High-confidence predictions under adversarial uncertainty

High-Confidence Predictions under Adversarial Uncertainty

Andrew Drucker*

Abstract

We study the setting in which the bits of an unknown infinite binary sequence x are revealed sequentially to an observer. We show that very limited assumptions about x allow one to make successful predictions about unseen bits of x. First, we study the problem of successfully predicting a single 0 from among the bits of x. In our model we have only one chance to make a prediction, but may do so at a time of our choosing. This model is applicable to a variety of situations in which we want to perform an action of fixed duration, and need to predict a "safe" time-interval to perform it.

Letting N_t denote the number of 1s among the first t bits of x, we say that x is "ε-weakly sparse" if $\liminf(N_t/t) \leq \varepsilon$. Our main result is a randomized algorithm that, given any ε-weakly sparse sequence x, predicts a 0 of x with success probability as close as desired to $1-\varepsilon$. Thus we can perform this task with essentially the same success probability as under the much stronger assumption that each bit of x takes the value 1 independently with probability ε.

We apply this result to show how to successfully predict a bit (0 or 1) under a broad class of possible assumptions on the sequence x. The assumptions are stated in terms of the behavior of a finite automaton M reading the bits of x. We also propose and solve a variant of the well-studied "ignorant forecasting" problem. For every $\varepsilon > 0$, we give a randomized forecasting algorithm S_ε that, given sequential access to a binary sequence x, makes a prediction of the form: "A p fraction of the next N bits will be 1s." (The algorithm gets to choose p, N, and the time of the prediction.) For any fixed sequence x, the forecast fraction p is accurate to within $\pm\varepsilon$ with probability $1-\varepsilon$.

1 Introduction

Suppose that the bits of an unknown infinite binary sequence $x = (x_1, x_2, \ldots)$ are revealed to us sequentially, and our goal is to make a nontrivial prediction about unseen bits. As a canonical example (which we will study closely), suppose we wish to make a single, successful prediction that some unseen bit of our choosing will be 0. This generic "0-prediction" task is applicable in many settings. In particular, it applies whenever we are trying to predict some "safe" time to perform some action of unit-duration, based on past observations: here, $[x_t = 0]$ represents safe conditions during the t^{th} possible time-slot for our action, while $[x_t = 1]$ represents dangerous (unacceptable) conditions. Note that we model time as discrete, and model "safety" as an all-or-nothing matter.

*Institute for Advanced Study, Princeton NJ. Email: andy.drucker@gmail.com. The author is currently supported by the National Science Foundation, under grants CCF-0832797, Sub-Contract No. 00001583, and DMS-0835373. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This research was conducted as a Ph.D. student at Massachusetts Institute of Technology, EECS Dept., and was supported by a DARPA YFA grant of Scott Aaronson.

If after observing $x_1, \ldots, x_{t-1}$ we predict "$x_t = 0$," and our prediction is false, we regard this as a catastrophic failure. Similarly, if we observe the entire infinite sequence, without ever announcing a 0-prediction, this is also regarded as a failure.

We ask: under what assumptions on the sequence x can we make a correct 0-prediction and perform our action safely? Obviously, if x is all-1s then we cannot, so we must make some assumption. One natural approach to this kind of situation is to assume the sequence x is generated according to some *probabilistic model*. For example, we might assume that each bit represents the outcome of an independent coin toss with some fixed bias p. More complicated probabilistic assumptions, involving dependence between the bits, can also be considered.

However, in applications we may be unlikely to have a detailed idea of how the bits of x are generated. It may be that rather than having a probabilistic model in mind, we merely know or conjecture some *constraint* obeyed by x. We then ask whether there exists a strategy which allows successful 0-prediction (at least, with sufficiently high probability), for *any* sequence obeying the constraint. This model will be our focus in the present paper.

For example, suppose that based on initial observations, the bits of x seem *individually* to equal 1 with probability at most .05, but that we suspect they are not fully independent. In such cases, we may make the weaker assumption that the limiting density of 1s is at most .05. Note that this condition holds with probability 1 if the bits are generated by independent .05-biased trials, so our limiting-density constraint can be considered a natural *relaxation* of this simple probabilistic model. We may then ask whether there exists a strategy that allows a successful 0-prediction with success probability nearly .95 under this relaxed assumption. (Happily, the answer is Yes; this will follow from our main result.)

1.1 Relation to previous work

Our work studies *prediction under adversarial uncertainty*. In such problems, an observer tries to make predictions about successive states of nature, without assuming that these states are governed by some known probability distribution. Instead, nature is regarded as an *adversary* who makes choices in an attempt to thwart the observer's prediction strategy. The focus is on understanding what kinds of predictions can be made under very limited assumptions about the behavior of nature.

Adversarial prediction is a broad topic, but two strands of research are particularly related to our work. The first strand is the study of *gales* and their relatives. Gales are a class of betting systems generalizing martingales; their study is fundamental for the theory of *effective dimension* in theoretical computer science (see [Hem05] for a survey). The basic idea is as follows. An infinite sequence x is chosen from some known subset A of the space $\{0,1\}^\omega$ of infinite binary sequences. A gambler is invited to gamble on predicting the bits of x as they are sequentially revealed; the gambler has a finite initial fortune and cannot go into debt. The basic question is, for which subsets A can the gambler be guaranteed long-term success in gambling, for any choice of $x \in A$? This question can be studied under different meanings of "success" for the gambler, and under more- or less-favorable classes of bets offered by the casino.

Intuitively, the difficulty of gambling successfully on an unknown $x \in A$ is a measure of the "largeness" of the set A. In fact, this perspective was shown to yield new characterizations of two important measures of fractal dimensionality. Lutz [Lut03a] gave a characterization of the Hausdorff dimension of subsets of $\{0,1\}^\omega$ in terms of gales, while Athreya, Hitchcock, Lutz, and Mayordomo [AHLM07] showed a gale characterization of the packing dimension. These works also investigated gales with a requirement that the gambler follows a computationally bounded betting

strategy; using such gales, the authors explored new notions of "effective dimension" for complexity classes in computational complexity theory.[1]

The second strand of related work is the so-called *forecasting problem* in decision theory (see [Daw82] for an early, influential discussion). In this problem, an infinite binary sequence $x \in \{0,1\}^\omega$ is once again revealed sequentially; we typically think of the t-th bit as indicating whether it rained on the t-th day at some location of interest. Each day a weather forecaster is asked to give, not an absolute prediction of whether it will rain tomorrow, but instead some estimate of the *probability* of rain tomorrow. In order to keep his job, the forecaster is expected to make forecasts which are *calibrated*: roughly speaking, this means that if we consider all the days for which the forecaster predicted some probability p of rain, about a p fraction turn out rainy (see [FV98] for more precise definitions).

In the adversarial setting, a forecaster must make such forecasts without knowledge of the probability distribution governing nature. An extreme case is the well-studied "ignorant forecaster" model, in which the forecaster is allowed *no assumptions whatsoever* about the sequence x. It is a remarkable fact, shown by Foster and Vohra [FV98], that there exists a randomized ignorant forecasting scheme whose forecasts are calibrated in the limit.

This result was extended by Sandroni [San03]. The calibration criterion is just one of many conceivable "tests" with which we might judge a forecaster's knowledge on the basis of his forecasts and the observed outcomes. Foster and Vohra's result showed that the calibration test can be passed even by an ignorant forecaster; but conceivably some other test of knowledge could be more meaningful. A reasonable class of tests to consider are those that can be passed with some high probability $1 - \varepsilon$ by a forecaster who knows the actual distribution $\mathcal{D}$ governing nature, for any possible setting of $\mathcal{D}$. However, Sandroni showed that any such test can *also* be passed with probability $1 - \varepsilon$ by an ignorant forecaster! Fortnow and Vohra [FV09] give evidence that the ignorant strategies provided by Sandroni's result cannot in general be computed in polynomial time, even if the test is polynomial-time computable.[2]

In both of the strands of research described above, researchers have typically looked for prediction schemes that have some desirable *long-term, aggregate* property. In the gale setting, the focus is on betting strategies that may lose money on certain bets, but that succeed in the limit; in the forecasting problem, an ignorant forecaster wants his forecasts to appear competent overall, but is not required to give definite predictions of whether or not it will rain on any given day. By contrast, in our 0-prediction problem, we want to perform an action successfully just once, and we stake everything on the outcome. Our focus is on making a *single prediction*, with success probability as close to 1 as possible.[3]

In a later section of the paper we will also study a variant of the ignorant forecasting scenario.

[1] Computationally bounded betting and prediction schemes have also been used to study *individual* sequences x, rather than sets of sequences. This approach has been followed using various resource bounds and measures of predictive success; see, e.g., [MF98, Lut03b].

[2] The tests considered in [San03, FV09] are required to halt with an answer in finite time. See [FV09] for references to work in which this restriction is relaxed.

[3] The distinction between our single-prediction setting and multiple-prediction models is not absolute, however. In particular, as will be apparent from our work, the algorithms we develop can also be used to make multiple predictions, with provable success guarantees on certain classes of sequences. The single-prediction model we study can also be re-interpreted as a particular type of gambling system making bets at every step. Namely, single-prediction strategies correspond, in a natural way, to gambling schemes in a casino where reinvestment of winnings is disallowed. This connection, and its relation to certain work in *regret-bounded* prediction algorithms, e.g. [KP11], will be discussed in Section 4.5.

Following [FV98, San03], we will make no assumption about the observation sequence x. Our goal will be to make a single forecast at a time of our choosing, of the following form: "A p fraction of the next N observations will take the value 1." We will seek to maximize the accuracy of our prediction, as well as the likelihood of falling within the desired accuracy. This forecasting variant is conceptually linked to our 0-prediction problem by its focus on making a single prediction with high confidence.

1.2 Our results on the 0-prediction problem

To appreciate the kinds of prediction-strategies that are possible, let us first consider a simple but instructive example. Suppose we know that *at most one* bit will ever equal 1. Even under this very-restrictive assumption, it is not hard to see that any *deterministic* 0-prediction strategy must fail on some sequences (either by making an incorrect prediction, or by waiting forever to see a 1 that never arrives). Thus, we are naturally led to consider a *randomized* strategy. Fixing some $\delta > 0$, consider the strategy which chooses a value $t^\star \in \{1, 2, \ldots, \lceil 1/\delta \rceil\}$ uniformly, and makes a 0-prediction at time $t^\star$. One can verify that this strategy fails with probability at most $\lceil 1/\delta \rceil^{-1} \leq \delta$.

Note that this error probability is over the randomness in the *algorithm*, not the sequence x; we regard the sequence as chosen by an adversary who knows the prediction strategy, but *not* the outcomes of the strategy's randomness. We are interested in strategies which succeed with high probability against any choice by the adversary (obeying the assumed constraint).

An easy modification of the above algorithm lets the us succeed with probability $1 - \delta$ against a sequence promised to contain at most M 1s, for any *fixed* $M < \infty$. However, it may come as a surprise that we can succeed in the 0-prediction task with probability $1 - \delta$ under much weaker assumptions. For example, we can do so under the assumption that the number of 1s is merely *finite*, with no upper bound M known in advance. In fact, we can handle an infinite number of 1s and still make a 0-prediction successfully, under the assumption that their *limiting density* is 0; that is, under the assumption that $\lim_{t\to\infty} \frac{1}{t} \sum_{1\leq i\leq t} x_i = 0$.

For any $\varepsilon > 0$, say that a sequence is ε-*weakly sparse* if

$$\lim_{s\to\infty} \inf_{t\geq s} \frac{1}{t} \sum_{1\leq i\leq t} x_i \leq \varepsilon.$$

Our main result on 0-prediction is that there is a prediction strategy $S = S_\varepsilon$ that makes a 0-prediction with successfully with probability as close as desired to $1-\varepsilon$, under the assumption that the sequence x given is ε-weakly sparse. (Simple examples show that this is optimal.) We state our result formally in Section 2, after setting up the necessary definitions. This result easily implies our claim that successful 0-predictions can be performed on sequences with finitely many 1s or with limiting density 0, although these special cases can also be handled more simply.

We feel that the techniques used to prove this result are of independent interest, and could find other applications. The basic idea of our 0-prediction strategy is easy to state. The prediction strategy maintains a stack of "chips;" observing 0s increases the stack height, while observing 1s decreases it. The height of the stack at a given time reflects the algorithm's "courage," and determines its likelihood to predict a 0. While this basic approach is intuitive, implementing it correctly and proving the strategy's correctness is a delicate task. Our analysis involves a careful study of individual chips' contributions to the success and failure probabilities.

Our result bears some resemblance to known results in dimension theory. Let $A_{\varepsilon-ws} \subseteq \{0,1\}^\omega$ denote the set of ε-weakly sparse infinite binary sequences. Eggleston [Egg49, Bil65] showed that

for $\varepsilon \leq 1/2$, the Hausdorff dimension of $A_{\varepsilon-ws}$ is equal to the binary entropy $H(\varepsilon)$. More recently, Lutz [Lut03a] gave an alternative proof using his gale characterization of Hausdorff dimension (Lutz also calculated the "effective dimension" of $A_{\varepsilon-ws}$ according to several definitions). Lutz upper-bounds the Hausdorff dimension of $A_{\varepsilon-ws}$ by giving a gale betting strategy that "succeeds" (in the appropriate sense) against all $x \in A_{\varepsilon-ws}$. This betting strategy, which is simple and elegant, does not appear to be applicable to our problem.

1.3 Further results

In Section 4, we prove a variant of our result on 0-prediction, in a a modified setting in which we are allowed to predict *either* a 0 or a 1. We define a condition on the binary sequence x that is significantly more general than ε-weak sparsity as defined in Section 1.2, and that still allows a bit to be predicted with high confidence. The condition is stated in terms of a finite automaton M that reads x: we assume that x causes M to enter a designated set of "bad" states B only infrequently. A certain "strong accessibility" assumption on the set B is needed for our result. Also, we caution that our algorithm's success guarantee in this problem is not as quantitatively strong as our result on 0-prediction; this is unavoidable, as will be shown.

Next, in Section 5, we study a problem closely related to the "ignorant forecasting" problem discussed earlier, where (as in the 0-prediction problem) a single prediction is to be made. In the "density prediction game," an arbitrary infinite binary sequence is chosen by Nature, and its bits are revealed to us sequentially. Our goal is to make a single forecast of the form

"A p fraction of the next N bits will be 1s."

We are allowed to choose p, N, and the time at which we make our forecast.

Fixing a binary sequence x, we say that a forecast described by (p, N), and made after viewing x_t, is *ε-successful* on x if the fraction of 1s among $x_{t+1}, \ldots, x_{t+N}$ is in the range $(p - \varepsilon, p + \varepsilon)$. For $\delta, \varepsilon > 0$, we say that a (randomized) forecasting strategy $\mathcal{S}$ is *(δ, ε)-successful* if for every $x \in \{0,1\}^\omega$,

$$\Pr[\mathcal{S} \text{ is } \varepsilon\text{-successful on } x] \geq 1 - \delta.$$

In Section 5 we show the following, perhaps surprising, result:

Theorem 1. *For any $\delta, \varepsilon > 0$, there exists a (δ, ε)-successful forecasting strategy.*

The proof uses a (seemingly folklore) technique from the analysis of martingales. My understanding of this technique benefited greatly from conversations with Russell Impagliazzo.

2 Preliminaries and the Main Theorem

First we develop a formal basis to state and prove our main result on 0-prediction. $\mathbb{N} = \{1, 2, \ldots\}$ denotes the positive whole numbers. For $N \in \mathbb{N}$, $[N]$ denotes the set $\{1, 2, \ldots, N\}$. $\{0,1\}^\omega$ denotes the set of all infinite bit-sequences $b = (b_1, b_2, \ldots)$.

A *0-prediction strategy* is a collection

$$\mathcal{S} = \{\pi_{\mathcal{S},b} : b \in \{0,1\}^\omega\},$$

where each $\pi_{\mathcal{S},b}$ is a probability distribution over $\mathbb{N} \cup \{\infty\}$. We require that for all $b = (b_1, b_2, \ldots), b' = (b'_1, b'_2, \ldots)$, and all $i \in \mathbb{N}$,

$$(b_1, \ldots, b_{i-1}) = (b'_1, \ldots, b'_{i-1}) \Rightarrow \pi_{\mathcal{S},b}(i) = \pi_{\mathcal{S},b'}(i). \tag{1}$$

That is, $\pi_{\mathcal{S},b}(i)$ depends only on $b_1, \ldots, b_{i-1}$.

Let us interpret the above definition. A 0-prediction strategy defines, for each sequence b and each $i \in \mathbb{N}$, a probability $\pi_{\mathcal{S},b}(i)$ that, when facing the sequence b, the predictor will make the prediction "$b_i = 0$." There is also some probability $\pi_{\mathcal{S},b}(\infty)$ that the predictor will wait forever without making a prediction. Whether it succeeds or fails, the strategy only makes a prediction at most once, so these probabilities sum to 1. Eq. (1) requires that the decision whether to predict a 0 at position i depends only upon what it has seen of the sequence during the first $(i-1)$ steps. The strategies we analyze in this paper will be defined in such a way that Eq. (1) obviously holds.

Given a 0-prediction strategy $\mathcal{S}$, define the *success probability*

$$\mathrm{Suc}(\mathcal{S}, b) := \sum_{i \in \mathbb{N}: b_i = 0} \pi_{\mathcal{S},b}(i)$$

as the probability that, facing b, the strategy makes a successful 0-prediction. Similarly, define the *false-guess probability*

$$\mathrm{False}(\mathcal{S}, b) := \sum_{i \in \mathbb{N}: b_i = 1} \pi_{\mathcal{S},b}(i) = 1 - \mathrm{Suc}(\mathcal{S}, b) - \pi_{\mathcal{S},b}(\infty)$$

as the probability that the strategy $\mathcal{S}$ leads to an incorrect 0-prediction on sequence b. For a subset $A \subseteq \{0,1\}^\omega$, define

$$\mathrm{Suc}(\mathcal{S}, A) := \inf_{b \in A} \mathrm{Suc}(\mathcal{S}, b).$$

We can now formally state our main result on 0-prediction:

Theorem 2. *Fix $\varepsilon \in (0,1)$ and let $A_{\varepsilon-ws} := \{b : b$ is $\varepsilon-$weakly sparse$\}$. Then for all $\gamma > 0$, there exists a strategy $\mathcal{S}_{\varepsilon,\gamma}$ such that*

$$\mathrm{Suc}(\mathcal{S}_{\varepsilon,\gamma}, A_{\varepsilon-ws}) > 1 - \varepsilon - \gamma.$$

Furthermore, $\mathcal{S}_{\varepsilon,\gamma}$ has the following "safety" property: for any *sequence $b \in \{0,1\}^\omega$, the false-guess probability* $\mathrm{False}(\mathcal{S}, b)$ *is at most $\varepsilon + \gamma$.*

It is not hard to see that Theorem 2 is optimal for 0-prediction strategies against $A_{\varepsilon-ws}$. For consider a randomly generated sequence $\boldsymbol{b}$ where the events $[\boldsymbol{b}_i = 1]$ occur independently, with $\mathbb{E}[\boldsymbol{b}_i] = \min\{1, \varepsilon + 2^{-i}\}$. Then $[\lim_{t\to\infty}(\boldsymbol{b}_1 + \ldots + \boldsymbol{b}_t)/t = \varepsilon]$ occurs with probability 1. On the other hand, any 0-prediction strategy $\mathcal{S}$ has success probability less than $1 - \varepsilon$ against $\boldsymbol{b}$. Thus, for any $\mathcal{S}$ we can find a particular sequence b for which $\lim_{t\to\infty}(b_1 + \ldots + b_t)/t = \varepsilon$ and which causes $\mathcal{S}$ to succeed with probability less than $1 - \varepsilon$.

3 Proof of Theorem 2

First we observe that, if we can construct a strategy $\mathcal{S}$ such that $\mathrm{Suc}(\mathcal{S}, A_{\varepsilon-ws}) > 1 - \varepsilon - \gamma$, then the "safety" property claimed for $\mathcal{S}$ in the theorem statement will follow immediately. For suppose

to the contrary that some sequence $b \in \{0,1\}^\omega$ satisfies $\mathrm{False}(\mathcal{S}, b) > \varepsilon + \gamma$. Then there exists $m \in \mathbb{N}$ such that $\sum_{i \le m : b_i = 1} \pi_{\mathcal{S},b}(i) > \varepsilon + \gamma$. If we define $b' \in \{0,1\}^\omega$ by

$$b'_i := \begin{cases} b_i & \text{if } i \le m, \\ 0 & \text{if } i > m, \end{cases}$$

then $b' \in A_{\varepsilon - ws}$ and $\mathrm{False}(\mathcal{S}, b') > \varepsilon + \gamma$, contradicting our assumption on $\mathcal{S}$.

To construct the strategy $\mathcal{S}$, we use a family of 0-prediction strategies for attempting to make a 0-prediction within a finite, bounded interval of time. The following lemma is our key tool, and is interesting in its own right.

Lemma 1. *For any $\delta \in (0,1)$ and integer $K > 1$, there exists a strategy $\mathcal{T} = \mathcal{T}_{K,\delta}$ such that for all $b \in \{0,1\}^\omega$:*

(i) The 0-prediction time of $\mathcal{T}$ is always in $[K] \cup \{\infty\}$. That is, for $K < i < \infty$, we have $\pi_{\mathcal{T},b}(i) = 0$;

(ii) If $(b_1 + \ldots + b_{K-1})/(K-1) \le \delta' < \delta$, then $\pi_{\mathcal{T},b}(\infty) \le 1 - \Omega((\delta - \delta')^2/\delta)$;

(iii) The false-guess probability satisfies

$$\mathrm{False}(\mathcal{T}, b) \le \frac{\delta}{1-\delta} \mathrm{Suc}(\mathcal{T}, b) + O\left(\frac{\delta}{(1-\delta)K}\right).$$

We defer the proof of Lemma 1, and use it to prove Theorem 2.

of Theorem 2. Fix settings of $\varepsilon, \gamma > 0$; we may assume $\varepsilon + \gamma < 1$, or there is nothing to prove. Let $\varepsilon_1 := \varepsilon + \gamma/3, \varepsilon_2 := \varepsilon + 2\gamma/3$. We also use a large integer $K > 1$, to be specified later. Divide $\mathbb{N}$ into a sequence of intervals $I_1 = \{1, 2, \ldots, K\}, I_2 = \{K+1, \ldots, 5K\}$, and so on, where I_r has length $r^2 K$.

Let $\mathcal{S} = \mathcal{S}_{\varepsilon,\gamma}$ be the 0-prediction strategy which does the following: first, follow the strategy $\mathcal{T}_{K,\varepsilon_2}$ (as given by Lemma 1) during the time interval I_1. If no 0-prediction is made during these steps, then run the strategy $\mathcal{T}_{4K,\varepsilon_2}$ on the interval I_2, after shifting the indices of I_2 appropriately (so that $\mathcal{T}_{4K,\varepsilon_2}$ considers its input sequence to begin on b_{K+1}). Similarly, for each $r > 0$, if we reach the interval I_r without a 0-prediction, we execute the strategy $\mathcal{T}_{r^2 K,\varepsilon_2}$ on the interval I_r, after shifting indices appropriately.

We will show that if K is sufficiently large, we have $\mathrm{Suc}(\mathcal{S}, A_{\varepsilon - ws}) > 1 - \varepsilon - \gamma$ as required. Fix any $b = (b_1, b_2, \ldots) \in A_{\varepsilon - ws}$. Let $\alpha_r := \left(\sum_{i \in I_r} b_i\right) / |I_r|$ be the fraction of 1-entries in b during interval I_r. We will use the following easy claim:

Claim 1. *For infinitely many r, $\alpha_r \le \varepsilon_1$.*

Proof. Suppose to the contrary that $\alpha_r > \varepsilon_1$ when $r \ge R$. Consider an interval $\{1, 2, \ldots, M\}$ large enough to properly contain $I_1, I_2, \ldots, I_R$. Let $t \ge R$ be such that $I_t \subseteq [M]$ but that $I_{t+1} \not\subseteq [M]$. Let $\alpha^\star$ be the fraction of 1-entries in $[M] \cap I_{t+1}$; we set $\alpha^\star := 0$ if $[M] \cap I_{t+1} = \emptyset$. With $N_M = (b_1 + \ldots + b_M)$, we have the expression

$$\frac{N_M}{M} = \sum_{r \le t} \frac{|I_r|}{M} \cdot \alpha_r + \frac{|[M] \cap I_{t+1}|}{M} \cdot \alpha^\star$$

which expresses the 1-density (fraction of 1s) in b in the positions $\{1,\ldots,M\}$ as a weighted average of the 1-densities in $I_1,\ldots,I_t$ and in $[M]\cap I_{t+1}$.

Note that

$$\frac{|[M]\cap I_{t+1}|}{M} \le \frac{|I_{t+1}|}{M} \le \frac{(t+1)^2K}{\sum_{r\le t} r^2K} = O(1/t) \to 0,$$

as $M\to\infty$. Now $\alpha_r > \varepsilon_1$ when $r \ge R$, so for sufficiently large M we have $N_M/M \ge (\varepsilon_1+\varepsilon)/2 > \varepsilon$. But this contradicts the fact that $b \in A_{\varepsilon-ws}$, proving the Claim. □

Fix $r>0$. If $I_r=\{j,\ldots,k\}$ and $\alpha_r = (b_j+\ldots+b_k)/(k-j+1) \le \varepsilon_1$, then we also have

$$(b_j+\ldots+b_{k-1})/(k-j) < \varepsilon+\gamma/2 \tag{2}$$

if r is large enough. Say that the interval I_r is *good* if Eq. (2) holds; it follows from Claim 1 that there are infinitely many good intervals.

Condition (ii) of Lemma 1 tells us that if our 0-prediction strategy $\mathcal{S}$ reaches a good interval I_r, it will make a 0-prediction during I_r with probability $\Omega((\gamma/6)^2/\varepsilon_2)$. Thus for $L>0$, the probability that $\mathcal{S}$ passes through L distinct good intervals $I_{r_1},\ldots,I_{r_L}$ without making a 0-prediction is at most $(1-\Omega((\gamma/6)^2/\varepsilon_2))^L$, which approaches 0 as $L\to\infty$. It follows that the strategy eventually makes a 0-prediction with probability 1, and that $\mathrm{False}(\mathcal{S},b) = 1-\mathrm{Suc}(\mathcal{S},b)$.

For $r>0$, let $P_r = P_r(b)$ be defined as the probability that $\mathcal{S}$ reaches I_r without making a 0-prediction earlier (about some position occurring before I_r). Let $b[I_r]$ denote the sequence b, shifted to begin at the first bit of I_r. Then we can reexpress the false-guess probability of $\mathcal{S}$ on b, and bound this quantity, as follows:

$$\begin{aligned}
\mathrm{False}(\mathcal{S},b) &= \sum_{r\ge1} P_r \cdot \mathrm{False}(\mathcal{T}_{r^2K,\varepsilon_2}, b[I_r]) \\
&\le \sum_{r\ge1} P_r \cdot \left(\frac{\varepsilon_2}{1-\varepsilon_2}\,\mathrm{Suc}(\mathcal{T}_{r^2K,\varepsilon_2}, b[I_r]) + O\left(\frac{\varepsilon_2}{(1-\varepsilon_2)r^2K}\right)\right) \\
&\quad \text{(by condition (iii) of Lemma 1)} \\
&= \frac{\varepsilon_2}{1-\varepsilon_2}\left(\sum_{r\ge1} P_r\cdot \mathrm{Suc}(\mathcal{T}_{r^2K,\varepsilon_2}, b[I_r])\right) + O\left(\frac{\varepsilon_2}{(1-\varepsilon_2)K}\right) \\
&\quad \text{(using the fact that } \sum_{r>0} r^{-2} < \infty) \\
&= \frac{\varepsilon_2}{1-\varepsilon_2}\,\mathrm{Suc}(\mathcal{S},b) + O\left(\frac{\varepsilon_2}{(1-\varepsilon_2)K}\right).
\end{aligned}$$

Thus, $\mathrm{False}(\mathcal{S},b) = 1-\mathrm{Suc}(\mathcal{S},b) \le \frac{\varepsilon_2}{1-\varepsilon_2}\,\mathrm{Suc}(\mathcal{S},b) + O\left(\frac{\varepsilon_2}{(1-\varepsilon_2)K}\right)$, which implies

$$\mathrm{Suc}(\mathcal{S},b) \ge 1-\varepsilon_2 - O\left(\frac{\varepsilon_2}{K}\right) = 1-(\varepsilon+2\gamma/3) - O\left(\frac{\varepsilon_2}{K}\right).$$

By setting $K \gg \varepsilon_2\gamma^{-1}$ sufficiently large, we can conclude $\mathrm{Suc}(\mathcal{S},b) > 1-\varepsilon-\gamma$, where the slack in the inequality is independent of the choice of $b\in A_{\varepsilon-ws}$. This proves Theorem 2. □

of Lemma 1. By an easy approximation argument, it suffices to prove the result for the case when δ is rational. So assume

$$\delta = p/d,$$

for some integers $0 < p < d$, and let

$$q := d - p.$$

The 0-prediction strategy $\mathcal{T}$ is as follows. First, pick a value $t^\star \in [K]$ uniformly at random. Do not make any 0-predictions for steps $1, 2, \ldots, t^\star - 1$. During this time, maintain an ordered stack of "chips," initially empty. For $1 \leq i < t^\star$, after viewing b_i, if $b_i = 0$ then add p chips to the top of the stack; if $b_i = 1$ then remove q chips from the top of the stack—or, if the stack contains fewer than q chips, remove all the chips. After this modification to the stack, we say that the bit b_i has been "processed."

For $0 \leq i \leq K-1$, let H_i denote the number of chips on the stack after processing $b_1, \ldots, b_i$ (so, $H_0 = 0$). After processing $b_{t^\star - 1}$, sample from a 0/1-valued random variable X, with expectation

$$\mathbb{E}[X] := \frac{H_{t^\star - 1}}{dK}.$$

(Note that this expectation is at most $\frac{p(K-1)}{dK} < 1$, so the definition makes sense.) Predict a 0 at step $t^\star$ if $X = 1$, otherwise make no prediction at any step.

Note that the variable H_t can be regarded as a measure of the strategy's "courage" after processing $b_1, \ldots, b_t$, as in our sketch-description in Section 1.2. We now verify that $\mathcal{T}$ has the desired properties. Condition (i) in Lemma 1 is clearly satisfied. Before verifying conditions (ii) and (iii), we first sketch why they hold. For (ii), the idea is that if much less than a δ fraction of $b_1, \ldots, b_{K-1}$ are 1s, then the stack of chips will be of significant height after processing these bits. Since the stack doesn't grow too quickly, we conclude that the *average* stack height during these steps is significant, which implies that the strategy makes a prediction with noticeable probability.

For (iii), the idea is that for any chip c, if c stays on the stack for a significant amount of time, then the fraction of 1s appearing during the interval in which c was on the stack must be not much more than δ. Thus c's contribution to the false-guess probability is not much more than $\delta/(1-\delta)$ times c's contribution to the success probability. On the other hand, chips c which don't stay on the stack very long make only a small contribution to the false-guess probability.

Now we formally verify condition (ii). Fix some sequence b. First note that the placement and removal of chips, and the height sequence $H_0, \ldots, H_{K-1}$, can be defined in terms of b alone, without reference to the algorithm's random choices. Throughout our analysis we consider the stack to continue to evolve as a function of the bits $b_1, \ldots, b_{K-1}$, regardless of the algorithm's choices.

Suppose $b_1 + \ldots + b_{K-1} \leq \delta'(K-1)$, where $\delta' < \delta$; we ask, how large can $\pi_{\mathcal{T},b}(\infty)$ be? From the definition of $\mathcal{T}$, we compute

$$\pi_{\mathcal{T},b}(\infty) = 1 - \frac{1}{K} \sum_{t \in [K]} \frac{H_{t-1}}{dK} = 1 - \frac{1}{dK^2} \sum_{0 \leq t < K} H_t. \tag{3}$$

Now, for a chip c, let $m_c \in \mathbb{N}$ denote the number of indices $i < K$ for which c was on the stack immediately after processing b_i. (We consider each chip to be "unique;" that is, it is added to the stack at most once.) We can reexpress the sum appearing in Eq. (3) as

$$\sum_{0 \leq t < K} H_t = \sum_c m_c.$$

We will lower-bound this sum by considering the contribution made by chips that are never removed from the stack—that is, chips which remain after processing b_{K-1}. We call such chips "persistent." First, we argue that there are many persistent chips. By our assumption, at least $p \cdot (1-\delta')(K-1)$ chips are added to the stack in total, while at most $q \cdot \delta'(K-1)$ chips are ever removed. Thus the number of persistent chips is at least

$$\begin{aligned} p(1-\delta')(K-1) - q\delta'(K-1) &= p(1-\delta+(\delta-\delta'))(K-1) - q(\delta+(\delta'-\delta))(K-1) \\ &= [\underbrace{p(1-\delta)-q\delta}_{=0} + \underbrace{(p+q)}_{=d}(\delta-\delta')](K-1) \\ &= (\delta-\delta')d(K-1), \end{aligned}$$

where we used $p/q = \delta/(1-\delta)$. Let $J := (\delta-\delta')d(K-1)$.

Pick any J persistent chips, and number them $c(1), \ldots, c(J)$ so that $j' < j \leq J$ implies $c(j')$ appears above $c(j)$ on the stack after processing b_{K-1}. This means $c(j')$ was added to the stack no earlier than $c(j)$, so that $m_{c(j')} \leq m_{c(j)}$. At most p chips are added for every processed bit of b, and if $c(j)$ was added while processing the $(K-i)$-th bit, then $m_{c(j)} = i$. Thus, by our indexing we conclude $m_{c(j)} \geq \lceil j/p \rceil \geq j/p$. Summing over j, we obtain

$$\sum_{\text{persistent } c} m_c \;\geq\; \sum_{j=1}^{J} j/p \;=\; \frac{J(J+1)}{2p} \;>\; \frac{(\delta-\delta')^2 d^2 (K-1)^2}{2p} \;=\; \frac{(\delta-\delta')^2 d (K-1)^2}{2\delta}.$$

Finally, returning to Eq. (3), we compute

$$\pi_{\mathcal{T},b}(\infty) = 1 - \frac{1}{dK^2}\sum_c m_c < 1 - \frac{1}{dK^2} \cdot \frac{(\delta-\delta')^2 d(K-1)^2}{2\delta} < 1 - \frac{(\delta-\delta')^2}{8\delta},$$

since $K > 1$. This establishes condition (ii).

Now we verify condition (iii). Fix any sequence b. From our definitions, we have the expressions

$$\mathrm{Suc}(\mathcal{S}, b) = \frac{1}{K} \sum_{t \in [K]: b_t = 0} \frac{H_{t-1}}{dK}, \quad \mathrm{False}(\mathcal{S}, b) = \frac{1}{K} \sum_{t \in [K]: b_t = 1} \frac{H_{t-1}}{dK}, \quad \text{and so}$$

$$\mathrm{False}(\mathcal{S}, b) - (p/q)\,\mathrm{Suc}(\mathcal{S}, b) = \frac{1}{dK^2}\left(\sum_{t \in [K]: b_t = 1} H_{t-1} - \sum_{t \in [K]: b_t = 0} (p/q) H_{t-1} \right). \tag{4}$$

We regard the quantity H_{t-1} as being composed of a contribution of 1 from each of the chips on the stack after processing b_{t-1}. We rewrite the right-hand side of Eq. (4) as a sum of the total contributions from each chip. For a chip c, and for $z \in \{0,1\}$, let

$$n_{c,z} := \left| \{ t \in [K] : b_t = z, \text{ and } c \text{ is on the stack immediately after processing } b_{t-1} \} \right|.$$

We then have

$$\mathrm{False}(\mathcal{S}, b) - (p/q)\,\mathrm{Suc}(\mathcal{S}, b) = \frac{1}{dK^2} \sum_c (n_{c,1} - (p/q) n_{c,0}). \tag{5}$$

Fix attention to some chip c, which was placed on the stack while processing the i_c-th bit, for some $i_c \in [K-1]$. First assume that c was later removed from the stack, and let $j_c \in [K-1]$ be

the index of the bit whose processing caused c to be removed (thus, $b_{j_c} = 1$). Then the stack was not empty after processing bits $i_c, \ldots, j_c - 1$, since in particular, the stack contained c. Thus each 1 appearing in $(b_{i_c+1}, \ldots b_{j_c-1})$ caused exactly q chips to be removed from the stack. The removal caused by $[b_{j_c} = 1]$ removes some number $r_c \leq q$ of chips. Also, each 0 appearing in the same range causes p chips to be added. Now $n_{c,0}, n_{c,1}$ count the number of 0s and 1s respectively among $(b_{i_c+1}, \ldots, b_{j_c})$. Thus we have

$$H_{j_c} - H_{i_c} = pn_{c,0} - q(n_{c,1} - 1) - r_c \leq pn_{c,0} - q(n_{c,1} - 1),$$

or rearranging,

$$n_{c,1} - (p/q)n_{c,0} \leq (H_{i_c} - H_{j_c})/q + 1. \tag{6}$$

The chip c is added to the stack with $p - 1$ other chips while processing bit i_c. Later, c is removed from the stack when processing bit j_c, along with at most $q - 1$ other chips. Thus we have

$$H_{i_c} - H_{j_c} \leq p + q - 1,$$

and combining this with Eq. (6) gives

$$n_{c,1} - (p/q)n_{c,0} \leq (p + q - 1)/q + 1 < p/q + 2. \tag{7}$$

Next suppose c was added after processing bit $i_c \in [K - 1]$, but never removed from the stack. Then the stack was nonempty after processing bit i_c and remained nonempty from then on, so each 1 in $b_{i_c+1}, \ldots, b_{K-1}$ caused exactly q chips to be removed. By reasoning similar to the previous case, we get

$$n_{c,1} - (p/q)n_{c,0} = (H_{i_c} - H_{K-1})/q.$$

Now, c was added along with $p-1$ other chips after processing b_{i_c}, and c remains on the stack after processing b_{K-1}. It follows that $H_{i_c} - H_{K-1} \leq p - 1$, so

$$n_{c,1} - (p/q)n_{c,0} \leq (p - 1)/q. \tag{8}$$

Plugging Eqs. (7) and (8) into Eq. (5), we bound

$$\mathrm{False}(\mathcal{S}, b) - (p/q)\,\mathrm{Suc}(\mathcal{S}, b) < \frac{1}{dK^2} \sum_c (p/q + 2) \;<\; \frac{p^2/q + 2p}{dK}$$

(since at most $p(K - 1)$ chips are ever used)

$$= \frac{1}{K}\left(\frac{p}{q} \cdot \frac{p}{d} + \frac{2p}{d}\right) \;=\; \frac{1}{K}\left(\frac{\delta}{1-\delta} \cdot \delta + 2\delta\right) \;=\; O\left(\frac{\delta}{(1-\delta)K}\right).$$

Since $(p/q) = \delta/(1 - \delta)$, this establishes condition (iii), and completes the proof of Lemma 1. □

4 Prediction under automata-based assumptions

In this section we present an variant of Theorem 2 that is able to predict single bits from classes of binary sequences that are modeled upon the 0-weakly sparse sequences (ε-weak sparsity is defined in Section 1.2; here we are setting $\varepsilon := 0$), but that are significantly more general.

4.1 Bit-prediction algorithms

Our result concerns a problem in which a predictor is asked to correctly predict a single bit of their choice from a sequence x. Unlike the 0-prediction problem, here the predictor is allowed to predict either a 0 or a 1. Thus we need to modify our definition of 0-prediction strategies (in the obvious way), as follows. A *bit-prediction strategy* is a collection

$$\mathcal{S} = \{\pi_{\mathcal{S},b} : b \in \{0,1\}^\omega\},$$

where each $\pi_{\mathcal{S},b}$ is now a probability distribution over $(\mathbb{N} \times \{0,1\}) \cup \{\infty\}$. We require that for all $b = (b_1, b_2, \ldots), b' = (b'_1, b'_2, \ldots)$, and all $i \in \mathbb{N}, z \in \{0,1\}$,

$$(b_1, \ldots, b_{i-1}) = (b'_1, \ldots, b'_{i-1}) \Rightarrow \pi_{\mathcal{S},b}((i,z)) = \pi_{\mathcal{S},b'}((i,z)).$$

That is, $\pi_{\mathcal{S},b}((i,z))$ depends only on $b_1, \ldots, b_{i-1}$. As in the 0-prediction setting, our bit-prediction strategies will be defined so that this constraint clearly holds.

Define the success probability

$$\mathrm{Suc}^{\text{bit-pred}}(\mathcal{S}, b) := \sum_{i \in \mathbb{N}} \pi_{\mathcal{S},b}((i, b_i))$$

as the probability that $\mathcal{S}$ correctly predicts a bit of b. For a subset $A \subseteq \{0,1\}^\omega$, define $\mathrm{Suc}^{\text{bit-pred}}(\mathcal{S}, A) := \inf_{b \in A} \mathrm{Suc}^{\text{bit-pred}}(\mathcal{S}, b)$.

4.2 Finite automata

To state our result, we need the familiar notion of a *finite automaton* over a binary alphabet. Formally, this is a 3-tuple $M = (Q, s, \Delta)$, where:

- Q is a finite set of states;
- $s \in Q$ is the designated *starting state*;
- $\Delta : Q \times \{0,1\} \to Q$ is the *transition function*.

For $q \in Q$, $B \subseteq Q$, say that B is *accessible from* q if there exists a sequence $y_1, \ldots, y_m$ of bits and a sequence $q_0 = q, q_1, \ldots, q_m$ of states, such that

1. $\Delta(q_i, y_{i+1}) = q_{i+1}$ for $i = 0, 1, \ldots, m-1$;
2. $q_m \in B$.

Say that B is *strongly accessible* if, for any state q that is accessible from the starting state s, B is accessible from q.

Finite automata operate on infinite sequences $x \in \{0,1\}^\omega$ as follows: we let $q_0(x) := s$, and inductively for $t \geq 1$ we define

$$q_t(x) := \Delta(q_{t-1}(x), x_t).$$

We say that $q_t(x)$ is the *state of M after t steps* on the sequence x.

For a state $q \in Q$ we define $V_q(x)$, the *visits to q on x*, as

$$V_q(x) := \{t \geq 0 : q_t(x) = q\}.$$

Similarly, for $B \subseteq Q$, define $V_B(x)$ as $V_B(x) := \{t \geq 0 : q_t(x) \in B\}$.

4.3 Statement of the result

Say we are presented with the bits of some unknown $x \in \{0,1\}^\omega$ sequentially. We assume that x is "nice" in the following sense: for some known finite automaton M, there is a proper subset $B \subset Q$ of "bad" states of M, which we assume M visits only infrequently when M is run on x. We show that, if B is strongly accessible, we can successfully predict a bit of x with high probability.

In this section we say that a sequence x is *weakly sparse* if it is 0-weakly sparse as defined in Section 1.2, i.e., if

$$\lim_{s\to\infty} \inf_{t\geq s} \frac{1}{t} \sum_{1\leq i\leq t} x_i = 0.$$

We say that a subset $S \subseteq \{0,1,2,\ldots\}$ is weakly sparse if its characteristic sequence is weakly sparse. We prove:

Theorem 3. *Let $M = (Q, s, \Delta)$ be a finite automaton, and let $B \subset Q$ be a strongly accessible proper subset of states. Define*

$$A_{B,ws} := \{x \in \{0,1\}^\omega : V_B(x) \text{ is weakly sparse}\} ,$$

and assume $A_{B,ws}$ is nonempty. Then for all $\varepsilon > 0$, there exists a bit-prediction strategy $\mathcal{S} = \mathcal{S}_\varepsilon$ such that

$$\mathrm{Suc}^{\text{bit-pred}}(\mathcal{S}, A_{B,ws}) > 1 - \varepsilon. \tag{9}$$

$\mathcal{S}$ also has the "safety" property that for any $x \in \{0,1\}^\omega$, the probability that $\mathcal{P}$ outputs an incorrect bit-prediction on x is less than ε.

We make a few remarks before proving Theorem 3. First, simple examples show that the conclusion of Theorem 3 can hold even in some cases where B is not strongly accessible. Finding necessary and sufficient conditions on B could be an interesting question for future study.

Second, it is natural to ask whether a more "quantitative" version of Theorem 3 can be given. Let $A_{B,\varepsilon-ws}$ be the set of sequences x for which the characteristic sequence of $V_B(x)$ is ε-weakly sparse. If B is strongly accessible then, by a slight modification of our proof of Theorem 3, one can derive a bit-prediction strategy $\mathcal{S}$ such that

$$\mathrm{Suc}^{\text{bit-pred}}(\mathcal{S}, A_{B,\varepsilon-ws}) > 1 - O\left(\ell\varepsilon^{1/\ell}\right),$$

where $\ell = |Q|$ is the number of states of the automaton M.

Something like this weak form of dependence on ε is essentially necessary, as can be seen from the following example. Let M be an automaton with states $Q = \{1,2,\ldots,\ell\}$, and define

$$\Delta(i,1) := \min\{i+1, \ell\}, \quad \Delta(i,0) := 1.$$

Let $B := \{\ell\}$, and consider running M on a sequence $\boldsymbol{b}$ of independent unbiased bits. Then with probability 1, $V_B(\boldsymbol{b})$ is $2^{-\ell+1}$-weakly sparse. On the other hand, no strategy can predict a bit of $\boldsymbol{b}$ with success probability greater than $1/2$.

4.4 Proof of Theorem 3

Let $A_{ws} \subseteq \{0,1\}^\omega$ denote the set of weakly sparse sequences. Given a sequence $x = (x_1, x_2, \ldots)$, define $\neg x := (\neg x_1, \neg x_2, \ldots)$. Say that x is *co-weakly sparse*, and write $x \in A_{co-ws}$, if $\neg x \in A_{ws}$. To prove Theorem 3, we need two lemmas. The following lemma follows easily from Theorem 2:

Lemma 2. *Given $\delta > 0$, there exists a bit-prediction strategy $\mathcal{P} = \mathcal{P}_\delta$ such that*

$$\mathrm{Suc}^{\text{bit-pred}}(\mathcal{P}, A_{ws} \cup A_{co-ws}) > 1 - \delta.$$

$\mathcal{P}$ also has the "safety" property that for any $x \in \{0,1\}^\omega$, the probability that $\mathcal{P}$ outputs an incorrect bit-prediction on x is at most δ.

Proof. First, note that a 0-prediction strategy (as defined in Section 2) can be regarded as a bit-prediction strategy that only ever predicts a 0. Let $\varepsilon = \gamma := \delta/4$. The bit-prediction strategy $\mathcal{P}$, given access to some sequence b, simulates the 0-prediction strategy $\mathcal{S}_{\varepsilon,\gamma}$ from Theorem 2 on b, and simultaneously simulates an independent copy of $\mathcal{S}_{\varepsilon,\gamma}$ on $\neg b$. If $\mathcal{S}_{\varepsilon,\gamma}(b)$ ever outputs a prediction (i.e., that the next bit of b will be 0), $\mathcal{P}$ immediately outputs the same prediction. On the other hand, if $\mathcal{S}_{\varepsilon,\gamma}(\neg b)$ ever outputs a prediction (that the next bit of $\neg b$ will be 0), then $\mathcal{P}$ predicts that the next bit of b will be 1. If both simulations output predictions simultaneously, $\mathcal{P}$ makes an arbitrary prediction for the next bit.

To analyze $\mathcal{P}$, say we are given input sequence $b \in A_{ws} \cup A_{co-ws}$. First suppose $b \in A_{ws}$. Then $\mathcal{S}_{\varepsilon,\gamma}(b)$ outputs a correct prediction with probability $> 1 - \varepsilon - \gamma$. Also, by the safety property of $\mathcal{S}_{\varepsilon,\gamma}$ shown in Theorem 2, the probability that $\mathcal{S}_{\varepsilon,\gamma}(\neg b)$ outputs an incorrect prediction about $\neg b$ is at most $\varepsilon + \gamma$. Thus the probability that $\mathcal{P}$ outputs a correct prediction on b is greater than $1 - 2\varepsilon - 2\gamma = 1 - \delta$.

The case where $b \in A_{co-ws}$ is analyzed similarly. Finally, the safety property of $\mathcal{P}$ follows from the safety property of $\mathcal{S}_{\varepsilon,\gamma}$. □

For the next lemma, we need some further definitions. Fix a finite automaton $M = (Q, s, \Delta)$. For $x \in \{0,1\}^\omega$, let

$$Q_{\inf}(x) := \{q \in Q : |V_q(x)| = \infty\}.$$

Of course, $Q_{\inf}(x)$ is nonempty since Q is finite. If $q \in Q_{\inf}(x)$, define a sequence $x^{(q)} \in \{0,1\}^\omega$ as follows. If $V_q(x) = \{t(1), t(2), \ldots,\}$ where $0 \leq t(1) < t(2) < \ldots$, we define

$$x_i^{(q)} := x_{t(i)+1}.$$

In words: if M is run on x, the i-th bit of $x^{(q)}$ records the bit of x seen immediately after the i-th visit to state q. If $q \notin Q_{\inf}(x)$, we define $x^{(q)} \in \{0,1\}^*$ similarly; in this case, $x_i^{(q)}$ is undefined if M visits state q fewer than i times while running on x.

The following lemma gives us a useful property obeyed by sequences x from the set $A_{B,ws}$ (defined in the statement of Theorem 3).

Lemma 3. *Given $M = (Q, s, \Delta)$, suppose $B \subseteq Q$ is strongly accessible. If $x \in A_{B,ws}$, then there exists a state $q \in Q_{\inf}(x)$ such that*

$$x^{(q)} \in A_{ws} \cup A_{co-ws}.$$

Proof. We prove the contrapositive. Assume that all $q \in Q_{\inf}(x)$ satisfy $x^{(q)} \notin A_{ws} \cup A_{co-ws}$; we will show that $x \notin A_{B,ws}$.

Say that a state $q \in Q$ is *frequent (on x)* if there exist $\alpha, \beta > 0$ such that for all $T \in \mathbb{N}$,

$$|V_q(x) \cap \{0, 1, \ldots, T-1\}| \geq \alpha T - \beta.$$

Let F denote the set of frequent states. Clearly $F \subseteq Q_{\inf}(x)$. We will show:

1. $F = Q_{\inf}(x)$;
2. F contains a state from B.

Item 2 will immediately imply that $x \notin A_{B,ws}$, as desired.

For each $q \in Q_{\inf}(x)$, our assumption $x^{(q)} \notin A_{ws} \cup A_{co-ws}$ implies that there is a $\delta_q \in (0, 1/2)$ and a $K_q > 0$ such that for $k \geq K_q$,

$$\delta_q < \frac{1}{k}\left(x_1^{(q)} + \ldots + x_k^{(q)}\right) < 1 - \delta_q. \tag{10}$$

Let $\delta := \min \delta_q$. Choose a value $T^\star > 0$ such that each $q \in Q_{\inf}(x)$ appears at least K_q times among $(q_0(x), q_1(x), \ldots, q_{T^\star-1}(x))$. Choose a second value $R > 0$, such that any $q \notin Q_{\inf}(x)$ occurs fewer than R times in the infinite sequence $(q_0(x), q_1(x), \ldots)$.

Let $\ell = |Q|$. Fix any $t \in \mathbb{N}$ satisfying

$$t \geq \max\left\{\frac{\ell R}{\delta^{2(\ell-1)}}, T^\star\right\}.$$

By simple counting, some $q^\star \in Q$ occurs at least t/ℓ times in $(q_0(x), q_1(x), \ldots, q_{t-1}(x))$. We have $t/\ell > R$, so this $q^\star$ must lie in $Q_{\inf}(x)$. Eq. (10) then implies that the states $\Delta(q^\star, 0), \Delta(q^\star, 1)$ each appear at least $\delta t/\ell - 1 > \delta^2 t/\ell$ times among $(q_0(x), q_1(x), \ldots, q_{t-1}(x))$. Now $\delta^2 t/\ell > R$, so we have $\Delta(q^\star, 0), \Delta(q^\star, 1) \in Q_{\inf}(x)$.

Iterating this argument $(\ell - 1)$ times, we conclude that every state q reachable from $q^\star$ by a sequence of $(\ell - 1)$ or fewer transitions lies in $Q_{\inf}(x)$, and appears at least $\delta^{2(\ell-1)} t/\ell = \Omega(t)$ times among $(q_0(x), q_1(x), \ldots, q_{t-1}(x))$. But *every* $q \in Q_{\inf}(x)$ is reachable from $q^\star$ by at most $(\ell - 1)$ transitions. Thus $F = Q_{\inf}(x)$, proving Item 1 above.

The argument above shows that if $q \in Q_{\inf}(x)$, then $\Delta(q, 0), \Delta(q, 1) \in Q_{\inf}(x)$ as well. Recall that B is strongly accessible; it follows that $Q_{\inf}(x) \cap B$ is nonempty, proving Item 2 above. This completes the proof of Lemma 3. □

We can now complete the proof of Theorem 3. Let $Q = \{p_1, \ldots, p_\ell\}$, where $\ell = |Q|$. We must have $\ell > 1$, since B is a nonempty proper subset of Q. Given $\varepsilon > 0$, let $\delta := \varepsilon/(2\ell)$. We define the algorithm $\mathcal{S} = \mathcal{S}_\varepsilon$ as follows. $\mathcal{S}$ runs in parallel ℓ different simulations

$$\mathcal{P}[1], \ldots, \mathcal{P}[\ell]$$

of the algorithm $\mathcal{P}_\delta$ from Lemma 2. $\mathcal{P}[j]$ is run, not on the input sequence x itself, but on the subsequence $x^{(p_j)}$. To determine which simulation receives each successive bit of x, the algorithm $\mathcal{S}$ simply simulates M on the bits of x seen so far. (Note that, if $p_j \notin Q_{\inf}(x)$, then the simulation $\mathcal{P}[j]$ may "stall" indefinitely without receiving any further input bits.)

Suppose that the simulation $\mathcal{P}[j]$ outputs a prediction $z \in \{0,1\}$ after seeing the i-th bit of $x^{(p_j)}$, and that we subsequently reach a time t such that $q_t(x) = p_j$ is the $(i+1)$-st visit to state p_j. The algorithm $\mathcal{S}$ then predicts that $x_{t+1} = x_{i+1}^{(p_j)} = z$.

We now analyze $\mathcal{S}$. Fix any $x \in A_{B-ws}$. By the safety property of Lemma 2, each $\mathcal{P}[j]$ outputs an incorrect prediction with probability at most δ on *any* input sequence, so the overall probability that $\mathcal{S}$ makes an incorrect prediction on any input sequence is at most $\ell\delta = \varepsilon/2$. This proves the safety property claimed for $\mathcal{S}$.

Now, on input sequences $x \in A_{B,ws}$, Lemma 3 tells us that there exists a $p_j \in Q_{\inf}(x)$ such that $x^{(p_j)} \in A_{ws} \cup A_{co-ws}$. Thus, if $\mathcal{P}[j]$ is run individually on $x^{(p_j)}$, $\mathcal{P}[j]$ outputs a correct prediction with probability greater than $1-\delta$. We conclude that

$$\mathrm{Suc}^{\text{bit-pred}}(\mathcal{S}, x) > (1-\delta) - \varepsilon/2 > 1 - \varepsilon,$$

using $\ell > 1$. This establishes Eq. (9), and completes the proof of Theorem 3.

4.5 Single-bit prediction strategies as gambling schemes

As mentioned in the Introduction, it is possible to interpret randomized bit-prediction strategies, as formalized in Section 4.1, as a certain kind of (deterministic) *gambling schemes* making repeated predictions, without reinvestment of winnings. Here we discuss this simple connection.

Our gambling takes place in a casino with two kinds of money: *blue money*, which can be used to gamble but is valueless outside the casino; and *red money*, which is of value outside the casino but cannot be used to gamble (or converted into blue money). A bit-prediction strategy, described by a family

$$\mathcal{S} = \{\pi_{\mathcal{S},b} : b \in \{0,1\}^\omega\},$$

can be viewed as a gambler who holds an initial fortune consisting of blue money. On input sequence $b = (b_1, b_2, \ldots)$, this gambler places a "stake" of blue money of size $\pi_{\mathcal{S},b}((i,z)) \in [0,1]$ on the prediction $[b_i = z]$ after viewing $b_1, \ldots, b_{i-1}$ and before viewing b_i. (Thus, the gambler may place a stake simultaneously on $[b_i = 0]$ and on $[b_i = 1]$. 0-prediction strategies correspond to gambling schemes where a stake is only ever placed on $[b_i = 0]$.) A successful prediction is rewarded with an amount of red money equal to the amount of blue money staked; an unsuccessful prediction gets no such reward. In either case, the blue money staked is taken by the casino.

From this viewpoint, the condition we imposed that $\pi_{\mathcal{S},b}((i,z))$ depends only on $b_1, \ldots, b_{i-1}$ still naturally expresses the gambler's lack of prescience about future bits. The requirement that $\sum_{i\in\mathbb{N}, z\in\{0,1\}} \pi_{\mathcal{S},b}((i,z)) \le 1$ is now interpreted as the condition that the gambler has a total initial stake of \$1 worth of blue money to invest. The quantity $\mathrm{Suc}^{\text{bit-pred}}(\mathcal{S}, b) = \sum_{i\in\mathbb{N}} \pi_{\mathcal{S},b}((i,b_i))$ we defined can, in this setting, be seen to equal the total amount of red money earned by the gambler over the entire sequence; this is the quantity our gambler would like to maximize. By this connection, for any set $A \subseteq \{0,1\}^\omega$, the value $\mathrm{Suc}^{\text{bit-pred}}(\mathcal{S}, A)$ equals the infimum over $b \in A$ of the amount of red money earned by the gambler on b.

Gambling schemes with variable stakes, determined adaptively by the gambler, have been intensively studied. The gambling schemes called *gales*, discussed in the introduction, correspond to such schemes in which reinvestment of winnings is allowed. Other works, e.g., [KP11], study a gambling setting in which the gambler may adaptively choose a stake in the bounded range $[0,1]$ at each step; in that work's setting, unlike with gales and our own setting, the gambler is allowed to go into debt. In [KP11] and related work, a main focus is to minimize a measure of the gambler's

regret with respect to a class of "expert" prediction strategies, over all bit-sequences. This goal is somewhat orthogonal to our goals in the present work.

5 The Density Prediction Game

In this section we prove Theorem 1 from Section 1.3.

For any fixed δ, ε, our prediction strategy will work entirely within a finite interval $(x_1, \ldots, x_T)$ of the sequence x. We note that, to derive a (δ, ε)-successful strategy over this interval, it suffices to show that for every distribution $\mathcal{D}$ over $\{0,1\}^T$, there exists a strategy $\mathcal{S}_{\mathcal{D}}$ that is (δ, ε)-successful when played against $\mathcal{D}$. (This observation follows from the minimax theorem of game theory, or from the result of Sandroni [San03] mentioned in Section 1.1.) However, using this idea would lead to a nonconstructive proof of Theorem 1, and in any case does not seem to make the proof any simpler. Thus we will not follow this approach.

Let $\delta, \varepsilon > 0$ be given; we give a forecasting strategy $\mathcal{S} = \mathcal{S}_{\delta,\varepsilon}$ for the density prediction game, and prove that $\mathcal{S}$ is (δ, ε)-successful. Set $n := \lceil 4/(\delta\varepsilon^2) \rceil$. Our strategy will always make a prediction about an interval $x_a, \ldots, x_b$ where $a \leq b \leq 2^n$. The strategy $\mathcal{S}$ is defined as follows:

1. Choose $R \in \{1, \ldots, n\}$ uniformly. Choose S uniformly from $\{1, \ldots, 2^{n-R}\}$.

2. Ignore the first $t = (S-1) \cdot 2^R$ bits of x. Observe bits $x_{t+1}, \ldots, x_{t+2^{R-1}}$, and let p be the fraction of 1s in this interval. Immediately after seeing $x_{t+2^{R-1}}$, predict:

   "Out of the next 2^{R-1} bits, a p fraction will be 1s."

We now analyze $\mathcal{S}$. To do so, it is helpful to describe $\mathcal{S}$ in a slightly different fashion. Let us re-index the first 2^n bits of our sequence x, considering each such bit to be indexed by a string $z \in \{0,1\}^n$. We use lexicographic order, so that the sequence is indexed $x_{0^n}, x_{0^{n-1}1}, x_{0^{n-2}10}$, and so on.

Let T be a directed binary tree of height n, whose vertices at depth i $(0 \leq i \leq n)$ are indexed by binary strings of length i; in particular, the root vertex is labeled by the empty string. If $i < n$ and $y \in \{0,1\}^i$, the vertex v_y has left and right children v_{y0}, v_{y1} respectively. Each leaf vertex is indexed by an n-bit string z, and any such vertex v_z is labeled with the bit x_z.

For $y \in \{0,1\}^*$, let T_y denote the subtree of T rooted at v_y. A direct translation of the strategy $\mathcal{S}$ into our current perspective gives the following equivalent description of $\mathcal{S}$:

1'. Choose $R \in \{1, \ldots, n\}$ uniformly. Starting at the root of T, take a directed, unbiased random walk of length $n - R$, reaching a vertex v_Y where $Y \in \{0,1\}^{n-R}$.

2'. Observe the bits of x that label leaf vertices in T_{Y0}, and let p be the fraction of 1s seen among these bits. Immediately after seeing the last of these bits, predict:

   "Out of the next 2^{R-1} bits of x (i.e., those labeling leaf vertices in T_{Y1}), a p fraction will be 1s."

To analyze $\mathcal{S}$ in this form, fix any binary sequence x. We consider the random walk performed in $\mathcal{S}$ to be extended to an unbiased random walk of length n. The walk terminates at some leaf vertex v_Z, where $Z = (z_1, \ldots, z_n)$ is uniform over $\{0,1\}^n$.

For $i \in [n]$ and $y \in \{0,1\}^i$, define

$$\rho(y) := 2^{i-n} \sum_{w \in \{0,1\}^{n-i}} x_{yw}$$

as the fraction of 1s among the labels of leaf vertices of T_y. We let $\rho(\emptyset) := 2^{-n} \sum_{w \in \{0,1\}^n} x_w$. Let $X(0) := \rho(\emptyset)$, and for $t \in [n]$, define the random variable

$$X(t) := \rho(z_1, \ldots, z_t),$$

defined in terms of Z. The sequence $X(0), \ldots, X(n)$ is a *martingale*; that is, for each $t \in [n]$ we have the identity $\mathbb{E}[X(t)|X(0), \ldots, X(t-1)] = X(t-1)$, which is easily verified. We follow a folklore technique by analyzing the squared differences between terms in the sequence. First, we have $X(t) \in [0,1]$, so that $(X(n) - X(0))^2 \leq 1$. On the other hand,

$$\begin{aligned}
\mathbb{E}[(X(n) - X(0))^2] &= \mathbb{E}\left[\left(\sum_{0 \leq t < n} (X(t+1) - X(t))\right)^2\right] \\
&= \mathbb{E}\left[\sum_{0 \leq t < n} (X(t+1) - X(t))^2\right] + \mathbb{E}\left[2 \sum_{0 \leq s < t < n} (X(s+1) - X(s))(X(t+1) - X(t))\right].
\end{aligned} \tag{11}$$

Now, for $0 \leq s < t < n$ and for any outcome of the bits $z_1, \ldots, z_t$ (which determine $X(s)$, $X(s+1)$, and $X(t)$), we have

$$\begin{aligned}
\mathbb{E}[(X(t+1) - X(t))|z_1, \ldots, z_t] &= \mathbb{E}_{z_{t+1} \in \{0,1\}}[(\rho(z_1, \ldots, z_{t+1})] - \rho(z_1, \ldots, z_t) \\
&= \frac{1}{2}\left[\rho(z_1, \ldots, z_t, 0) + \rho(z_1, \ldots, z_t, 1)\right] - \rho(z_1, \ldots, z_t) \\
&= 0.
\end{aligned}$$

Thus the second right-hand term in Eq. (11) is 0, and

$$\mathbb{E}[(X(n) - X(0))^2] = \sum_{0 \leq t < n} \mathbb{E}\left[(X(t+1) - X(t))^2\right]. \tag{12}$$

Next we relate this to the accuracy of our guess p. Let p^* be the fraction of 1s in T_{Y1}, i.e., the quantity $\mathcal{S}$ attempts to predict; note that p^* and p are both random variables. From the definitions, we have

$$p = \rho(Y0), \quad p^* = \rho(Y1), \quad X(n-R) = \frac{1}{2}(p + p^*).$$

Also,

$$X(n-R+1) = \begin{cases} p & \text{if } z_{n-R+1} = 0, \\ p^* & \text{if } z_{n-R+1} = 1. \end{cases}$$

Thus we have the identity

$$(X(n-R+1) - X(n-R))^2 = \frac{1}{4}(p - p^*)^2.$$

Now, $n-R$ is is uniform over $\{0,1,\ldots,n-1\}$, and independent of Z. It follows from Eq. (12) that

$$\mathbb{E}[(X(n-R+1)-X(n-R))^2] = \frac{1}{n}\mathbb{E}[(X(n)-X(0))^2] \leq 1/n.$$

Combining, we have

$$\mathbb{E}[(p-p^*)^2] \leq 4/n. \tag{13}$$

On the other hand,

$$\mathbb{E}[(p-p^*)^2] \geq \Pr[|p-p^*| \geq \varepsilon] \cdot \varepsilon^2. \tag{14}$$

Combining Eqs. (13) and (14), we obtain

$$\Pr[|p-p^*| \geq \varepsilon] \leq 4/(n\varepsilon^2) \leq \delta,$$

by our setting $n = \lceil 4/(\delta\varepsilon^2) \rceil$. This proves Theorem 1.

6 Questions for Future Work

1. Fix some $p \in [1/2,1]$; is there a satisfying characterization of the sets $A \subseteq \{0,1\}^\omega$ for which some bit-prediction strategy (as defined in Section 4.1) succeeds with probability $\geq p$ against all $x \in A$? Perhaps there is a characterization in terms of some appropriate notion of dimension, analogous to the gale characterizations of Hausdorff dimension [Lut03a] and packing dimension [AHLM07].

2. Could the study of computationally bounded bit-prediction strategies be of value to the study of complexity classes, by analogy to the study of computationally bounded gales in [Lut03a, AHLM07] and in related work?

3. Find necessary and sufficient conditions on the set B of "infrequently visited" states, for the conclusion of Theorem 3 (in Section 4.3) to hold.

References

[AHLM07] K. B. Athreya, J. M. Hitchcock, J. H. Lutz, and E. Mayordomo. Effective strong dimension in algorithmic information and computational complexity. *SIAM Journal on Computing*, 37(3):671–705, 2007.

[Bil65] P. Billingsley. *Ergodic Theory and Information*. John Wiley and Sons, 1965.

[Daw82] A. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

[Egg49] H. Eggleston. The fractional dimension of a set defined by decimal properties. *Quarterly Journal of Mathematics*, 20:31–36, 1949.

[FV98] D. P. Foster and R. V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

[FV09] L. Fortnow and R. V. Vohra. The complexity of forecast testing. *Econometrica*, 77:93–105, 2009.

[Hem05] L. A. Hemaspaandra. Sigact news complexity theory column 48. *SIGACT News*, 36(3):24–38, 2005. Guest Column: The Fractal Geometry of Complexity Classes, by J. M. Hitchcock, J. H. Lutz, and E. Mayordomo.

[KP11] Michael Kapralov and Rina Panigrahy. Prediction strategies without loss. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, pages 828–836, 2011.

[Lut03a] J. H. Lutz. Dimension in complexity classes. *SIAM Journal on Computing*, 32(5):1236–1259, 2003.

[Lut03b] J. H. Lutz. The dimensions of individual strings and sequences. *Information and Computation*, 187(1):49–79, 2003.

[MF98] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

[San03] A. Sandroni. The reproducible properties of correct forecasts. *International Journal of Game Theory*, 32(1):151–159, December 2003.