Theorem 3 established a convergence result given a pretrained score function. We will now see that similar results hold in for the online setting as long as the model converges in some sense. In many settings, we might expect s_t to converge to some score function s—for example, if our fitted regression functions, $\hat{f}_t$, converge to some "true" model f^*, then $s_t(x,y) = |y - \hat{f}_t(x)|$ converges to $s(x,y) = |y - f^*(x)|$. As before, we let $\mathsf{Coverage}(q) = \mathbb{P}_P\left(s(X,Y) \leq q\right)$, and write q^* to denote the $(1-\alpha)$-quantile of this distribution. We now extend the convergence results of Theorem 3 to this setting.

Theorem 4. *Let $(X_t, Y_t) \overset{\text{iid}}{\sim} P$ for some distribution P, and assume the score functions s_t are trained online. Assume that η_t is a fixed nonnegative step size sequence satisfying* (7)*. Let $s : \mathcal{X} \times \mathcal{Y} \to [0, B]$ be a fixed score function, and assume that q^* is unique as in* (6)*.*

Then online conformal prediction satisfies the following statement almost surely:[1]

$$\text{If } s_t \overset{\text{d}}{\to} s, \text{ then } q_t \to q^*.$$

As in the previous section, an additional assumption implies convergence of the coverage, $\mathsf{Coverage}_t(q_t)$:

Corollary 2. *Under the setting and assumptions of Theorem 4, assume also that $s(X,Y)$ has a continuous distribution (under $(X,Y) \sim P$). Then online conformal prediction satisfies the following statement almost surely:*

$$\text{If } s_t \overset{\text{d}}{\to} s, \text{ then } \mathsf{Coverage}_t(q_t) \to 1-\alpha.$$

To summarize, the results of this section show that the coverage of each prediction set $\mathcal{C}_t(X_t)$, given by $\mathsf{Coverage}_t(q_t)$, will converge even in a setting where the model is being updated in a streaming fashion, as long as the fitted model itself converges over time.

In particular, if we choose $\eta_t \propto t^{-1/2-\epsilon}$ for some $\epsilon \in (0, 1/2)$, then in the adversarial setting the long-run coverage error is bounded as $\mathcal{O}(\frac{1}{T^{1/2-\epsilon}})$ by Theorem 1, while in the I.I.D. setting, Theorem 4 guarantees convergence. In other words, this choice of η_t simultaneously achieves both types of guarantees.

While the results of this section have assumed I.I.D. data, the proof techniques used here can be extended to handle broader settings—for example, a stationary time series,

[1] We use $s_t \overset{\text{d}}{\to} s$ in the sense of convergence in distribution under $(X,Y) \sim P$, while treating the s_t's as fixed. Specifically, we are assuming $\mathsf{Coverage}_t(q) \to \mathsf{Coverage}(q)$, for all $q \in \mathbb{R}$ at which $\mathsf{Coverage}(q)$ is continuous.

where despite dependence we may still expect to see convergence over time. We leave these extensions to future work.

4. Experiments

We include two experiments: an experiment on the Elec2 dataset (Harries et al., 1999) where the data shows significant distribution shift over time, and an experiment on Imagenet (Deng et al., 2009) where the data points are exchangeable.[2]

The experiments are run with two different choices of step size for online conformal: first, a fixed step size ($\eta_t \equiv 0.05$); and second, a decaying step size ($\eta_t = t^{-1/2-\epsilon}$ with $\epsilon = 0.1$). We also compare to an oracle method, where online conformal is run with q^* in place of q_t at each time t, and q^* is chosen to be the value that gives $1-\alpha$ average coverage over the entire sequence $t = 1, \ldots, T$. All methods are run with $\alpha = 0.1$.

4.1. Results

Figures 1 and 2 display the results of the experiment for the Elec2 data and the Imagenet data, respectively. We now discuss our findings.

The thresholds q_t. The first panel of each figure plots the value of the threshold q_t over time t. We can see that the procedure with a fixed step size has significantly larger fluctuations in the quantile value as compared to the decaying step size procedure.

The instantaneous coverage $\mathsf{Coverage}_t(q_t)$. The second panel of each figure plots the value of the instantaneous coverage $\mathsf{Coverage}_t(q_t)$ over time t. For each dataset, since the true data distribution is unknown, we estimate $\mathsf{Coverage}_t(q_t)$ using a holdout set. We observe that $\mathsf{Coverage}_t(q_t)$ is substantially more stable for the decaying step size as compared to fixed step size in both experiments. While $\mathsf{Coverage}_t(q_t)$ concentrates closely around the nominal level $1-\alpha$ for decaying η_t, for fixed η_t the coverage level oscillates and does not converge.

Long-run coverage and rolling coverage. The third panel of each figure plots the value of the long-run coverage, $\frac{1}{t}\sum_{r=1}^{t} \mathbb{1}_{Y_r \in \mathcal{C}_r(X_r)}$, over time t. We see that the long-run coverage converges quickly to $1-\alpha$ for all methods, and we cannot differentiate between them in this plot.

Consequently, in the fourth panel of each figure, we also plot the "rolling" coverage, which computes coverage rate averaged over a rolling window of 1000 time points. We can see that this measure is tighter around $1-\alpha$ for the fixed step size procedure; for the decaying step size procedure, rolling coverage fluctuates more, but is not larger than the fluctuations for the oracle method. At first glance, it might appear that having lower variance in the rolling coverage indicates that the fixed step size procedure is actually performing *better* than decaying step size—but this is not the case. The low variance with fixed $\eta_t \equiv \eta$ is due to *overcorrecting*. For example, if we have several miscoverage events in a row (which can happen by random chance, even with the oracle intervals), then the fixed-step-size method will necessarily return an overly wide interval (e.g., $\mathcal{C}_t(X_t) = \mathbb{R}$) to give certain coverage at the next time step. Thus, the fixed-step-size method ensures low variance in rolling coverage at the cost of extremely high variance in the width and instantaneous coverage of the interval $\mathcal{C}_t(X_t)$ at each time t. This type of overcorrection is undesirable.

4.2. Implementation details for Elec2 Time Series

The Elec2 (Harries et al., 1999) dataset is a time-series of 45312 hourly measurements of electricity demand in New South Wales, Australia. We use even-numbered time points as the time series, and odd-numbered time points as a holdout set for estimating $\mathsf{Coverage}_t(q_t)$. The demand measurements are normalized to lie in the interval $Y_t \in [0,1]$. The covariate vector $X_t = (Y_1, \ldots, Y_{t-1})$ is the sequence of all previous demand values. The forecast $\hat{Y}_t$ is one-day-delayed moving average of Y_t (i.e., at time t, our predicted value $\hat{Y}_t$ is given by the average of observations taken between 24 and 48 hours earlier), and the score is $s_t(X_t, Y_t) = |Y_t - \hat{Y}_t|$.

4.3. Implementation details for Imagenet

The Imagenet (Deng et al., 2009) is a standard computer vision dataset of natural images. We take the 50000 validation images of Imagenet 2012 and treat them as a time series for the purpose of evaluating our methods. Because the validation split of Imagenet is shuffled, this comprises an exchangeable time series. We use 45000 points for the time series, and the remaining 5000 points as a holdout set for estimating $\mathsf{Coverage}_t(q_t)$. As the score function, we use $s_t(X_t, Y_t) = 1 - \max_{y \in \mathcal{Y}} \hat{f}(X_t)_y$ (here $\max_{y \in \mathcal{Y}} \hat{f}(X_t)_y$ is the softmax score of the pretrained ResNet-152 model).

4.4. Additional experiments

As discussed in Section 2.2, in applications where the distribution of the data may drift or may have changepoints, it might be beneficial to allow η_t to increase at times to allow for updates in the learning process. To study this empirically, in the Appendix, we include additional experiments in a broader range of settings—we test over 3000

[2] Code to reproduce these experiments is available at `https://github.com/aangelopoulos/online-conformal-decaying`.

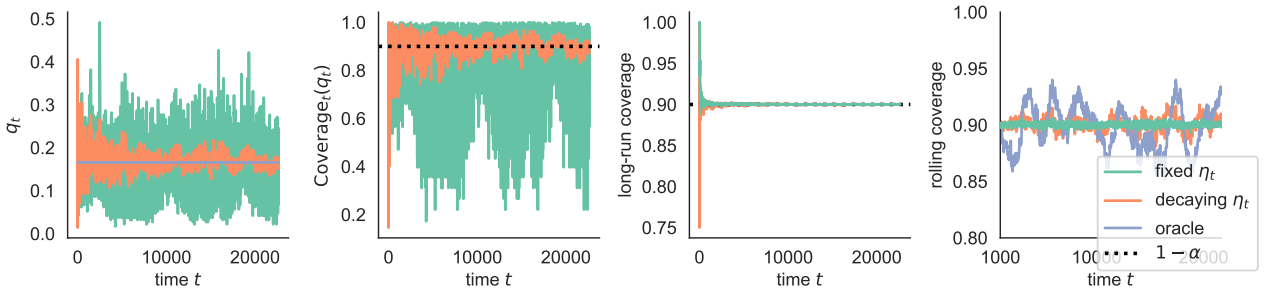

Figure 1. **Elec2 results.** From left to right, the panels display the following (over all times t): first, the value of the threshold q_t; second, the instantaneous coverage $\text{Coverage}_t(q_t)$; third, the long-run coverage $\frac{1}{t}\sum_{r=1}^{t} \mathbb{1}_{Y_r \in \mathcal{C}_r(X_r)}$; and fourth, the rolling coverage, averaged over a rolling window of 1000 time points.

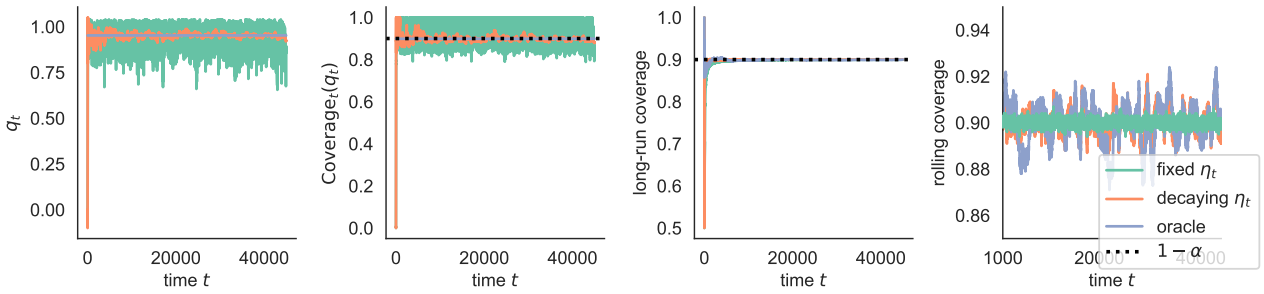

Figure 2. **Imagenet results.** Same details as for Figure 1.

real datasets, and compare the fixed step size method, the decaying step size method, and a "decay+adapt" version of our method where the sequence η_t adapts to trends in the data (decaying if the distribution of the data appears stationary, but increasing if distribution shift is detected).

5. Proofs

In this section, we prove Theorems 1, 2, 3, and 4, and Propositions 1 and 3. All other results are proved in the Appendix.

5.1. Proof of Theorems 1 and 2

First, we need a lemma to verify that the values q_t are uniformly bounded over all t. This result is essentially the same as that in Lemma 4.1 of Gibbs & Candes (2021), except extended to the setting of decaying, rather than constant, step size. The proof is given in the Appendix.

Lemma 1. *Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be an arbitrary sequence of data points, and let $s_t : \mathcal{X} \times \mathcal{Y} \to [0, B]$ be arbitrary functions. Let η_t be an arbitrary nonnegative sequence, and fix an initial threshold $q_1 \in [0, B]$.*

Then online conformal prediction satisfies

$$-\alpha M_{t-1} \leq q_t \leq B + (1-\alpha) M_{t-1} \text{ for all } t \geq 1, \quad (8)$$

where $M_0 = 0$, and $M_t = \max_{1 \leq r \leq t} \eta_r$ for each $t \geq 1$.

We are now ready to prove the theorems. As discussed earlier, Theorem 1 is simply a special case, so we only prove the more general result Theorem 2.

By definition of Δ, we have $\eta_t^{-1} = \sum_{r=1}^{t} \Delta_r$ for all $t \geq 1$. We then calculate

$$\begin{aligned}
&\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{Y_t \in \mathcal{C}_t(X_t)} - (1-\alpha) \right| = \left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{Y_t \notin \mathcal{C}_t(X_t)} - \alpha \right| \\
&= \left| \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{r=1}^{t} \Delta_r \right) \cdot \eta_t \left(\mathbb{1}_{Y_t \notin \mathcal{C}_t(X_t)} - \alpha \right) \right| \\
&= \left| \frac{1}{T} \sum_{r=1}^{T} \Delta_r \left(\sum_{t=r}^{T} \eta_t \left(\mathbb{1}_{Y_t \notin \mathcal{C}_t(X_t)} - \alpha \right) \right) \right| \\
&= \left| \frac{1}{T} \sum_{r=1}^{T} \Delta_r \left(q_{T+1} - q_r \right) \right| \text{ by (4)} \\
&\leq \frac{1}{T} \sum_{r=1}^{T} |\Delta_r| \cdot \max_{1 \leq r \leq T} |q_{T+1} - q_r| \\
&\leq \frac{1}{T} \cdot \|\Delta_{1:T}\|_1 \cdot (B + \max_{1 \leq t \leq T} \eta_t),
\end{aligned}$$

where the last step holds since q_r, q_{T+1} are bounded by Lemma 1.

5.2. Proof of Proposition 3

First we prove that $\limsup_{t \to \infty} \text{Coverage}_t(q_t) = 1$ almost surely. Equivalently, for any fixed $\epsilon > 0$, we need to prove

that $\mathbb{P}\left(\limsup_{t\to\infty} \mathsf{Coverage}_t(q_t) < 1-\epsilon\right) = 0$.

We begin by constructing a useful coupling between the online conformal process, and a sequence of I.I.D. uniform random variables. For each $t \geq 1$, define

$$U_t \sim \begin{cases} \mathsf{Uniform}[0, \mathsf{Coverage}_t(q_t)], & \text{if } Y_t \in \mathcal{C}_t(X_t), \\ \mathsf{Uniform}[\mathsf{Coverage}_t(q_t), 1], & \text{if } Y_t \notin \mathcal{C}_t(X_t), \end{cases}$$

drawn independently for each t after conditioning on all the data, $\{(X_t, Y_t)\}_{t\geq 1}$. Since $\mathbb{P}\left(Y_t \in \mathcal{C}_t(X_t) \mid \{(X_r, Y_r)\}_{r<t}\right) = \mathsf{Coverage}_t(q_t)$ by construction, we can verify that $U_t \overset{\text{iid}}{\sim} \mathsf{Uniform}[0,1]$.

Next fix any integer $N \geq \frac{B+\eta\alpha}{\eta(1-\alpha)}$. Let A_i be the event that

$$U_t > 1-\epsilon \text{ for all } (i-1)N < t \leq iN.$$

Since the U_t's are I.I.D. uniform random variables, we have $\mathbb{P}(A_i) = \epsilon^N$ for each i, and the events A_i are mutually independent. Therefore, by the second Borel–Cantelli lemma, $\mathbb{P}\left(\sum_{i\geq 1} \mathbb{1}_{A_i} = \infty\right) = 1$. Now we claim that

$$\text{If } A_i \text{ occurs then } \max_{(i-1)N < t \leq iN+1} \mathsf{Coverage}_t(q_t) > 1-\epsilon. \tag{9}$$

Suppose that A_i holds and that $\mathsf{Coverage}_t(q_t) \leq 1-\epsilon$ for all t in the range $(i-1)N < t \leq iN$. Then by construction of the U_t's, we have $Y_t \notin \mathcal{C}_t(X_t)$ for all $(i-1)N < t \leq iN$. Therefore by (4),

$$\begin{aligned} q_{iN+1} &= q_{(i-1)N+1} + \sum_{t=(i-1)N+1}^{iN} \eta(\mathbb{1}_{Y_t \notin \mathcal{C}_t(X_t)} - \alpha) \\ &= q_{(i-1)N+1} + N \cdot \eta(1-\alpha) \geq B, \end{aligned}$$

where the last step holds by our choice of N, together with the fact that $q_{(i-1)N+1} \geq -\alpha\eta$ by Lemma 1. But since the score function s_{iN+1} takes values in $[0, B]$ by assumption, we therefore have $\mathsf{Coverage}_{iN+1}(q_{iN+1}) \geq \mathsf{Coverage}_{iN+1}(B) = 1$. Therefore, we have verified the claim (9).

Since A_i occurs for infinitely many i, almost surely, by (9) we therefore have $\limsup_{t\to\infty} \mathsf{Coverage}_t(q_t) \geq 1-\epsilon$, almost surely, as desired. Since $\epsilon > 0$ is arbitrary, this completes the proof that $\limsup_{t\to\infty} \mathsf{Coverage}_t(q_t) = 1$ almost surely.

Finally, a similar argument verifies $\liminf_{t\to\infty} \mathsf{Coverage}(q_t) = 0$ almost surely.

5.3. Proof of Proposition 1

Since $s_t \equiv s$, we have $\mathsf{Coverage}_t(q_t) = \mathsf{Coverage}(q_t)$, for each t. By Proposition 3, $\liminf_{t\to\infty} \mathsf{Coverage}(q_t) = 0$ and $\limsup_{t\to\infty} \mathsf{Coverage}(q_t) = 1$, almost surely. Since we have assumed that α is a rational number, by the definition of the procedure (4), all values q_t must lie on a discrete grid (i.e., if $\alpha = k/K$ for some integers k, K then, for all t, $q_t - q_1$ must be an integer multiple of η/K). Moreover, by Lemma 1, q_t is uniformly bounded above and below for all t, so q_t can only take finitely many values. This implies $\mathsf{Coverage}(q_t)$ also can take only finitely many values, and in particular, this means that if $\liminf_{t\to\infty} \mathsf{Coverage}(q_t) = 0$ (respectively, if $\limsup_{t\to\infty} \mathsf{Coverage}(q_t) = 1$) then $\mathsf{Coverage}(q_t) = 0$ (respectively, $\mathsf{Coverage}(q_t) = 1$) for infinitely many t.

5.4. Proofs of Theorems 3 and 4

We observe that Theorem 3 is simply a special case of Theorem 4 (obtained by taking $s_t \equiv s$ for all t), so we only need to prove Theorem 4.

First, consider the sequence

$$Z_t = \sum_{r=1}^{t} \eta_r(\mathbb{1}_{Y_r \in \mathcal{C}_r(X_r)} - \mathsf{Coverage}_r(q_r)).$$

Define events $\mathcal{E}_Z$, the event that $\lim_{t\to\infty} Z_t$ exists, and $\mathcal{E}_s$, the event that $s_t \overset{\text{d}}{\to} s$. In the Appendix, we will verify that

$$\lim_{t\to\infty} Z_t \text{ exists, almost surely,} \tag{10}$$

i.e., $\mathbb{P}(\mathcal{E}_Z) = 1$, using martingale theory.

To establish the theorem, then, it suffices for us to verify that on the event $\mathcal{E}_Z \cap \mathcal{E}_s$, it holds that $q_t \to q^*$. From this point on, we assume that $\mathcal{E}_Z$ and $\mathcal{E}_s$ both hold.

Fix any $\epsilon > 0$. Since $q \mapsto \mathsf{Coverage}(q)$ is monotone, it can have at most countably infinitely many discontinuities. Without loss of generality, then, we can assume that this map is continuous at $q = q^* - \epsilon/3$ and at $q = q^* + \epsilon/3$ (by taking a smaller value of ϵ if needed).

First, since Z_t converges, we can find some finite time T_1 such that

$$\begin{aligned} &\sup_{t' \geq t \geq T_1} \left| \sum_{r=t}^{t'} \eta_r(\mathbb{1}_{Y_r \in \mathcal{C}_r(X_r)} - \mathsf{Coverage}_r(q_r)) \right| \\ &\qquad = \sup_{t' \geq t \geq T_1} |Z_{t'} - Z_{t-1}| \leq \frac{\epsilon}{3}. \end{aligned} \tag{11}$$

Moreover, since $\sum_t \eta_t^2 < \infty$, we have $\eta_t \to 0$ and so we can find some finite time T_2 such that $\eta_t \leq \frac{\epsilon}{3}$ for all $t \geq T_2$. Furthermore, on $\mathcal{E}_s$, we have $\mathsf{Coverage}_t(q) \to \mathsf{Coverage}(q)$, at each $q = q^* \pm \epsilon/3$. Thus we can find some finite time T_3 and some $\delta > 0$ such that

$$\mathsf{Coverage}_t(q^* - \epsilon/3) \leq 1 - \alpha - \delta \tag{12}$$

for all $t \geq T_3$ (we are using the fact that $\mathsf{Coverage}(q^* - \epsilon/3) < 1 - \alpha$ by (6)). Similarly we can find a finite T_4 and some $\delta' > 0$ such that $\mathsf{Coverage}_t(q^* + \epsilon/3) \geq 1 - \alpha + \delta'$ for all $t \geq T_4$. Let $T = \max\{T_1, T_2, T_3, T_4\}$.

We will now split into cases. If it does not hold that $q_t \in q^* \pm \epsilon$ for all sufficiently large t, then one of the following cases must hold:

- **Case 1a:** $q_t < q^* - \epsilon/3$ for all $t \geq T$.
- **Case 1b:** $q_t > q^* + \epsilon/3$ for all $t \geq T$.
- **Case 2a:** for some $t' \geq t \geq T$, it holds that $q_t \geq q^* - \epsilon/3$ and $q_{t'} < q^* - \epsilon$.
- **Case 2b:** for some $t' \geq t \geq T$, it holds that $q_t \leq q^* + \epsilon/3$ and $q_{t'} > q^* + \epsilon$.

We now verify that each case is impossible.

Case 1a is impossible. We have

$$\begin{aligned}
&q^* - \frac{\epsilon}{3} - q_T \geq \sup_{t>T} q_t - q_T \\
&= \sup_{t>T} \sum_{r=T}^{t-1} \eta_r (\mathbb{1}_{Y_r \notin \mathcal{C}_r(X_r)} - \alpha) \text{ by (4)} \\
&\geq \sup_{t>T} \sum_{r=T}^{t-1} \eta_r \left((1-\alpha) - \mathsf{Coverage}_r(q_r)\right) - \frac{\epsilon}{3} \text{ by (11)} \\
&\geq \sup_{t>T} \left\{ \left[\sum_{r=T}^{t-1} \eta_r \cdot \delta \right] - \frac{\epsilon}{3} \right\},
\end{aligned}$$

where the last step holds since $q_r < q^* - \epsilon/3$ for $r \geq T$, and $q \mapsto \mathsf{Coverage}_r(q)$ is nondecreasing, and so we have

$$\mathsf{Coverage}_r(q_r) \leq \mathsf{Coverage}_r(q^* - \epsilon/3) \leq 1 - \alpha - \delta, \quad (13)$$

by (12). Since $\sum_r \eta_r = \infty$, we therefore have that $q^* - \frac{\epsilon}{3} - q_T \geq \infty$, which is a contradiction.

Case 1b is impossible. This proof is analogous to the proof for Case 1a.

Case 2a is impossible. First, by assumption for this case, we can find a unique time $t'' \geq T$ such that

$$\begin{cases} q_{t''} \geq q^* - \epsilon/3, \\ q_r < q^* - \epsilon/3 \text{ for all } t'' < r < t', \\ q_{t'} < q^* - \epsilon. \end{cases}$$

In other words, t'' is the last time before time t' when the threshold is $\geq q^* - \epsilon/3$. Then we have

$$\begin{aligned}
&-\frac{2\epsilon}{3} > q_{t'} - q_{t''} = \sum_{r=t''}^{t'-1} \eta_r \left(\mathbb{1}_{Y_r \notin \mathcal{C}_r(X_r)} - \alpha\right) \text{ by (4)} \\
&\geq \left[\sum_{r=t''}^{t'-1} \eta_r \left((1-\alpha) - \mathsf{Coverage}_r(q_r)\right) \right] - \frac{\epsilon}{3} \text{ by (11)} \\
&\geq -\eta_{t''} + \left[\sum_{r=t''+1}^{t'-1} \eta_r \left((1-\alpha) - \mathsf{Coverage}_r(q_r)\right) \right] - \frac{\epsilon}{3} \\
&\geq -\eta_{t''} - \frac{\epsilon}{3} \text{ by (13).}
\end{aligned}$$

But since $\eta_{t''} \leq \epsilon/3$ (because $t'' \geq T$), we have therefore reached a contradiction.

Case 2b is impossible. This proof is analogous to the proof for Case 2a.

We have verified that all four cases are impossible. Therefore, $q_t \in q^* \pm \epsilon$ for all sufficiently large t. Since $\epsilon > 0$ is arbitrarily small, this completes the proof.

6. Discussion

Our paper analyzes online conformal prediction that with a decaying step size, enabling simultaneous guarantees of convergence for I.I.D. sequences and long-run coverage for adversarial ones. Moreover, it helps further unify online conformal prediction with online learning and online convex optimization, since decaying step sizes are known to have desirable properties and hence standard for the latter. Of course, the usefulness of the method will rely on choosing score functions that are well suited to the (possibly time-varying) data distribution, and choosing step sizes that decay at an appropriate rate and perhaps adapt to the level of distribution shift—building a better understanding of how to make these choices in practice is crucial for achieving informative and stable prediction intervals. Many additional open questions about extending the methodology to broader settings and understanding connections to other tools remain. In particular, we expect fruitful avenues of future inquiry would be: (1) to extend this analysis to online risk control, as in Feldman et al. (2021); (2) to adapt our analysis of Theorem 3 to deal with stationary or slowly moving time-series which may not be I.I.D. but are slowly varying enough to permit estimation; and (3) to further understand the connection between this family of techniques and the theory of online learning.

Acknowledgements

The authors thank Isaac Gibbs, Ryan Tibshirani, and Margaux Zaffran for helpful feedback on this work. R.F.B. was partially supported by the National Science Foundation via grant DMS-2023109, and by the Office of Naval Research via grant N00014-20-1-2337.

Impact Statement

This paper presents work whose goal is to advance the field of conformal prediction and, more broadly, reliable machine learning. These tools may be used for machine learning methods across a broad range of domains (including applications with societal consequences), but do not relate directly to any specific issues of impact and ethics, and therefore we do not highlight any particular examples here.

References

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

Angelopoulos, A. N., Candes, E., and Tibshirani, R. Conformal PID control for time series prediction. In *Neural Information Processing Systems*, 2023.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *arXiv:2202.13415*, 2022.

Bastani, O., Gupta, V., Jung, C., Noarov, G., Ramalingam, R., and Roth, A. Practical adversarial multivalid conformal prediction. *Advances in Neural Information Processing Systems*, 35:29362–29373, 2022.

Bhatnagar, A., Wang, H., Xiong, C., and Bai, Y. Improved online conformal prediction via strongly adaptive online learning. *arXiv preprint arXiv:2302.07869*, 2023.

Biau, G. and Patra, B. Sequential quantile prediction of time series. *IEEE Transactions on Information Theory*, 57(3):1664–1674, 2011.

Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 42–1. JMLR Workshop and Conference Proceedings, 2012.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Chen, S., Tu, W.-W., Zhao, P., and Zhang, L. Optimistic online mirror descent for bridging stochastic and adversarial online convex optimization. In *International Conference on Machine Learning*, pp. 5002–5035. PMLR, 2023.

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On Learning Theory*, pp. 732–749. PMLR, 2018.

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., Gneiting, T., House, K. H., Huang, Y., et al. Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the united states. *Proceedings of the National Academy of Sciences*, 119(15):e2113561119, 2022.

Daniely, A., Gonen, A., and Shalev-Shwartz, S. Strongly adaptive online learning. In *International Conference on Machine Learning*, pp. 1405–1411. PMLR, 2015.

Dann, C., Wei, C.-Y., and Zimmert, J. Best of both worlds policy optimization. In *International Conference on Machine Learning*, pp. 6968–7008. PMLR, 2023.

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

Feldman, S., Bates, S., and Romano, Y. Improving conditional coverage via orthogonal quantile regression. In *Advances in Neural Information Processing Systems*, 2021.

Feldman, S., Ringel, L., Bates, S., and Romano, Y. Achieving risk control in online learning settings. *Transactions on Machine Learning Research*, 2023.

Foreman-Mackey, D., Agol, E., Ambikasaran, S., and Angus, R. Fast and scalable Gaussian process modeling with applications to astronomical time series. *The Astronomical Journal*, 154(6):220, 2017.

Foster, D. P. and Vohra, R. V. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Gibbs, I. and Candes, E. Adaptive conformal inference under distribution shift. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672. Curran Associates, Inc., 2021.

Gibbs, I. and Candès, E. Conformal inference for online prediction with arbitrary distribution shifts. *arXiv preprint arXiv:2208.08401*, 2022.

Gradu, P., Hazan, E., and Minasyan, E. Adaptive regret for control of time-varying dynamics. In *Learning for Dynamics and Control Conference*, pp. 560–572. PMLR, 2023.

Harries, M., Wales, N. S., et al. Splice-2 comparative evaluation: Electricity pricing. 1999.

Jin, T., Huang, L., and Luo, H. The best of both worlds: stochastic and adversarial episodic MDPs with unknown transition. *Advances in Neural Information Processing Systems*, 34:20491–20502, 2021.

Koenker, R. and Bassett Jr, G. Regression quantiles. *Econometrica: Journal of the Econometric Society*, 46 (1):33–50, 1978.

Koolen, W. M., Grünwald, P., and Van Erven, T. Combining adversarial guarantees and stochastic fast rates in online learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Lin, Z., Trivedi, S., and Sun, J. Conformal prediction with temporal quantile adjustments. *Advances in Neural Information Processing Systems*, 35:31017–31030, 2022.

Lindemann, L., Cleaveland, M., Shim, G., and Pappas, G. J. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 2023.

Mykland, P. A. Financial options and statistical prediction intervals. *The Annals of Statistics*, 31(5):1413–1438, 2003.

Noarov, G., Ramalingam, R., Roth, A., and Xie, S. High-dimensional unbiased prediction for sequential decision making. In *OPT 2023: Optimization for Machine Learning*, 2023.

Robinson, J. Sequential choice of an optimal dose: A prediction intervals approach. *Biometrika*, 65(1):75–78, 1978.

Tian, Q., Nordman, D. J., and Meeker, W. Q. Methods to compute prediction intervals: A review and new results. *Statistical Science*, 37(4):580–597, 2022.

Vovk, V. On-line confidence machines are well-calibrated. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science*, pp. 187–196. IEEE, 2002.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Xu, C. and Xie, Y. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569. PMLR, 2021.

Xu, C. and Xie, Y. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pp. 38707–38727. PMLR, 2023.

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pp. 25834–25866. PMLR, 2022.

Zimmert, J. and Seldin, Y. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936, 2003.

A. Additional proofs

A.1. Proof of Lemma 1

The proof of this result is similar to the proof of Lemma 4.1 of Gibbs & Candes (2021). We prove this by induction. First, $q_1 \in [0, B]$ by assumption, so (8) is satisfied at time $t = 1$. Next fix any $t \geq 1$ and assume q_t lies in the range specified in (8), and consider q_{t+1}. We now split into cases:

- If $q_t \in [0, B]$, then we have
$$q_{t+1} = q_t + \eta_t \left(\mathbb{1}_{Y_t \notin \mathcal{C}_t(X_t)} - \alpha\right) \in [q_t - \eta_t\alpha, q_t + \eta_t(1-\alpha)] \subseteq [-\alpha M_t, B + (1-\alpha)M_t].$$
- If $q_t \in (B, B + (1-\alpha)M_{t-1}]$, then we must have $\mathcal{C}_t(X_t) = \mathcal{Y}$. Then $\mathbb{1}_{Y_t \notin \mathcal{C}_t(X_t)} = 0$, and so
$$q_{t+1} = q_t - \eta_t\alpha \in [B - \eta_t\alpha, B + (1-\alpha)M_{t-1}] \subseteq [-\alpha M_t, B + (1-\alpha)M_t].$$
- If $q_t \in [-\alpha M_{t-1}, 0)$, then we must have $\mathcal{C}_t(X_t) = \emptyset$. Then $\mathbb{1}_{Y_t \notin \mathcal{C}_t(X_t)} = 1$, and so
$$q_{t+1} = q_t + \eta_t(1-\alpha) \in [-\alpha M_{t-1}, \eta_t(1-\alpha)] \subseteq [-\alpha M_t, B + (1-\alpha)M_t].$$

In all cases, then, (8) holds for $t + 1$ in place of t, which completes the proof.

A.2. Proof of 10

We need to prove that Z_t converges almost surely (note that the limit of Z_t may be a random variable). For each $t \geq 1$, we have
$$\mathbb{P}\left(Y_t \in \mathcal{C}_t(X_t) \mid \{(X_r, Y_r)\}_{r<t}\right) = \mathbb{P}\left(s_t(X_t, Y_t) \leq q_t \mid \{(X_r, Y_r)\}_{r<t}\right) = \mathsf{Coverage}_t(q_t),$$
since q_t and s_t are functions of $\{(X_r, Y_r)\}_{r<t}$ and are therefore independent of $(X_t, Y_t) \sim P$. This proves that Z_t is a martingale with respect to the filtration generated by the sequence of data points. We also have $\sup_{t\geq 1} \mathrm{Var}(Z_t) < \infty$, since we have assumed $\sum_{t=1}^{\infty} \eta_t^2 < \infty$. This means that Z_t is a uniformly integrable martingale, and therefore, Z_t converges almost surely (to some random variable), by Doob's second martingale convergence theorem.

A.3. Proofs of Corollaries 1 and 2

As for the theorems, it suffices to prove Corollary 2, since Corollary 1 is simply a special case.

Using the notation defined in the proof of Theorem 4, suppose that events $\mathcal{E}_Z$ and $\mathcal{E}_s$ both hold. Now we need to show that $\mathsf{Coverage}_t(q_t) \to 1 - \alpha$ holds as well. Fix any $\epsilon > 0$. Since $s(X, Y)$ has a continuous distribution, the map $q \mapsto \mathsf{Coverage}(q)$ is continuous, and so we can find some $\delta > 0$ such that
$$|\mathsf{Coverage}(q) - \mathsf{Coverage}(q^*)| \leq \epsilon/2 \text{ for all } q \in q^* \pm \delta.$$
Moreover, $\mathsf{Coverage}(q^*) = 1-\alpha$, since the distribution of $s(X, Y)$ is continuous and q^* is its $(1-\alpha)$-quantile, so we have
$$|\mathsf{Coverage}(q) - (1-\alpha)| \leq \epsilon/2 \text{ for all } q \in q^* \pm \delta.$$
Next, by Theorem 4, for all sufficiently large t, we have
$$|q_t - q^*| \leq \delta.$$
By definition of the event $\mathcal{E}_s$, for all sufficiently large t we have
$$|\mathsf{Coverage}_t(q^* - \delta) - \mathsf{Coverage}(q^* - \delta)| \leq \epsilon/2$$
and
$$|\mathsf{Coverage}_t(q^* + \delta) - \mathsf{Coverage}(q^* + \delta)| \leq \epsilon/2.$$
Then, combining all of these calculations, for all sufficiently large t we have
$$\mathsf{Coverage}_t(q_t) \geq \mathsf{Coverage}_t(q^* - \delta) \geq \mathsf{Coverage}(q^* - \delta) - \epsilon/2 \geq (1 - \alpha - \epsilon/2) - \epsilon/2 = 1 - \alpha - \epsilon,$$
where the first step holds since $q_t \geq q^* - \delta$, and $q \mapsto \mathsf{Coverage}_t(q)$ is nondecreasing. Similarly, for all sufficiently large t it holds that
$$\mathsf{Coverage}_t(q_t) \leq 1 - \alpha + \epsilon.$$
Since $\epsilon > 0$ is arbitrary, this completes the proof.

A.4. Proof of Proposition 2

By Lemma 1, $q_t \in [-\alpha c, B+(1-\alpha)c]$ for all t (since $\eta_t \leq c$ for all t). Since $q^* \in [0, B]$, we therefore have $|q_t - q^*| \leq B+c$ almost surely for all t. We also have $2\gamma\delta \leq 1$ and $\delta \leq B \leq B + c$, since the density of $s(X, Y)$ is supported on $[0, B]$ and must integrate to 1.

Next, by the assumptions of the proposition, for any $q_t \geq q^*$, if $q_t \leq q^* + \delta$ then

$$\mathsf{Coverage}(q_t) \geq 1 - \alpha + \gamma(q_t - q^*)$$

while if $q_t > q^* + \delta$ then

$$\mathsf{Coverage}(q_t) \geq \mathsf{Coverage}(q^* + \delta) \geq 1 - \alpha + \gamma\delta.$$

Either way, then, if $q_t \geq q^*$ then

$$\mathsf{Coverage}(q_t) \geq (1 - \alpha) + (q_t - q^*) \cdot \frac{\gamma\delta}{B + c}.$$

A similar calculations shows that if $q_t \leq q^*$, then

$$\mathsf{Coverage}(q_t) \leq (1 - \alpha) - (q^* - q_t) \cdot \frac{\gamma\delta}{B + c}.$$

Defining $a = \frac{\gamma\delta}{B+c} > 0$, we therefore have

$$\frac{\mathsf{Coverage}(q_t) - (1 - \alpha)}{q_t - q^*} \geq a \tag{14}$$

whenever $q_t \neq q^*$. Note that we must have $a \leq 1/c$, by construction.

Next, from the update step (4), we have

$$q_{t+1} - q^* = (q_t - q^*) + \eta_t\big((1 - \alpha) - \mathbb{1}_{Y_t \in \mathcal{C}_t(X_t)}\big).$$

Since $\mathbb{P}\left(Y_t \in \mathcal{C}_t(X_t) \mid \{(X_r, Y_r)\}_{r<t}\right) = \mathsf{Coverage}(q_t)$, we then calculate

$$\mathbb{E}\left[q_{t+1} - q^* \mid \{(X_r, Y_r)\}_{r<t}\right] = (q_t - q^*) + \eta_t\big((1 - \alpha) - \mathsf{Coverage}(q_t)\big),$$

and

$$\mathrm{Var}\big(q_{t+1} - q^* \mid \{(X_r, Y_r)\}_{r<t}\big) = \eta_t^2 \cdot \mathsf{Coverage}(q_t) \cdot (1 - \mathsf{Coverage}(q_t)) \leq \eta_t^2/4.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[(q_{t+1} - q^*)^2 \;\middle|\; \{(X_r, Y_r)\}_{r<t}\right] &\leq \big((q_t - q^*) + \eta_t\big((1 - \alpha) - \mathsf{Coverage}(q_t)\big)\big)^2 + \eta_t^2/4 \\ &\leq (q_t - q^*)^2 \cdot (1 - a\eta_t)^2 + \eta_t^2/4,\end{aligned}$$

where the last step holds by (14) above. After marginalizing, then,

$$\mathbb{E}\left[(q_{t+1} - q^*)^2\right] \leq \mathbb{E}\left[(q_t - q^*)^2\right] \cdot (1 - a\eta_t)^2 + \eta_t^2/4.$$

Next recall $\eta_t = ct^{-1/2-\epsilon}$ for each t. Fix some $T \geq 1$ that satisfies $T^{1/2-\epsilon} \geq \frac{1/2+\epsilon}{ac}$. First, since $|q_t - q^*| \leq B + c$ for all t as above, by choosing $b \geq (B + c)^2 T^{1/2+\epsilon}$ we must have $(q_t - q^*)^2 \leq bt^{-1/2-\epsilon}$ for all $t \leq T$, almost surely. Next, for each $t \geq T$, we proceed by induction. Assume $\mathbb{E}\left[(q_t - q^*)^2\right] \leq bt^{-1/2-\epsilon}$. Then

$$\begin{aligned}\mathbb{E}\left[(q_{t+1} - q^*)^2\right] &\leq \mathbb{E}\left[(q_t - q^*)^2\right] \cdot (1 - a\eta_t)^2 + \eta_t^2/4 \\ &= \mathbb{E}\left[(q_t - q^*)^2\right](1 - 2a\eta_t) + \left(\mathbb{E}\left[(q_t - q^*)^2\right] \cdot a^2 + 1/4\right)\eta_t^2 \\ &\leq bt^{-1/2-\epsilon}\left(1 - 2act^{-1/2-\epsilon}\right) + ((B + c)^2a^2 + 1/4)c^2t^{-1-2\epsilon} \\ &= bt^{-1/2-\epsilon} - 2abct^{-1-2\epsilon} + ((B + c)^2a^2 + 1/4)c^2t^{-1-2\epsilon} \\ &\leq b\left(t^{-1/2-\epsilon} - act^{-1-2\epsilon}\right),\end{aligned}$$

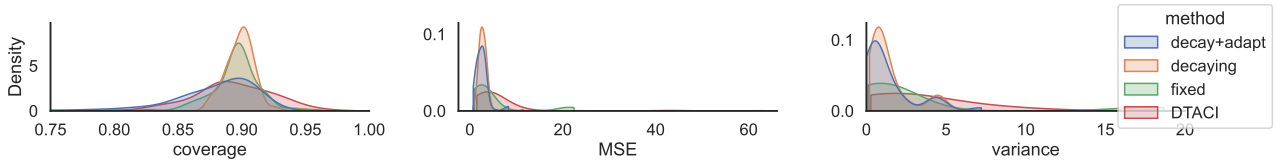

Figure 3. **Density plots of results on M4 datasets.** These plots show the same quantities as in Table 1, but now as histograms over the time-series in M4.

method	coverage	variance	MSE	infinite sets
DTACI	0.895491	4.107740	5.439935	0.036584
decay+adapt	0.885174	1.144337	1.967552	0.005011
decaying	0.900495	1.320579	2.366297	0.006883
fixed	0.901636	1.580243	2.922989	0.011179

Table 1. **Table of results on M4 datasets.** The table shows average results over all time series in the dataset—thus, all columns should be interpreted on average *over time-series in M4*. The coverage column displays the long-run coverage. The variance column shows the variance of the quantile normalized by the variance of the score sequence. The MSE column shows the squared error of the quantile normalized by the variance of the score sequence. Finally, the infinite sets column shows the fraction of time steps in the sequence for which the output is an infinite-width prediction set.

where the last step holds as long as we choose $b \geq \frac{((B+c)^2a^2+1/4)c}{a}$. And, since $t \geq T$, we have

$$act^{-1-2\epsilon} = act^{1/2-\epsilon} \cdot t^{-3/2-\epsilon} \geq acT^{1/2-\epsilon} \cdot t^{-3/2-\epsilon} \geq (1/2+\epsilon)t^{-3/2-\epsilon} \geq t^{-1/2-\epsilon} - (t+1)^{-1/2-\epsilon},$$

where the last step holds since $t \mapsto t^{-1/2-\epsilon}$ is convex, with derivative $-(1/2+\epsilon)t^{-3/2-\epsilon}$. Therefore, we have verified that $\mathbb{E}\left[(q_{t+1}-q^*)^2\right] \leq b(t+1)^{-1/2-\epsilon}$, as desired.

B. Additional experiments

We compare against two additional methods: first, "decay+adapt", a variant of our procedure that decays until it detects a change point, then resets the learning rate. Change points are identified when at least $N_{\text{miscoverage}}$ consecutive miscoverage events or N_{coverage} events are observed in a row (we set these constants to 10 and 30 by default, respectively). When a change point is identified, the learning rate is reset to $\frac{\hat{B}}{(t-T_{\text{changepoint}})^{1/2+\epsilon}}$, where $T_{\text{changepoint}}$ is the time at which the changepoint is detected and $\epsilon \in (0, 1/2)$. In these experiments, like in the main text, we set $\epsilon = 0.1$.

We additionally compare against DTACI (Gibbs & Candès, 2022), an adaptive-learning-rate variant of ACI that uses multiplicative weights to perform the updates (see (Gibbs & Candès, 2022) for further details.)

We compare these methods on a dataset of over 3000 time series subsampled from the M4 time series dataset. This dataset is a diverse array of time series with varying numbers of samples and distribution shifts. Code is available in our GitHub repository to run on all 100,000 time series in M4; here, we show results on the first 3000.

Finally, to showcase the conceptual differences between the standard decaying learning rate sequence and the 'decay+adapt' method, we display a simulated score sequence in Figure 4. Here, the scores are simulated from $\mathcal{N}(\mu_t, 1)$, where $\mu_t = 0$ for the first thousand time steps, $\mu_t = 2$ for the second thousand, $\mu_t = 4$ for the third thousand, and $\mu_t = 6$ for the final thousand. Especially towards the end of the time series, 'decay+adapt' can more quickly adjust to the change points.

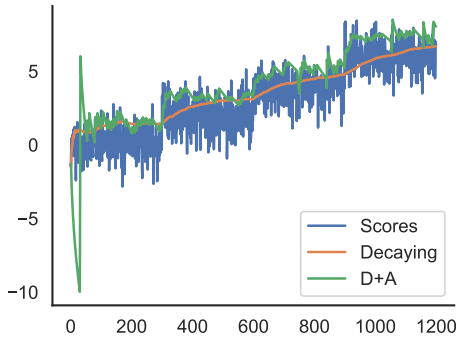

Figure 4. **Simulation comparison of decaying step size and 'decay+adapt'.** The raw score sequence is shown in blue, the decaying step size sequence is in orange, and 'decay+adapt' is in green.