Online Conformal Prediction with Efficiency Guarantees

Vaidehi Srinivas

Abstract

We study the problem of conformal prediction in a novel online framework that directly optimizes efficiency. In our problem, we are given a target miscoverage rate $\alpha > 0$, and a time horizon T. On each day $t \leq T$ an algorithm must output an interval $I_t \subseteq [0, 1]$, then a point $y_t \in [0, 1]$ is revealed. The goal of the algorithm is to achieve coverage, that is, $y_t \in I_t$ on (close to) a $(1-\alpha)$-fraction of days, while maintaining efficiency, that is, minimizing the average volume (length) of the intervals played. This problem is an online analogue to the problem of constructing efficient confidence intervals.

We study this problem over arbitrary and exchangeable (random order) input sequences. For exchangeable sequences, we show that it is possible to construct intervals that achieve coverage $(1 - \alpha) - o(1)$, while having length upper bounded by the best fixed interval that achieves coverage in hindsight. For arbitrary sequences however, we show that any algorithm that achieves a μ-approximation in average length compared to the best fixed interval achieving coverage in hindsight, must make a multiplicative factor more mistakes than αT, where the multiplicative factor depends on μ and the aspect ratio of the problem. Our main algorithmic result is a matching algorithm that can recover all Pareto optimal settings of μ and number of mistakes. Furthermore, our algorithm is deterministic and therefore robust to an adaptive adversary.

This gap between the exchangeable and arbitrary settings is in contrast to the classical online learning problem. In fact, we show that no single algorithm can simultaneously be Pareto optimal for arbitrary sequences and optimal for exchangeable sequences. On the algorithmic side, we give an algorithm that achieves the optimal tradeoff between the two cases.

Contents

1 Introduction

Conformal prediction is the problem of generating confidence sets. Given a target *miscoverage* rate $\alpha \in [0,1]$, and past data $Y_1, \ldots, Y_t$ from a ground set $\mathcal{Y}$, we must output a confidence set C that contains an unseen test point Y_{t+1} with probability $\geq 1-\alpha$. In addition, subject to our confidence set being valid, we would like it to be *efficient*, in that it is not too large. The most commonly studied settings by far are where $\mathcal{Y}$ is either a discrete label space, or $\mathcal{Y} = \mathbb{R}$. We will focus on the setting where $\mathcal{Y} = [0,1]$, which already has many applications. Conformal prediction is a well-studied problem in statistics, and we refer the reader to the book of Angelopoulos et al. (2025) for a full introduction.

The goal of conformal prediction is to give statistically valid guarantees for *real world data*. Thus we want to make as few assumptions about the data as possible: for example, we do not want to assume that it is i.i.d.. A large body of work adapts conformal methods to a variety of settings, where data may exhibit complex dependencies on time, location, and other factors. A textbook application of conformal prediction is the problem of uncertainty quantification. Consider the following scenario. We have a set of candidate pretrained machine learning models, that we deploy to real data. This data may not be generated i.i.d., in fact we may expect it to be time-varying. After seeing some data, we want to compare these models by measuring their uncertainty: on the next (unseen) datapoint, what is the expected error of each model? For a each model M, we can let the $Y_i^{(M)}$ be the error of M on datapoint i. Then, we could use conformal prediction to generate a confidence set for $Y_{t+1}^{(M)}$. Comparing the sizes of these confidence sets for the different choices of M gives us a way to compare the models.

What does it mean to be statistically valid on real data? Typically, this is modeled as $Y_1, \ldots, Y_{t+1}$ being *exchangeable* random variables. This means that the multiset of values $\{y_1, \ldots, y_{t+1}\}$ taken on by $Y_1, \ldots, Y_{t+1}$ can be distributed arbitrarily. However, the order in which the values arrive is uniformly random. That is, for any permutation $\sigma : [t+1] \to [t+1]$ and values $y_1, \ldots, y_{t+1} \in \mathcal{Y}$, we have

$$\mathbb{P}\left[Y_1 = y_1, \ldots, Y_{t+1} = y_{t+1}\right] = \mathbb{P}\left[Y_{\sigma(1)} = y_1, \ldots, Y_{\sigma(t+1)} = y_{t+1}\right].$$

This is also referred to as the *random order model* in the algorithms community. Note that i.i.d. sequences are exchangeable, but in fact exchangeability is a strictly more general condition than requiring the Y_i to be i.i.d.. A conformal prediction guarantee says that an algorithm outputs a set $C(Y_1, \ldots, Y_t) \subseteq \mathcal{Y}$ (a function of the training data), such that

$$\mathbb{P}[Y_{t+1} \in C(Y_1, \ldots, Y_t)] \geq 1 - \alpha, \tag{1}$$

where the probability is over the exchangeability of the $(Y_i)_{i=1}^{t+1}$. We say that a method that satisfies (1) achieves coverage.

Achieving coverage alone is trivial– a predictor could simply output all of $\mathcal{Y}$. Thus we also care about efficiency, which is minimizing the volume of the prediction set. Conformal methods are required to always be *statistically valid*, that is, they must provably achieve coverage (1). Typically, they have efficiency empirically validated on real and synthetic data. However, without rigorous efficiency guarantees, a conformal predictor could be meaningless, or even misleading. Consider the uncertainty quantification example we introduced earlier. If a method tells us that one model exhibits higher uncertainty than a second, we would like to know that it is because the second model really is better, and not because the conformal predictor is inefficient in its evaluation of the first model. We note that there is some work that establishes rigorous efficiency guarantees, and we discuss this further in Section 1.1.

Many conformal methods also restrict to outputting confidence sets that are a single interval, which is the focus of this work. This has many justifications, including explainability and sample efficiency (Gao et al., 2025).

The reason these results are built on exchangeability is that it is considered to be the weakest possible assumption that still allows one to make non-trivial predictions that are statistically valid. However, while exchangeability is much more general than, say, assuming the data are drawn i.i.d., it still rules out many scenarios that arise in practice. For example, if the data are time-varying, with even a minor distribution shift, the Y_i are no longer exchangeable, and standard conformal guarantees are no longer valid. This motivates our main question.

Is it possible to design a non-trivial conformal predictor for fully arbitrary data?

It is clear that the type of "one step" guarantee that is typically studied in conformal prediction is impossible. If the Y_is are fully arbitrary, the only set that provably achieves coverage on Y_{t+1} is the entirety of $\mathcal{Y}$. However, we show that in an online setting, it is indeed possible to achieve coverage *on average* over time.

Problem statement. We consider the problem in the following online setting. Fix a time horizon T. On each day $t \le T$, an algorithm $\mathcal{A}$ must output a set $C_t \subseteq [0,1]$. Then, $y_t \in \mathcal{Y}$ is revealed to $\mathcal{A}$. We will refer to the input sequence as $S = (y_1, \ldots, y_T)$. The goal is to achieve coverage close to $1-\alpha$, where we now define the coverage of $\mathcal{A}$ on S in an average sense as

$$\text{coverage}_{\mathcal{A}}(S) = \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\left[y_t \in C_t\right], \tag{2}$$

while minimizing the average volume (or Lebesgue measure) of the intervals played

$$\text{volume}_{\mathcal{A}}(S) = \frac{1}{T}\sum_{i=1}^{T} \text{volume}(C_t).$$

We will also refer to the number of points that $\mathcal{A}$ does not cover as the number of *mistakes* that $\mathcal{A}$ makes (where we want this to be close to αT). We will adopt the framework of competitive analysis, and aim to design an algorithm that, subject to achieving coverage close to $1-\alpha$, achieves average volume that is close to the best fixed interval in hindsight. That is, we define

$$\textsc{Opt}_S(\alpha) = \min_{\text{intervals } I} \text{volume}(I) \quad \text{s.t. } \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\left[y_t \in I\right] \ge 1-\alpha,$$

and our goal is to design an algorithm with low *volume approximation* (competitive ratio) μ, where

$$\mu \ge \frac{\text{volume}_{\mathcal{A}}(S)}{\textsc{Opt}_S(\alpha)}, \quad \text{for all sequences } S.$$

In this paper we will refer to the typical "one step" problem as *standard conformal prediction*, and this online learning framework as *online conformal prediction*. We note that there are other works that consider conformal prediction in an online setting, notably the line of work started by Gibbs and Candès (2025). However, they do not provide any guarantees for efficiency. We provide a more in-depth comparison in the related work (Section 1.1).

Contrast to online learning. This problem is reminiscent of classical online learning. Think of each interval I as an "expert." In fact, there are infinitely many intervals, and this poses a genuine issue, but for now, assume there $n < \infty$ experts, and we will justify this assumption later. On day t, the expert corresponding to interval I has loss 0 if $y_t \in I$ and loss 1 if $y_t \notin I$. The classical online learning problem is to pick an expert on each day, so as to achieve average loss that competes with the loss of the best fixed expert in hindsight. For our particular setting, this is trivial, as there is an expert in our set that corresponds to an interval that covers the whole space, and achieves loss 0 on every day, but the online learning problem is of course much more general.

In our problem, instead of loss being an objective that we are trying to minimize, loss becomes a constraint: we want to achieve average loss $\leq \alpha$ over the sequence. The objective that we are trying to minimize is an average over some other function over the experts that we choose, for us volume of the corresponding interval. In this sense, our problem becomes an online *constrained optimization* problem, where the set of experts that are feasible in hindsight with respect to the constraint depends on the input sequence.

One of the beautiful properties of classical online learning is that it is possible to design algorithms for arbitrary input sequences that have vanishing regret, i.e., they make $1 + O(\sqrt{\log n/T})$ times as many errors as the best fixed expert in hindsight (see e.g., Blum et al., 2020). This vanishing regret term matches the sampling error that one would face even for i.i.d. sequences. In essence, this says that going from i.i.d. (and therefore exchangeable) data to arbitrary data is without loss, as we are able to achieve the best possible guarantee even in the arbitrary case. Hence there is one algorithm that is able to achieve the "best of both worlds." Such results are also known in bandit problems and other settings (see e.g., Bubeck and Slivkins, 2012). This raises the following question for our setting:

> *Is it possible to design an algorithm for fully arbitrary data that recovers the best known bounds for exchangeable data?*

It turns out that, in contrast to classical online learning, our problem exhibits a large gap between what is possible in the arbitrary setting and the i.i.d. setting. For exchangeable data, we show that it is possible to design an algorithm that has volume approximation $\mu = 1$, while having a vanishing fraction of additional mistakes, i.e., the algorithm has coverage $(1 - \alpha) - \widetilde{O}(\sqrt{\log n/T})$, where we will make what n means in this setting precise shortly, and $\widetilde{O}$ hides subpolynomial factors in T. In the arbitrary setting however, we show that any algorithm, even one that is randomized, only robust to an oblivious adversary, and improper (outputs sets that are not intervals), faces a stark tradeoff between volume approximation μ and how many times more than αT mistakes it makes. In fact, we show that even subject to this lower bound, there can be no algorithm that is simultaneously optimal for the exchangeable setting and Pareto optimal for the arbitrary setting. This rules out any "best of both worlds" guarantee.

On the positive side, we design a matching algorithm that is able to achieve any Pareto optimal tradeoff. This algorithm is deterministic, which is also in contrast to the online learning problem which exhibits a separation between deterministic and randomized algorithms. Furthermore, our algorithms for exchangeable and arbitrary sequences fit in a unified framework that allows us to interpolate between being optimal for one and optimal for the other.

Challenges for arbitrary sequences. Designing an algorithm for arbitrary input sequences presents some challenges. Consider the following example. Our goal is to design a conformal predictor for miscoverage rate $\alpha = \frac{1}{2}$. There are two sequences S_1 and S_2 which have the first $T/2$ days drawn uniformly from $[0, \varepsilon]$, for some small parameter $\varepsilon > 0$. S_1 then has the last $T/2$ days

drawn uniformly from $[0,1]$ and S_2 has the last $T/2$ days drawn from $[0,\varepsilon^2]$. Any algorithm $\mathcal{A}$ that captures a constant fraction of points must incur a large multiplicative error in volume. Since S_1 and S_2 are the same for the first $T/2$ days, $\mathcal{A}$ must capture a constant fraction c of points in the first half of both sequences, or in the second half of both sequences. If $\mathcal{A}$ captures a constant fraction c of points on the first half of the sequences, then on sequence S_2, $\mathcal{A}$ has average volume $\geq c\varepsilon/2$, when the optimal interval that captures $\geq 1-\alpha$ fraction of points has volume $\leq \varepsilon^2$, so the volume approximation $\mu \geq \frac{c}{2\varepsilon}$. However, if $\mathcal{A}$ captures a constant fraction of c of points on second half of the sequences, then on sequence S_1, $\mathcal{A}$ has average volume $\geq c/2$, when the optimal interval that captures $\geq 1-\alpha$ fraction of points has volume $\leq \varepsilon$, so again the volume approximation $\mu \geq \frac{c}{2\varepsilon}$.

This example highlights two issues. The first is a recurring issue in online learning. As $\varepsilon \to 0$, our volume approximation becomes unbounded. This arises because the aspect ratio of our problem is unbounded. Because the problem is scale-free, a non-trivial algorithm that outputs intervals can be forced to make an unbounded number of mistakes. We address this by introducing a parameter $\mathsf{minwidth} > 0$, and require an algorithm to compete with the smallest volume interval that achieves coverage, among intervals that have volume at least $\mathsf{minwidth}$. This, along with the fact that $\mathcal{Y}$ is the bounded interval $[0,1]$, fixes the aspect ratio of the problem. In essence, this tells us that we may as well think of a grid of size $1/\mathsf{minwidth}$ over the interval $[0,1]$, and restrict the problem to considering intervals that are between two gridpoints. This reduces the number of intervals we have to consider to $O(1/\mathsf{minwidth}^2)$, which we should think of as analogous to the number of experts n in the earlier analogy to online learning. Similarly to online learning, we think of $n \approx (1/\mathsf{minwidth})^2$ as very large, and aim for guarantees that depend on $\log n \approx \log(1/\mathsf{minwidth})$.

However, there is a more significant issue. Even if $\mathsf{minwidth}$ is fixed, the above example still forces any algorithm that captures a constant fraction of points to incur a multiplicative error $\Omega(1/\mathsf{minwidth})$ in volume, which one should think of as being extremely large. This issue arises because we require the algorithm to achieve coverage very close to the target, i.e., a non trivial guarantee for $\alpha = \frac{1}{2}$ requires the algorithm to make $< T = 2\alpha T$ mistakes. Surprisingly, the problem actually becomes tractable for smaller values of α if we consider a biciteria approximation, that competes with the best interval that makes at most αT mistakes, while making a multiplicative factor more than αT mistakes.

Algorithmically, this example illustrates the difference between online learning and our problem. In online learning, on day t an algorithm can choose an expert that does well locally, that is, does well around time t. While this may not be the best expert over the whole sequence, the fact that it is doing as well as the best expert is doing right now, means that in every local time window, the algorithm is only performing as well or better than the best expert in hindsight. However, for our problem, choosing a different expert could incur a vastly different cost than the best expert. For example, if $\mathcal{A}$ captures a constant fraction of days in the first half of the sequence, on S_2 it only has coverage even higher than the best expert in hindsight, so it satisfies the coverage constraint. In exchange, however, it incurs much higher cost than the best expert. Thus on some days, an efficient algorithm must choose to abstain, and sacrifice coverage to minimize cost.

Results for arbitrary sequences. Our first result is an algorithm for arbitrary sequences, which achieves a given volume approximation μ while making a bounded number of mistakes.

Theorem 1.1 (Informal version of Corollary 2.2)**.** *For a given scale lower bound* $\mathsf{minwidth} > 0$*, multiplicative volume approximation* $\mu > 3$*, target miscoverage rate* $\alpha \geq 0$*, and time horizon* T*, we give an algorithm that on any sequence* S *of length* T *plays intervals of maximum volume (and therefore average volume)*

$$\leq \mu \max\{\textsc{Opt}_S(\alpha), \mathsf{minwidth}\},$$

Algorithm 1 Meta-algorithm for Online Conformal Prediction

Require: scale lower bound $\mathsf{minwidth} > 0$, multiplicative volume approximation $\mu \geq 1$, allowable error rate function $R(t) : \{0, \ldots, T-1\} \to [0, 1]$

$I_{\text{current}} \leftarrow [0, 0]$
for day t **do**
  if I_{current} has empirical coverage $< 1 - R(t-1)$ over $\{y_{t'} : t' < t\}$ **then**
    $\mathcal{F}_t :=$ set of intervals that achieve coverage $\geq 1 - R(t-1)$ over $\{y_{t'} : t' < t\}$
      // set of *feasible* intervals on day t
    $I_t :=$ arbitrary minimum volume interval $\in \mathcal{F}_t$
    $\widehat{\mu} := \mu \max\{1, \ \frac{\mathsf{minwidth}}{\text{vol}(I_t)}\}$
      // ensure the interval will have volume $\geq \mathsf{minwidth}$
    $I_{\text{current}} \leftarrow \widehat{\mu} I_t$
      // for $s > 0$, interval $I = [a, b]$, define $sI = \{x \subseteq \mathbb{R} : |x - \frac{a+b}{2}| \leq s \cdot \frac{b-a}{2}\}$
  play $I_{\text{current}} \cap [0, 1]$
  y_t is revealed

and makes number of mistakes bounded by

$$O\left(\frac{\log(1/\mathsf{minwidth})}{\log(\mu)}(\alpha T + 1)\right).$$

Moreover, the algorithm is deterministic and therefore robust to an adaptive adversary.

This tells us that, for example, it is possible to achieve a constant volume approximation μ, and have number of mistakes bounded by $O(\log(1/\mathsf{minwidth})\alpha T)$. The algorithm is simple, and is in fact an instantiation of a meta-algorithm, Algorithm 1, that we can apply to the exchangeable setting as well. Algorithm 1 maintains an interval I_{current} that has achieved error rate at most $R(t)$ by day t, where for the arbitrary order setting we choose $R(t) = \alpha\frac{T}{t}$. If on any day, I_{current} no longer meets this requirement, Algorithm 1 resets I_{current} to be a μ-scaling of the smallest volume interval that does achieve the error requirement. The analysis for the arbitrary order case shows that, while this is not explicitly enforced by the algorithm, this procedure ends up doing a doubling search for the scale of the optimal interval.

We highlight that the analysis is not specific to the choice of $R(t)$, and different choices of $R(t)$ achieve different trade-offs between μ and the number of mistakes. In the introduction, we present the result for the optimal choice of $R(t)$ for clarity, and we refer the reader to Theorem 2.1 for full details.

Our algorithm faces a stark tradeoff in the multiplicative volume approximation and the multiplicative approximation in number of mistakes. We show that this tradeoff is essentially tight.

Theorem 1.2 (Informal version of Theorem 4.1)**.** *For any scale lower bound* $\mathsf{minwidth} > 0$*, miscoverage rate* $\alpha > 0$*, and time horizon* T*, there is a set* $\mathcal{S}$ *of input sequences of length* T *chosen oblivious to any algorithm, such that for any potentially randomized algorithm* $\mathcal{A}$ *that outputs confidence sets over* $[0, 1]$ *that are not necessarily intervals,*

1. *If* $\mathcal{A}$ *plays sets with expected* average *volume* $\leq \mu_{\text{avg}} \max\{\text{OPT}_{S_i}(\alpha), \ \mathsf{minwidth}\}$ *on every sequence* $S_i \in \mathcal{S}$*, for some value* $\mu_{\text{avg}} > 0$*, then* $\mathcal{A}$ *must make*

$$\Omega\left(\min\left\{\frac{\log(1/\mathsf{minwidth})}{\log(\mu_{\text{avg}})}\alpha^{1+\varepsilon'}T, \ T\right\}\right), \quad \text{for any } \varepsilon' > 0$$

mistakes in expectation on some sequence $S_j \in \mathcal{S}$.

2. *If* $\mathcal{A}$ *plays sets with expected* maximum *volume* $\leq \mu_{\max} \max\{\textsc{Opt}_{S_i}(\alpha),\ \mathsf{minwidth}\}$ *on any sequence* $S_i \in \mathcal{S}$*, for some value* $\mu_{\max} > 0$*, then* $\mathcal{A}$ *must make*

$$\Omega\left(\min\left\{\frac{\log(1/\mathsf{minwidth})}{\log(\mu_{\max})} \cdot \alpha T,\ T\right\}\right)$$

mistakes in expectation on some sequence $S_j \in \mathcal{S}$.

Recall that our algorithm is deterministic, robust to an adaptive adversary, proper (it outputs sets that are intervals), and it has a bound on the volume of the largest interval it ever outputs. This lower bound tells us that it is essentially optimal, even among algorithms that are randomized, only robust to an oblivious adversary, improper (output sets that are not intervals), and have a bound only on the *average* volume of sets that they play.

Results for exchangeable sequences. Since conformal prediction is typically studied in the setting where the data is exchangeable, it is natural to consider this in our online formulation as well. Note that the guarantees we establish for this setting are incomparable to the standard "one step" guarantees of conformal prediction: we will establish coverage on average over time (2), as opposed to statistical validity on day T alone (1).

We show that the same meta-algorithm Algorithm 1, instantiated with $\mu = 1$ and a different choice of $R(t) = \alpha - O(\sqrt{\log T/T})$ can achieve an optimal guarantee for exchangeable data.

Theorem 1.3 (Informal version of Corollary 3.2)**.** *For a given scale lower bound* $\mathsf{minwidth} > 0$*, miscoverage rate* α*, and time horizon* T*, we have an algorithm that, on an exchangeable sequence* $\mathbf{S}$*, with probability* $\geq 1 - \frac{1}{100}$ *over the exchangeability of* $\mathbf{S}$*, plays intervals of maximum volume*

$$\leq \max\{\textsc{Opt}_{\mathcal{S}}(\alpha),\ \mathsf{minwidth}\},$$

and achieves expected coverage $\geq (1-\alpha) - O(\sqrt{\log T/T})$.

This bound is essentially optimal, as the vanishing error in coverage $O(\sqrt{\log T/T})$ is what we would get due to sampling error when the sequence is drawn i.i.d., which is a special case of exchangeability.

The analysis of the algorithm is tricky for exchangeable sequences because the interval I_t that Algorithm 1 plays on day t is dependent on the first $t-1$ days of the sequence $S_1, \ldots, S_{t-1}$. This is in turn *not* independent of S_t. Thus the coverage on day t

$$\mathbb{P}[S_t \in I_t(S_1, \ldots, S_{t-1})],$$

where the probability is over the exchangeability of S, is challenging to bound. This is actually the exact setting of standard conformal prediction. However, existing methods only reason about coverage and do not come with provable volume guarantees, which are necessary for this problem.[1]

To understand the challenge, consider the following example. Let S be a uniformly random permutation of a multiset that has $(1-2\alpha)T$ copies of $1/2$, αT copies of 0 and αT copies of 1. Clearly S is exchangeable. On day T, we have seen all but one element, which could have been $1/2$ with probability $1-2\alpha$, 0 with probability α, or 1 with probability α.

[1] To the best of our knowledge, all known volume optimality guarantees for standard conformal prediction require stronger assumptions than just exchangeability on the data. For example, they typically require the data to be drawn i.i.d.. We refer the reader to Section 1.1 for more discussion.

Algorithm 1, for $\mu = 1$, $R(t) = \alpha - O(\sqrt{\log T/T})$, will arbitrarily choose a minimum volume interval that achieves coverage $1 - R(T-1)$ on the first $T-1$ points. The intervals $[0, 1/2]$ or $[1/2, 1]$ will both meet the coverage requirement and have equal (minimal) volume, so say that Algorithm 1 tiebreaks by choosing the interval that achieved higher coverage, and choosing $[0, 1/2]$ if they have equal coverage. Then, if $S_T = 0$, $[1/2, 1]$ will achieve higher coverage on the prefix and S_T will not be in the output interval. Similarly, if $S_T = 1$, $[0, 1/2]$ will achieve higher coverage on the prefix and S_T will not be in the output interval. Thus S_T is covered with probability $\leq 1 - 2\alpha$, when we are aiming for coverage $1 - \alpha$. Our analysis handles this subtle issue by showing that there cannot be many days t on which this issue arises. This allows us to amortize the loss in coverage by allowing for higher error rates both at the beginning and at the end of the sequence. Carefully accounting for this allows us to show that the standard analysis for i.i.d. sequences goes through even for sequences that are only exchangeable and not necessarily i.i.d..

Similarly to the arbitrary case, our analysis is not specific to the choice of $R(t)$, and different choices of $R(t)$ achieve different trade-offs between μ and the number of mistakes. In the introduction, we present the result for the optimal choice of $R(t)$ for clarity, and we refer the reader to Theorem 3.1 for full details.

As a corollary to this analysis, we show that the interval that Algorithm 1 plays on day $T/2$ is actually a statistically valid and volume optimal conformal set for y_T (Corollary 3.3). We note that this guarantee can also be seen as a consequence of the result of Gao et al. (2025) along with uniform convergence for exchangeable sequences (Lemma 6.1). We view this observation as evidence that the online conformal prediction problem is more general than the offline version, in a similar spirit to how online learning is a generalization to the offline learning problem.

"Best of both worlds." Since our optimal algorithms for both the arbitrary order setting and the exchangeable setting are instantiations of the same meta-algorithm Algorithm 1, it is natural to ask whether there is a single algorithm that simultaneously achieves optimal bounds in both settings. That is, is there a single algorithm $\mathcal{A}$ such that, on arbitrary inputs $\mathcal{A}$ achieves a given Pareto optimal tradeoff in μ and number of mistakes, and on exchangeable inputs $\mathcal{A}$ achieves the better tradeoff of $\mu = 1$ and coverage $(1 - \alpha) - o(1)$? We show that this is indeed not the case.

Theorem 1.4 (Informal version of Theorem 5.1)**.** *Fix a scale lower bound* $\mathsf{minwidth} > 0$*, target miscoverage rate* $\alpha < \frac{1}{4}$*, and multiplicative volume approximation* $\mu > 1$*. Take any (potentially randomized) algorithm* $\mathcal{A}$ *that on any arbitrary order sequence* S *plays intervals of expected average volume*

$$\leq \mu \max\{\textsc{Opt}_{S^{\text{arb}}}(\alpha),\ \mathsf{minwidth}\}.$$

Then, $\mathcal{A}$ *must make*

$$\Omega\left(\min\left\{\frac{\log(1/\mathsf{minwidth})}{\log(\mu)},\ \log(1/\alpha)\right\}\alpha T\right)$$

mistakes in expectation on some i.i.d. sequence S'*.*

That is, an algorithm that achieves a non-trivial guarantee on every arbitrary sequence must incur a multiplicative overhead in the number of mistakes it makes on i.i.d., and therefore exchangeable sequences.

On the positive side, we show that Algorithm 1 with $R(t) = \frac{\alpha T}{t}$ meets this mistake bound on exchangeable sequences (Theorem 3.4). Recall that this is the same setting of $R(t)$ that gives an optimal algorithm for arbitrary sequences. Thus, this algorithm is optimal for arbitrary sequences, and gets the best possible mistake bound subject to achieving a guarantee for arbitrary sequences. We remark that Algorithm 1 does incur the same multiplicative volume approximation of μ for

both arbitrary and exchangeable sequences, whereas the lower bound does not rule out the case that a single algorithm can achieve a better volume approximation for exchangeable sequences than arbitrary sequences.

Takeaways. We study conformal prediction in a new online framework that directly optimizes the volume of the prediction sets. We show that the problem exhibits very different tradeoffs compared to classical online learning. We give near-optimal upper and lower bounds for algorithms over arbitrary order and exchangeable input sequences. For arbitrary sequences, our algorithm recovers all Pareto-optimal bicriteria approximations. Our algorithms are instantiations of the same simple deterministic meta-algorithm, Algorithm 1, with different settings of parameters. We show that the difference in parameters is necessary, as no algorithm can be simultaneously optimal on both arbitrary order and exchangeable sequences, and we show that Algorithm 1 achieves the optimal tradeoff.

1.1 Related Work

Conformal prediction is a very well-studied area in statistics with a wealth of recent work. The theoretical work alone in this space is the topic of a recent textbook (Angelopoulos et al., 2025). We provide a brief comparison to some of the work most relevant to the setting we consider, and refer the reader to the textbook for a more in-depth treatment.

Efficiency guarantees for standard conformal prediction. Typically, conformal prediction methods are required to provably achieve coverage (1). Efficiency, however, is often empirically validated on real and simulated data. There is a line of work that tries to establish efficiency guarantees subject to provably achieving coverage. The work of Sadinle et al. (2019), which considers the problem when the set $\mathcal{Y}$ is a discrete label space. The line of work including Lei et al. (2013); Izbicki et al. (2020a,b) work in settings where the data is drawn i.i.d. from a distribution that allows for a good p.d.f. estimator, and then achieving a volume optimal prediction set corresponds to taking a level set of that estimator. However, this approach does not work for arbitrary distributions, that may in general be hard to approximate. The work of Gao et al. (2025) shows that when the data is drawn i.i.d., if one chooses a prediction set from a family of potential sets that has bounded VC-dimension, it is sufficient to find the most efficient prediction set that achieves coverage over the samples. To the best of our knowledge, all known volume optimality bounds for standard conformal prediction must make stronger assumptions on the data than just exchangeability.

Beyond exchangeability. A line of work has tried to relax the assumption of exchangeability for standard conformal prediction. A line of work including Tibshirani et al. (2019); Barber et al. (2022) designs conformal methods that work when the joint distribution of the data is "approximately" exchangeable. Prinster et al. (2024) design conformal methods when the joint distribution of the data is not exchangeable, but requires that the joint distribution is known in advance.

Online conformal prediction. For exchangeable sequences, a foundational result of Vovk (2002) shows that a standard conformal method when run repeatedly on prefixes of an online sequence, errs with probability that is independent on each day, and is therefore valid on every day. For arbitrary sequences, the problem of computing confidence sets that achieve coverage on average was initiated by Gibbs and Candes (2021), and studied further by Zaffran et al. (2022); Gibbs and Candès (2025); Bhatnagar et al. (2023). Feldman et al. (2023); Angelopoulos et al. (2023) give an algorithm that has coverage provably tending to $1-\alpha$ as $T \to \infty$. To the best of our knowledge, ours is the first work to give provable efficiency guarantees for conformal prediction in the online setting, even for exchangeable data. In fact, our lower bound (Theorem 4.1) proves that

any method that achieves coverage tending to $1-\alpha$ on arbitrary sequences, must achieve a very large multiplicative volume approximation in the worst case.

The algorithm of Angelopoulos et al. (2023) works by estimating the bottom $\alpha/2$ and top $\alpha/2$ quantile of the data in an online fashion. The confidence interval will then be the interval between these two estimated quantiles. This is possible since a given quantile is the minimizer of a convex function, so it is possible to get a low regret estimator using online convex optimization. In our setting of optimizing volume, this runs into a couple of issues. The first is that it is subject to an "equal tails" assumption: it always outputs intervals that have approximately equal mass lying outside of the interval on each side. This may not be volume-optimal if the data distribution is skewed.

However, even if the data is subject to an equal tails constraint, where an input sequence must have each y_t drawn from a distribution $\mathcal{D}_t$ that is unimodal and symmetric around median $1/2$, the lower bound in Theorem 4.1 still holds. We can interpret this as saying that the quantile-based algorithm is not conservative enough to achieve a strong volume approximation. Put differently, suppose that on each day t, the algorithm was told the distribution $\mathcal{D}_t$ of y_t before outputting a confidence set for y_t. In this setting, choosing the volume optimal set that achieves coverage $1-\alpha$ on each day does not necessarily imply that the algorithm will compete in volume with the best set that achieves coverage $1-\alpha$ over the sequence of y_ts. The optimum in hindsight may have lower coverage on days that have spread out distributions and higher coverage on days that have concentrated distributions. This fact that the optimal loss over the sequence could be lower than the sum of the optimal losses on each day poses a roadblock to any strategy that goes through regret minimization with respect to a convex loss function.

2 Algorithmic Guarantees for Arbitrary Sequences

We begin by analyzing the performance of Algorithm 1 on arbitrary sequences, for general settings of the allowable error rate $R(t)$. In the analysis, we show that, even though it is not explicitly enforced by the algorithm, most times that I_{current} is reset (Algorithm 1, Line 3), its size increases multiplicatively by a factor that is linear in μ (the core of this argument is illustrated in Figure 1). Since the smallest width that I_{current} is reset to is $\mathsf{minwidth}$, and the largest it can be is 1, the number of times it is resets essentially depends on $\log_\mu(1/\mathsf{minwidth}) = \log(1/\mathsf{minwidth})/\log(\mu)$.

Call the time window between two subsequent resets of I_{current} a "phase." Over a phase, Algorithm 1 plays one fixed interval that is feasible over the whole phase. This allows us to bound the number of mistakes made in a particular phase by the total number of mistakes that any feasible interval is allowed to have made at that time.

Theorem 2.1 (Meta-algorithm Guarantee for Arbitrary Sequences)**.** *Fix a scale lower bound* $\mathsf{minwidth} > 0$*, a multiplicative volume approximation* $\mu > 3$*, and an allowable error rate function* $R(t) : \{0, \ldots, T-1\} \to [0,1]$ *such that* $R(t) \leq \frac{1}{4}$ *for all* $t \geq t^\star$*.*

On any sequence S *of length* T*, Algorithm 1 plays intervals of maximum volume*

$$\leq \mu \max\left\{ \textsc{Opt}_S\left(\min_{1\leq t\leq T} \frac{t}{T} R(t) \right),\ \mathsf{minwidth} \right\},$$

and makes number of mistakes

$$\leq t^\star + O\left(\frac{\log(1/\mathsf{minwidth}) \log T}{\log(\mu)} \left(1 + \max_{0\leq t\leq T-1} R(t)t \right) \right).$$

Moreover, the algorithm is deterministic and therefore robust to an adaptive adversary.

Proof. Any interval that has made $\leq R(t-1)(t-1)$ mistakes before day $t \geq 2$ is feasible for I_t (Algorithm 1, line 6). Therefore, any interval that makes $\leq R(t-1)(t-1)$ mistakes over the whole sequence, for every $t-1$, with error rate (miscoverage) over the full sequence S

$$\leq \min_{1 \leq t \leq T-1} \frac{t}{T} R(t)$$

is always feasible for I_t. Lines 6 and 9 ensure that the interval that is played has volume at most μ times the maximum of $\mathsf{minwidth}$ and the volume of any feasible interval. Therefore the volume played by Algorithm 1 is bounded by

$$\leq \mu \max \left\{ \textsc{Opt}_S \left(\min_{1 \leq t < T} \frac{t}{T} R(t) \right), \ \mathsf{minwidth} \right\}.$$

Now we bound the number of mistakes. Let t_1, t_2 be two subsequent times that I_{current} is reset. That is, the same I_{current} is played on all days $[t_1, t_2)$. Call $[t_1, t_2)$ a *phase*. We begin by bounding the number of mistakes in a phase. Since I_{current} is feasible on day $t_2 - 1$, I_{current} makes $\leq R(t_2-2)(t_2-2)$ mistakes on $\{x_1, \ldots, x_{(t_2-1)}\}$. Therefore I_{current} makes $\leq R(t_2-2)(t_2-2)+1$ mistakes on this phase: $\{x_{t_1}, \ldots, x_{t_2-1}\}$. Thus the number of mistakes that the algorithm makes on any phase is

$$\leq 1 + \max_{0 \leq t \leq T-1} R(t)t. \tag{3}$$

Now we bound the number of phases. Consider any two days t_1, t_2 such that $t^\star < t_1 < t_2$, so $R(t_1-1), R(t_2-2) \leq \frac{1}{4}$. Let I_1 be any interval that is feasible at t_1 and I_2 be any interval that is feasible at t_2. I_1 has captured $\geq \frac{3}{4}(t_1-1)$ points out of $\{x_1, \ldots, x_{t_1-1}\}$. I_2 has missed $\leq \frac{1}{4}(t_2-1)$ points out of $\{x_1, \ldots, x_{t_1-1}\} \subseteq \{x_1, \ldots, x_{t_2-1}\}$. Therefore as long as $\frac{t_2-1}{t_1-1} < 3$, I_1 and I_2 must capture some point in common and must overlap.

Partition the sequence after day $t^\star$ into $O(\log T)$ *epochs*, where each epoch $[t_{\text{start}}, t_{\text{end}}]$ has $\frac{t_{\text{end}}-1}{t_{\text{start}}-1} < 3$. We bound the number of mistakes that Algorithm 1 makes on each epoch. Fix an epoch e, and consider the first time in this epoch that Algorithm 1 resets I_{current}. By line 9, I_{current} is set to have volume $\geq \mu\mathsf{minwidth}$.

Now consider a subsequent time t in epoch e when I_{current} is reset. Let $t_{\text{prev}} < t$ be the previous time in epoch e when I_{current} was reset. Let $I_{t_{\text{prev}}}$ and I_t be the minimum volume feasible intervals on day t_{prev} and t, respectively (as chosen in line 6). Let $I_{\text{current}}^{(t_{\text{prev}})}$ denote the state of I_{current} before day t (after day t_{prev}).

Since I_{current} is reset on day t, $I_{\text{current}}^{(t_{\text{prev}})}$ is not feasible on day t, and in particular $I_t \not\subseteq I_{\text{current}}^{(t_{\text{prev}})}$. By the previous argument $I_{t_{\text{prev}}}$ and I_t must overlap. Any interval that contains a point in $I_{t_{\text{prev}}}$ and also contains a point that is not in $I_{\text{current}}^{(t_{\text{prev}})} = \widehat{\mu} I_{t_{\text{prev}}}$, must have volume $\geq \frac{\mu-1}{2} \cdot \text{vol}\left(I_{\text{current}}^{(t_{\text{prev}})}\right)$. This argument is illustrated in Figure 1. Denote the state of I_{current} at the end of day t as $I_{\text{current}}^{(t)}$. By line 9, we have that $I_{\text{current}}^{(t)} \supseteq \mu I_t$, so

$$\text{vol}\left(I_{\text{current}}^{(t)}\right) \geq \frac{\mu-1}{2} \text{vol}\left(I_{\text{current}}^{(t_{\text{prev}})}\right).$$

Since $\text{vol}(I_t) \leq 1$ for all days t, we have that $\text{vol}(I_{\text{current}}^{(t)}) \leq \mu$ for all days t. Thus, within an epoch, as long as $\mu > 3$, the number of times I_{current} can be reset is

$$\leq 1 + \log_{\frac{\mu-1}{2}} \left(\frac{\mu}{\mu\mathsf{minwidth}} \right) = O\left(\frac{\log(1/\mathsf{minwidth})}{\log(\mu)} \right), \tag{4}$$

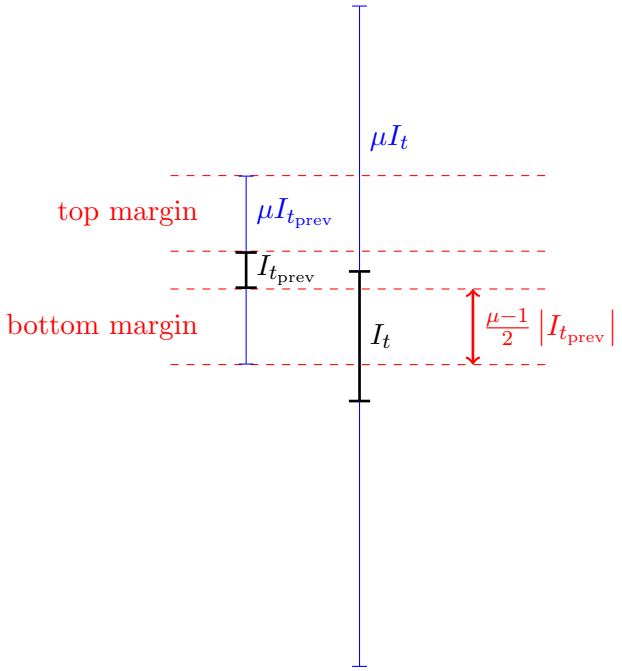


Figure 1: I_t achieved more coverage than $I_{\text{current}}^{(t_{\text{prev}})} = \mu I_{t_{\text{prev}}}$ on the sequence so far, so I_t is not contained in $\mu I_{t_{\text{prev}}}$. We also know that $I_{t_{\text{prev}}}$ and I_t must overlap. Thus, I_t must fully contain one of the margins that we add to $I_{t_{\text{prev}}}$ to form $\mu I_{t_{\text{prev}}}$ (denoted here as the regions between the red dashed lines). This gives a lower bound on the size of I_t in terms of the size of $I_{t_{\text{prev}}}$, and therefore a lower bound on the size of $I_{\text{current}}^{(t)}$ in terms of the size of $I_{\text{current}}^{(t_{\text{prev}})}$

and the total number of times that I_{current} is reset after day $t^\star$ is $O\left(\frac{\log(1/\mathsf{minwidth})\log T}{\log(\mu)}\right)$.

Our previous bound on the number of mistakes in a phase (3) allows us to bound the total number of mistakes that Algorithm 1 makes as

$$= t^\star + O\left(\frac{\log(1/\mathsf{minwidth})\log T}{\log(\mu)}\left(1 + \max_{0\le t\le T-1} R(t)t\right)\right).$$

□

The previous analysis allows us to choose an $R(t)$ to optimize the bounds. The following corollary observes that $R(t) = \frac{\alpha T}{t}$ gives a guarantee in terms of $\textsc{Opt}(\alpha)$. This gives a guarantee that exhibits a tradeoff between μ, the multiplicative volume approximation of the intervals played by Algorithm 1, and the number of mistakes that Algorithm 1 makes. Later, in Theorem 4.1 we give a lower bound that shows that the μ-mistake tradeoff that Algorithm 1 with $R(t) = \frac{\alpha T}{t}$ achieves is optimal. That is, for any setting of μ, this algorithm makes a near-optimal number of mistakes subject to achieving multiplicative approximation $\le \mu$.

$R(t) = \frac{\alpha T}{t}$ ensures that the intervals that are feasible on day t are exactly those that have made $\le \alpha T$ mistakes so far, where αT is the total number of mistakes that $\textsc{Opt}(\alpha)$ is allowed to make over the whole sequence. We can think of this algorithm as playing conservatively, and not ruling out any interval until it is demonstrably not $\textsc{Opt}(\alpha)$.

Corollary 2.2 (Algorithm for Arbitrary Sequences)**.** *Fix a scale lower bound* $\mathsf{minwidth} > 0$*, multiplicative volume approximation* $\mu > 3$*, target miscoverage rate* $\alpha \ge 0$*, and time horizon* T *(all*

known to the algorithm). Set $R(t)$ such that $R(0) = 1$ and $R(t) = \alpha\frac{T}{t}$ for $t > 0$. With these parameters, on any sequence S of length T, Algorithm 1 plays intervals of maximum volume

$$\leq \mu \max\{\textsc{Opt}_S(\alpha),\ \mathsf{minwidth}\},$$

and makes number of mistakes bounded by

$$O\left(\frac{\log(1/\mathsf{minwidth})}{\log(\mu)}(\alpha T + 1)\right).$$

Moreover, the algorithm is deterministic and therefore robust to an adaptive adversary.

Proof. The bound on maximum volume follows directly from the volume guarantee in Theorem 2.1.

For the bound on number of mistakes, this choice of $R(t)$ ensures that an interval is feasible on day t if and only if it has made $\leq \alpha T$ mistakes up to day t. Consider t_1, t_2 such that $2\alpha T + 1 < t_1 < t_2$. Let I_1 be any feasible interval on day t_1 and I_2 be any feasible interval on day t_2. Since I_1 and I_2 each make $\leq \alpha T$ mistakes over the first $2\alpha T + 1$ days, they must capture at least one day in common. Therefore I_1 and I_2 overlap. Thus we can think of all days $t > 2\alpha T + 1$ as forming one epoch, and the proof of Theorem 2.1 (Equation (4)) tells us that the number of phases in this epoch is

$$= O\left(\frac{\log(1/\mathsf{minwidth})}{\log(\mu)}\right).$$

The proof of Theorem 2.1 (Equation (3)) also tells us that, for this choice of $R(t)$, the number of mistakes in each phase is bounded by $\leq 1 + \alpha T$. Thus the total number of mistakes made is

$$\leq (2\alpha T + 1) + O\left(\frac{\log(1/\mathsf{minwidth})}{\log(\mu)}\right)(\alpha T + 1) \quad = O\left(\frac{\log(1/\mathsf{minwidth})}{\log(\mu)}(\alpha T + 1)\right).$$

□

3 Algorithmic Guarantees for Exchangeable Sequences

In this section we analyze the performance of Algorithm 1 on exchangeable sequences. The analysis differs significantly from the arbitrary order setting, because we want to take advantage of concentration. That is, an exchangeable sequence $\mathbf{S}$ has the property that the coverage of any interval $I \subseteq [0, 1]$ over one subsequence of $\mathbf{S}$ should be similar to the coverage of I over a different subsequence of $\mathbf{S}$.

Quantitatively, we can access this property through *uniform convergence*. In the most commonly used form, uniform convergence tells us that for a set family $\mathcal{F}$ of bounded VC-dimension, for us the family of intervals over $[0, 1]$, and a set Y of n samples drawn i.i.d. from some distribution $\mathcal{D}$, the coverage of every set in $\mathcal{F}$ over Y simultaneously converges to its true coverage over $\mathcal{D}$. That is, treating the VC-dimension of $\mathcal{F}$ and failure probability as constants, the coverage of every set in $\mathcal{F}$ over Y is within $\pm O(1/\sqrt{n})$ of its coverage over $\mathcal{D}$.

A similar statement holds even for sets of samples that are exchangeable but not necessarily i.i.d.. Let $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_t)$ be an exchangeable sequence of length T, and let Y be some (not necessarily contiguous) subsequence of size $t \leq T$ of $\mathbf{S}$.

Uniform convergence for exchangeable sequences tells us that, treating the VC-dimension of $\mathcal{F}$ and failure probability as constants, the coverage of every set in $\mathcal{F}$ over Y is within $\pm O(1/\sqrt{t})$ of its coverage over $\mathbf{S}$. That is, the coverage of sets in $\mathcal{F}$ over subsequences of $\mathbf{S}$ behaves essentially

as though the elements of Y were drawn i.i.d. from $\text{Unif}\{\mathbf{S}_1, \ldots, \mathbf{S}_T\}$. A formal version of this statement can be found in Lemma 6.1.

In analyzing the algorithm on exchangeable sequences, we must bound the number of mistakes by arguing that on each day t, the interval I_t chosen by the algorithm has high coverage. However, this requires us to have a tight bound on the coverage of I_t over the subsequence of length 1 that only contains day t, and the previous argument does not give tight bounds for short subsequences. We address this issue by instead bounding the coverage of I_t on the *suffix* of the input sequence that contains all days between t and T. Considering a longer subsequence allows us to get tighter bounds. Then, we observe that the coverage of I_t on the suffix is a property only of the unordered (multi)set of values in the suffix, and not the order that the values appear. Thus, conditioned on the coverage bound holding, the suffix of the sequence is still exchangeable within itself, so the coverage on day t is actually equal to the coverage over the whole suffix.

Theorem 3.1 (Meta-algorithm Guarantee for Exchangeable Sequences)**.** *Fix a scale lower bound* $\mathsf{minwidth} > 0$*, a multiplicative volume approximation* $\mu \geq 1$*, and an allowable error rate function* $R(t) : \{0, \ldots, T-1\} \to [0, 1]$*. Let* $\mathbf{S}$ *be an* exchangeable *input sequence of length* T*. On input* $\mathbf{S}$*, Algorithm 1 plays intervals of expected maximum volume*

$$\leq \mu \max\left\{ \textsc{Opt}_{\mathbf{S}}\left(\min_{0 \leq t < T} \left[R(t) - O\left(\sqrt{\log T / t}\right)\right]\right), \ \mathsf{minwidth} \right\} + \frac{1}{T^2},$$

and achieves expected coverage

$$\geq 1 - \frac{1}{T}\sum_{t=1}^{T} R(t-1) - O\left(\sqrt{\log T / T}\right).$$

Proof. Fix an exchangeable input sequence $\mathbf{S} = (Y_1, \ldots, Y_T)$, where the random variable Y_i is the input on day i. Let the random variable $M^{\mathbf{S}}$ be the number of mistakes that Algorithm 1 (for $\mathsf{minwidth}$, μ, $R(t)$) makes on $\mathbf{S}$. Let $M_t^{\mathbf{S}}$ be an indicator random variable of whether Algorithm 1 makes a mistake on day t. That is, let I_t be the interval that Algorithm 1 plays on day t. Then

$$M_t^{\mathbf{S}} = \mathbf{1}\left[Y_t \notin I_t\right].$$

We analyze the algorithm's performance on one particular day t. We will refer to the days before t, i.e. the window $[1, t-1]$, as the *prefix*, and the days t and after, i.e. the window $[t, T]$, as the *suffix*. For an interval $I \subseteq [0, 1]$ and a window (contiguous subsequence) $W = [t_1, t_2] \subseteq [1, T]$, define $\text{err}_{\mathbf{S}}(I, W)$ to be the fraction of days in W of $\mathbf{S}$ that are not covered by I. Since the set of intervals over $[0, 1]$ has VC-dimension 2, Lemma 6.1 tells us that the following statements hold, each with probability $\geq 1 - \frac{1}{T^3}$, for some absolute constant C:

$$\forall I \quad \text{err}_{\mathbf{S}}(I, [1, T]) - C\sqrt{\frac{\log T}{t-1}} \ \leq \ \text{err}_{\mathbf{S}}(I, [1, t-1]) \ \leq \ \text{err}_{\mathbf{S}}(I, [1, T]) + C\sqrt{\frac{\log T}{t-1}}, \tag{5}$$

$$\forall I \quad \text{err}_{\mathbf{S}}(I, [1, T-t+1]) - C\sqrt{\frac{\log T}{t-1}} \ \leq \ \text{err}_{\mathbf{S}}(I, [t, T]) \ \leq \ \text{err}_{\mathbf{S}}(I, [1, T]) + C\sqrt{\frac{\log T}{T-t+1}}. \tag{6}$$

We will refer to (5) and (6) as the *uniform convergence property*. Note that (5) and (6) are properties only of the set of values that appear in the prefix and suffix, and does not imply anything about the order that elements appear in the suffix. That is, conditioned on (5) and (6), the suffix is still exchangeable, and we have

$$\forall I, t' \in [t, T] \quad \mathop{\mathbb{P}}_{\mathbf{S}}\left[Y_{t'} \in I \mid (5), (6)\right] = \mathop{\mathbb{E}}_{\mathbf{S}}\left[\text{err}_{\mathbf{S}}(I, [t, T]) \mid (5), (6)\right]. \tag{7}$$

This allows us to bound the expected value of $M_i^{\mathbf{S}}$ conditioned on (5) and (6). We have deterministically that $\mathrm{err}_{\mathbf{S}}(I_t, [1, t-1]) \le R(t-1)$ (Algorithm 1, Line 6).

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{S}}\left[M_t^{\mathbf{S}} \mid (5), (6)\right] &= \mathop{\mathbb{P}}_{\mathbf{S}}\left[Y_t \in I_t \mid (5), (6)\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{S}}\left[\mathrm{err}_{\mathbf{S}}(I, [t, T]) \mid (5), (6)\right] && \text{by (7)} \\
&\le \mathop{\mathbb{E}}_{\mathbf{S}}\left[\mathrm{err}_{\mathbf{S}}(I, [1, T]) + C\sqrt{\frac{\log T}{T-t+1}} \;\middle|\; (5), (6)\right] && \text{by (6)} \\
&\le \mathop{\mathbb{E}}_{\mathbf{S}}\left[\mathrm{err}_{\mathbf{S}}(I, [1, t-1]) + C\sqrt{\frac{\log T}{T-t+1}} + C\sqrt{\frac{\log T}{t-1}} \;\middle|\; (5), (6)\right] && \text{by (5)} \\
&\le R(t-1) + C\sqrt{\frac{\log T}{T-t+1}} + C\sqrt{\frac{\log T}{t-1}}.
\end{aligned}$$

We also have that $M_t^{\mathbf{S}} \le 1$. So in total this gives us that

$$\mathop{\mathbb{E}}_{\mathbf{S}}\left[M_t^{\mathbf{S}}\right] \le R(t-1) + C\sqrt{\frac{\log T}{T-t+1}} + C\sqrt{\frac{\log T}{t-1}} + \frac{2}{T^3}. \tag{8}$$

This allows us to bound $M^{\mathbf{S}}$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{S}}\left[M^{\mathbf{S}}\right] &\le \sum_{t=1}^{T}\left[M_t^{\mathbf{S}}\right] \\
&\le \sum_{t=1}^{T}\left[R(t-1) + C\sqrt{\frac{\log T}{T-t+1}} + C\sqrt{\frac{\log T}{t-1}} + \frac{2}{T^3}\right] \\
&\le O\left(\sqrt{T\log T}\right) + \sum_{t=1}^{T} R(t-1).
\end{aligned}$$

Thus the expected coverage of Algorithm 1 can be bounded as

$$1 - \frac{1}{T}\mathop{\mathbb{E}}_{\mathbf{S}}\left[M^{\mathbf{S}}\right] = 1 - \frac{1}{T}\sum_{t=1}^{T} R(t-1) - O\left(\sqrt{\log T/T}\right).$$

Now we bound the volume of intervals played by Algorithm 1. We condition on (5) holding simultaneously for all $t \in [1, T]$. Since this holds with probability $\ge 1 - \frac{1}{T^3}$ for each t, it holds for all t with probability $\ge 1 - \frac{1}{T^2}$.

Consider any interval I that has coverage $\ge 1 - \min_{0\le t\le T}\left[R(t) + C\sqrt{\log T/t}\right]$ over $\mathbf{S}$ in hindsight. Then, conditioned on (5) holding for all t, I must be a feasible choice for I_t on every day t (Algorithm 1, Line 6). Thus, conditioned on (5) holding for all t, Algorithm 1 plays intervals of maximum volume

$$\le \mu \max\left\{\textsc{Opt}_{\mathbf{S}}\left(\min_{0\le t<T}\left[R(t) - O\left(\sqrt{\log T/t}\right)\right]\right),\ \mathsf{minwidth}\right\}.$$

Even if (5) does not hold for all t, Algorithm 1 plays intervals of maximum length 1. So Algorithm 1 plays intervals of expected maximum volume

$$\le \mu \max\left\{\textsc{Opt}_{\mathbf{S}}\left(\min_{0\le t<T}\left[R(t) - O\left(\sqrt{\log T/t}\right)\right]\right),\ \mathsf{minwidth}\right\} + \frac{1}{T^2}.$$

□

From the previous guarantee, it is simple to derive a setting of $R(t)$ that achieves volume essentially that of $\textsc{Opt}(\alpha)$, and coverage that is within $O(\sqrt{\log T/T})$ of the target $1-\alpha$. The $O(\sqrt{\log T/T})$ term is a standard sampling error that should arise since exchangeable sequences are only more general than i.i.d. sequences.

Corollary 3.2 (Algorithm for Exchangeable Sequences)**.** *Fix a scale lower bound* $\mathsf{minwidth} > 0$*, miscoverage rate* α*, and time horizon* T *(all known to the algorithm). Set* $\mu = 1$*, and* $R(t) = \alpha + O\left(\sqrt{\log T/t}\right)$*. Let* $\mathbf{S}$ *be an exchangeable input sequence of length* T*. Algorithm 1, with the above parameters, on input* $\mathbf{S}$ *plays intervals of expected maximum volume*

$$\leq \max\{\textsc{Opt}_{\mathbf{S}}(\alpha),\ \mathsf{minwidth}\} + \frac{1}{T^2},$$

and achieves expected coverage

$$\geq (1-\alpha) - O\left(\sqrt{\log T/T}\right).$$

Proof. Follows directly from Theorem 3.1. □

We observe that this near-optimal algorithm for online conformal prediction on exchangeable sequences implies a near-optimal algorithm for standard conformal prediction with exchangeable data. We view this simple observation as evidence that online conformal prediction on exchangeable sequences can be seen as a generalization of standard conformal prediction with exchangeable data, in a similar way to how online learning generalizes PAC learning.

We note this statement is also a direct consequence of the result of Gao et al. (2025), which states this for Y_i drawn i.i.d. from some unknown distribution $\mathcal{D}$, along with uniform convergence for exchangeable sequences (Lemma 6.1), via a similar sample splitting argument.

Corollary 3.3 (Efficiency Guarantee for Standard Conformal Prediction)**.** *Fix a scale lower bound* $\mathsf{minwidth} > 0$*, and miscoverage rate* α*. Let* $\mathbf{S} = (Y_1, \ldots, Y_T)$ *be exchangeable random variables valued in* $[0,1]$*.*

Given $Y_1, \ldots, Y_{T-1}$*, we construct the following conformal predictor for* Y_T*. We run Algorithm 1 with parameters* $\mathsf{minwidth}, \mu = 1, \alpha, T, R(t) = \alpha + O(\sqrt{\log T/t})$ *for* $\lceil t/2 \rceil$ *days using* $Y_1, \ldots, Y_{\lceil t/2 \rceil}$ *as inputs. Let* $I_{\lceil t/2 \rceil}$ *be the interval that Algorithm 1 predicts on day* $\lceil t/2 \rceil$*.*

$I_{\lceil t/2 \rceil}$ *is a conformal predictor for* Y_T*, achieving coverage*

$$\mathop{\mathbb{P}}_{\mathbf{S}}\left[Y_T \in I_{\lceil t/2 \rceil}\right] \geq (1-\alpha) - O\left(\sqrt{\frac{\log T}{T}}\right),$$

and $I_{\lceil t/2 \rceil}$ *has volume bounded by*

$$\leq \max\{\textsc{Opt}_{\mathbf{S}}(\alpha),\ \mathsf{minwidth}\} + \frac{1}{T^2}.$$

Proof. This follows from the proof of Theorem 3.1. Let $I_{\lceil t/2 \rceil}$ be the interval that *Algorithm 1* plays on day $\lceil t/2 \rceil$ of input $\mathbf{S}$. Let $M^{\mathbf{S}}_{\text{final}}$ be an indicator random variable of whether $Y_T \notin I_{\lceil t/2 \rceil}$. By (7) we have that

$$\mathop{\mathbb{E}}_{\mathbf{S}}\left[M^{\mathbf{S}}_{\text{final}} \mid (5) \text{ and } (6)\right] = \mathop{\mathbb{P}}_{\mathbf{S}}\left[Y_T \in I_{\lceil t/2 \rceil} \mid (5) \text{ and } (6)\right] = \mathop{\mathbb{P}}_{\mathbf{S}}\left[Y_{\lceil t/2 \rceil} \in I_{\lceil t/2 \rceil} \mid (5) \text{ and } (6)\right].$$

Thus (8) ensures that

$$\begin{aligned}\mathop{\mathbb{P}}_{\mathbf{S}}\left[Y_T \notin I_{\lceil t/2\rceil}\right] &= \mathop{\mathbb{E}}_{\mathbf{S}}\left[M^{\mathbf{S}}_{\text{final}}\right] \\ &\le R(\lceil t/2\rceil - 1) + C\sqrt{\frac{\log T}{T - \lceil t/2\rceil + 1}} + C\sqrt{\frac{\log T}{\lceil t/2\rceil - 1}} + \frac{2}{T^3} \\ &\le \alpha + O\left(\sqrt{\frac{\log T}{T}}\right),\end{aligned}$$

and $I_{\lceil t/2\rceil}$ achieves coverage

$$\ge (1-\alpha) - O\left(\sqrt{\frac{\log T}{T}}\right)$$

on Y_T.

For the efficiency bound, Corollary 3.2 ensures that the expected maximum volume of any interval played by Algorithm 1, and therefore the expected volume of $I_{\lceil t/2\rceil}$ is bounded by

$$\le \max\{\textsc{Opt}_{\mathbf{S}}(\alpha),\ \mathsf{minwidth}\} + \frac{1}{T^2}.$$

□

Finally, we note that for the optimal algorithm for exchangeable sequences we use Algorithm 1 with $R(t) = \alpha - O(\sqrt{\log T/T})$ (Corollary 3.2). However, for the optimal algorithm for arbitrary sequences we use Algorithm 1 with $R(t) = \alpha T/t$ (Corollary 2.2). This raises the natural question of whether we can design a single algorithm (perhaps but not necessarily by using Algorithm 1 with a unified choice of $R(t)$) that simultaneously achieves the optimal guarantee on both exchangeable and arbitrary sequences.

Later in Theorem 5.1, we give a lower bound that shows that it is indeed not possible to design any algorithm that is simultaneously optimal in both of these settings. Theorem 5.1 quantifies this bound by showing that for a given choice of μ, $\mathsf{minwidth}$, and α, any algorithm that has volume bounded by $\mu \max\{\textsc{Opt}_S(\alpha), \mathsf{minwidth}\}$ on every *arbitrary* input sequence S, must make

$$\widetilde{\Omega}\left(\min\left\{\ln(1/\alpha),\ \frac{\ln(1/\mathsf{minwidth})}{\ln \mu}\right\}\alpha T\right)$$

mistakes in expectation on some *i.i.d.* sequence S'. Here $\widetilde{\Omega}$ hides subpolynomial factors in α.

Algorithmically, we show that Algorithm 1 with the conservative choice of $R(t) = \alpha T/t$ is able to meet this mistake bound on exchangeable data, up to a sampling error term. We note that this algorithm has the same multiplicative volume approximation μ on both arbitrary and exchangeable data, while our lower bound does not rule out the possibility of an algorithm having a better approximation ratio on exchangeable sequences than on arbitrary sequences.

To match the lower bound, we must do a finer analysis of Algorithm 1 on exchangeable sequences that accounts for the number of phases of the algorithm. In the analysis of Algorithm 1 on arbitrary sequences (Theorem 2.1), we used the fact that the algorithm plays one fixed interval for a whole phase. This allowed us to amortize the error of the interval that the algorithm plays over the whole phase, and get a mistake bound that depends on the number of phases which is $O(\log(1/\mathsf{minwidth})/\log\mu)$. We observe that when the number of phases is small, a similar amortizing argument can get a tighter bound on the number of mistakes that Algorithm 1 makes, even in the exchangeable case.

Theorem 3.4 (Simultaneous Guarantee for Arbitrary and Exchangeable Sequences)**.** *Fix a scale lower bound* $\mathsf{minwidth} > 0$*, target miscoverage rate* α*, multiplicative approximation* μ*, and time horizon* T *(all known to the algorithm). Set* $R(t) = \frac{\alpha T}{t}$*. Let* $\mathbf{S}$ *be an exchangeable input sequence of length* T*. On input* $\mathbf{S}$*, Algorithm 1 with the above parameters makes*

$$O\left(\min\left\{\sqrt{T\log T}\sqrt{\frac{\log(1/\mathsf{minwidth})}{\log\mu}} + \frac{\log(1/\mathsf{minwidth})}{\log\mu}\alpha T, \quad \sqrt{T\log T} + \log(1/\alpha)\alpha T\right\}\right)$$

mistakes in expectation, and plays intervals of maximum volume

$$\leq \mu \ \max\left\{\textsc{Opt}_{\mathbf{S}}(\alpha),\ \mathsf{minwidth}\right\}.$$

Proof. We condition on uniform convergence for every window (contiguous subsequence) of $\mathbf{S}$. For a window $W = [t_1, t_2] \subseteq [1, T]$, and an interval $I \subseteq [0, 1]$, define $\mathrm{err}_{\mathbf{S}}(I, W)$ to be the fraction of days in W of $\mathbf{S}$ that are not covered by I. For each window W, we have that with probability $\geq 1 - \frac{1}{T^4}$, the coverage of every interval I over W is within

$$\varepsilon_W = C\sqrt{\frac{\log T}{|W|}}$$

its coverage over all of $\mathbf{S}$, for some universal constant C. Taking a union bound over all $O(T^2)$ possible windows, we have that

$$\forall W, I \quad \mathrm{err}_S(I, [1, T]) - \varepsilon_W \leq \mathrm{err}_S(I, W) \leq \mathrm{err}_S(I, [1, T]) + \varepsilon_W. \tag{9}$$

with probability $\geq 1 - \frac{1}{T^2}$. We refer to the property (9) as the *uniform convergence property.*

Now let $\widehat{S}$ be any realization of $\mathbf{S}$ that satisfies the uniform convergence property. Let K be the number of times sets I_{current} Algorithm 1 (for μ, $\mathsf{minwidth}$, $R(t) = \frac{\alpha T}{t}$) on $\widehat{S}$. Let t_i for $1 \leq i \leq K$ be the days when I_{current} is set, where $t_1 = 1$, and we define $t_{K+1} = T + 1$ for convenience. We refer to the window $[t_i, t_{i-1})$ as *phase* i.

Let $M^{\widehat{S}}$ be the number of mistakes that Algorithm 1 makes on $\widehat{S}$. Let $M_i^{\widehat{S}}$ be the miscoverage rate of Algorithm 1 over phase i of $\widehat{S}$. Let I_i be the value of I_{current} on phase i. If Algorithm 1 resets I_{current} on day t_{i+1}, it means that I_i has error rate $\leq \frac{\alpha T}{t_{i+1}-1}$ over the window $[1, t_{i+1} - 1]$ (Algorithm 1, Line 3). By the uniform convergence property (9),

$$\begin{aligned}
M_i^{\widehat{S}} &= \mathrm{err}_{\widehat{S}}(I_i,\ [t_i, t_{i+1} - 1]) \\
&\leq \mathrm{err}_{\widehat{S}}(I_i,\ [1, T]) + \varepsilon_{[t_i, t_{i+1}-1]} \\
&\leq \mathrm{err}_{\widehat{S}}(I_i, [1, t_{i+1} - 1]) + \varepsilon_{[1, t_{i+1}-1]} + \varepsilon_{[t_i, t_{i+1}-1]} \\
&\leq \frac{\alpha T}{t_{i+1} - 1} + C\sqrt{\frac{\log T}{t_{i+1} - 1}} + C\sqrt{\frac{\log T}{t_{i+1} - t_i}} \\
&\leq \frac{\alpha T}{t_{i+1} - 1} + 2C\sqrt{\frac{\log T}{t_{i+1} - t_i}}.
\end{aligned}$$

We bound $M^{\widehat{S}}$ in terms of the $M_i^{\widehat{S}}$s. Since $R(t) = \frac{\alpha T}{t} \geq 1$ for $t \leq \alpha T$, and the initial interval has coverage 0, we have that $t_2 = \alpha T + 1$ and $M_1 = 1$.

$$M^{\widehat{S}} = \sum_{i=1}^{K} (t_{i+1} - t_i) M_i^{\widehat{S}}$$

$$
\begin{aligned}
&= \alpha T + \sum_{i=2}^{K} (t_{i+1} - t_i) M_i^{\widehat{S}} \\
&\leq \alpha T + \sum_{i=2}^{K} (t_{i+1} - t_i) \left(\frac{\alpha T}{t_{i+1} - 1} + 2C \sqrt{\frac{\log T}{t_{i+1} - t_i}} \right) \\
&\leq \alpha T + \sum_{i=2}^{K} \left[(t_{i+1} - t_i) \frac{\alpha T}{t_{i+1} - 1} + 2C \sqrt{\log T} \sqrt{(t_{i+1} - t_i)} \right] \\
&\leq \alpha T + 2C \sqrt{\log T} \sqrt{K-1} \sqrt{T} + \alpha T \sum_{i=2}^{K} \left[\frac{t_{i+1} - 1}{t_{i+1} - 1} - \frac{t_i - 1}{t_{i+1} - 1} \right] \qquad \text{by concavity of } \sqrt{\cdot} \\
&\leq \alpha T + 2C \sqrt{(K-1) T \log T} + (K-1)\alpha T - \alpha T \sum_{i=2}^{K} \frac{t_i - 1}{t_{i+1} - 1}. \qquad (10)
\end{aligned}
$$

This upper bound on mistakes is maximized when $\sum_{i=2}^{K} \frac{t_i - 1}{t_{i+1} - 1}$ is minimized. Recall that $t_2 = \alpha T + 1$ and $t_{K+1} = T + 1$. Assume without loss of generality that αT is an integer. So we have

$$
\begin{aligned}
\prod_{i=2}^{K} \frac{t_i - 1}{t_{i+1} - 1} &= \frac{t_2 - 1}{t_{K+1} - 1} = \alpha \\
\sum_{i=2}^{K} \ln \left(\frac{t_i - 1}{t_{i+1} - 1} \right) &= \ln \alpha.
\end{aligned}
$$

For convenience of arithmetic, define $q_i = (t_i - 1)/(t_{i+1} - 1)$ where every $q_i > 0$, so we have $\sum_{i=2}^{K} \ln q_i = \ln \alpha$. By concavity of ln we have

$$
\begin{aligned}
\frac{1}{K-1} \sum_{i=2}^{K} \ln q_i &\leq \ln \left(\frac{1}{K-1} \sum_{i=2}^{K} q_i \right) \\
\frac{\ln \alpha}{K-1} + \ln(K-1) &\leq \left(\sum_{i=2}^{K} q_i \right) \\
(K-1)\alpha^{1/(K-1)} &\leq \sum_{i=2}^{K} q_i = \sum_{i=2}^{K} \frac{t_i - 1}{t_{i+1} - 1}.
\end{aligned}
$$

Substituting back into (10), we have

$$
\begin{aligned}
M^{\widehat{S}} &\leq \alpha T + \sqrt{(K-1) T \log T} + (K-1)\alpha T - (K-1)\alpha^{1/(K-1)} \alpha T \\
&\leq \alpha T + \sqrt{(K-1) T \log T} + (K-1) \left(1 - \alpha^{1/(K-1)} \right) \alpha T.
\end{aligned}
$$

Since $\alpha \geq 0$, we have that $(K-1)\left(1 - \alpha^{1/(K-1)}\right) \leq K - 1$. We can also bound this term by $\ln(1/\alpha)$:

$$
\begin{aligned}
\frac{\ln \alpha}{K-1} &= -\frac{\ln(1/\alpha)}{K-1} \\
\alpha^{1/(K-1)} &= e^{-\ln(1/\alpha)/(K-1)} \geq 1 - \frac{\ln(1/\alpha)}{K-1} \\
\frac{\ln(1/\alpha)}{K-1} &\geq 1 - \alpha^{1/(K-1)}
\end{aligned}
$$

$$\ln(1/\alpha) \geq (K-1)\left(1-\alpha^{1/(K-1)}\right).$$

Putting this together we have

$$M^{\widehat{S}} \leq \sqrt{(K-1)T\log T} + \left(1+\min\left\{K-1,\ \ln(1/\alpha)\right\}\right)\alpha T.$$

We can upper bound the number of phases of Algorithm 1 on any input as

$$K-1 \leq \log_{\frac{\mu-1}{2}}(1/\mathsf{minwidth}) = O\left(\frac{\log(1/\mathsf{minwidth})}{\log\mu}\right).$$

So for every realization $\widehat{S}$ of $\mathbf{S}$ that satisfies the uniform convergence property (9), we have

$$M^{\widehat{S}} = O\left(\sqrt{T\log T}\sqrt{\frac{\log(1/\mathsf{minwidth})}{\log\mu}} + \min\left\{\frac{\log(1/\mathsf{minwidth})}{\log\mu},\ \log(1/\alpha)\right\}\alpha T\right).$$

We have that $\mathbf{S}$ satisfies the uniform convergence property with probability $\geq 1-\frac{1}{T^2}$. If $\mathbf{S}$ does not satisfy the uniform convergence property, the number of mistakes is trivially upper bounded by T. So we have that the expected number of mistakes by Algorithm 1 (for μ, $\mathsf{minwidth}$, $R(t)=\frac{\alpha T}{t}$) is

$$O\left(\sqrt{T\log T}\sqrt{\frac{\log(1/\mathsf{minwidth})}{\log\mu}} + \min\left\{\frac{\log(1/\mathsf{minwidth})}{\log\mu},\ \log(1/\alpha)\right\}\alpha T\right). \tag{11}$$

When the number of phases $\frac{\log(1/\mathsf{minwidth})}{\log\mu}$ is large, the first "sampling error" term will dominate. In this setting, the original bound in Theorem 3.1 which does not do a phase-by-phase analysis gives a tighter bound of

$$\begin{aligned}
M^{\mathbf{S}} &\leq T\left(\frac{1}{T}\sum_{t=1}^{T} R(t-1) + O\left(\sqrt{\log T/T}\right)\right)\\
&\leq \sum_{t=1}^{T}\max\left\{1,\frac{\alpha T}{t}\right\} + O\left(\sqrt{T\log T}\right)\\
&\leq \alpha T + \alpha T\sum_{t=\alpha T+1}^{T}\frac{1}{t} + O\left(\sqrt{T\log T}\right)\\
&\leq O\left(\sqrt{T\log T} + \log(1/\alpha)\alpha T\right).
\end{aligned} \tag{12}$$

(11) and (12) together bound the expected number of mistakes by

$$O\left(\min\left\{\sqrt{T\log T}\sqrt{\frac{\log(1/\mathsf{minwidth})}{\log\mu}} + \frac{\log(1/\mathsf{minwidth})}{\log\mu}\alpha T,\quad \sqrt{T\log T} + \log(1/\alpha)\alpha T\right\}\right).$$

The volume guarantee follows directly from Corollary 2.2. □

4 Lower Bound for Arbitrary Order Sequences

We provide a lower bound that shows that our algorithm for arbitrary order sequences is optimal. That is, for any given choice of miscoverage rate α, scale lower bound $\mathsf{minwidth}$, and

multiplicative volume approximation μ, we prove that any algorithm that achieves volume $\leq \mu \min\{\textsc{Opt}_S(\alpha),\ \mathsf{minwidth}\}$ on every arbitrary order sequence S of length T must make a multiplicative factor more mistakes than the target αT. This mistake bound shows that our upper bound of running Algorithm 1 with $R(t) = \frac{\alpha T}{t}$ is near-optimal in all parameters. In particular, we note that Algorithm 1 actually plays intervals with maximum volume bounded by $\mu \min\{\textsc{Opt}_S(\alpha),\ \mathsf{minwidth}\}$, which is even stronger than playing intervals with average volume bounded by $\mu \min\{\textsc{Opt}_S(\alpha),\ \mathsf{minwidth}\}$. Among algorithms that have this property, we show that the expected number of mistakes that Algorithm 1 makes:

$$O\left(\min\left\{\frac{\log(1/\mathsf{minwidth})}{\log(\mu)} \cdot \alpha T,\ T\right\}\right),$$

is optimal (see Corollary 2.2 for the upper bound).

For algorithms that play intervals that have average volume bounded by $\mu \min\{\textsc{Opt}_S(\alpha),\ \mathsf{minwidth}\}$, but do not necessarily maximum volume bounded by this, we show that they must make

$$\Omega\left(\left\{\frac{\log(1/\mathsf{minwidth})}{\log(\mu) + \log(1/\alpha)} \cdot \alpha T,\ T\right\}\right)$$

mistakes in expectation. This differs from the earlier bound by a subpolynomial factor in α, and thus Algorithm 1 is near-optimal even among algorithms that meet this weaker condition.

The lower bound construction is directly inspired by a worst-case instance for Algorithm 1 with $R(t) = \frac{\alpha T}{t}$. In the analysis (Theorem 2.1), we bound the number of times that Algorithm 1 resets I_{current}. We then say between two subsequent resets of I_{current}, Algorithm 1 plays a fixed interval that has accrued less than the target error. So that interval must make at most αT mistakes over the time that it is feasible, and therefore at most αT mistakes while it is being played. In the lower bound, we present the algorithm with a series of phases of increasing "scales," each lasting αT days. See Figure 2 for an illustration. In each phase, the algorithm does not yet know if this phase contains points that are captured by Opt, or this is an "outlier" phase. If the algorithm chooses to capture a large fraction of the points, it will accrue high volume, which leads to high multiplicative ratio in the case that this was an outlier phase. If the algorithm chooses not to capture a large fraction of points, it will accrue a large number of mistakes.

Theorem 4.1 (Lower Bound for Arbitrary Sequences)**.** *For any scale lower bound* $\mathsf{minwidth} > 0$*, miscoverage rate* $\alpha > 0$*, and time horizon* T*, such that for any, potentially randomized, algorithm* $\mathcal{A}$ *that outputs confidence sets over* $[0, 1]$ *that are not necessarily intervals we have:*

1. *If* $\mathcal{A}$ *plays sets with expected* average *volume* $\leq \mu_{\text{avg}} \max\{\textsc{Opt}_{S_i}(\alpha),\ \mathsf{minwidth}\}$ *on every sequence* S*, for some value* $\mu_{\text{avg}} > 0$*, then* $\mathcal{A}$ *must make*

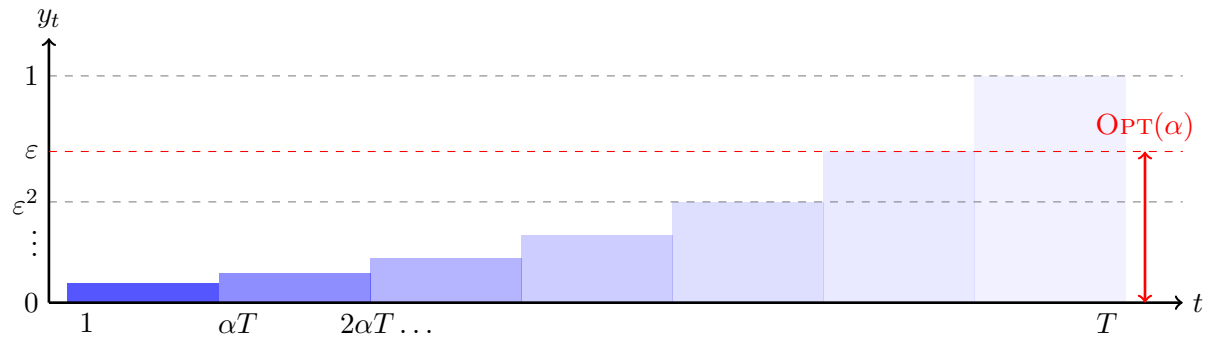
$$\Omega\left(\min\left\{\frac{\log(1/\mathsf{minwidth})}{\log(\mu_{\text{avg}}) + \log(1/\alpha)} \cdot \alpha T,\ T\right\}\right)$$

mistakes in expectation on some sequence S'*. This is* $\Omega\left(\min\left\{\frac{\log(1/\mathsf{minwidth})}{\log(\mu_{\text{avg}})}\alpha^{1+\varepsilon'}T,\ T\right\}\right)$ *for any* $\varepsilon' > 0$*.*

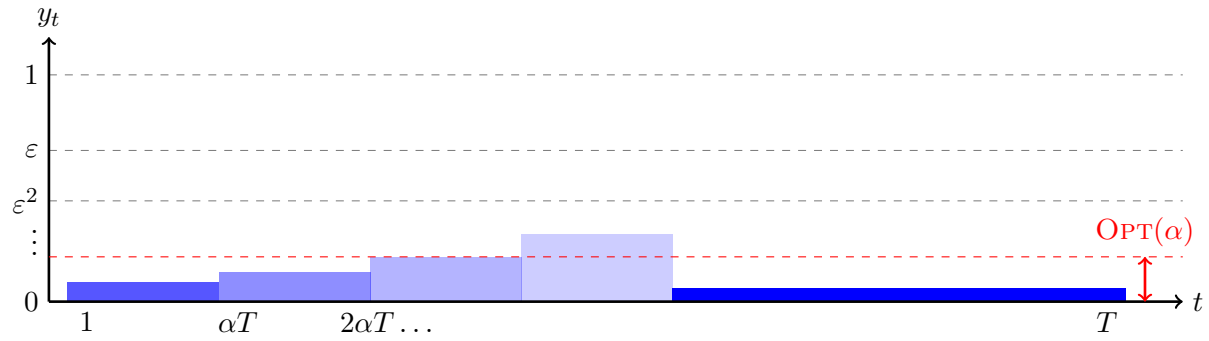
2. *If* $\mathcal{A}$ *plays sets with expected* maximum *volume* $\leq \mu_{\max} \max\{\textsc{Opt}_{S_i}(\alpha),\ \mathsf{minwidth}\}$ *on every sequence* S*, for some value* $\mu_{\max} > 0$*, then* $\mathcal{A}$ *must make*

$$\Omega\left(\min\left\{\frac{\log(1/\mathsf{minwidth})}{\log(\mu_{\max})} \cdot \alpha T,\ T\right\}\right)$$

mistakes in expectation on some sequence S'*.*



(a) If an algorithm $\mathcal{A}$ has low expected coverage ($\leq \frac{1}{2}$) on every phase of this sequence, it must make $\Omega(T)$ mistakes over the whole sequence.



(b) If an algorithm $\mathcal{A}$ has high expected coverage ($\geq \frac{1}{2}$) on some phase i of sequence (a), say phase 4, then it also has high expected coverage on phase 4 of this alternate sequence, as $\mathcal{A}$ does not yet know which case it is in. Therefore, $\mathcal{A}$ must incur a high multiplicative volume approximation on this alternate sequence, as the scale of phase 4 is multiplicatively larger than the scales of the other phases.

Figure 2: We illustrate the lower bound construction in Theorem 4.1. The horizontal axis is the day t of the sequence and the vertical axis is the value of y_t. y_t is drawn according to a uniform distribution, and we illustrate this as a box over the support of y_t with depth of color corresponding to the density of the distribution. We divide the input sequence into "phases" of length αT. We then construct an input sequence S that has the scale of y_t increasing multiplicatively from one phase to the next (Figure 2a). We also consider a set of alternate sequences S_i, where S_i is the same as S through phase i, and then drops to a very small scale for the rest of the sequence (Figure 2b). On phase i, the algorithm does not know if it is in case (a), where the scale will continue to grow, or case (b), where the sequence will drop off.

Furthermore, this lower bound holds against algorithms that only have to be competitive on sequences where the input on each day is drawn from a symmetric unimodal distribution with median $1/2$.

Proof. Assume without loss of generality that $1/\alpha$ is an integer, and T is a multiple of $1/\alpha$. Divide the input sequence into $1/\alpha$ "phases" of length αT. Let $K \in \{1, \ldots, 1/\alpha\}$ and $\varepsilon > 0$ be parameters to be chosen later. For each $1 \leq i \leq K$ we generate a sequence $S_i^{(K)}$ as follows:

- For phases $j \leq i$, all days in phase j are drawn i.i.d. from Unif $\left[0, \varepsilon^{K-j}\right]$.
- For phases $j > i$, all days in phase j are drawn i.i.d. from Unif $\left[0, \varepsilon^{K}\right]$.

Fix any potentially randomized algorithm $\mathcal{A}$. Consider what $\mathcal{A}$ does on input sequence $S_K^{(K)}$. We say that $\mathcal{A}$ "hits" phase j if it captures $\geq \frac{\alpha T}{2}$ points in phase j, in expectation. If $\mathcal{A}$ captures $< \frac{\alpha T}{2}$ points in phase j in expectation, we say it "misses." Either:

(a) $\mathcal{A}$ misses every phase $1 \le j \le K$. Then $\mathcal{A}$ makes $\ge K\frac{\alpha T}{2}$ mistakes in expectation.

(b) $\mathcal{A}$ hits some phase $1 \le j \le K$. Since phase j is drawn i.i.d. Unif $\left[0, \varepsilon^{K-j}\right]$, the expected average volume that $\mathcal{A}$ plays on phase j must be $\ge \frac{\varepsilon^{K-j}}{2}$.

Now consider what $\mathcal{A}$ would do on sequence $S_j^{(K)}$. Since $S_j^{(K)}$ and $S_K^{(K)}$ are identical through phase j, this means that $\mathcal{A}$ must have expected average volume $\ge \frac{\varepsilon^{K-j}}{2}$ on phase j of $S_j^{(K)}$ as well. Thus, on $S_j^{(K)}$, $\mathcal{A}$ must play average volume $\ge \frac{\alpha\varepsilon^{K-j}}{2}$ and maximum volume $\ge \frac{\varepsilon^{K-j}}{2}$ in expectation.

For $S_j^{(K)}$, we know that the interval $[0, \varepsilon^{K-j+1}]$ achieves coverage 1 on every phase other than phase j. Thus, $\textsc{Opt}_{S_j^{(K)}}(\alpha) \le \varepsilon^{K-j+1}$. This means that on $S_j^{(K)}$, $\mathcal{A}$ plays intervals of average volume

$$\ge \frac{\alpha}{2\varepsilon}\textsc{Opt}_{S_j^{(K)}}(\alpha)$$

in expectation, and intervals of maximum volume

$$\ge \frac{1}{2\varepsilon}\textsc{Opt}_{S_j^{(K)}}(\alpha)$$

in expectation.

We will set ε, K such that $\mathsf{minwidth} = \frac{1}{2}\varepsilon^K$. If $\mathcal{A}$ plays intervals with expected average volume $\le \mu_{\text{avg}} \max\{\textsc{Opt}_{S_j^{(K)}}(\alpha),\ \mathsf{minwidth}\}$ on $S_j^{(K)}$ we have that, either $\mathcal{A}$ makes $\ge K\frac{\alpha T}{2}$ mistakes in expectation, or

$$\begin{aligned}
\frac{\alpha}{2\varepsilon}\textsc{Opt}_{S_j^{(K)}}(\alpha) &\le \mu_{\text{avg}} \max\{\textsc{Opt}_{S_j^{(K)}}(\alpha),\ \mathsf{minwidth}\} \\
\min\{\frac{\alpha}{2\varepsilon}, \frac{\alpha}{2\varepsilon^{K+2}}\textsc{Opt}_{S_j^{(K)}}(\alpha)\} &\le \mu_{\text{avg}} \\
\frac{\alpha}{2(\mathsf{minwidth}/2)^{1/K}} = \frac{\alpha}{2\varepsilon} &\le \mu_{\text{avg}}, \qquad (13)
\end{aligned}$$

where the last line follows because $\textsc{Opt}_{S_j^{(K)}}(\alpha) \ge \varepsilon^{K+1}$. Similarly, if $\mathcal{A}$ plays intervals with expected maximum volume $\le \mu_{\max} \max\{\textsc{Opt}_{S_j^{(K)}}(\alpha),\ \mathsf{minwidth}\}$ on $S_j^{(K)}$ we have that, either $\mathcal{A}$ makes $\ge K\frac{\alpha T}{2}$ mistakes in expectation, or

$$\frac{1}{2(\mathsf{minwidth}/2)^{1/K}} = \frac{1}{2\varepsilon} \le \mu_{\max}. \qquad (14)$$

Setting $K = \min\{\left\lfloor \frac{\log(2/\mathsf{minwidth})}{\log(1/\alpha)+\log(2\mu_{\text{avg}})} \right\rfloor, 1/\alpha\}$ ensures that (13) is satisfied. Thus if $\mathcal{A}$ has expected average volume bounded as $\le \mu_{\text{avg}} \max\{\textsc{Opt}_{S_j^{(K)}}(\alpha),\ \mathsf{minwidth}\}$ for all $S_j^{(K)} \in \mathcal{S}$, then $\mathcal{A}$ must make

$$\begin{aligned}
&\ge \min\left\{\left(\frac{\log(2/\mathsf{minwidth})}{\log(1/\alpha)+\log(2\mu_{\text{avg}})} - 1\right)\alpha T,\ \frac{T}{2}\right\} \\
&= \Omega\left(\min\left\{\frac{\log(1/\mathsf{minwidth})}{\log(\mu_{\text{avg}})+\log(1/\alpha)}\alpha T,\ T\right\}\right) \\
&= \Omega\left(\min\left\{\frac{\log(1/\mathsf{minwidth})}{\log(\mu_{\text{avg}})}\alpha^{1+\varepsilon'} T,\ T\right\}\right) \qquad \text{for any } \varepsilon' > 0,
\end{aligned}$$

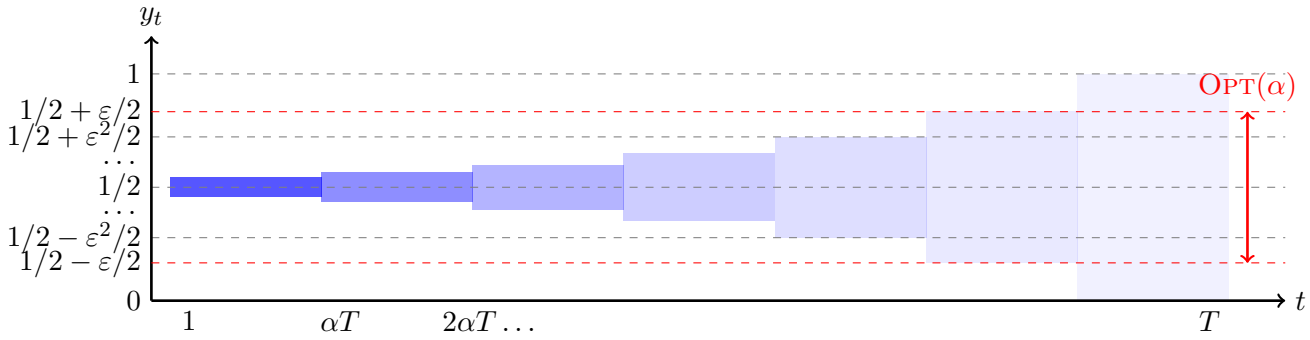


Figure 3: Symmetric version of the lower bound construction in Figure 2

mistakes in expectation on $S_K^{(K)}$. Similarly, Setting $K = \min\{\left\lfloor \frac{\log(2/\mathsf{minwidth})}{\log(2\mu_{\max})} \right\rfloor, 1/\alpha\}$ ensures that (14) is satisfied and $K \le 1/\alpha$. Thus if $\mathcal{A}$ has expected maximum volume bounded as $\le \mu_{\max} \max\{\textsc{Opt}_{S_j}(\alpha), \mathsf{minwidth}\}$ for all $S_j^{(K)} \in \mathcal{S}^{(K)}$, then $\mathcal{A}$ must make

$$\begin{aligned} &\ge \min\left\{\left(\frac{\log(2/\mathsf{minwidth})}{\log(2\mu_{\max})} - 1\right)\alpha T,\ \frac{T}{2}\right\} \\ &= \Omega\left(\min\left\{\frac{\log(1/\mathsf{minwidth})}{\log(\mu_{\max})}\alpha T,\ T\right\}\right) \end{aligned}$$

mistakes in expectation on $S_K^{(K)}$.

Finally, we note that while this argument considered the performance of $\mathcal{A}$ on intervals with one end at 0 for ease of notation, we could have instead considered a symmetric version of the sequences, where we map each phase that is drawn

$$\text{Unif}[0, b] \to \text{Unif}\left[\frac{1}{2} - \frac{1}{2}b,\ \frac{1}{2} + \frac{1}{2}b\right],$$

(see Figure 3). This leads to the same lower bounds on volume, while ensuring that the input distribution on each day is unimodal and symmetric around 1/2. This gives a lower bound against algorithms that are competitive on such sequences. □

5 Lower Bound Against "Best of Both Worlds" Algorithms

We provide a lower bound that says that it is impossible to for a single algorithm to achieve the best possible guarantee both for arbitrary sequences and exchangeable sequences. Recall that for an arbitrary sequence S, Corollary 2.2 shows that for choices of the multiplicative volume approximation μ, scale lower bound $\mathsf{minwidth}$, and target miscoverage α, Algorithm 1 with $R(t) = \frac{\alpha T}{t}$ plays intervals of average volume $\le \mu \max\{\textsc{Opt}_S(\alpha), \mathsf{minwidth}\}$, and makes $O(\log(1/\mathsf{minwidth})/\log(\mu)\alpha T)$ mistakes. Theorem 4.1 tells us that this is optimal up to subpolynomial factors in α.

For exchangeable sequences however, we can do better. Corollary 3.2 tells us that Algorithm 1 with $R(t) = (1-\alpha) + O(\sqrt{\log T/t})$ plays intervals of average volume $\lesssim \textsc{Opt}_S(\alpha)$ and has expected coverage $\ge (1-\alpha) - O(\sqrt{\log T/t})$ (that is, $\le (1 + O(\sqrt{\log T/t}))\alpha T$ mistakes in expectation).

This leads to the natural question: is there a single algorithm that can achieve both guarantees? It turns out that these two bounds are indeed at odds. That is, achieving an algorithm that achieves the optimal tradeoff for arbitrary sequences must make a large number of mistakes on

some exchangeable sequence, and an algorithm that achieves the optimal number of mistakes for exchangeable sequences must have a large multiplicative volume approximation.

The following lower bound gives a tradeoff between the two cases. Theorem 3.4 tells us that Algorithm 1 with $R(t) = \frac{\alpha T}{t}$ actually achieves this tradeoff, and thus it is optimal in the sense that it gets as close as possible to a "best of both worlds" guarantee. That is, for a fixed choice of μ, $\mathsf{minwidth}$, and α, subject to playing intervals of average volume $\leq \mu \max\{\textsc{Opt}_S(\alpha), \mathsf{minwidth}\}$ on every arbitrary sequence S, Algorithm 1 with $R(t) = \frac{\alpha T}{t}$ makes the lowest possible number of mistakes on both arbitrary and exchangeable sequences. We remark, however, that the lower bound does not say anything about whether the multiplicative factor of μ that Algorithm 1 incurs in the exchangeable case is necessary.

The construction of this lower bound, like the lower bound for arbitrary sequences, is inspired by observing the performance of Algorithm 1 with $R(t) = \frac{\alpha T}{t}$ on i.i.d. sequences. The optimal algorithm for exchangeable sequences will play intervals that have coverage very close to the target $1-\alpha$ from very early on. Algorithm 1 with $R(t) = \frac{\alpha T}{t}$ however, will be more conservative and play intervals that have substantially lower coverage, to hedge against the possibility the scale of the problem will shrink in the future, and that many of the points seen so far were actually "outliers." Intuitively, such a strategy should be optimal subject to a worst case guarantee.

The lower bound proof essentially directly argues that this conservative strategy is optimal. That is, we design a distribution $\mathcal{D}$, and then argue that any algorithm $\mathcal{A}$ must fall into one of two categories.

(a) On every day t, $\mathcal{A}$ plays an interval that has expected miscoverage $\geq \frac{1}{2} \cdot \frac{\alpha T}{t}$. Summing over t, we see that this leads to $\approx \ln(1/\alpha)\alpha T$ mistakes in expecation overall.

(b) On some day t, $\mathcal{A}$ plays an interval with expected miscoverage $\leq \frac{1}{2} \cdot \frac{\alpha T}{t}$. Then, we show that $\mathcal{A}$ has not been conservative, and there exists an alternate sequence on which $\mathcal{A}$ incurs a large volume approximation.

A core part of the argument is in designing a distribution $\mathcal{D}$ that achieves the desired tradeoff between the miscoverage of the algorithm and the multiplicative volume approximation μ. We provide an illustration of the argument in Figure 4.

Theorem 5.1 (No "Best of Both Worlds" Algorithm)**.** *Fix a scale lower bound* $\mathsf{minwidth} > 0$ *and target miscoverage rate* $\alpha > 0$*. Let* $\mathcal{A}$ *be any algorithm, deterministic or randomized.*

1. *If* $\mathcal{A}$ *plays sets with expected* average *volume* $\leq \mu_{\text{avg}} \max\{\textsc{Opt}_S(\alpha), \mathsf{minwidth}\}$ *on every arbitrary order sequence* S*, for some* $\mu_{\text{avg}} > 0$*, then* $\mathcal{A}$ *must make*

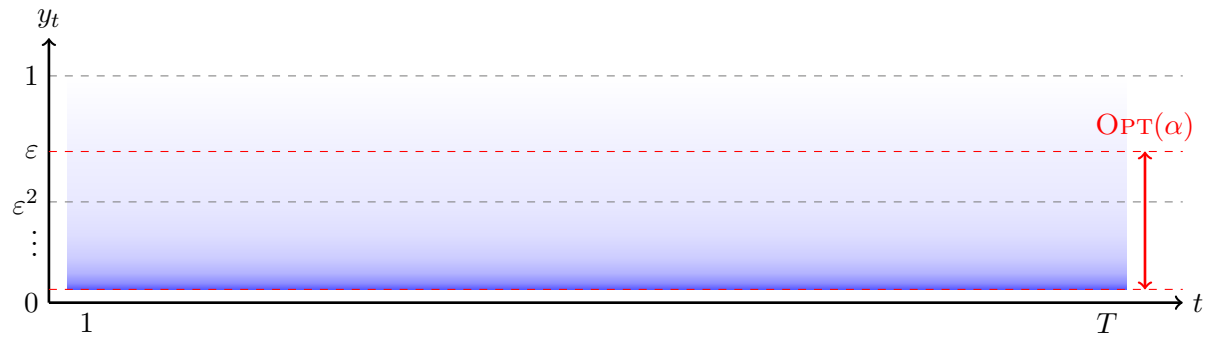
$$\Omega\left(\min\left\{\ln(1/\alpha), \frac{\ln(1/\mathsf{minwidth})}{\ln \mu_{\text{avg}} + \ln(1/\alpha)}\right\}\alpha T\right)$$

mistakes in expectation on some i.i.d. sequence.

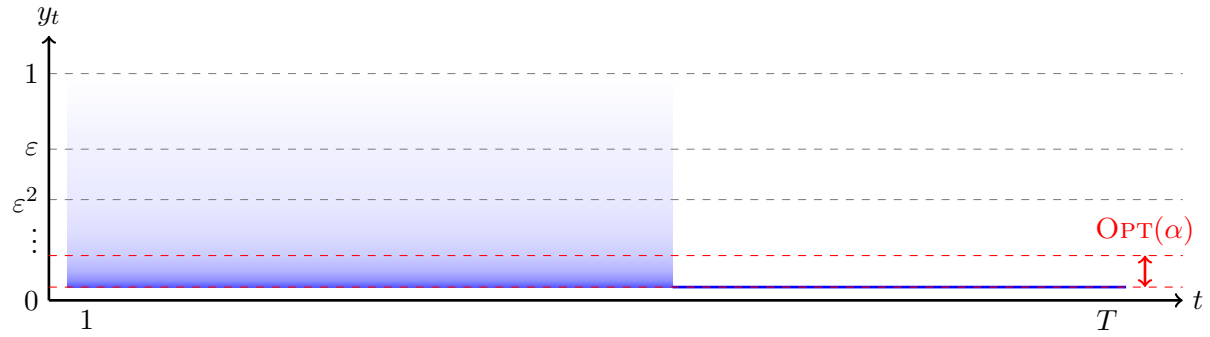
2. *If* $\mathcal{A}$ *plays sets with expected* maximum *volume* $\leq \mu_{\max} \max\{\textsc{Opt}_S(\alpha), \mathsf{minwidth}\}$ *on every arbitrary order sequence* S*, for some* $\mu_{\text{avg}} > 0$*, then* $\mathcal{A}$ *must make*

$$\Omega\left(\min\left\{\ln(1/\alpha), \frac{\ln(1/\mathsf{minwidth})}{\ln \mu_{\max}}\right\}\alpha T\right)$$

mistakes in expectation on some i.i.d. sequence.



(a) In this sequence, every day's input is drawn i.i.d. from $\mathcal{D}$. If an algorithm $\mathcal{A}$ has miscoverage $\geq \frac{1}{2} \cdot \frac{\alpha T}{t}$ on every day, it will make a large number ($\geq \frac{1}{2}\ln(1/\alpha)\alpha T$) of mistakes on this sequence.



(b) In this sequence, the prefix is drawn i.i.d. from $\mathcal{D}$, but later the sequence drops off to be drawn deterministically one value. If the algorithm $\mathcal{A}$ does not play conservatively on the prefix of the sequence, it risks accumulating average volume much larger than $\textsc{Opt}$.

Figure 4: We illustrate the lower bound construction in Theorem 5.1. The horizontal axis is the day t of the sequence and the vertical axis is the value of y_t. y_t is drawn according to a distribution $\mathcal{D}$ that we illustrate with depth of color corresponding to the density of the distribution. $\mathcal{D}$ has the property that the minimum volume interval that achieves miscoverage $\leq \alpha e^i$ over $\mathcal{D}$ is multiplicatively larger than the minimum volume interval that achieves miscoverage $\leq \alpha e^{i+1}$.

Furthermore, this lower bound holds against algorithms that only have to be competitive on sequences where the input on each day is drawn from a symmetric unimodal distribution with median $1/2$.

Proof. Let $0 < \varepsilon \leq 1/2$ and $0 \leq K \leq \ln(1/\alpha)$ be parameters that we will set later. We define a distribution $\mathcal{D}^{(K)}$ over $[0,1]$, such that

$$\begin{aligned} \mathop{\mathbb{P}}_{x\sim\mathcal{D}^{(K)}}\left[x < \varepsilon^{K+1}\right] &= 0, \\ \mathop{\mathbb{P}}_{x\sim\mathcal{D}^{(K)}}\left[x \leq \varepsilon^{1+K-i}\right] &= 1 - \alpha e^{K-i} && \text{for } 0 \leq i \leq K, \\ \mathop{\mathbb{P}}_{x\sim\mathcal{D}^{(K)}}\left[x \leq \varepsilon + t(1-\varepsilon)\right] &= 1 - \alpha(1-t) && \text{for } 0 \leq t \leq 1, \end{aligned} \tag{15}$$

where e is the base of the natural logarithm. Note that $1 - \alpha e^K$ is a valid probability because $\ln(\alpha e^K) = K - \ln(1/\alpha) \leq 0$ so $\alpha e^K \leq 1$.

We compute $v^\star(c)$, which is the minimum volume of any set that achieves coverage c over $\mathcal{D}^{(K)}$.

We take the c.d.f. of $\mathcal{D}^{(K)}$,

$$F_{\mathcal{D}^{(K)}}(x) = \begin{cases} 0 & \text{for } 0 \le x \le \varepsilon^{K+1} \\ 1 - \frac{\alpha}{e} x^{1/\ln \varepsilon} & \text{for } \varepsilon^{K+1} \le x \le \varepsilon \\ 1 - \alpha \frac{1-x}{1-\varepsilon} & \text{for } \varepsilon \le x \le 1 \end{cases}.$$

From the c.d.f. we can compute the p.d.f. of $\mathcal{D}^{(K)}$. For $0 \le x < \varepsilon^{K+1}$, $\mathcal{D}^{(K)}$ has density 0. At $x = \varepsilon^{K+1}$, $\mathcal{D}^{(K)}$ has a point mass of probability $1 - \alpha e^K$. For $\varepsilon^{K+1} < x \le 1$, we compute the density of $\mathcal{D}^{(K)}$ by taking the derivative of the c.d.f. with respect to x:

$$f_{\mathcal{D}^{(K)}}(x) = \begin{cases} \frac{\alpha}{e \ln(1/\varepsilon)} x^{-\frac{1}{\ln(1/\varepsilon)} - 1} & \text{for } \varepsilon^{K+1} < x < \varepsilon \\ \frac{\alpha}{1-\varepsilon} & \text{for } \varepsilon < x \le 1 \end{cases}.$$

This p.d.f. is non-increasing with x for $\varepsilon^{K+1} \le x \le 1$. At $x = \varepsilon^{K+1}$ the p.d.f. is infinite. The first piece is proportional to x to a constant negative power, and the second piece is constant. So it suffices to verify that the density is non-increasing where the two pieces meet. At $x = \varepsilon$, the first piece would give density $\alpha \frac{1/\varepsilon}{\ln(1/\varepsilon)}$. Since $(\ln z) + 1 \le z$ for all $z > 0$,

$$\frac{1}{\varepsilon} - 1 \ge \ln(1/\varepsilon) \quad \Longrightarrow \quad \frac{1/\varepsilon}{\ln(1/\varepsilon)} \ge \frac{1}{1-\varepsilon} \quad \Longrightarrow \quad \alpha \frac{1/\varepsilon}{\ln(1/\varepsilon)} \ge \frac{\alpha}{1-\varepsilon}.$$

Since the density is non-increasing for $x \ge \varepsilon^{K+1}$, and $\mathcal{D}^{(K)}$ has no mass on $x < \varepsilon^{K+1}$, for any coverage level $0 \le c \le 1$, a minimum volume set achieving coverage c over $\mathcal{D}^{(K)}$ is of the form $[\varepsilon^{K+1}, F^{-1}_{\mathcal{D}^{(K)}}(c)]$, so

$$v^\star(c) = F^{-1}_{\mathcal{D}^{(K)}}(c) - \varepsilon^{K+1},$$

(where we define $F^{-1}_{\mathcal{D}^{(K)}}(0) = \varepsilon^{K+1}$). Since the derivative (p.d.f.) of $F_{\mathcal{D}^{(K)}}$ is non-increasing for $\varepsilon^{K+1} \le x \le 1$, $F_{\mathcal{D}^{(K)}}(x)$ is concave over $x \in [\varepsilon^{K+1}, 1]$. So $F^{-1}_{\mathcal{D}^{(K)}}(c)$ is convex over $c \in [0, 1]$, and $v^\star(c)$ is also convex.

The convexity of $v^\star(c)$ along with Jensen's inequality imply the following.

Fact 5.2. *Let* $\mathbf{I}$ *be a (randomized) set of intervals over* $[0, 1]$ *that have average expected coverage* $\ge c$ *over* $\mathcal{D}^{(K)}$*. Then, the average expected volume of the intervals* $\mathbf{I}$ *is*

$$\ge v^\star(c) = F^{-1}_{\mathcal{D}^{(K)}}(c) - \varepsilon^{K+1}.$$

We use $\mathcal{D}^{(K)}$ to design a lower bound instance. Let $S^{(K)}$ be the sequence of T i.i.d. draws from $\mathcal{D}^{(K)}$. Assume also that $\mathsf{minwidth} \le \varepsilon^{K+1}$. We will analyze the behavior of an algorithm $\mathcal{A}$ based on its expected performance on $S^{(K)}$. Fix an algorithm $\mathcal{A}$. Let M be the expected number of mistakes that $\mathcal{A}$ makes on input $S^{(K)}$. Let M_t be the expected miscoverage of $\mathcal{A}$ on day t on input $S^{(K)}$. That is, M_t is the marginal probability that $\mathcal{A}$ does not capture a point on day t of $S^{(K)}$, and $M = M_1 + \cdots + M_t$.

Let $1 \le w \le T$ be a window length that we will set later, and let W_t denote the window (sequence of consecutive days) of length w that ends on day t. If $\mathcal{A}$ has average expected miscoverage $\le \frac{1}{e}\frac{\alpha T}{t}$ over W_t on input $S^{(K)}$, then we say that $\mathcal{A}$ "hits" W_t. Otherwise, $\mathcal{A}$ has average expected miscoverage $> \frac{1}{e}\frac{\alpha T}{t}$ over W_t on input $S^{(K)}$, and we say that $\mathcal{A}$ "misses" W_t.

We analyze by cases. Either $\mathcal{A}$ hits some window W_t for $T/e^{K-1} \le t \le T/e$, or $\mathcal{A}$ misses every window W_t for $T/e^{K-1} \le t \le T/e$.

(a) $\mathcal{A}$ hits some window W_t for $T/e^{K-1} \le t \le T/e$. By definition, this means that $\mathcal{A}$ has average expected coverage $\ge 1 - \frac{1}{e}\frac{\alpha T}{t}$ over W_t. Fact 5.2 tells us that we can bound the expected volume of the intervals that $\mathcal{A}$ plays over W_t using the inverse c.d.f.. We use the original definition of $\mathcal{D}^{(K)}$ (15) for convenience. For $T/e^{K-1} \le t \le T/e$, expected coverage of $1 - \frac{1}{e}\frac{\alpha T}{t}$ corresponds to $i = K + 1 - \ln(T/t)$. Thus, the expected average volume of $\mathcal{A}$ over W_t is

$$\ge \varepsilon^{1+K-i} - \varepsilon^{K+1} \quad = \varepsilon^{\ln(T/t)} - \varepsilon^{K+1}.$$

Consider an alternate input sequence $S_t^{(K)}$, that has the first t days drawn i.i.d. from $\mathcal{D}^{(K)}$, and has future days deterministically set to ε^{K+1}. As $T \to \infty$ (and the sampling error disappears, since $t \ge T/e^{K+1}$ is also going to infinity), $\textsc{Opt}_{S_t^{(K)}}(\alpha)$ will converge to the volume of the smallest interval that achieves coverage $\frac{\alpha T}{t}$ over $\mathcal{D}^{(K)}$. For $T/e^K \le t \le T/e$, we fall into the middle case of (15) and we have

$$\textsc{Opt}_{S_t^{(K)}}(\alpha) \to \varepsilon^{1+\ln(T/t)} - \varepsilon^{K+1} \qquad \text{as } T \to \infty. \tag{16}$$

Let $\text{vol}_{\max}$ be the expected maximum volume of any interval played by $\mathcal{A}$ on $S_t^{(K)}$, and let vol_{avg} be the expected average volume played by $\mathcal{A}$ on $S_t^{(K)}$. By definition, we have that

$$\text{vol}_{\max} \le \mu_{\max} \max\{\textsc{Opt}_{S_t^{(K)}}(\alpha), \mathsf{minwidth}\}, \qquad \text{vol}_{\text{avg}} \le \mu_{\text{avg}} \max\{\textsc{Opt}_{S_t^{(K)}}(\alpha), \mathsf{minwidth}\}.$$

To lower bound the multiplicative ratio, we need to be in the case where $\textsc{Opt}$ is larger than $\mathsf{minwidth}$. Using our assumption that $\mathsf{minwidth} \le \varepsilon^{K+1}$, it suffices to show that

$$\varepsilon^{1+\ln(T/t)} - \varepsilon^{K+1} \ge \varepsilon^{K+1} \quad \iff \quad \varepsilon^{1+\ln(T/t)} \ge 2\varepsilon^{K+1}.$$

Since $\varepsilon \le 1/2$, we have $2\varepsilon^{K+1} \le \varepsilon^K$, and the above is implied by $\varepsilon^{1+\ln(T/t)} \ge \varepsilon^K$. Finally $\ln(T/t) \le K - 1 \iff T/e^{K-1} \le t$ ensures that $\textsc{Opt}_{S_t^{(K)}}(\alpha) \ge \mathsf{minwidth}$. Thus we have

$$\text{vol}_{\max} \le \mu_{\max} \textsc{Opt}_{S_t^{(K)}}(\alpha), \qquad \text{vol}_{\text{avg}} \le \mu_{\text{avg}} \textsc{Opt}_{S_t^{(K)}}(\alpha). \tag{17}$$

Now we analyze the performance of $\mathcal{A}$ on $S_t^{(K)}$. Since $S_t^{(K)}$ is identical to $S^{(K)}$ for the first t days, $\mathcal{A}$ must have the same expected performance over W_t on $S_t^{(K)}$ as it does on $S^{(K)}$. Thus by our earlier bound, we have that the expected average volume of $\mathcal{A}$ over W_t of $S_t^{(K)}$ is $\ge \varepsilon^{\ln(T/t)} - \varepsilon^{K+1}$. We have that

$$\text{vol}_{\max} \ge \varepsilon^{\ln(T/t)} - \varepsilon^{K+1}, \qquad \text{vol}_{\text{avg}} \ge \frac{w}{T}(\varepsilon^{\ln(T/t)} - \varepsilon^{K+1}).$$

Combining with (17) and our bound on $\textsc{Opt}$ (16) we get

$$\mu_{\max} \ge \frac{\varepsilon^{\ln(T/t)} - \varepsilon^{K+1}}{\varepsilon^{1+\ln(T/t)} - \varepsilon^{K+1}} \ge \frac{1}{\varepsilon}, \qquad \mu_{\text{avg}} \ge \frac{w}{T} \cdot \frac{\varepsilon^{\ln(T/t)} - \varepsilon^{K+1}}{\varepsilon^{1+\ln(T/t)} - \varepsilon^{K+1}} \ge \frac{w}{T} \cdot \frac{1}{\varepsilon}. \tag{18}$$

(b) $\mathcal{A}$ misses every window W_t for $T/e^{K-1} \le t \le T/e$. We lower bound the expected number of mistakes $\mathcal{A}$ makes on $S^{(K)}$.

$$M = \sum_{t=1}^{T} M_t \ \ge \sum_{t=w}^{T} \frac{1}{w} \sum_{j=0}^{w} M_{t-j}$$

$$
\begin{aligned}
&\geq \sum_{t=\max\{w,T/e^{K-1}\}}^{T/e} \frac{1}{e}\frac{\alpha T}{t} \\
&= \frac{\alpha T}{e}\left(H_{T/e} - H_{\max\{w,T/e^{K-1}\}-1}\right) && H_n \text{ the } n\text{th harmonic number} \\
&\geq \frac{\alpha T}{e}\left(\ln(T/e) - (\ln(\max\{w, T/e^{K-1}) + 1)\right) \\
&= \frac{1}{e}\left(\min\{\ln(T/ew),\ K-2\}\right) - 1)\,\alpha T. && (19)
\end{aligned}
$$

We set $w/T = e^{-K}$. Then, assuming $\mathsf{minwidth} \leq \varepsilon^{K+1}$, since $K \leq \ln(1/\alpha)$, we have that $\mathcal{A}$ must either have

$$\mu_{\max} \geq \frac{1}{\varepsilon} \quad \text{and } \mu_{\text{avg}} \geq \frac{\alpha}{\varepsilon}$$

from (18), or it must make

$$\geq \frac{1}{e}(K-3)\alpha T$$

mistakes in expectation on the i.i.d. sequence $S^{(K)}$, from (19).

To get a bound based on $\mu_{\max}$, we set $\varepsilon = \frac{1}{\mu_{\max}}$, which is possible as long as $\mu_{\max} \geq 2$. We set K such that $\mathsf{minwidth} = \varepsilon^{K+1}$, which is possible for $\mathsf{minwidth} \geq (1/\mu_{\max})^{\ln(1/\alpha)+1}$. This gives us that $K = \Omega(\ln(1/\mathsf{minwidth})/\ln\mu_{\max})$. As $\mathsf{minwidth}$ gets smaller, the problem only becomes more general, so we have that for any $\mathsf{minwidth} > 0$ and $\mu_{\max} \geq 2$, either $\mathcal{A}$ has maximum multiplicative approximation $\geq \mu_{\max}$ or $\mathcal{A}$ makes

$$\Omega\left(\min\left\{\ln(1/\alpha),\ \frac{\ln(1/\mathsf{minwidth})}{\ln\mu_{\max}}\right\}\alpha T\right)$$

mistakes in expectation on some i.i.d. sequence.

To get a bound based on μ_{avg}, we set $\varepsilon = \frac{\alpha}{\mu_{\text{avg}}}$, which is possible as long as $\mu_{\text{avg}} \geq 2\alpha$. We set K such that $\mathsf{minwidth} = \varepsilon^{K+1}$, which is possible for $\mathsf{minwidth} \geq (\alpha/\mu_{\text{avg}})^{\ln(1/\alpha)+1}$. This gives us that $K = \Omega(\ln(1/\mathsf{minwidth})/(\ln\mu_{\text{avg}} + \ln(1/\alpha)))$. As $\mathsf{minwidth}$ gets smaller, the problem only becomes more general, so we have that for any $\mathsf{minwidth} > 0$ and $\mu_{\text{avg}} \geq 2\alpha$, either $\mathcal{A}$ has average multiplicative approximation $\geq \mu_{\text{avg}}$ or $\mathcal{A}$ makes

$$\Omega\left(\min\left\{\ln(1/\alpha),\ \frac{\ln(1/\mathsf{minwidth})}{\ln\mu_{\max} + \ln(1/\alpha)}\right\}\alpha T\right)$$

mistakes in expectation on some i.i.d. sequence.

Finally, we note that while this argument considered the performance of $\mathcal{A}$ on intervals with one end at ε^{K+1} for ease of notation, we could have instead considered a symmetric version of the sequences, where we map each day ts input that is drawn $y_t \sim \mathcal{D}_t$ to

$$y'_t = \frac{1}{2} + \frac{\text{Unif}\{\pm 1\}}{2}\left(y_t - \varepsilon^{K-1}\right).$$

see for example Figure 5. This leads to the same lower bounds on volume. This also ensures that the input distribution on each day is unimodal and symmetric around median 1/2, because the distributions $\mathcal{D}_t$ in this construction are not supported below ε^{K+1}, and have decreasing density with y increasing from ε^{K+1}. Thus the same argument gives a lower bound against algorithms that are competitive on such sequences. □

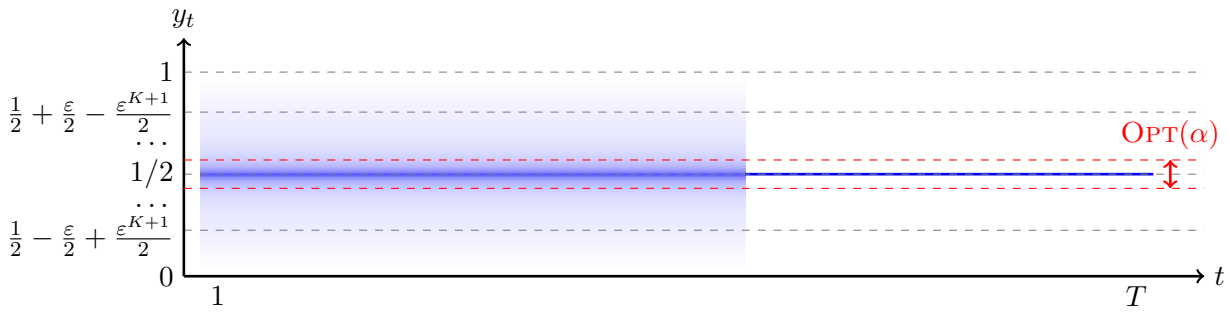


Figure 5: Symmetric version of i.i.d. lower bound construction in Figure 4

6 Uniform Convergence for Exchangeable Sequences

In this section, we show that exchangeability suffices for standard uniform convergence bounds based on VC dimension. We will consider a domain $\mathcal{X}$, and a hypothesis class $\mathcal{H}$ comprised of functions of the form $h : \mathcal{X} \to \{0,1\}$. Given a set of sample points $\{S_1, \ldots, S_T\} \in \mathcal{X}$ and a hypothesis $h \in \mathcal{H}$, we define the loss $\ell(h, \{S_1, \ldots, S_T\})$ of hypothesis h on the samples $\{S_1, \ldots, S_T\}$ to be the fraction of points in $\{S_1, \ldots, S_T\}$ that are mapped to 1 i.e.,

$$\ell\big(h, \{S_1, \ldots, S_T\}\big) = \frac{1}{T}\sum_{i=1}^{T} h(S_i).$$

For an exchangeable sequence of samples $\mathbf{S} = (S_1, \ldots, S_T)$, we will overload notation and define $\ell(h, \mathbf{S}) = \ell(h, \{S_1, \ldots, S_T\})$.

Lemma 6.1 (Uniform Convergence for Exchangeable Sequences)**.** *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ of VC-dimension d. Let $\mathbf{S} = (S_1, \ldots, S_T)$ be an exchangeable sequence of random variables over elements of $\mathcal{X}$. Then there exists a universal constant C such that for any fixed setting of $t \le T, \delta > 0, \varepsilon > 0$ such that*

$$t \ge C\Big(\frac{d\log(d/\varepsilon) + \log(1/\delta)}{\varepsilon^2}\Big),$$

we have that with probability $\ge 1-\delta$ over the exchangeability of $\mathbf{S}$, for all $h \in \mathcal{H}$ simultaneously,

$$\ell(h, \mathbf{S}) - \varepsilon \;\le\; \ell(h, \{S_1, \ldots, S_t\}) \;\le\; \ell(h, \mathbf{S}) + \varepsilon.$$

We follow the standard symmetrization argument for uniform convergence bounds, with some extra care to make them work with exchangeability. We use the following Hoeffding concentration bound for the sum of exchangeable random variables.

Lemma 6.2. *(Hoeffding, 1963) Let $X_1, \ldots, X_n$ be an exchangeable sequence of random variables with $X_i \in [0,1]$ and mean $\mathbb{E}[X_1] = \mu$. Then we have the following upper tail bound*

$$\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \ge \lambda\Big] \le \exp\Big(-n\lambda^2/2\Big),$$

and we get an identical bound the lower tail.

The above lemma is a consequence of Hoeffding's concentration bounds for sampling with replacement (Hoeffding, 1963; Serfling, 1974). This also uses the fact that any finite sequence of exchangeable variables being expressible as a mixture of *urn* sequences (i.e., sampling with replacement) (Diaconis and Freedman, 1980). (See also (Barber, 2024) for concentration bounds for general weighted sums of exchangeable r.v.s.)

We now proceed to the proof of Lemma 6.1.

Proof. First we note from the projection property of exchangeable sequences, the subsequence $\mathbf{S} = (S_1, \dots S_t)$ is also exchangeable. The proof of uniform convergence follows the standard symmetrization approach as given in Bousquet et al. (2004).

Let $\mathbf{S}' = (S'_1, \dots, S'_t)$ be an independent sample from the same exchangeable distribution as $(S_1, \dots, S_t)$. Recall that for a hypothesis $h \in \mathcal{H}$, $\ell(h, \mathbf{S}) = \mathbb{E}_{\mathbf{S}}[\ell(h, (S_1, \dots, S_T))]$ is the expected loss, while $\ell(h, \mathbf{S})$ (and $\ell(h, \mathbf{S}')$) is the empirical loss on the samples $S_1, \dots, S_t$ (respectively $S'_1, \dots, S'_t$).

First from Hoeffding's bound for exchangeable sequences (Lemma 6.2), we have

$$\mathbb{P}\Big[\ell(h, \mathbf{S}) - \ell(h, \mathbf{S}_t) \geq \lambda\Big] \leq \exp\Big(-t\lambda^2/2\Big), \text{ and } \mathbb{P}\Big[\ell(h, \mathbf{S}'_t) - \ell(h, \mathbf{S}) \geq \lambda\Big] \leq \exp\Big(-t\lambda^2/2\Big). \quad (20)$$

We first prove the following claim through symmetrization. The claim allows us to relate the error on one sample to the discrepancy between two independent samples from the distribution.

Claim 6.3. *For any $\lambda > 0$, such that $t\lambda^2 \geq 2$,*

$$\mathbb{P}\Big[\sup_{h \in \mathcal{H}} \ell(h, \mathbf{S}) - \ell(h, \mathbf{S}_t) > \lambda\Big] \leq 2\,\mathbb{P}\Big[\sup_{h \in \mathcal{H}} \ell(h, \mathbf{S}'_t) - \ell(h, \mathbf{S}_t) > \lambda/2\Big] \quad (21)$$

Proof. Let $h^\star \in \mathcal{H}$ be a hypothesis that achieves the supremum on the left side of (21). We will lower bound the probability of the event on the right by showing that it is satisfied for $h^\star$. For any $h \in \mathcal{H}$,

$$\mathbf{1}\Big[\ell(h, \mathbf{S}'_t) - \ell(h, \mathbf{S}_t) > \lambda/2\Big] \geq \mathbf{1}\Big[\big(\ell(h, \mathbf{S}) - \ell(h, \mathbf{S}_t) > \lambda\big) \ \wedge \ \big(\ell(h, \mathbf{S}) - \ell(h, \mathbf{S}'_t) < \lambda/2\big)\Big].$$

Note that $\mathbf{S}'_t$ is independent of $\mathbf{S}_t$. Taking expectation on both sides w.r.t. $\mathbf{S}'_t$,

$$\mathop{\mathbb{P}}_{\mathbf{S}'_t}\Big[\ell(h, \mathbf{S}'_t) - \ell(h, \mathbf{S}_t) > \lambda/2\Big] \geq \mathbf{1}\big[\ell(h, \mathbf{S}_t) - \ell(h, \mathbf{S}) > \lambda\big] \cdot \mathop{\mathbb{P}}_{\mathbf{S}'_t}\Big[\ell(h, \mathbf{S}'_t) - \ell(h, \mathbf{S}) > \lambda/2\Big]. \quad (22)$$

The above inequality holds for any fixed hypothesis $h \in \mathcal{H}$, including any $h^\star$ that achieves the supremum of the left side of (21). Note that $\mathbb{E}_{\mathbf{S}'_t}[\ell(h, \mathbf{S}'_t)] = \ell(h, \mathbf{S})$. The event on the right does not depend on $\mathbf{S}_t$, and follows by applying the Chebyshev inequality on the exchangeable sequence $\mathbf{S}'_t = (S'_1, \dots, S'_t)$. Hence for any fixed hypothesis h (including $h^\star$)

$$\mathop{\mathbb{P}}_{\mathbf{S}'_t}\Big[\ell(h, \mathbf{S}'_t) - \ell(h, \mathbf{S}) > \lambda/2\Big] \leq \frac{\mathrm{Var}[\ell(h, \mathbf{S}'_t)]}{\lambda^2} \leq \frac{1}{4t\lambda^2} \leq \frac{1}{2}.$$

Note that we have a variance upper bound for $\ell(h, \mathbf{S}'_t)$ is due to exchangeability (sampling with replacement). Substituting in (22) and taking expectation w.r.t. $\mathbf{S}_t$, we have

$$\mathbb{P}\Big[\sup_{h \in \mathcal{H}} \ell(h, \mathbf{S}'_t) - \ell(h, \mathbf{S}_t) > \lambda/2\Big] \geq \mathop{\mathbb{P}}_{\mathbf{S}_t}\Big[\sup_{h \in \mathcal{H}} \ell(h, \mathbf{S}_t) - \ell(h, \mathbf{S}) > \lambda\Big] \times \frac{1}{2}.$$

By rearranging terms, we get the claim. □

We now upper bound the right side of (21). For a fixed $h \in \mathcal{H}$, we have from (20)

$$\mathop{\mathbb{P}}_{\mathbf{S}_t, \mathbf{S}'_t}\Big[\ell(h, \mathbf{S}'_t) - \ell(h, \mathbf{S}_t) \geq \frac{\lambda}{2}\Big] \leq \mathbb{P}\Big[\ell(h, \mathbf{S}'_t) - \ell(h, \mathbf{S}) \geq \frac{\lambda}{4}\Big] + \mathbb{P}\Big[\ell(h, \mathbf{S}) - \ell(h, \mathbf{S}_t) \geq \frac{\lambda}{4}\Big] \leq 2\exp\Big(-\frac{t\lambda^2}{32}\Big).$$

Now we observe that for a given hypothesis $h \in \mathcal{H}$, $\ell(h, \mathbf{S}'_t) - \ell(h, \mathbf{S}_t)$ is an empirical sum and only depends on the $2n$ samples $S_1, \ldots, S_t, S'_1, \ldots, S'_t$. Hence the number of distinct hypothesis that we need to union bound over is at most the shattering number $N(2t, d) = \sum_{j=1}^{d} \binom{2t}{d}$. Hence

$$\begin{aligned}\mathbb{P}\Big[\sup_{h\in\mathcal{H}} \ell(h, \mathbf{S}) - \ell(h, \mathbf{S}_t) > \lambda\Big] &\leq 2 \mathop{\mathbb{P}}_{\mathbf{S}_t, \mathbf{S}'_t}\Big[\sup_{h\in\mathcal{H}} \ell(h, \mathbf{S}'_t) - \ell(h, \mathbf{S}_t) \geq \frac{\lambda}{2}\Big] \\ &\leq 4N(2t, d) \cdot \exp\Big(-\frac{t\lambda^2}{32}\Big) \leq \delta,\end{aligned}$$

for our choice of t and λ. This finishes the proof. □

Acknowledgements

I thank my advisor Aravindan Vijayaraghavan for helpful discussions, and for the proof that uniform convergence extends to exchangeable sequences.

References

Anastasios Angelopoulos, Emmanuel Candes, and Ryan J Tibshirani. Conformal pid control for time series prediction. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 23047–23074. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/47f2fad8c1111d07f83c91be7870f8db-Paper-Conference.pdf.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction, 2025. URL https://arxiv.org/abs/2411.11824.

Rina Foygel Barber. Hoeffding and bernstein inequalities for weighted sums of exchangeable random variables. *Electronic Communications in Probability*, 2024. URL https://api.semanticscholar.org/CorpusID:269009938.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 2022. URL https://api.semanticscholar.org/CorpusID:247158820.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, 2020.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. *Introduction to Statistical Learning Theory*, pages 169–207. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-28650-9. doi: 10.1007/978-3-540-28650-9_8. URL https://doi.org/10.1007/978-3-540-28650-9_8.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 42.1–42.23, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL https://proceedings.mlr.press/v23/bubeck12b.html.

Persi Diaconis and David A. Freedman. Finite exchangeable sequences. *Annals of Probability*, 8: 745–764, 1980. URL https://api.semanticscholar.org/CorpusID:119580521.

Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Achieving risk control in online learning settings. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=5Y04GWvoJu.

Chao Gao, Liren Shan, Vaidehi Srinivas, and Aravindan Vijayaraghavan. Volume optimality in conformal prediction with structured prediction sets, 2025. URL https://arxiv.org/abs/2502.16658.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0d441de75945e5acbc865406fc9a2559-Paper.pdf.

Isaac Gibbs and Emmanuel Candès. Conformal inference for online prediction with arbitrary distribution shifts. *J. Mach. Learn. Res.*, 25(1), March 2025. ISSN 1532-4435.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830.

Rafael Izbicki, Gilson Shimizu, and Rafael Stern. Flexible distribution-free conditional predictive bands using density estimators. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3068–3077. PMLR, 26–28 Aug 2020a. URL https://proceedings.mlr.press/v108/izbicki20a.html.

Rafael Izbicki, Gilson T. Shimizu, and Rafael Bassi Stern. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *J. Mach. Learn. Res.*, 23:87:1–87:32, 2020b. URL https://api.semanticscholar.org/CorpusID:238354408.

Jing Lei, James Robins, and Larry Wasserman and. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013. doi: 10.1080/01621459.2012.751873.

Drew Prinster, Samuel Stanton, Anqi Liu, and Suchi Saria. Conformal validity guarantees exist for any data distribution (and how to find them). In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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

Robert Serfling. Probability inequalities for the sum in sampling without replacement. *Annals of Statistics*, 2:39–48, 1974. URL https://api.semanticscholar.org/CorpusID:120916609.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. *Conformal prediction under covariate shift*. Curran Associates Inc., Red Hook, NY, USA, 2019.

V. Vovk. On-line confidence machines are well-calibrated. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 187–196, 2002. doi: 10.1109/SFCS.2002.1181895.

Margaux Zaffran, Olivier Feron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25834–25866. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/zaffran22a.html`.