High-Confidence Predictions under Adversarial Uncertainty

Andrew Drucker

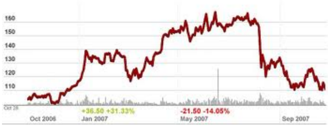
MIT

Feb. 13, 2012

Setting: prediction on binary sequences

$$x = (x_1, x_2, x_3, \ldots) \quad \in \quad \{0,1\}^\omega$$

- Bits of x revealed sequentially.
- **Goal:** make some nontrivial prediction about unseen bits of sequence x, given bits seen so far.



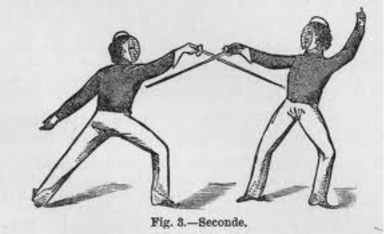
Setting: prediction on binary sequences

$$x = (x_1, x_2, x_3, \ldots)$$

- **Question:** What kinds of assumptions on x are needed to make interesting predictions?
- **Our message:** Surprisingly weak ones.

Modeling questions

- Prediction: a game between the Predictor and Nature.
- What kind of opponent is nature?



Probabilistic models

$$x = (x_1, x_2, x_3, \ldots)$$

$$x \sim \mathcal{D},$$

where $\mathcal{D}$ is some known probability distribution.

Probabilistic models

$$x = (x_1, x_2, x_3, \ldots)$$

$$x \sim \mathcal{D},$$

where $\mathcal{D}$ is some known probability distribution.

- **Problem:** how to choose correct $\mathcal{D}$ for realistic applications?

Classes of assumptions

$$x = (x_1, x_2, x_3, \ldots)$$

- **Adversarial models:**

$$x \in A,$$

  where $A \subseteq \{0,1\}^\omega$ is some known set.

- Interested in worst-case performance.

Classes of assumptions

$$x = (x_1, x_2, x_3, \ldots)$$

- **Adversarial models:**

  $$x \in A,$$

  where $A \subseteq \{0,1\}^\omega$ is some known set.
- Interested in worst-case performance.
- These assumptions can be quite "safe"...
- Our focus today.

Prior work on adversarial prediction

Gales and fractal dimension
[Lutz ‘03; Athreya, Hitchcock, Lutz, Mayordomo ‘07]

- Gales: a class of betting strategies, to bet on unseen bits of $x \in A$.
- Goal: reach a fortune of ∞, on any $x \in A$.
- The “handicap” we need can be related to measures of fractal dimension for A...

Prior work on adversarial prediction

Ignorant forecasting

- What is the chance of rain tomorrow?

- Basic test of a meteorologist: "calibration."
- If governing distribution $\mathcal{D}$ is known, easy to achieve with Bayes' rule...
- But: calibration can also be achieved by an ignorant forecaster! **[Foster, Vohra '98]**

Prior work on adversarial prediction

$$x = (x_1, x_2, x_3, \ldots)$$

- These works' goal: long-term, overall predictive success.
- Our focus: make a single prediction with high confidence.

0-prediction

- **Our main scenario:** want to predict a single 0 among the bits of x.

- **Interpretation:** choose a time to "safely" perform some action;

  $[x_t = 0]$ means "time t is safe."

Possible assumptions

- ε**-biased arrivals assumption:** bits of x independent, with

$$\Pr[x_t = 1] \ = \ \varepsilon.$$

- Best strategy succeeds with prob. $1 - \varepsilon$.

Possible assumptions

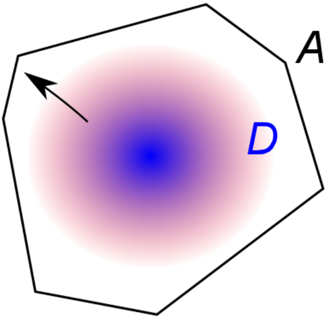
- Very strong assumption...
- **Idea** (not new)**:** study adversarial "relaxations" of ε-biased model.

Possible assumptions

- Very strong assumption...
- **Idea** (not new)**:** study adversarial "relaxations" of ε-biased model.

Possible assumptions

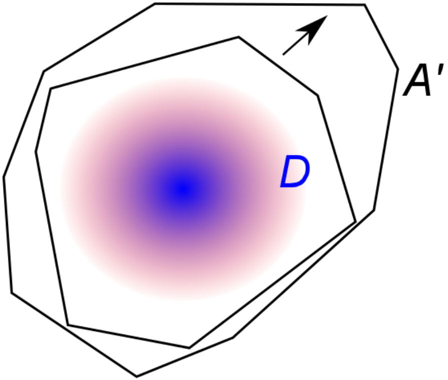
- Very strong assumption...
- **Idea** (not new)**:** study adversarial "relaxations" of ε-biased model.



Possible assumptions

- Very strong assumption...

- **Idea** (not new)**:** study adversarial "relaxations" of ε-biased model.



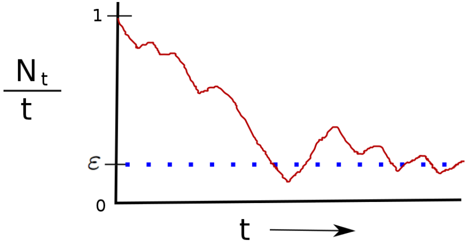
Possible assumptions

- Let

$$N_t := x_1 + \ldots + x_t.$$

- ε**-sparsity assumption:** say that x is ε-sparse if

$$\lim_{t \to \infty} N_t / t \ \leq \ \varepsilon.$$



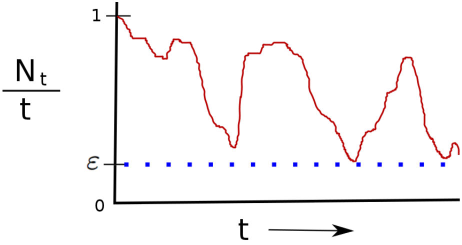
Possible assumptions

- Let

$$N_t := x_1 + \ldots + x_t.$$

- ε**-weak sparsity assumption:** say that x is ε-weakly sparse if

$$\lim_{s\to\infty} \inf_{t\geq s} N_t/t \;\leq\; \varepsilon.$$



Our main result

Theorem

For any $\varepsilon, \gamma > 0$, there is a (randomized) 0-prediction strategy $\mathcal{S}_{\varepsilon,\gamma}$ that succeeds with prob. $\geq 1 - \varepsilon - \gamma$, on any ε-weakly sparse sequence.

- Can do nearly as well as under ε-biased arrivals!

Our main result

Theorem

For any $\varepsilon, \gamma > 0$, there is a (randomized) 0-prediction strategy $\mathcal{S}_{\varepsilon,\gamma}$ that succeeds with prob. $\geq 1 - \varepsilon - \gamma$, on any ε-weakly sparse sequence.

- Can do nearly as well as under ε-biased arrivals!
- (Adversary's sequence gets fixed before randomness in $\mathcal{S}_{\varepsilon,\gamma}$...)

Proof ideas

- Divide sequence into "epochs:"

x = 0 1 1 1 0 0 1 0 1 1 0 0 0 0 . . .

E_1 E_2 E_3

- (r-th epoch of length $K_r = \Theta(r^2)$.)
- Run a separate 0-prediction algorithm for each individual epoch.

Proof ideas

x = 0 1 1 1 0 0 1 0 1 1 0 0 0 0 . . .

E_1 E_2 E_3

- **Easy claim:** x is ε-weakly sparse

$$\Downarrow$$

$\exists$ a subsequence of **"nice"** epochs,
whose 1-densities are $\alpha < \varepsilon + \gamma$.

Proof ideas

$$x = \underbrace{0}_{E_1}\,\underbrace{1\,1\,1\,0}_{E_2}\,\underbrace{0\,1\,0\,1\,1\,0\,0\,0\,0}_{E_3}\,\ldots$$

Idea: give an algorithm $\mathcal{S}$ with the properties:

1. Makes a 0-prediction with noticeable prob. on each nice epoch;
2. On every epoch,

$$\Pr\left[\text{true prediction}\right] \geq \left(\frac{1-\alpha}{\alpha}\right) \cdot \Pr\left[\text{false prediction}\right].$$

Proof ideas

x = 0 1 1 1 0 0 1 0 1 1 0 0 0 0 . . .

E_1 E_2 E_3

Idea: give an algorithm $\mathcal{S}$ with the properties:

1. Makes a 0-prediction with noticeable prob. on each nice epoch;
2. On every epoch,

$$\Pr\Big[\text{true prediction}\Big] \geq \left(\frac{1-\alpha}{\alpha}\right) \cdot \Pr\Big[\text{false prediction}\Big].$$

(Would achieve our goal!)

Proof ideas

$$\Pr\left[\text{true prediction}\right] \overset{?}{\geq} \left(\frac{1-\alpha}{\alpha}\right) \cdot \Pr\left[\text{false prediction}\right]$$

Proof ideas

$$\Pr\left[\text{true prediction}\right] \overset{?}{\geq} \left(\frac{1-\alpha}{\alpha}\right) \cdot \Pr\left[\text{false prediction}\right]$$

- Whoops—can't achieve this!

Proof ideas

$$\Pr\left[\text{true prediction}\right] \overset{?}{\geq} \left(\frac{1-\alpha}{\alpha}\right) \cdot \Pr\left[\text{false prediction}\right]$$

- Whoops—can't achieve this!
- **Modified goal:** an upper bound

$$\left(\frac{1-\alpha}{\alpha}\right) \cdot \Pr\left[\text{false prediction}\right] - \Pr\left[\text{true prediction}\right]$$

$$\leq \textbf{(small)}$$

Proof ideas

$$\Pr\left[\text{true prediction}\right] \overset{?}{\geq} \left(\frac{1-\alpha}{\alpha}\right) \cdot \Pr\left[\text{false prediction}\right]$$

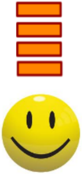
- Whoops—can't achieve this!
- **Modified goal:** an upper bound

$$\left(\frac{1-\alpha}{\alpha}\right) \cdot \Pr\left[\text{false prediction}\right] - \Pr\left[\text{true prediction}\right]$$

$$\leq O(1/|K_r|)$$

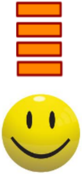
Proof ideas

- During the r-th epoch, alg. maintains a stack of **"chips"** (initially empty);



Proof ideas

- During the r-th epoch, alg. maintains a stack of **"chips"** (initially empty);



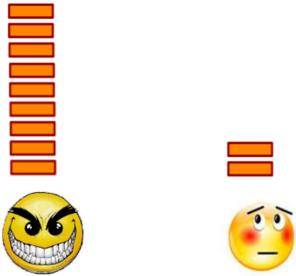
stack's height

$\updownarrow$

algorithm's **"courage"** to predict next bit of x will be a 0.

Proof ideas

- During the r-th epoch, alg. maintains a stack of **"chips"** (initially empty);



stack's height

$\updownarrow$

algorithm's **"courage"** to predict next bit of x will be a 0.

Stack dynamics

- Assume

$$\alpha \ < \ \frac{p}{d} \ = \ 1 - \frac{q}{d} \ < \ \varepsilon + \gamma.$$

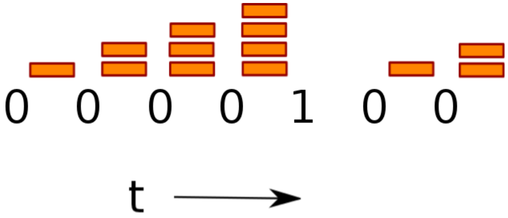
Stack dynamics

- Assume

$$\alpha < \frac{p}{d} = 1 - \frac{q}{d} < \varepsilon + \gamma.$$

- Observe a 0: add p "courage chips."
- Observe a 1: remove q chips.

e.g., $p = 1$:



Making predictions

- Let $H_t =$ stack height after observing first t bits of r-th epoch.

Making predictions

- Let H_t = stack height after observing first t bits of r-th epoch.
- **Overall algorithm for epoch** r:

Making predictions

- Let H_t = stack height after observing first t bits of r-th epoch.

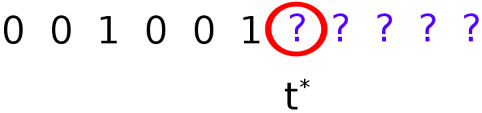
- **Overall algorithm for epoch** r:

    1. Choose t^* uniformly from $\{1, 2, \ldots, K_r\}$;

Making predictions

- Let H_t = stack height after observing first t bits of r-th epoch.

- **Overall algorithm for epoch** r:
  1. Choose t^* uniformly from $\{1, 2, \ldots, K_r\}$;
  2. Observe first $t^* - 1$ bits;

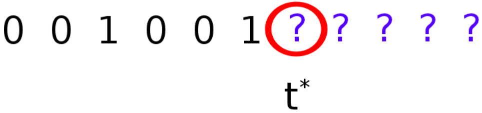


Making predictions

- Let H_t = stack height after observing first t bits of r-th epoch.

- **Overall algorithm for epoch** r:

  1. Choose t^* uniformly from $\{1, 2, \ldots, K_r\}$;
  2. Observe first $t^* - 1$ bits;



  3. Predict a 0 on step t^* with probability

$$\frac{H_{t^*-1}}{d \cdot K_r},$$

else make no prediction this epoch.

Analysis ideas

1. If fraction of 1s in r-th epoch is $< \alpha < p/d$,

   a 0-prediction is made in epoch r with $\Omega(1)$ prob.

Analysis ideas

1. If fraction of 1s in r-th epoch is $< \alpha < p/d$,

   a 0-prediction is made in epoch r with $\Omega(1)$ prob.

$\Longrightarrow$ Eventually (in some epoch), a prediction is made.

Analysis ideas

2. To compare odds of correct and incorrect 0-predictions,

Analysis ideas

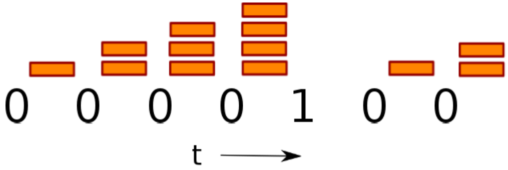
2. To compare odds of correct and incorrect 0-predictions,

   analyze each chip's contribution.

Analysis ideas

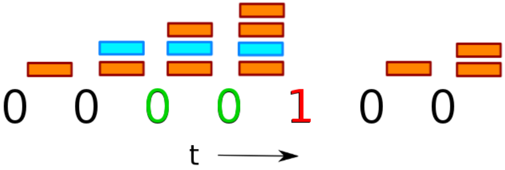
2. To compare odds of correct and incorrect 0-predictions,

 <u>analyze each chip's contribution.</u>



Analysis ideas

2. To compare odds of correct and incorrect 0-predictions, analyze each chip's contribution.



Analysis ideas

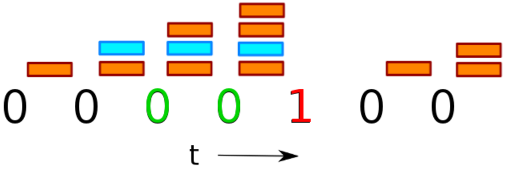
- **Intuition:**
  - If a chip remains on the stack long enough, fraction of 1s while it's on is $\lesssim p/d$.

    GOOD!

  - Total contribution to failure probability of other ("bad") chips is small.
- We can analyze all chips in a simple, unified way...

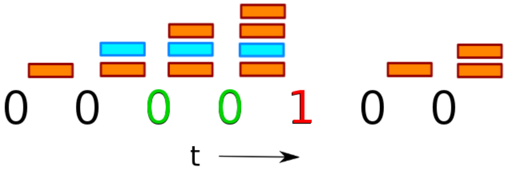
Analysis ideas

- Fix attention to a chip c on input x.



- Let zeros(c) (ones(c)) denote the number of zeros (ones) appearing after steps where c is on the stack.

Analysis ideas

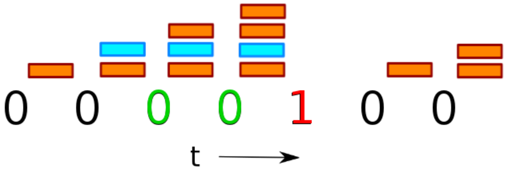


- Chip c's contribution to **success** and **failure** probabilities:

$$\downarrow \qquad \downarrow$$

$$\frac{\text{zeros}(c)}{d \cdot K_r^2}, \qquad \frac{\text{ones}(c)}{d \cdot K_r^2}.$$

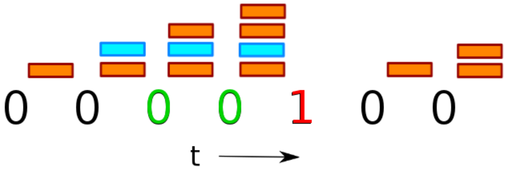
Analysis ideas



Claim:

$$q \cdot \text{ones}(c) \;-\; p \cdot \text{zeros}(c) \;\leq\; q.$$

Analysis ideas

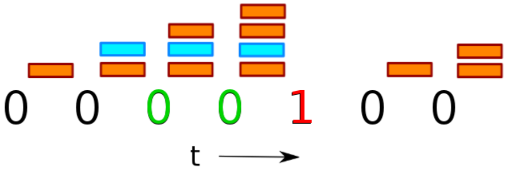


Claim:

$$q \cdot \text{ones}(c) \; - \; p \cdot \text{zeros}(c) \; \leq \; q.$$

Proof: LHS bounded by the **net loss** in stack height between first appearance of c and (possible) removal...

Analysis ideas



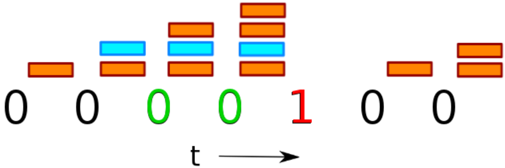
Claim:

$$q \cdot \mathrm{ones}(c) \; - \; p \cdot \mathrm{zeros}(c) \; \leq \; q.$$

Proof: LHS bounded by the **net loss** in stack height between first appearance of c and (possible) removal...

c removed along with $\leq q$ other chips!

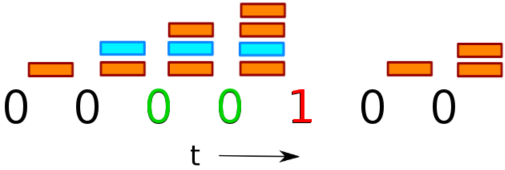
Analysis ideas



Let's sum over all c (at most $p \cdot K_r$ chips total):

$$q \cdot \sum_c \text{ones}(c) \; - \; p \cdot \sum_c \text{zeros}(c) \; \leq \; pqK_r.$$

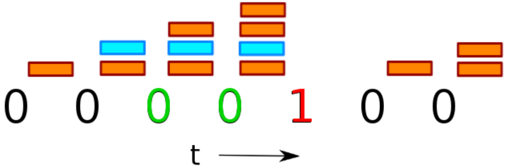
Analysis ideas



Let's sum over all c (at most $p \cdot K_r$ chips total):

$$(q/p) \cdot \sum_{c} \text{ones}(c) \; - \; \sum_{c} \text{zeros}(c) \; \leq \; qK_r.$$

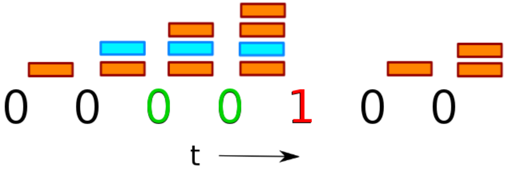
Analysis ideas



Let's sum over all c (at most $p \cdot K_r$ chips total):

$$(q/p) \cdot \frac{\sum_c \text{ones}(c)}{d \cdot K_r^2} - \sum_c \frac{\text{zeros}(c)}{d \cdot K_r^2} \leq \frac{q}{dK_r}.$$

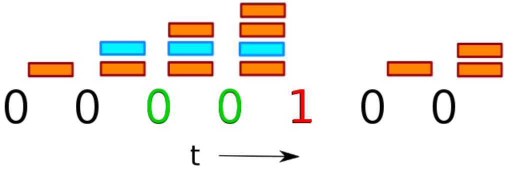
Analysis ideas



$$(q/p) \cdot \Pr[\text{failure in epoch } r] \; - \; \Pr[\text{success in epoch } r]$$

$$\leq \; O(K_r) \; = \; O(r^{-2}).$$

Analysis ideas



$$(q/p) \cdot \Pr[\text{failure in epoch } r] \; - \; \Pr[\text{success in epoch } r]$$

$$\leq \; O(K_r) \; = \; O(r^{-2}).$$

Q.E.D.

Also in the paper

- Bit prediction for broader classes of assumptions:
- E.g., predict a bit (0 or 1), under the assumption that a certain word appears only rarely.

Also in the paper

- Bit prediction for broader classes of assumptions:
- E.g., predict a bit (0 or 1), under the assumption that a certain word appears only rarely.
- Results are quantitatively weaker (by necessity)...

Also in the paper

- Bit prediction for broader classes of assumptions:
- E.g., predict a bit (0 or 1), under the assumption that a certain word appears only rarely.
- Results are quantitatively weaker (by necessity)...
- General statement involves finite automata.

Also in the paper

- Also: high-confidence predictions under no assumptions on x!

Ignorant interval-forecasting

- Sequence $x \in \{0, 1\}^{\omega}$: now completely arbitrary.
- **Goal:** predict the fraction of 1s in an unseen interval, with high accuracy and high confidence.

Ignorant interval-forecasting

- Sequence $x \in \{0, 1\}^\omega$: now completely arbitrary.
- **Goal:** predict the fraction of 1s in an unseen interval, with high accuracy and high confidence.

(Huh?)

Ignorant interval-forecasting

- **Our "hook"**—we get to choose the position and size of the prediction-interval.

- **Interval-forecaster alg.:** makes a prediction of form:

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

Ignorant interval-forecasting

Theorem

For any $\varepsilon, \delta > 0$, there is a ignorant interval-forecaster $\mathcal{S}_{\varepsilon,\delta}$ that is accurate to $\pm\varepsilon$, with success probability $1 - \delta$.

Ignorant interval-forecasting

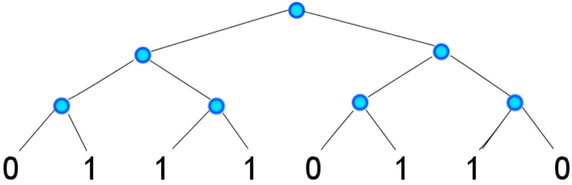
Theorem

For any $\varepsilon, \delta > 0$, there is a ignorant interval-forecaster $\mathcal{S}_{\varepsilon,\delta}$ that is accurate to $\pm\varepsilon$, with success probability $1-\delta$.

Runtime *of $\mathcal{S}_{\varepsilon,\delta}$ is finite: $= 2^{O(\varepsilon^{-2}\delta^{-1})}$.*

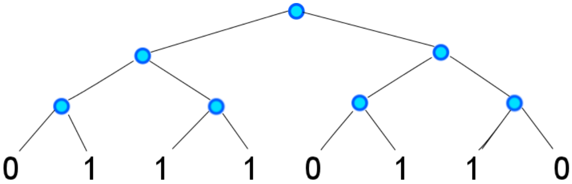
The approach

- Consider $x \in \{0,1\}^{2^n}$, $n = \lceil 4/(\varepsilon^2\delta) \rceil$.
- Arrange bits of x on leaves of a binary tree $\mathcal{T}$.



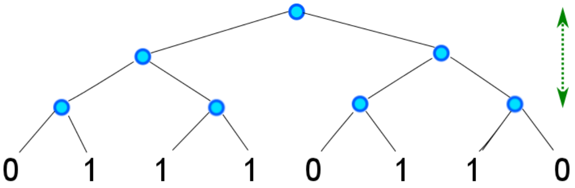
The approach

Forecasting algorithm:



The approach

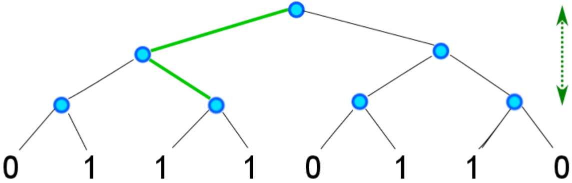
Forecasting algorithm:



1. Pick a uniform $t^* \in \{0, 1, \ldots, n-1\}$;

The approach

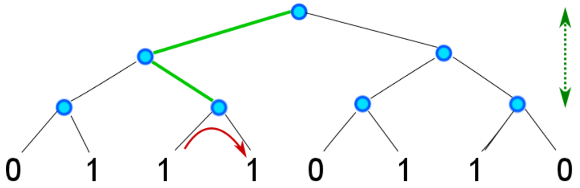
Forecasting algorithm:



1. Pick a uniform $t^* \in \{0, 1, \ldots, n-1\}$;
2. Take a random walk in $\mathcal{T}$ of length t^*;

The approach

Forecasting algorithm:



1. Pick a uniform $t^* \in \{0, 1, \ldots, n-1\}$;
2. Take a random walk in $\mathcal{T}$ of length t^*;
3. **Predict** that

$$\text{(fraction of 1s in right subtree)} = \text{(fraction of 1s in left subtree)}.$$

Analysis idea

- Extend to a length-n random walk; for $0 \le t \le n$ let

$$X(t) \quad \in [0, 1]$$

denote the fraction of 1s below the t-th step vertex.

Analysis idea

- Extend to a length-n random walk; for $0 \leq t \leq n$ let

$$X(t) \ \in [0, 1]$$

denote the fraction of 1s below the t-th step vertex.

- **Fact:**

$$X(0), X(1), \ldots, X(n)$$

is a martingale (for any x), from which:

$$\mathbb{E}[(X(t+1) - X(t))(X(s+1) - X(s))] \ = \ 0,$$

for all $s < t$.

Analysis idea

- Extend to a length-n random walk; for $0 \le t \le n$ let

$$X(t) \in [0, 1]$$

denote the fraction of 1s below the t-th step vertex.

- **Fact:**

$$X(0), X(1), \ldots, X(n)$$

is a martingale (for any x), from which:

$$\mathbb{E}[(X(t+1) - X(t))(X(s+1) - X(s))] = 0,$$

for all $s < t$. Thus:

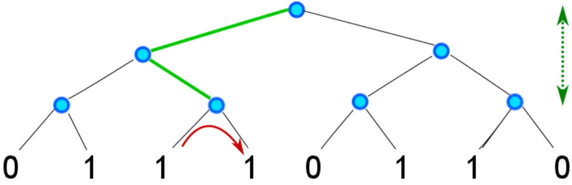
$$\sum_{0 \le t < n} \mathbb{E}[(X(t+1) - X(t))^2] = \mathbb{E}[(X(n) - X(0))^2] \le 1.$$

Analysis idea

$$\sum_{0 \le t < n} \mathbb{E}[(X(t+1) - X(t))^2] \le 1$$

- $(X(t+1) - X(t))^2$ small $\implies$ t is a good choice for t^*!

  (i.e., left and right subtrees have similar 1-densities).



Questions

- Characterize the sets $A \subset \{0, 1\}^{\omega}$ for which confident 0-prediction is possible?

  Connect prediction to geometry, a la (Lutz et al.)?

Questions

- Characterize the sets $A \subset \{0,1\}^\omega$ for which confident 0-prediction is possible?

  Connect prediction to geometry, a la (Lutz et al.)?

- Find an interval-forecaster running in time $\text{poly}(\frac{1}{\delta}, \frac{1}{\varepsilon})$?

Questions

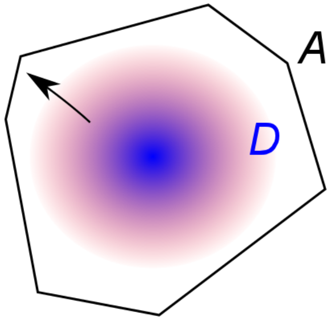
- For which distributions $\mathcal{D}$ on $\{0,1\}^\infty$ can we extend to a "supporting set" A, preserving easiness of prediction?

Questions

- For which distributions $\mathcal{D}$ on $\{0,1\}^\infty$ can we extend to a "supporting set" A, preserving easiness of prediction?

Questions

- For which distributions $\mathcal{D}$ on $\{0,1\}^\infty$ can we extend to a "supporting set" A, preserving easiness of prediction?

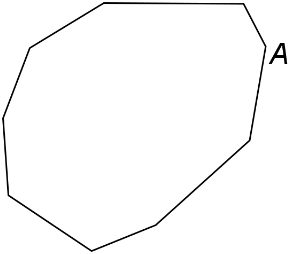


Questions

- Is there a minimax theorem for 0-prediction?
- Hard set A for 0-prediction $\Rightarrow$ hard distribution D over A?

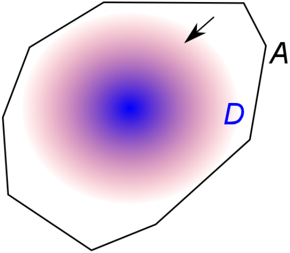
Questions

- Is there a minimax theorem for 0-prediction?
- Hard set A for 0-prediction $\Rightarrow$ hard distribution D over A?



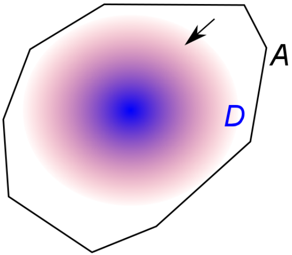
Questions

- Is there a minimax theorem for 0-prediction?
- Hard set A for 0-prediction $\Rightarrow$ hard distribution D over A?



Questions

- Is there a minimax theorem for 0-prediction?
- Hard set A for 0-prediction $\Rightarrow$ hard distribution D over A?



- Would give alternate (non-constructive) proof of our main result...

Thanks!