An Elementary Predictor Obtaining $2\sqrt{T}+1$ Distance to Calibration

Mirah Shi (Penn)

Joint work with
Eshwar Ram Arunachaleswaran, Natalie Collina, Aaron Roth (Penn)

Sequential prediction

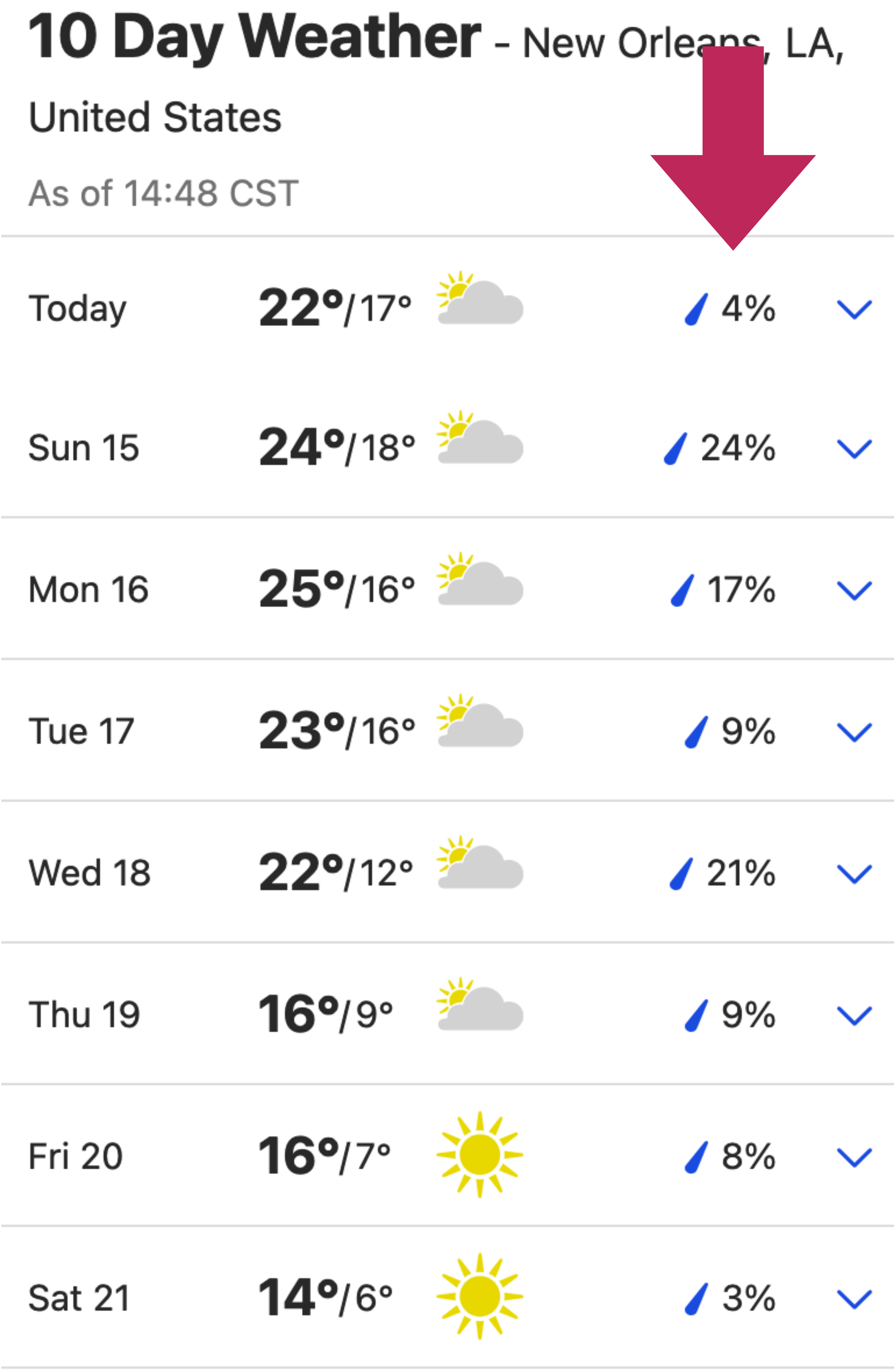
Sequential prediction

10 Day Weather - New Orleans, LA, United States

As of 14:48 CST

Today	22°/17°	4%
Sun 15	24°/18°	24%
Mon 16	25°/16°	17%
Tue 17	23°/16°	9%
Wed 18	22°/12°	21%
Thu 19	16°/9°	9%
Fri 20	16°/7°	8%
Sat 21	14°/6°	3%

Sequential prediction

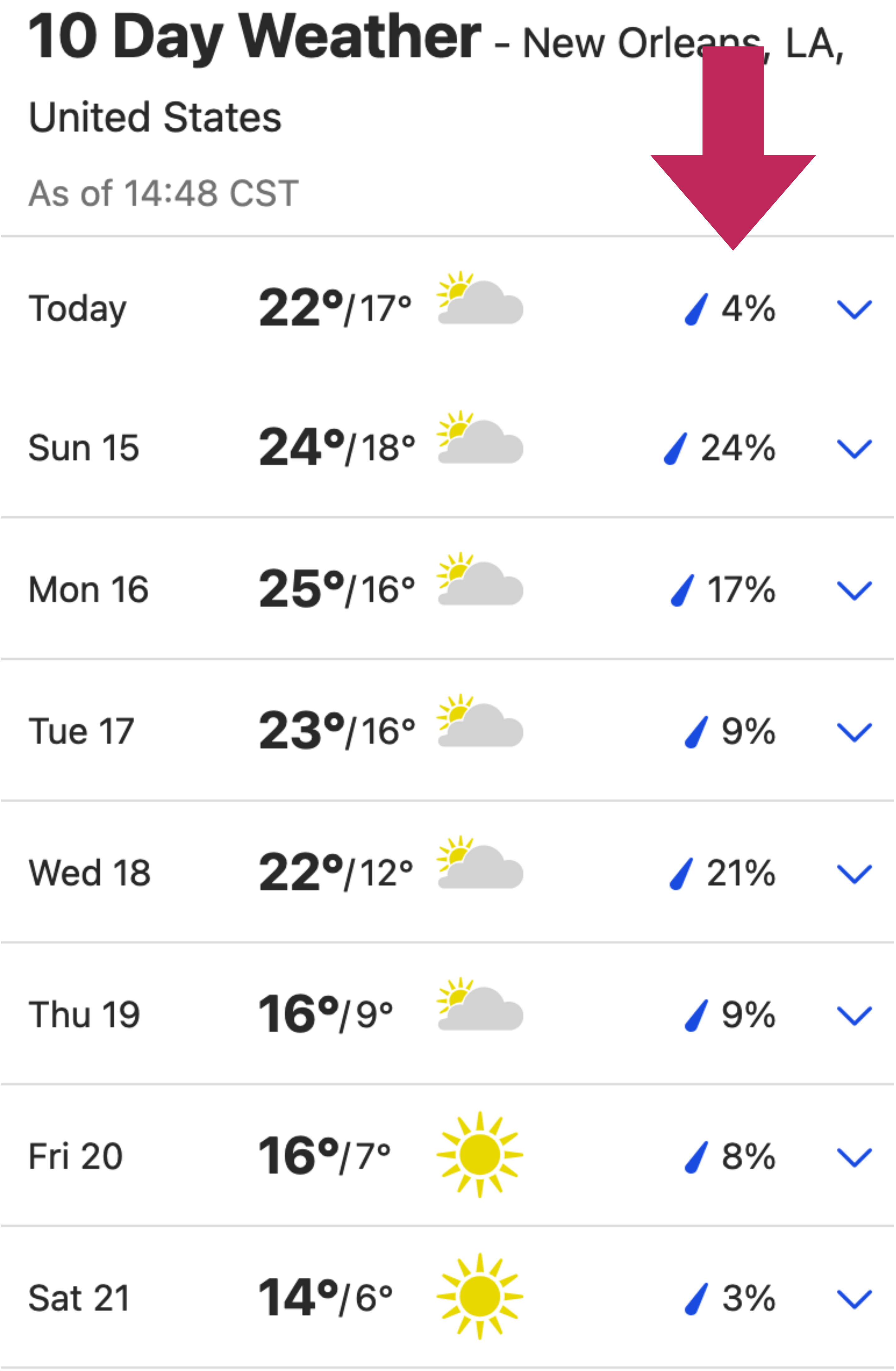


On every day $t = 1,\ldots,T$:

Predict probability $p^t \in [0,1]$

Observe binary outcome $y^t \in \{0,1\}$ (adversarially chosen)

Sequential prediction


10 Day Weather - New Orleans, LA, United States

As of 14:48 CST

Today	22°/17°	4%
Sun 15	24°/18°	24%
Mon 16	25°/16°	17%
Tue 17	23°/16°	9%
Wed 18	22°/12°	21%
Thu 19	16°/9°	9%
Fri 20	16°/7°	8%
Sat 21	14°/6°	3%

On every day $t = 1,\ldots,T$:

Predict probability $p^t \in [0,1]$

Observe binary outcome $y^t \in \{0,1\}$ (adversarially chosen)

How good are the forecasts?

Probabilities should mean something…

Probabilities should mean something…

Calibrated forecasts: “On the days I predicted a 20% chance of rain, it rained 20% of the time.”

Probabilities should mean something...

Calibrated forecasts: "On the days I predicted a 20% chance of rain, it rained 20% of the time."

For all predictions $p \in [0,1]$:

$$\underbrace{\sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t)}_{\textit{bias of } p} = 0$$

Probabilities should mean something…

Calibrated forecasts: "On the days I predicted a 20% chance of rain, it rained 20% of the time."

For all predictions $p \in [0,1]$:

$$\underbrace{\sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t)}_{\textit{bias of } p} = 0$$

How do we measure calibration error?

Measuring calibration error

Expected Calibration Error (ECE):

summed absolute bias of predictions

$$\text{ECE} = \sum_{p \in [0,1]} \left| \underbrace{\sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t)}_{\textit{bias of } p} \right|$$

Measuring calibration error

Expected Calibration Error (ECE):
summed absolute bias of predictions

$$\text{ECE} = \sum_{p \in [0,1]} \left| \underbrace{\sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t)}_{\textit{bias of } p} \right|$$

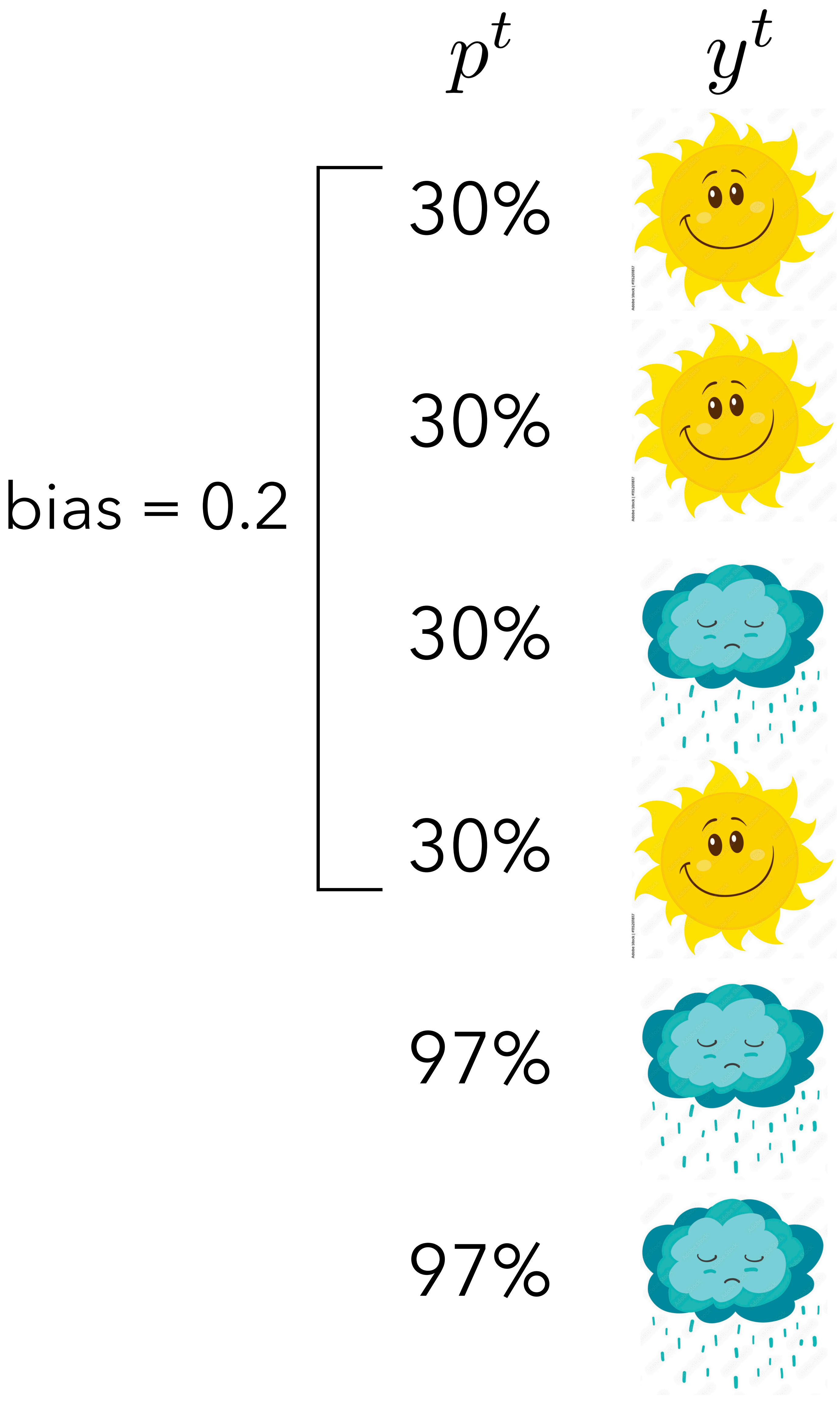
p^t	y^t
30%	
30%	
30%	
30%	
97%	
97%	

Measuring calibration error

Expected Calibration Error (ECE):

summed absolute bias of predictions

$$\mathrm{ECE} = \sum_{p\in[0,1]} \left| \underbrace{\sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t)}_{\text{bias of } p} \right|$$

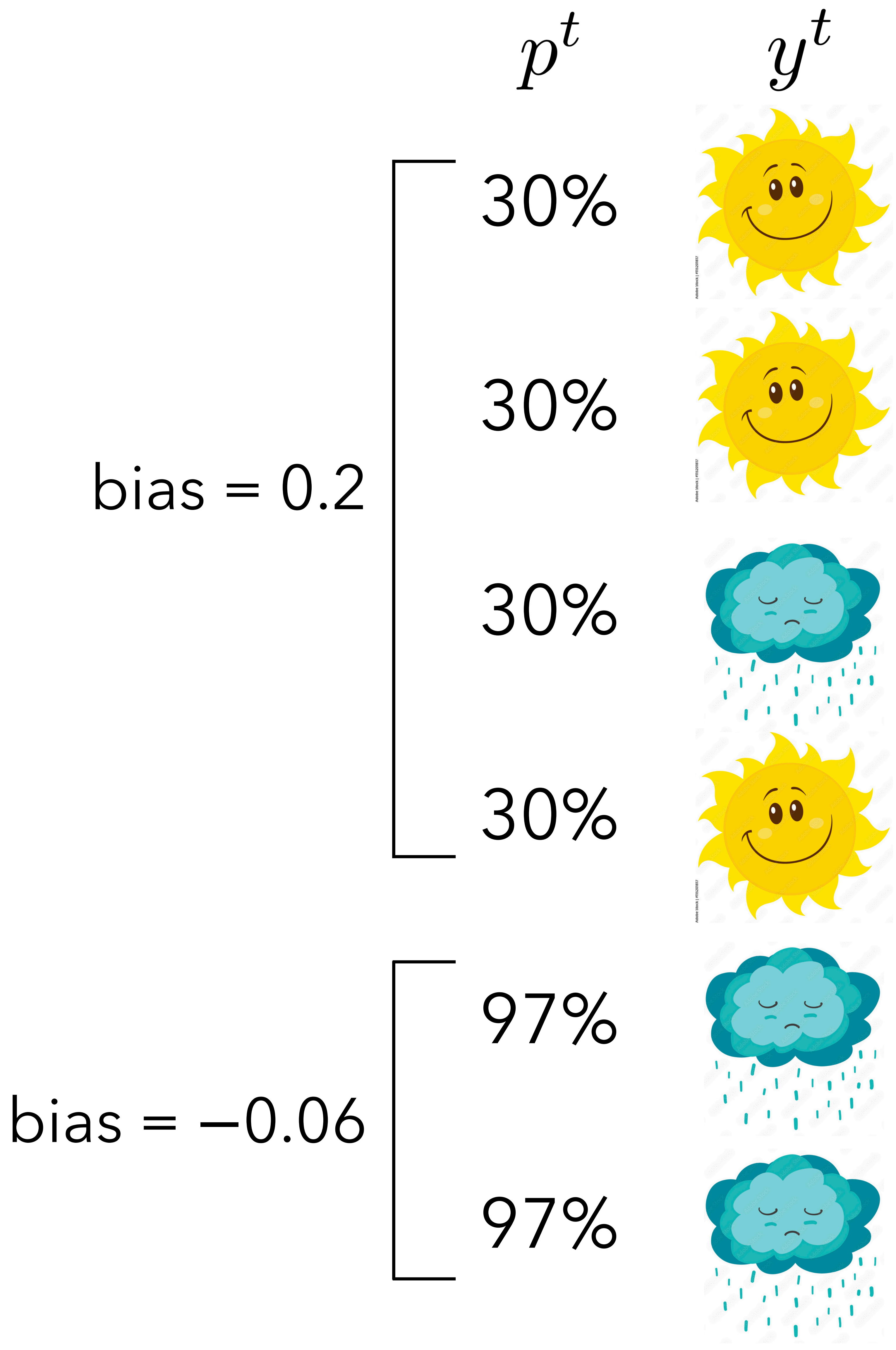


Measuring calibration error

Expected Calibration Error (ECE):

summed absolute bias of predictions

$$\mathrm{ECE} = \sum_{p\in[0,1]} \left| \underbrace{\sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t)}_{\textit{bias of } p} \right|$$



Measuring calibration error

Expected Calibration Error (ECE):
summed absolute bias of predictions

$$\text{ECE} = \sum_{p \in [0,1]} \left| \underbrace{\sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t)}_{\textit{bias of } p} \right|$$

	p^t	y^t
bias = 0.2	30%	
	30%	
	30%	
	30%	
bias = –0.06	97%	
	97%	

ECE = 0.26

Measuring calibration error

Expected Calibration Error (ECE):

summed absolute bias of predictions

$$\text{ECE} = \sum_{p \in [0,1]} \Bigg| \underbrace{\sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t)}_{\textit{bias of } p} \Bigg|$$

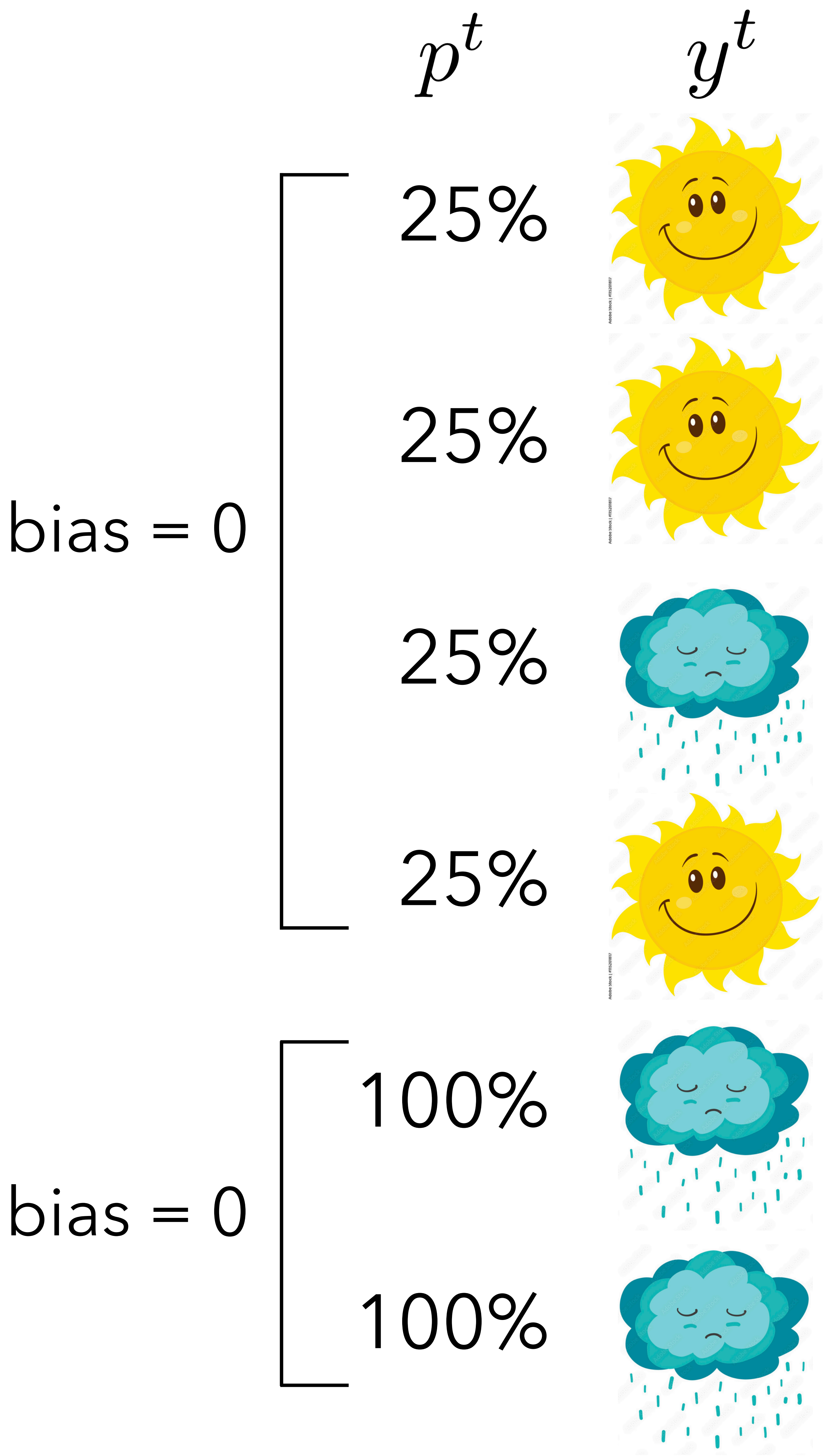
p^t	y^t
25%	
25%	
25%	
25%	
100%	
100%	

Measuring calibration error

Expected Calibration Error (ECE):

summed absolute bias of predictions

$$\text{ECE} = \sum_{p \in [0,1]} \left| \underbrace{\sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t)}_{\textit{bias of } p} \right|$$



Measuring calibration error

Expected Calibration Error (ECE):

summed absolute bias of predictions

$$\text{ECE} = \sum_{p \in [0,1]} \left| \underbrace{\sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t)}_{\textit{bias of } p} \right|$$

	p^t
bias = 0	25%
	25%
	25%
	25%
bias = 0	100%
	100%

y^t

ECE = 0

Measuring calibration error

Expected Calibration Error (ECE):

summed absolute bias of predictions

$$\text{ECE} = \sum_{p \in [0,1]} \left| \underbrace{\sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t)}_{\textit{bias of } p} \right|$$

	p^t	y^t
bias = 0	50%	
	50%	
	50%	
	50%	
	50%	
	50%	

Measuring calibration error

Expected Calibration Error (ECE):

summed absolute bias of predictions

$$\text{ECE} = \sum_{p\in[0,1]} \left| \underbrace{\sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t)}_{\textit{bias of } p} \right|$$

p^t	y^t
50%	
50%	
50%	
50%	
50%	
50%	

bias = 0

$\boxed{\text{ECE} = 0}$

Why ECE?

[Foster-Vohra '96]

Why ECE?

[Foster-Vohra '96]

Trustworthy for decision makers: if predictions satisfy ECE $\leq \epsilon$,

- best responding to predictions is an ϵ-approx dominant strategy, no matter what utility

Why ECE?

[Foster-Vohra '96]

Trustworthy for decision makers: if predictions satisfy ECE $\leq \epsilon$,

- best responding to predictions is an ϵ-approx dominant strategy, no matter what utility
- players best responding in a repeated game converge to an ϵ-approx correlated equilibrium

Why *not* ECE?

Why *not* ECE?

Discontinuous in predictions

Why *not* ECE?

Discontinuous in predictions

p^t	y^t
50%	
50%	
50%	
50%	
50%	
50%	

ECE = 0

Why *not* ECE?

Discontinuous in predictions

p^t	y^t
49%	
48%	
51%	
47%	
53%	
52%	

Why *not* ECE?

Discontinuous in predictions

p^t	y^t
49%	
48%	
51%	
47%	
53%	
52%	

ECE = $\Omega(T)$

Why *not* ECE?

Discontinuous in predictions

Cannot minimize ECE at “good” rates

Why *not* ECE?

Discontinuous in predictions

Cannot minimize ECE at “good” rates

Why *not* ECE?

Discontinuous in predictions

Cannot minimize ECE at “good” rates

Why *not* ECE?

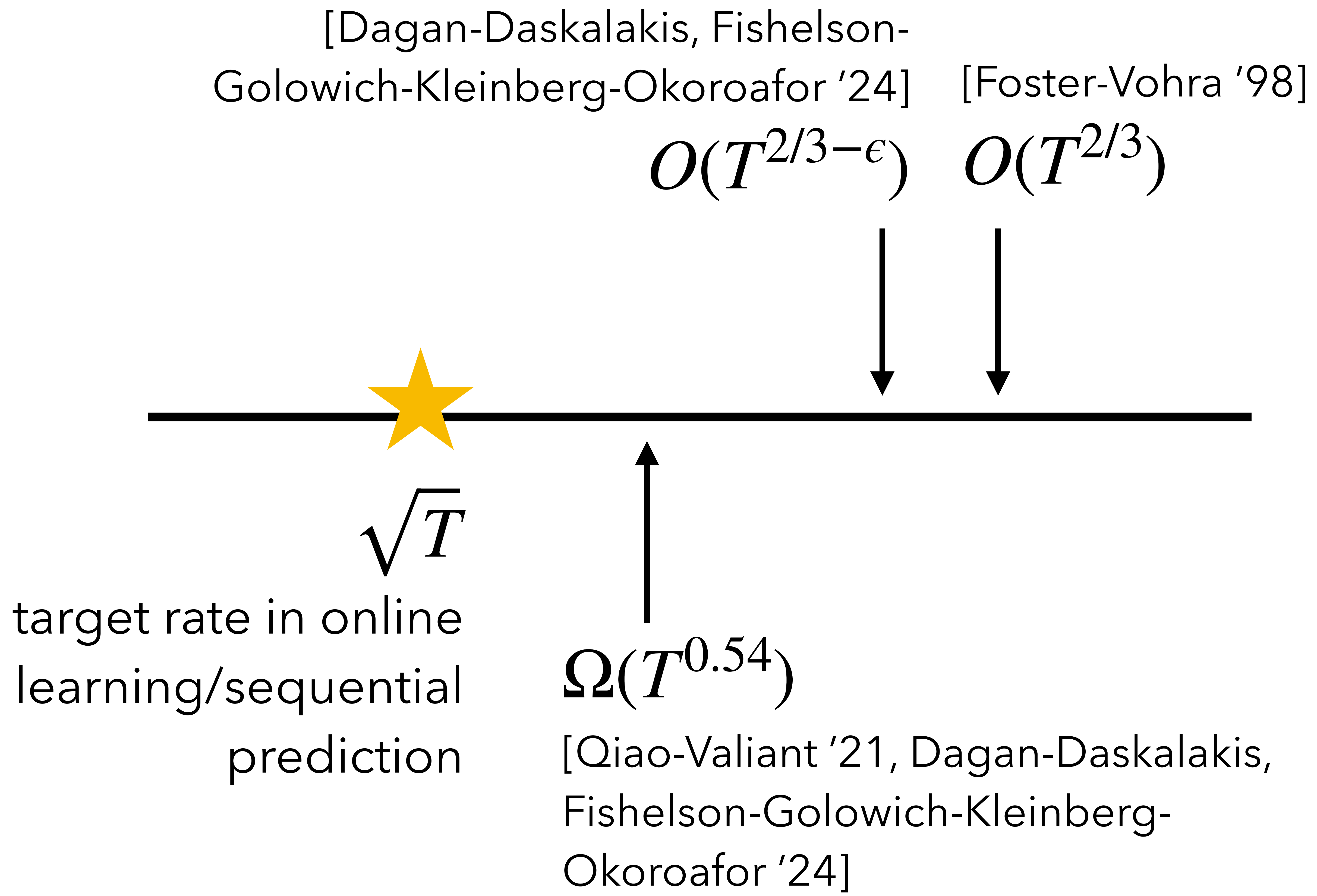
Discontinuous in predictions

Cannot minimize ECE at “good” rates

Why *not* ECE?

Discontinuous in predictions

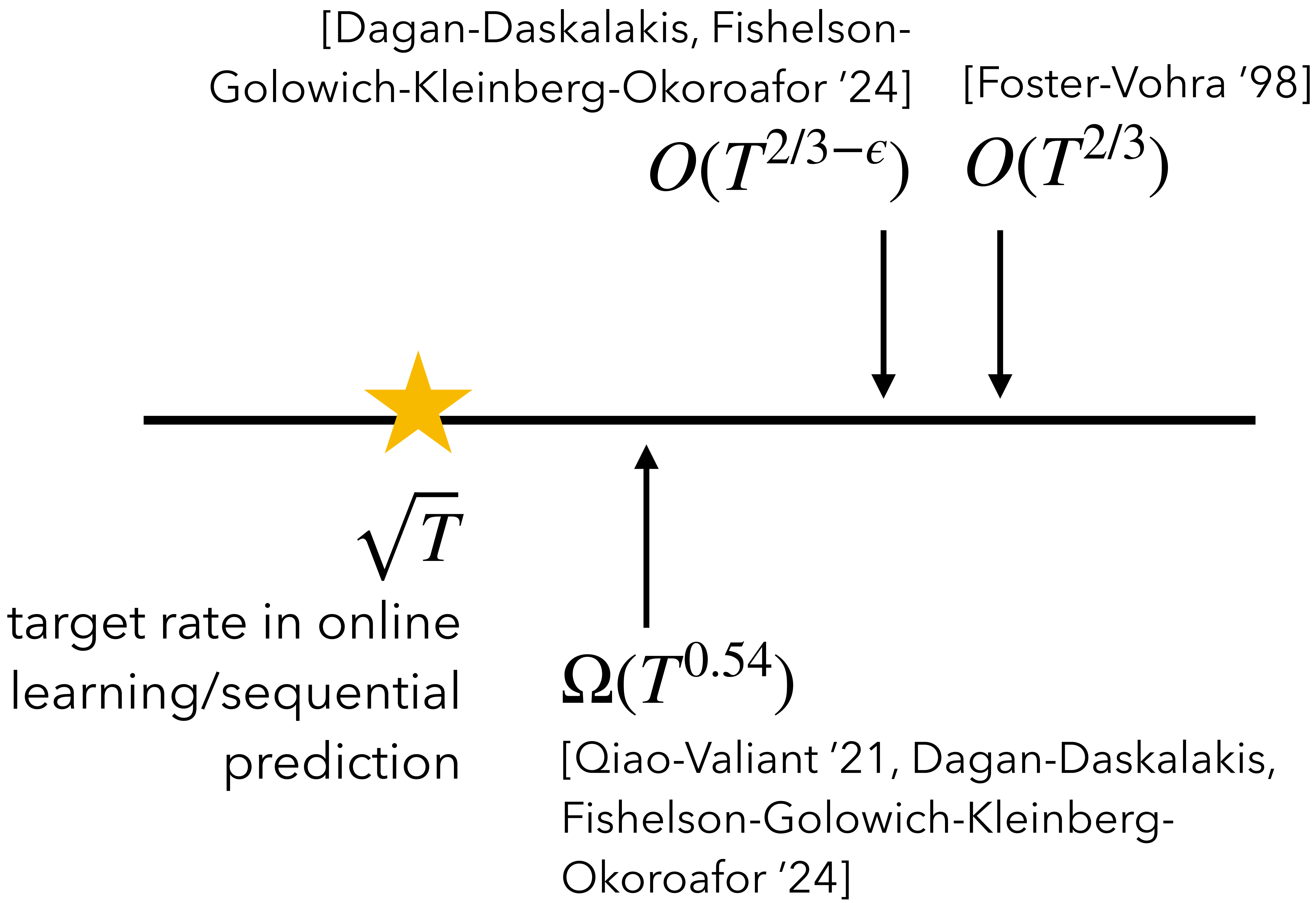
Cannot minimize ECE at "good" rates



Why *not* ECE?

Discontinuous in predictions

Cannot minimize ECE at “good” rates



Punchline: Hard to have low ECE, but easy to be “close”

Distance to calibration

[Blasiok-Gopalan-Hu-Nakkiran '23, Qiao-Zheng '24]

Distance to Calibration (CalDist):

min ℓ_1 distance to any *perfectly calibrated* sequence of predictions

$$\text{CalDist} = \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1$$

where $\mathcal{C}(y^{1:T})$ is the set of predictions with ECE = 0 against outcomes $y^{1:T}$

Distance to calibration

[Blasiok-Gopalan-Hu-Nakkiran '23, Qiao-Zheng '24]

Distance to Calibration (CalDist):

min ℓ_1 distance to any *perfectly calibrated* sequence of predictions

$$\text{CalDist} = \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1$$

where $\mathcal{C}(y^{1:T})$ is the set of predictions with ECE = 0 against outcomes $y^{1:T}$

p^t	y^t
50%	
50%	
50%	
50%	
50%	
50%	

Distance to calibration

[Blasiok-Gopalan-Hu-Nakkiran '23, Qiao-Zheng '24]

Distance to Calibration (CalDist):

min ℓ_1 distance to any *perfectly calibrated* sequence of predictions

$$\text{CalDist} = \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1$$

where $\mathcal{C}(y^{1:T})$ is the set of predictions with ECE = 0 against outcomes $y^{1:T}$

p^t	y^t
50%	
50%	
50%	
50%	
50%	
50%	

CalDist = 0

Distance to calibration

[Blasiok-Gopalan-Hu-Nakkiran '23, Qiao-Zheng '24]

Distance to Calibration (CalDist):

min ℓ_1 distance to any *perfectly calibrated* sequence of predictions

$$\text{CalDist} = \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1$$

where $\mathcal{C}(y^{1:T})$ is the set of predictions with ECE = 0 against outcomes $y^{1:T}$

p^t	y^t
49%	
48%	
51%	
47%	
53%	
52%	

Distance to calibration

[Blasiok-Gopalan-Hu-Nakkiran '23, Qiao-Zheng '24]

Distance to Calibration (CalDist):

min ℓ_1 distance to any *perfectly calibrated* sequence of predictions

$$\text{CalDist} = \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1$$

where $\mathcal{C}(y^{1:T})$ is the set of predictions with ECE = 0 against outcomes $y^{1:T}$

p^t	y^t
49%	
48%	
51%	
47%	
53%	
52%	

CalDist = $O(1)$

Distance to calibration

[Blasiok-Gopalan-Hu-Nakkiran '23, Qiao-Zheng '24]

Distance to Calibration (CalDist):

min ℓ_1 distance to any *perfectly calibrated* sequence of predictions

$$\text{CalDist} \;=\; \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1$$

Continuous in predictions!

where $\mathcal{C}(y^{1:T})$ is the set of predictions with ECE = 0 against outcomes $y^{1:T}$

p^t y^t

49%

48%

51%

47%

53%

52%

CalDist = $O(1)$

Distance to calibration

[Blasiok-Gopalan-Hu-Nakkiran '23, Qiao-Zheng '24]

Distance to Calibration (CalDist):

min ℓ_1 distance to any *perfectly calibrated* sequence of predictions

$$\text{CalDist} = \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1$$

Continuous in predictions!

where $\mathcal{C}(y^{1:T})$ is the set of predictions with ECE = 0 against outcomes $y^{1:T}$

Fact: CalDist $\leq$ ECE

p^t	y^t
49%	
48%	
51%	
47%	
53%	
52%	

CalDist = $O(1)$

Distance to calibration

[Blasiok-Gopalan-Hu-Nakkiran '23, Qiao-Zheng '24]

Distance to Calibration (CalDist):

min ℓ_1 distance to any *perfectly calibrated* sequence of predictions

Continuous in predictions!

$$\text{CalDist} = \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1$$

where $\mathcal{C}(y^{1:T})$ is the set of predictions with ECE = 0 against outcomes $y^{1:T}$

$\boxed{\textbf{Fact}\text{: CalDist} \leq \text{ECE}}$

p^t $\quad$ y^t

30%

30%

30%

30%

97%

97%

Distance to calibration

[Blasiok-Gopalan-Hu-Nakkiran '23, Qiao-Zheng '24]

Distance to Calibration (CalDist):

min ℓ_1 distance to any *perfectly calibrated* sequence of predictions

Continuous in predictions!

$$\text{CalDist} = \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1$$

where $\mathcal{C}(y^{1:T})$ is the set of predictions with ECE = 0 against outcomes $y^{1:T}$

Fact: CalDist $\leq$ ECE

p^t	y^t		
30%		$\longrightarrow$	25%
30%		$\longrightarrow$	25%
30%		$\longrightarrow$	25%
30%		$\longrightarrow$	25%
97%		$\longrightarrow$	100%
97%		$\longrightarrow$	100%

Why distance to calibration?

Continuous in predictions

Why distance to calibration?

Continuous in predictions

Maintains trustworthiness properties for Lipschitz utilities [e.g. Collina-Goel-Gupta-Roth ’24]

Why distance to calibration?

Continuous in predictions

Maintains trustworthiness properties for Lipschitz utilities [e.g. Collina-Goel-Gupta-Roth ’24]

And… much more tractable!

Why distance to calibration?

Continuous in predictions

Maintains trustworthiness properties for Lipschitz utilities [e.g. Collina-Goel-Gupta-Roth '24]

And… much more tractable!

We give a predictor that achieves $2\sqrt{T} + 1$ distance to calibration.

Bonus: it's deterministic!

Why distance to calibration?

Continuous in predictions

Maintains trustworthiness properties for Lipschitz utilities [e.g. Collina-Goel-Gupta-Roth '24]

And… much more tractable!

Beats $T^{0.54}$ lower bound for ECE

We give a predictor that achieves $2\sqrt{T} + 1$ distance to calibration.

Bonus: it's deterministic!

Why distance to calibration?

Continuous in predictions

Maintains trustworthiness properties for Lipschitz utilities [e.g. Collina-Goel-Gupta-Roth '24]

And… much more tractable!

Beats $T^{0.54}$ lower bound for ECE

We give a predictor that achieves $2\sqrt{T}+1$ distance to calibration.

Bonus: it's deterministic!

Before: existence proof of randomized predictor achieving $O(\sqrt{T})$ distance to calibration [Qiao-Zheng '24]

The algorithm is extremely simple.

The algorithm is extremely simple.

And so is the analysis.

The algorithm is extremely simple.

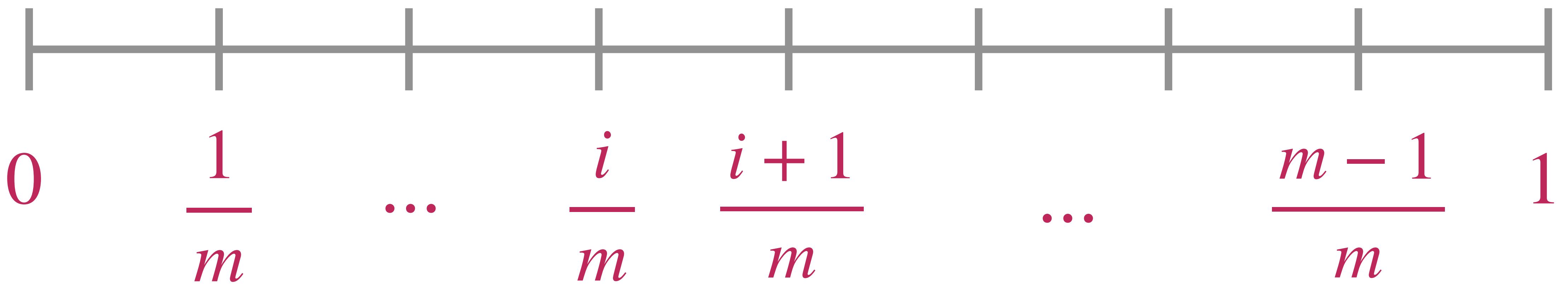
And so is the analysis.

Let’s go.

First, a fictitious algorithm: One Step Ahead

First, a fictitious algorithm: One Step Ahead

Discretize predictions:

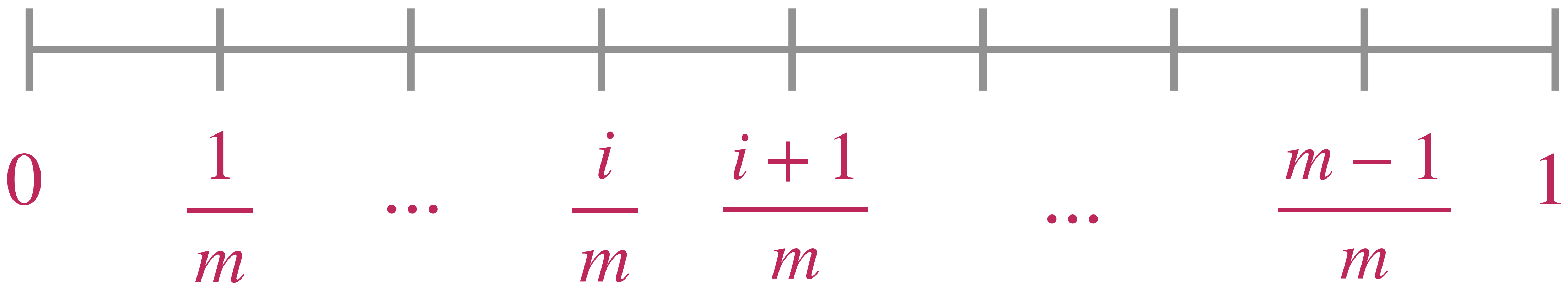


First, a fictitious algorithm: One Step Ahead

On day $t = 1,\ldots,T$:

1. Fix two adjacent points i/m and $(i+1)/m$ with negative and positive bias so far (guaranteed to exist!)
2. Look at outcome y^t
3. Predict i/m if $y^t = 0$, $(i+1)/m$ if $y^t = 1$

Discretize predictions:

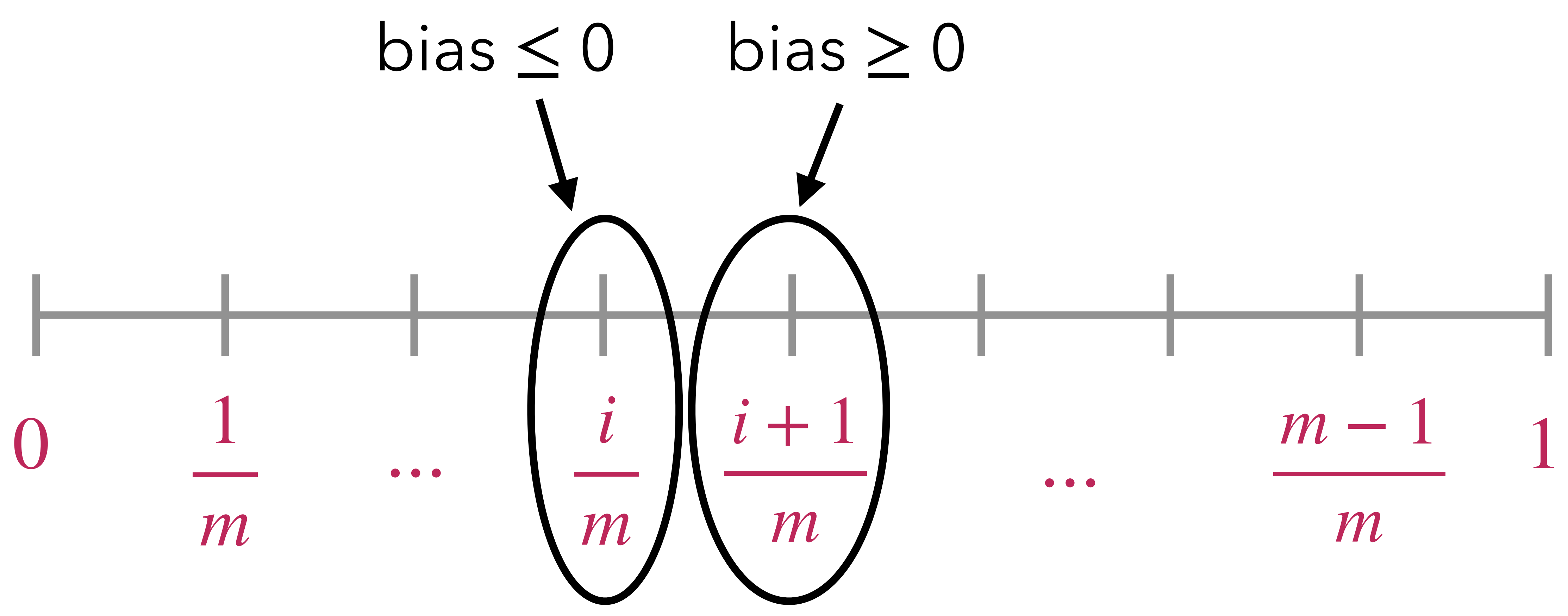


First, a fictitious algorithm: One Step Ahead 🔮

On day $t = 1,\ldots,T$:

1. Fix two adjacent points i/m and $(i+1)/m$ with negative and positive bias so far (guaranteed to exist!)
2. Look at outcome y^t
3. Predict i/m if $y^t = 0$, $(i+1)/m$ if $y^t = 1$

Discretize predictions:

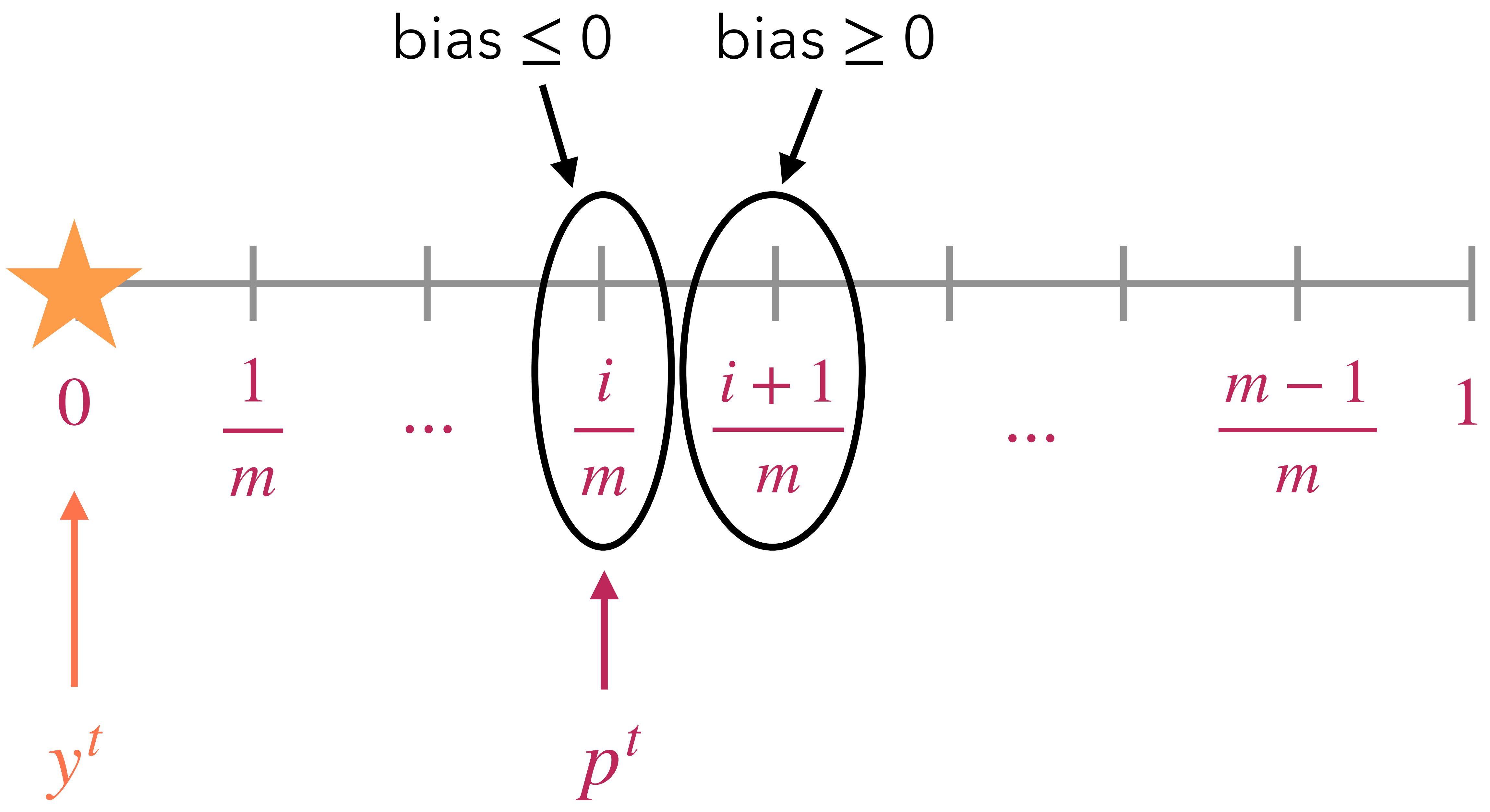


First, a fictitious algorithm: One Step Ahead 🔮

On day $t = 1,\ldots,T$:

1. Fix two adjacent points i/m and $(i+1)/m$ with negative and positive bias so far (guaranteed to exist!)
2. Look at outcome y^t
3. Predict i/m if $y^t = 0$, $(i+1)/m$ if $y^t = 1$

Discretize predictions:

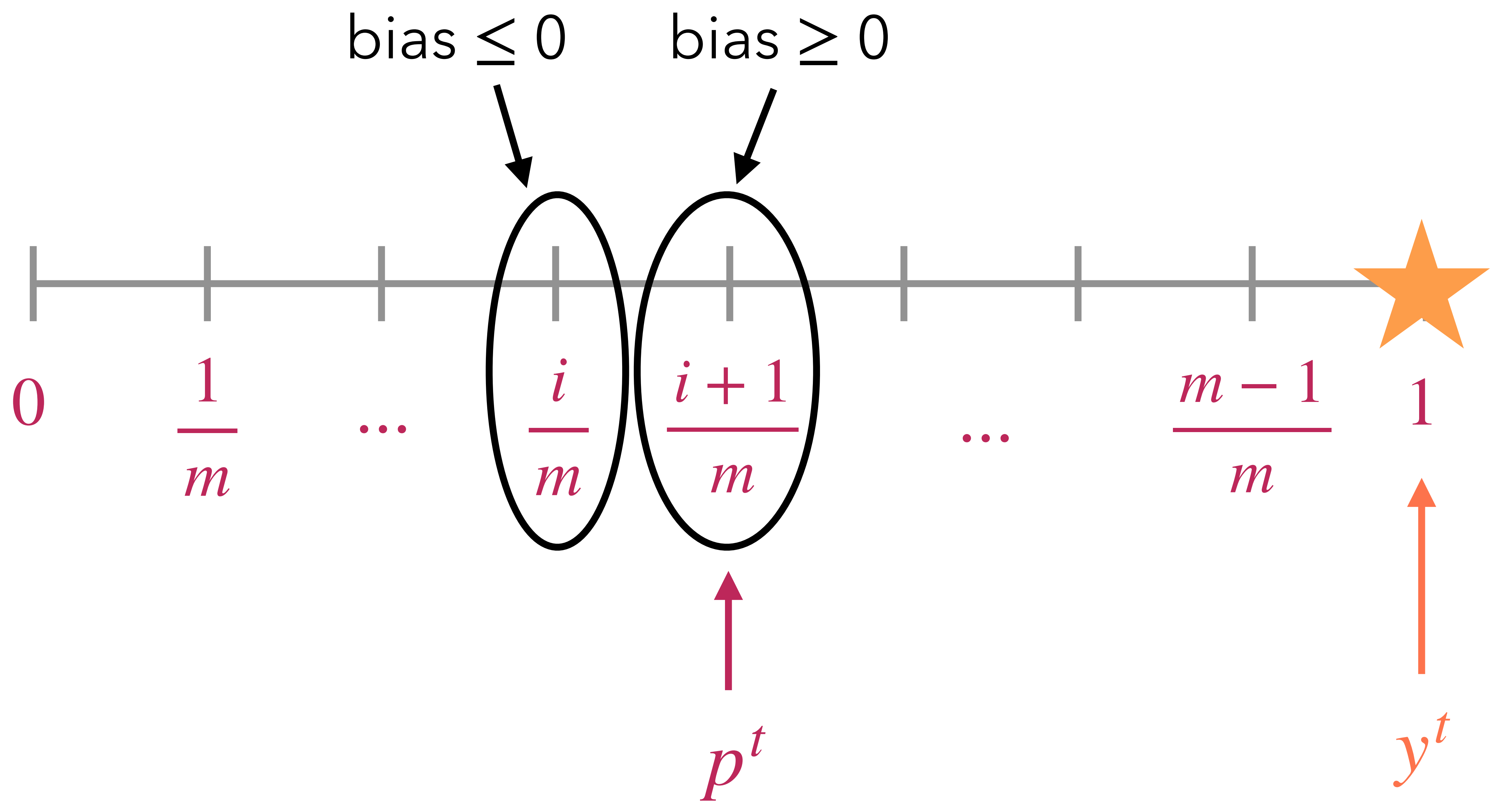


First, a fictitious algorithm: One Step Ahead

On day $t = 1, \ldots, T$:

1. Fix two adjacent points i/m and $(i+1)/m$ with negative and positive bias so far (guaranteed to exist!)
2. Look at outcome y^t
3. Predict i/m if $y^t = 0$, $(i+1)/m$ if $y^t = 1$

Discretize predictions:

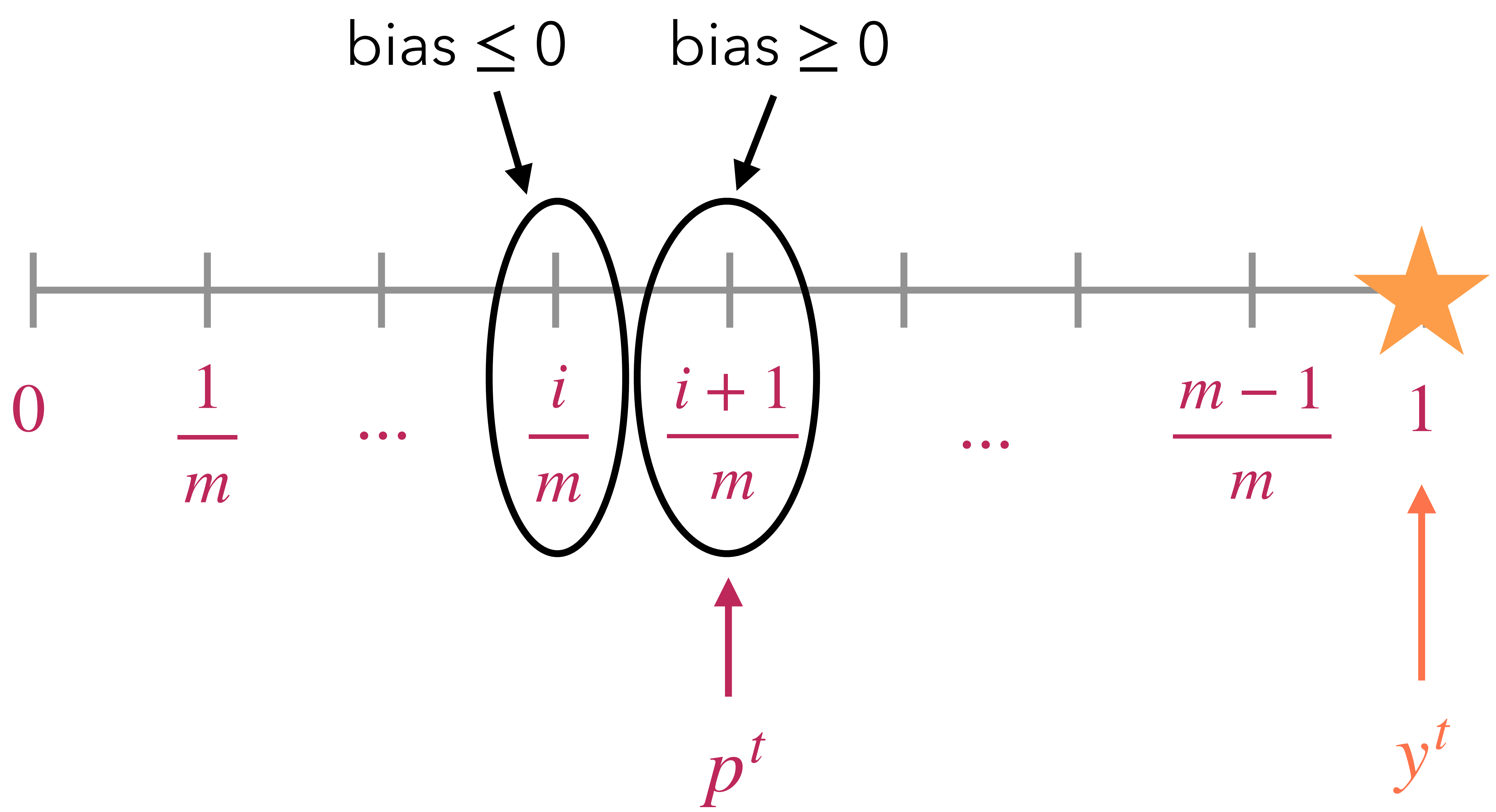


First, a fictitious algorithm: One Step Ahead 🔮

On day $t = 1,...,T$:

1. Fix two adjacent points i/m and $(i+1)/m$ with negative and positive bias so far (guaranteed to exist!)
2. Look at outcome y^t
3. Predict i/m if $y^t = 0$, $(i+1)/m$ if $y^t = 1$

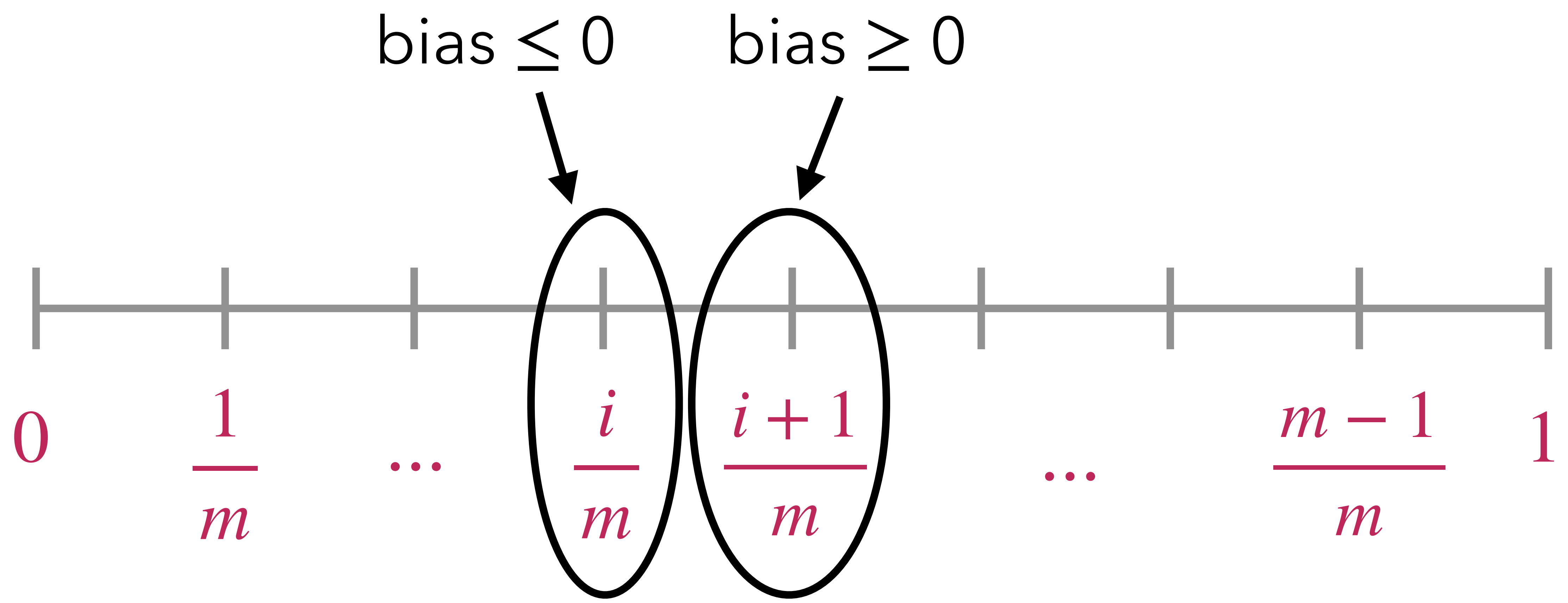
Discretize predictions:



Q: What is CalDist of One Step Ahead?

First, a fictitious algorithm: One Step Ahead

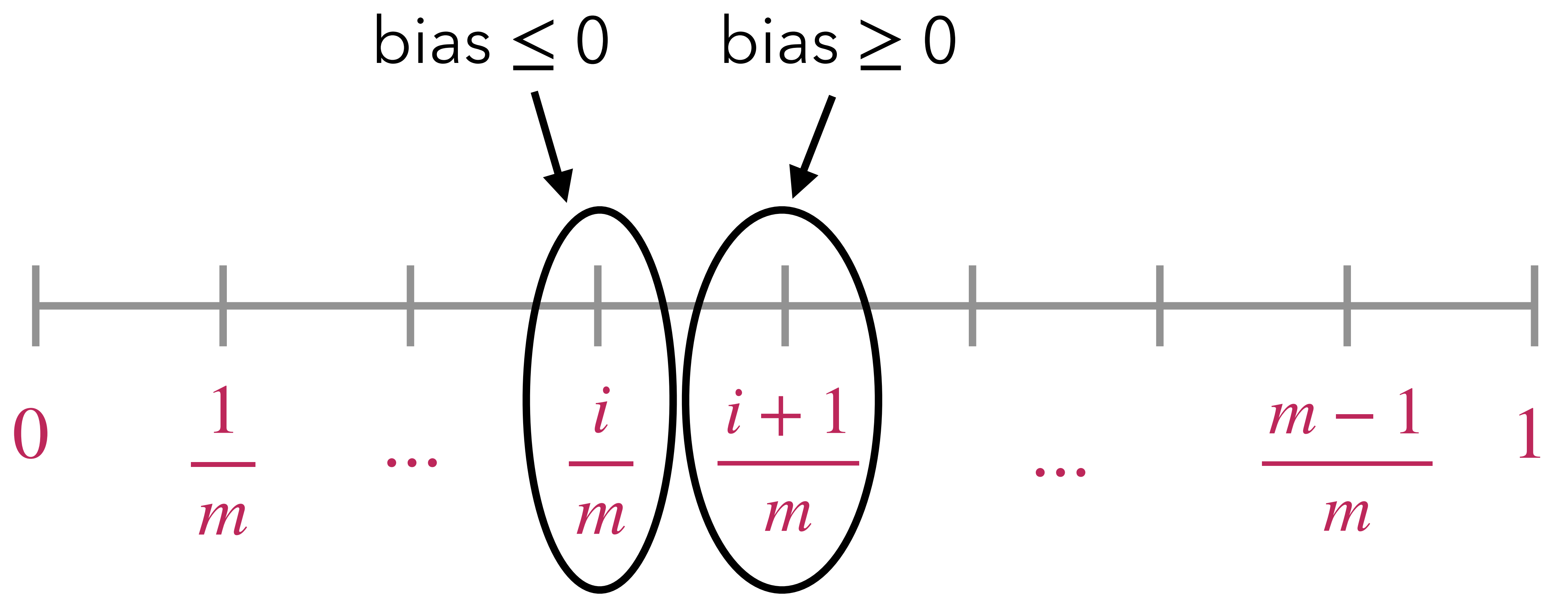
Lemma: One Step Ahead achieves
CalDist $\leq m + 1$



First, a fictitious algorithm: One Step Ahead

Lemma: One Step Ahead achieves
CalDist $\leq m + 1$

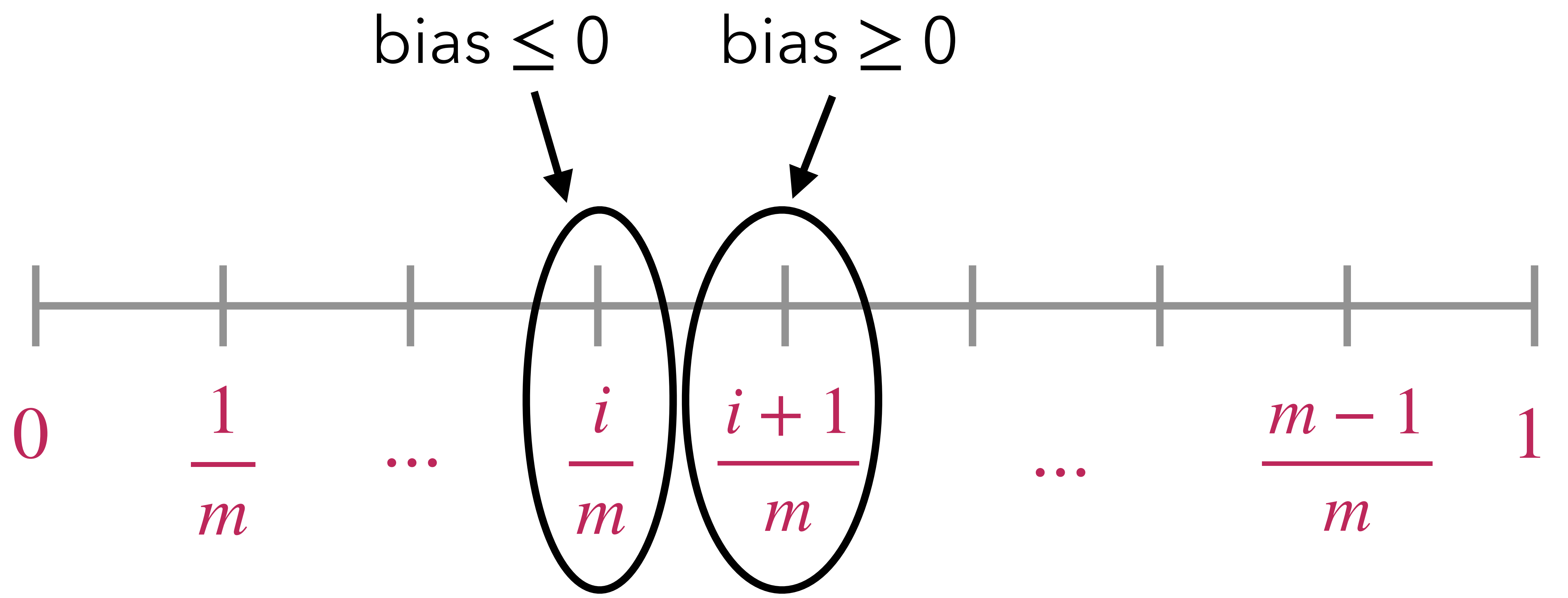
Proof:



First, a fictitious algorithm: One Step Ahead 🔮

Lemma: One Step Ahead achieves
CalDist $\leq m + 1$

Proof:
CalDist $\leq$ ECE



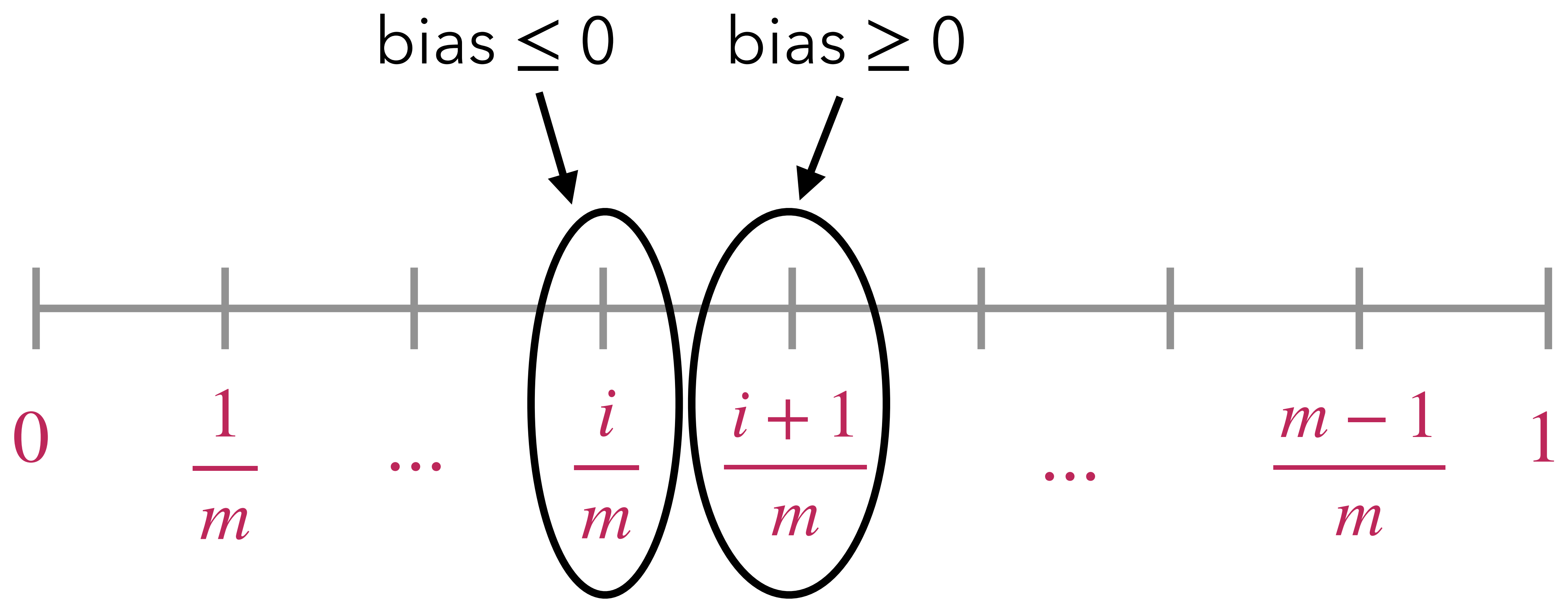
First, a fictitious algorithm: One Step Ahead

Lemma: One Step Ahead achieves
CalDist $\leq m + 1$

Proof:

CalDist $\leq$ ECE

$$= \sum_{p \in [0,1]} \left| \sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t) \right|$$



First, a fictitious algorithm: One Step Ahead

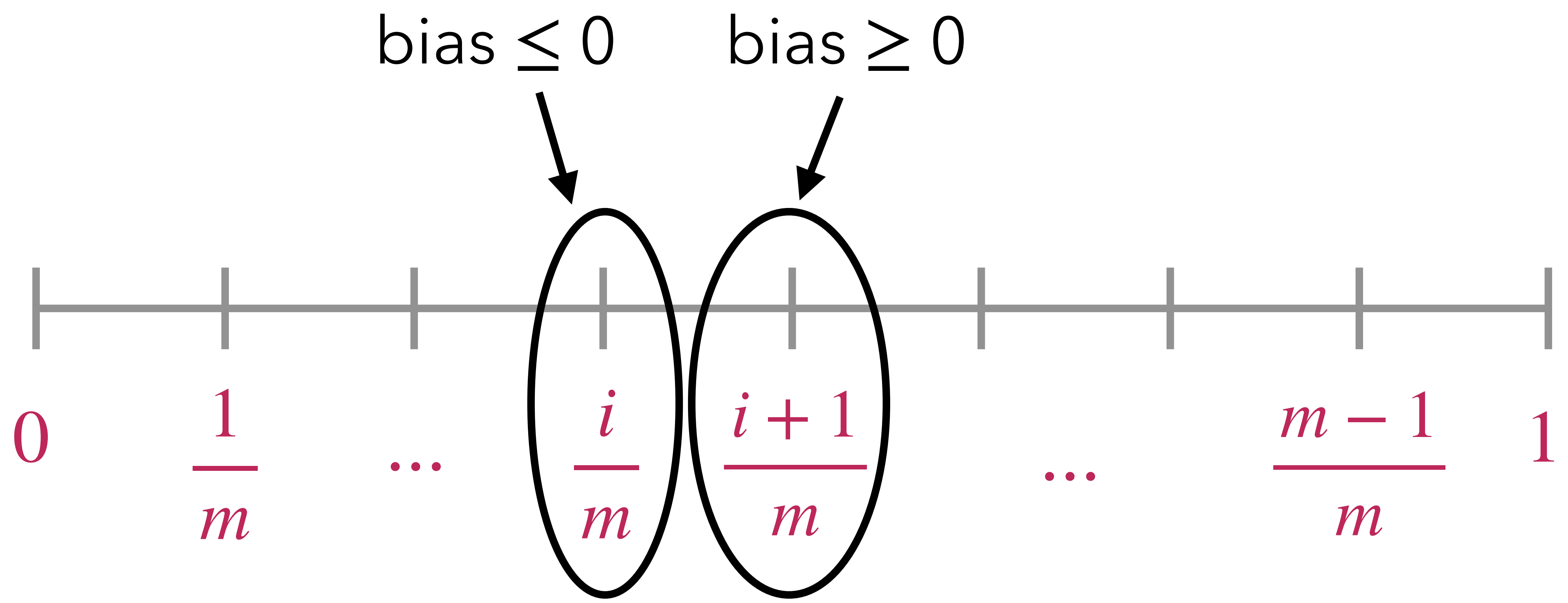
Lemma: One Step Ahead achieves
CalDist $\leq m + 1$

Proof:

CalDist $\leq$ ECE

$$= \sum_{p \in [0,1]} \boxed{\left| \sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t) \right|}$$

bias moves in opposite direction every day $\longrightarrow$ absolute value always ≤ 1



First, a fictitious algorithm: One Step Ahead

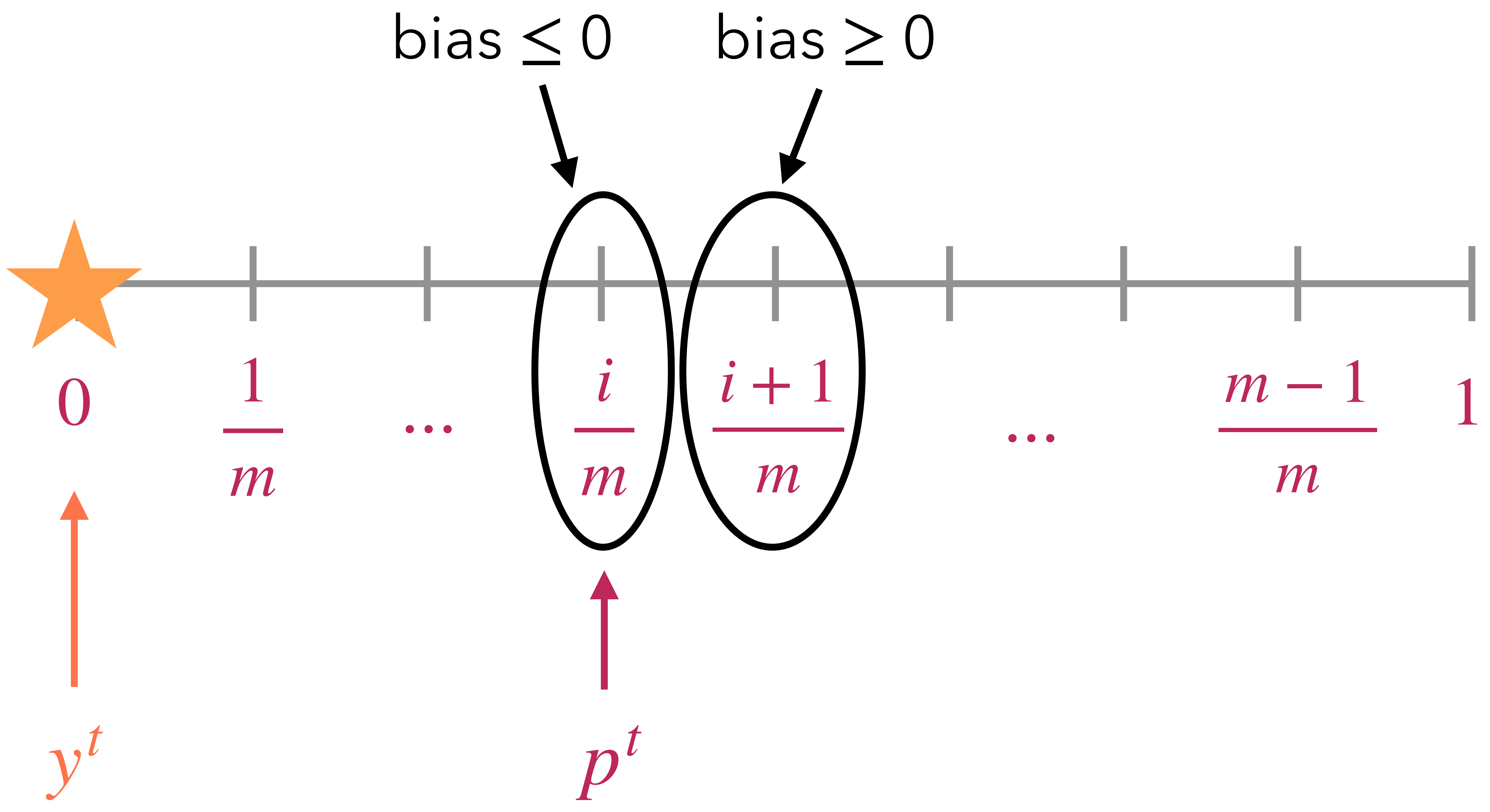
Lemma: One Step Ahead achieves
CalDist $\leq m + 1$

Proof:

CalDist $\leq$ ECE

$$= \sum_{p\in[0,1]} \boxed{\left|\sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t)\right|}$$

bias moves in opposite direction every day $\longrightarrow$ absolute value always ≤ 1



First, a fictitious algorithm: One Step Ahead

Lemma: One Step Ahead achieves

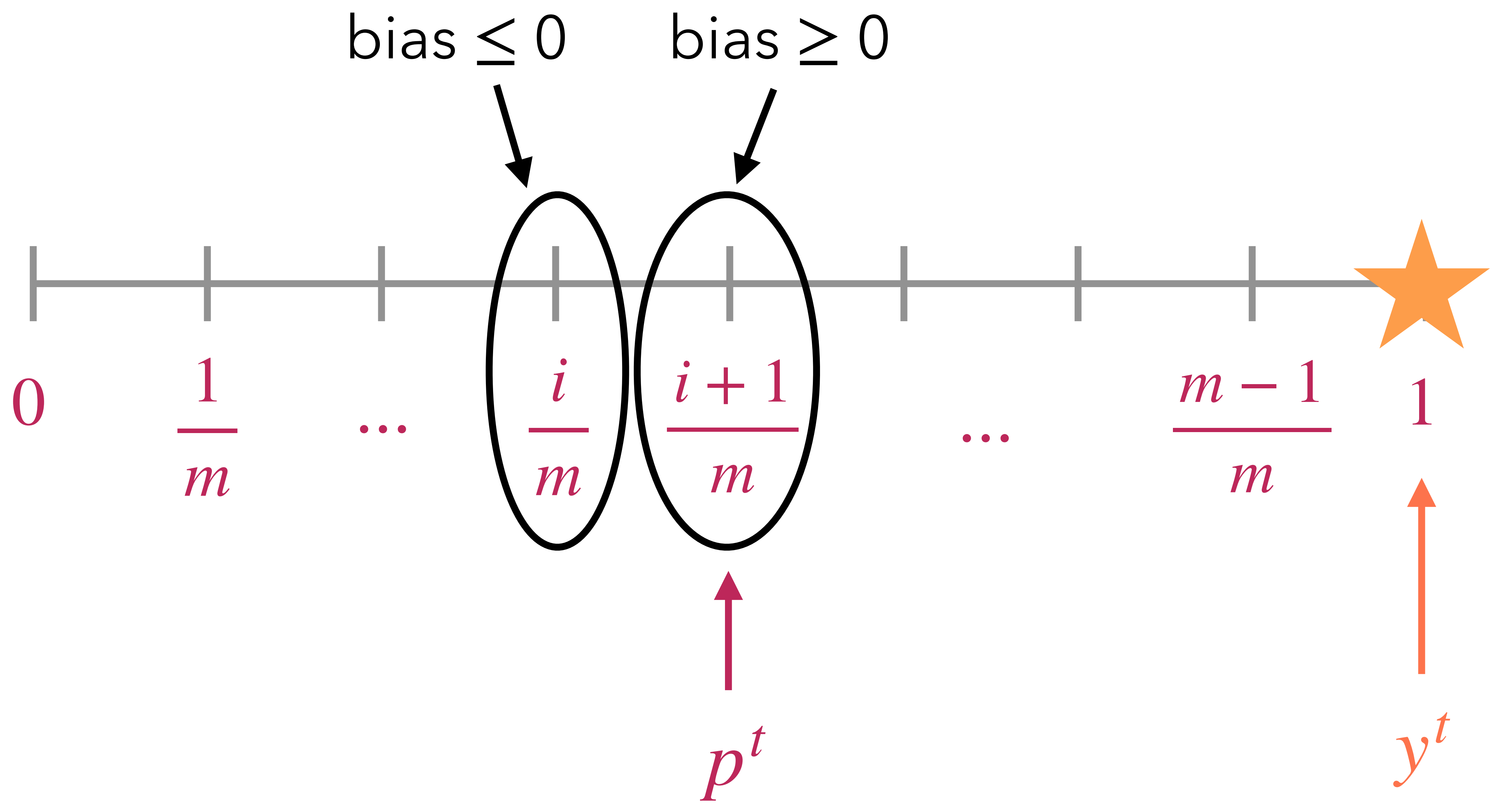
CalDist $\leq m + 1$

Proof:

CalDist $\leq$ ECE

$$= \sum_{p \in [0,1]} \boxed{\left| \sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t) \right|}$$

bias moves in opposite direction every day $\longrightarrow$ absolute value always $\leq$ 1



First, a fictitious algorithm: One Step Ahead

Lemma: One Step Ahead achieves
CalDist $\leq m + 1$

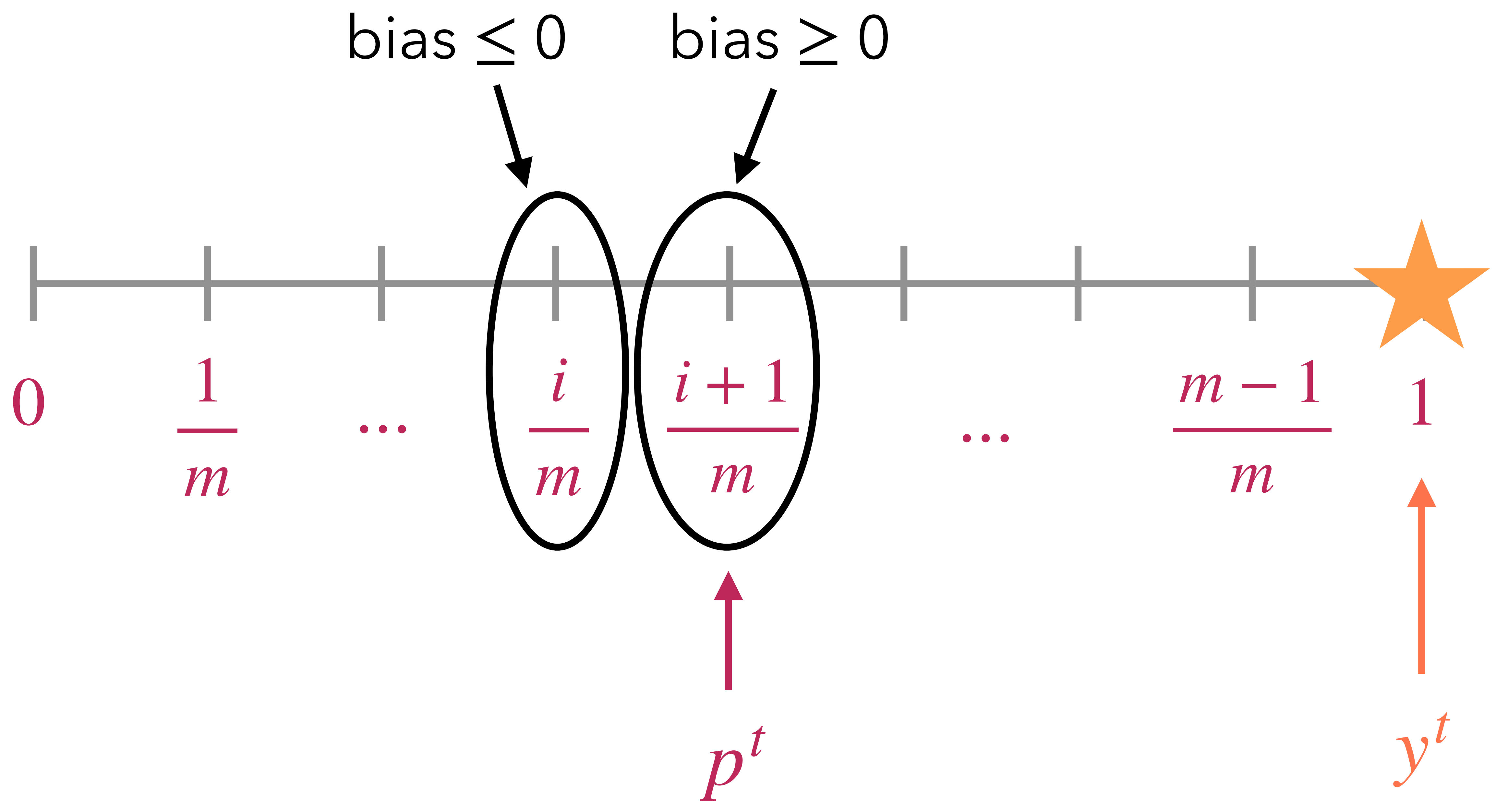
Proof:

CalDist $\leq$ ECE

$$= \sum_{p \in [0,1]} \boxed{\left| \sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t) \right|}$$

bias moves in opposite direction every day $\longrightarrow$ absolute value always ≤ 1

$$\leq m + 1$$



First, a fictitious algorithm: ~~One Step Ahead~~ 

Can’t look into the future…

First, a fictitious algorithm: ~~One Step Ahead~~

Can't look into the future…

…but can be *almost* one step ahead

Almost One Step Ahead

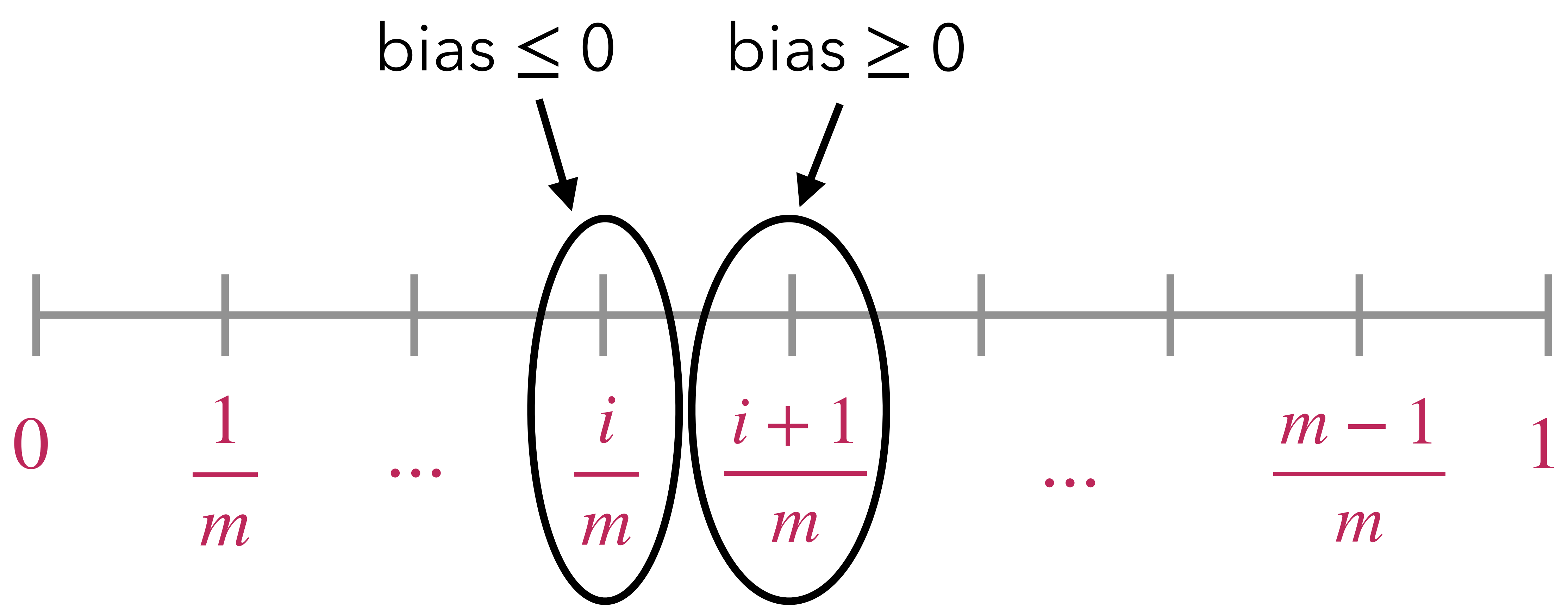
Idea: Mimic One Step Ahead without looking into future

Almost One Step Ahead

Idea: Mimic One Step Ahead without looking into future

On day $t = 1,\ldots,T$:

1. Predict (arbitrarily) one of two points i/m and $(i+1)/m$ that One Step Ahead would commit to on day t
2. Observe outcome y^t
3. Keep track of bias of predictions that One Step Ahead *would have made*

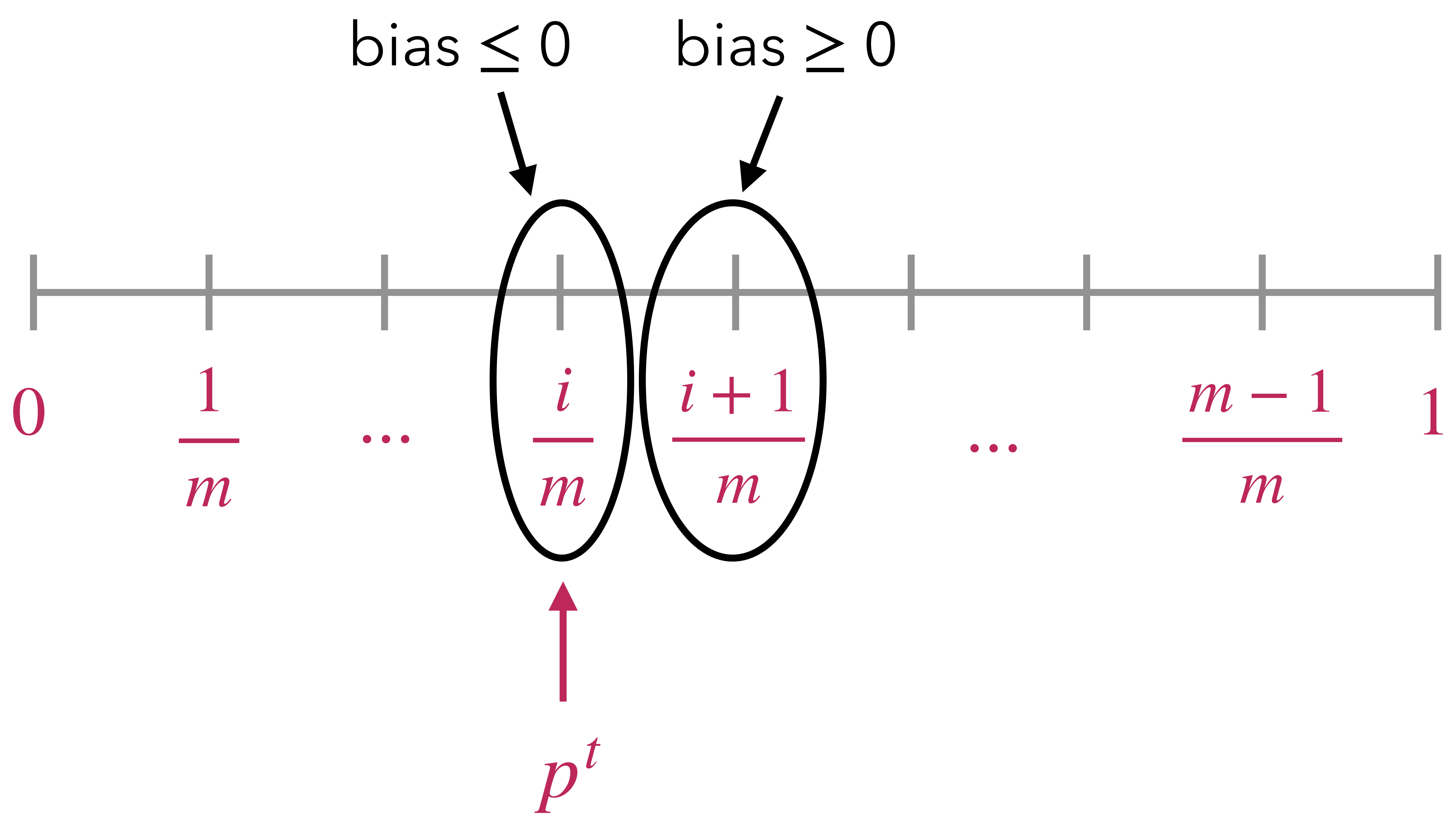


Almost One Step Ahead

Idea: Mimic One Step Ahead without looking into future

On day $t = 1,\ldots,T$:

1. Predict (arbitrarily) one of two points i/m and $(i+1)/m$ that One Step Ahead would commit to on day t
2. Observe outcome y^t
3. Keep track of bias of predictions that One Step Ahead *would have made*

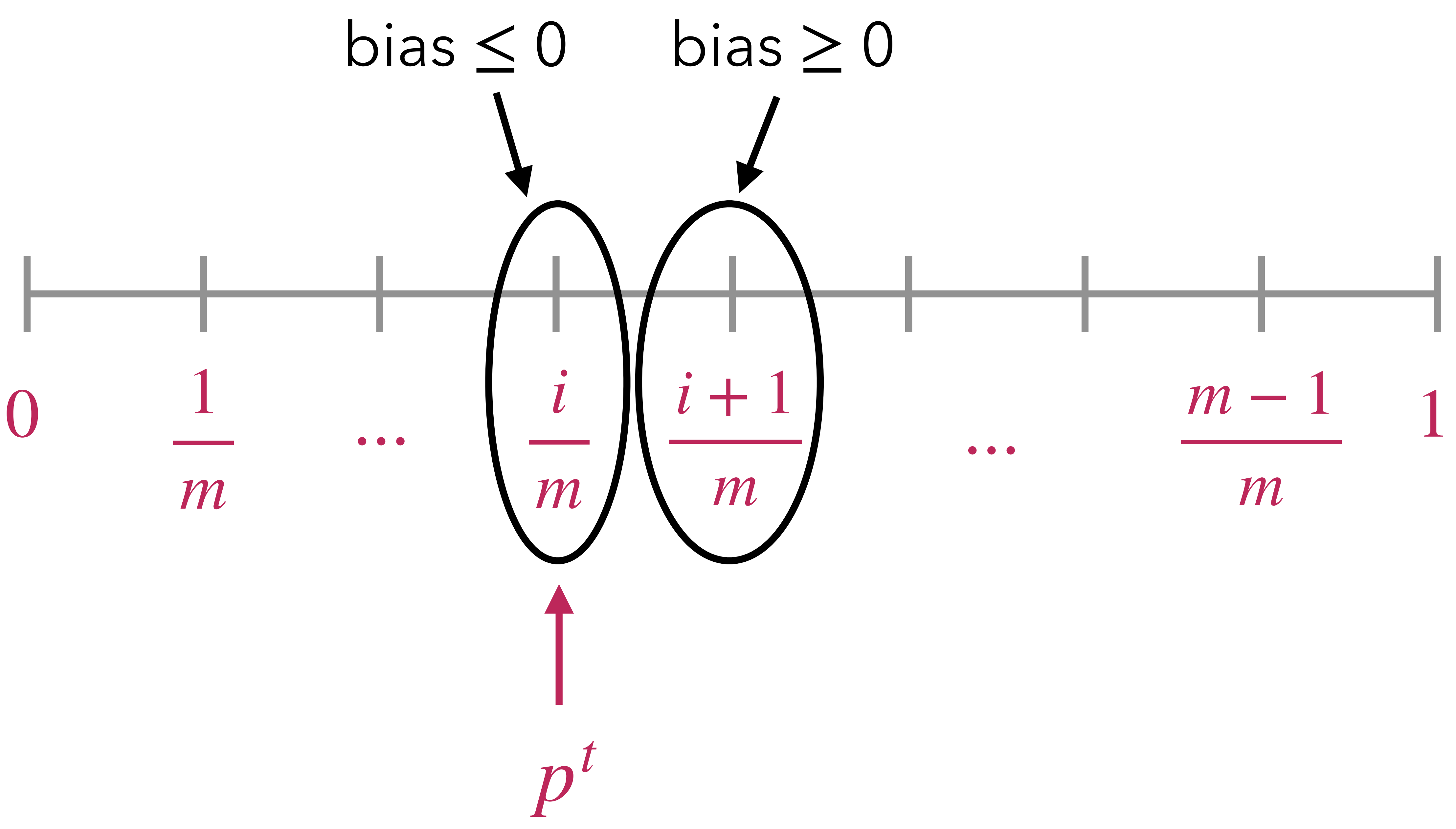


Almost One Step Ahead

Idea: Mimic One Step Ahead without looking into future

On day $t = 1,\ldots,T$:

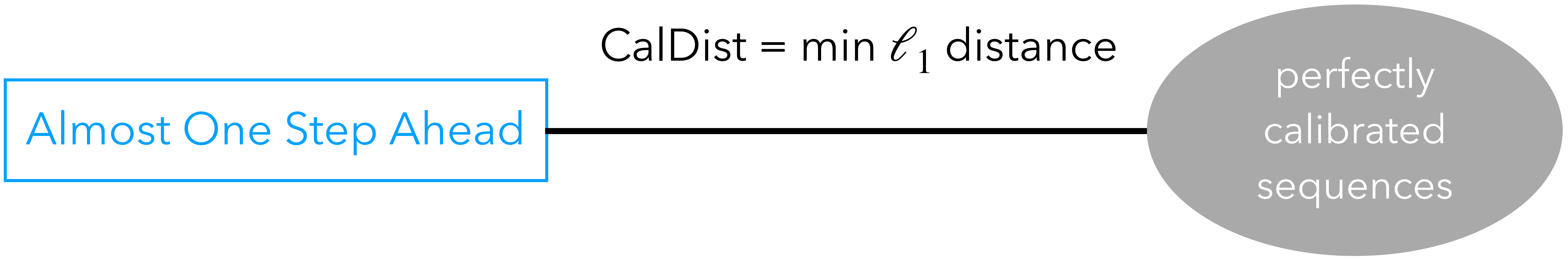
1. Predict (arbitrarily) one of two points i/m and $(i+1)/m$ that One Step Ahead would commit to on day t
2. Observe outcome y^t
3. Keep track of bias of predictions that One Step Ahead *would have made*



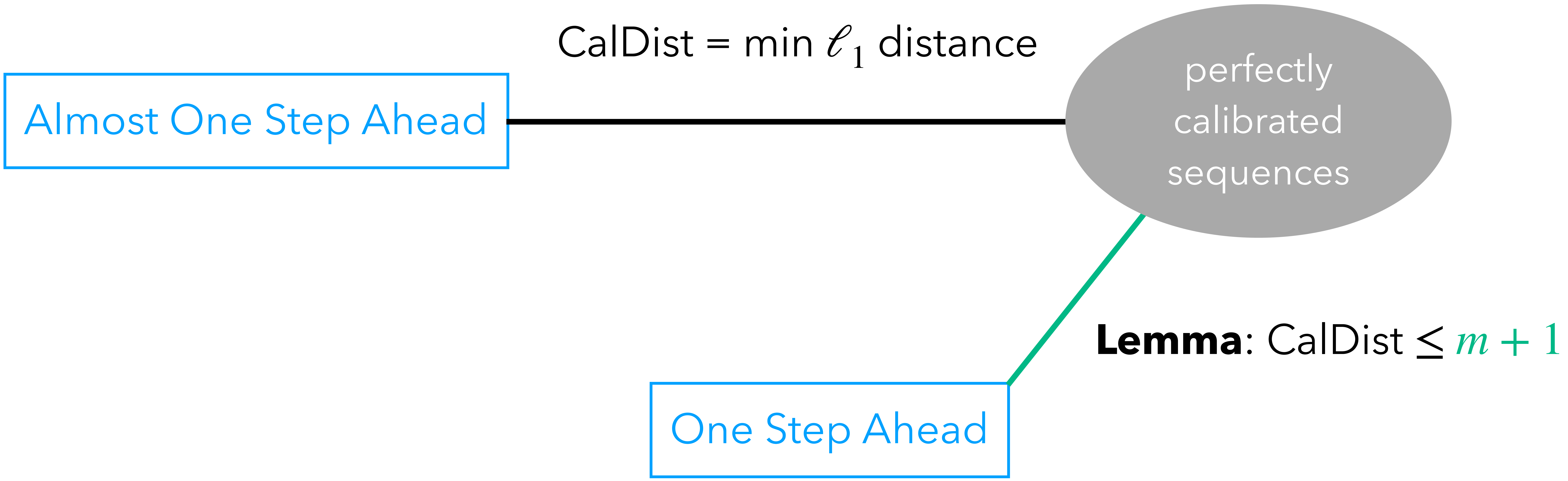
Let's analyze CalDist of Almost One Step Ahead

Almost One Step Ahead

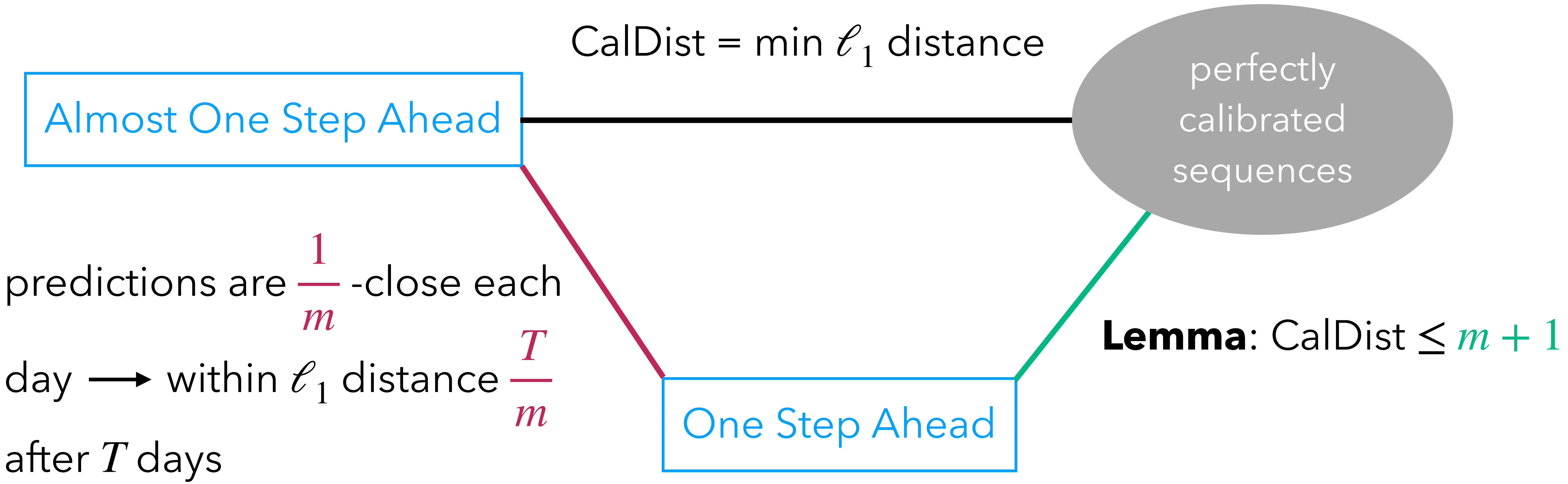
Almost One Step Ahead



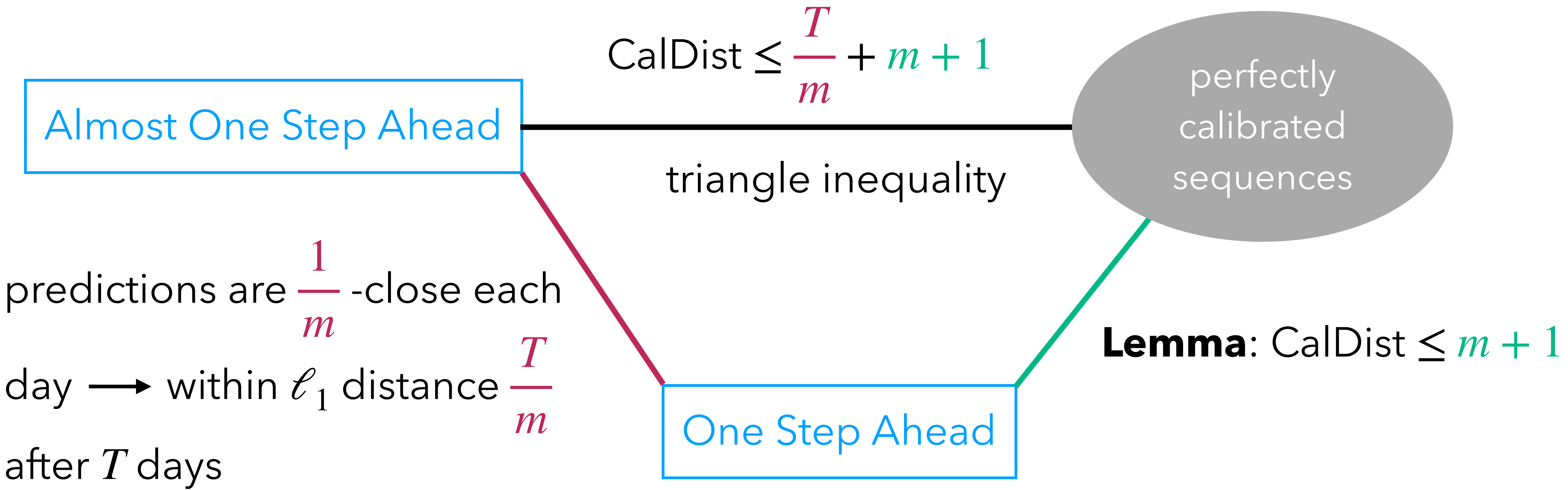
Almost One Step Ahead



Almost One Step Ahead



Almost One Step Ahead



Almost One Step Ahead

Theorem: Almost One Step Ahead achieves CalDist $\leq 2\sqrt{T} + 1$ (Set $m = \sqrt{T}$)

CalDist $\leq \frac{T}{m} + m + 1$

Almost One Step Ahead

triangle inequality

perfectly calibrated sequences

predictions are $\frac{1}{m}$-close each day $\longrightarrow$ within ℓ_1 distance $\frac{T}{m}$ after T days

Lemma: CalDist $\leq m + 1$

One Step Ahead

to summarize

to summarize

How should we measure forecast quality? One answer: calibration.

to summarize

How should we measure forecast quality? One answer: calibration.

Expected Calibration Error (ECE) is a classic measure of miscalibration, but has disadvantages (discontinuous in predictions, cannot get good rates, etc).

to summarize

How should we measure forecast quality? One answer: calibration.

Expected Calibration Error (ECE) is a classic measure of miscalibration, but has disadvantages (discontinuous in predictions, cannot get good rates, etc).

Distance to Calibration (CalDist) resolves some of these shortcomings.

to summarize

How should we measure forecast quality? One answer: calibration.

Expected Calibration Error (ECE) is a classic measure of miscalibration, but has disadvantages (discontinuous in predictions, cannot get good rates, etc).

Distance to Calibration (CalDist) resolves some of these shortcomings.

In particular, unlike ECE, it is incredibly tractable: we give a simple, efficient, and deterministic algorithm.

to summarize

to summarize

Fictitious lookahead algorithm (One Step Ahead) obtains low distance to calibration

to summarize

Fictitious lookahead algorithm (One Step Ahead) obtains low distance to calibration

Can't look ahead, but…

Can make a prediction within small distance every day (*Almost* One Step Ahead)

to summarize

Fictitious lookahead algorithm (One Step Ahead) obtains low distance to calibration

Can't look ahead, but…

Can make a prediction within small distance every day (*Almost* One Step Ahead)

↓

Not much difference between looking ahead and not looking ahead

to summarize

Fictitious lookahead algorithm (One Step Ahead) obtains low distance to calibration

Can't look ahead, but…

Can make a prediction within small distance every day (*Almost* One Step Ahead)

↓

Not much difference between looking ahead and not looking ahead

↓

Almost One Step Ahead obtains low distance to calibration

An Elementary Predictor Obtaining $2\sqrt{T}+1$ Distance to Calibration

Eshwar Ram Arunachaleswaran, Natalie Collina, Aaron Roth, and Mirah Shi

1 Introduction

Probabilistic predictions of binary outcomes are said to be *calibrated*, if, informally, they are unbiased conditional on their own predictions. For predictors that are not perfectly calibrated, there are a variety of ways to measure calibration error. Perhaps the most popular measure is Expected Calibration Error (ECE), which measures the average bias of the predictions, weighted by the frequency of the predictions. ECE has a number of difficulties as a measure of calibration, not least of which is that it is discontinuous in the predictions. Motivated by this, Błasiok et al. [2023] propose a different measure: distance to calibration, which measures how far a predictor is in ℓ_1 distance from the nearest perfectly calibrated predictor. In the online adversarial setting, it has been known since Foster and Vohra [1998] how to make predictions with ECE growing at a rate of $O(T^{2/3})$. Qiao and Valiant [2021] show that obtaining $O(\sqrt{T})$ rates for ECE is impossible. Recently, in a COLT 2024 paper, Qiao and Zheng [2024] showed that it was possible to make sequential predictions against an adversary guaranteeing expected distance to calibration growing at a rate of $O(\sqrt{T})$. Their algorithm is the solution to a minimax problem of size doubly-exponential in T. They leave as an open problem finding an explicit, efficient, deterministic algorithm for this problem. In this paper we resolve this problem, by giving an extremely simple such algorithm with an elementary analysis.

Algorithm 1: Almost-One-Step-Ahead

Input: Sequence of outcomes $y^{1:T} \in \{0,1\}^T$
Output: Sequence of predictions $p^{1:T} \in \{0, \frac{1}{m}, ..., 1\}^T$ for some discretization parameter $m > 0$
for $t = 1$ **to** T **do**

Given look-ahead predictions $\tilde{p}^{1:t-1}$, define the look-ahead bias conditional on a prediction p as:

$$\alpha_{\tilde{p}^{1:t-1}}(p) := \sum_{s=1}^{t-1} \mathbb{1}[\tilde{p}^s = p](\tilde{p}^s - y^s)$$

Choose two adjacent points $p_i = \frac{i}{m}, p_{i+1} = \frac{i+1}{m}$ satisfying:

$$\alpha_{\tilde{p}^{1:t-1}}(p_i) \le 0 \text{ and } \alpha_{\tilde{p}^{1:t-1}}(p_{i+1}) \ge 0$$

Arbitrarily predict $p^t = p_i$ or $p^t = p_{i+1}$;
Upon observing the (adversarially chosen) outcome y^t, set look-ahead prediction

$$\tilde{p}^t = \text{argmin}_{p \in \{p_i, p_{i+1}\}} |p - y^t|$$

2 Setting

We study a sequential binary prediction setting: at every round t, a forecaster makes a prediction $p^t \in [0,1]$, after which an adversary reveals an outcome $y^t \in \{0,1\}$. Given a sequence of predictions $p^{1:T}$ and outcomes $y^{1:T}$, we measure expected calibration error (ECE) as follows:

$$\text{ECE}(p^{1:T}, y^{1:T}) = \sum_{p \in [0,1]} \left| \sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t) \right|$$

Following Qiao and Zheng [2024], we define *distance to calibration* to be the minimum ℓ_1 distance between a sequence of predictions produced by a forecaster and any *perfectly calibrated* sequence of predictions:

$$\text{CalDist}(p^{1:T}, y^{1:T}) = \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1$$

where $\mathcal{C}(y^{1:T}) = \{q^{1:T} : \text{ECE}(q^{1:T}, y^{1:T}) = 0\}$ is the set of predictions that are perfectly calibrated against outcomes $y^{1:T}$. First we observe that distance to calibration is upper bounded by ECE.

Lemma 1 (Qiao and Zheng [2024]). *Fix a sequence of predictions $p^{1:T}$ and outcomes $y^{1:T}$. Then,* $\text{CalDist}(p^{1:T}, y^{1:T}) \le \text{ECE}(p^{1:T}, y^{1:T})$.

Proof. For any prediction $p \in [0,1]$, define

$$\bar{y}^T(p) = \sum_{t=1}^{T} \frac{\mathbb{1}[p^t = p]}{\sum_{t=1}^{T} \mathbb{1}[p^t = p]} y^t$$

to be the average outcome conditioned on the prediction p. Consider the sequence $q^{1:T}$ where $q^t = \bar{y}^T(p^t)$. Observe that $q^{1:T}$ is perfectly calibrated. Thus, we have that

$$\begin{aligned}
\text{CalDist}(p^{1:T}, y^{1:T}) &\le \|p^{1:T} - q^{1:T}\|_1 \\
&= \sum_{t=1}^{T} |p^t - q^t| \\
&= \sum_{p \in [0,1]} \sum_{t=1}^{T} \mathbb{1}[p^t = p] |p - \bar{y}^T(p)| \\
&= \sum_{p \in [0,1]} |p - \bar{y}^T(p)| \sum_{t=1}^{T} \mathbb{1}[p^t = p] \\
&= \sum_{p \in [0,1]} \left| p \sum_{t=1}^{T} \mathbb{1}[p^t = p] - \bar{y}^T(p) \sum_{t=1}^{T} \mathbb{1}[p^t = p] \right| \\
&= \sum_{p \in [0,1]} \left| \sum_{t=1}^{T} \mathbb{1}[p^t = p](p - y^t) \right| \\
&= \text{ECE}(p^{1:T}, y^{1:T})
\end{aligned}$$

□

The upper bound is not tight, however. The best known sequential prediction algorithm obtains ECE bounded by $O(T^{2/3})$ [Foster and Vohra, 1998], and it is known that there is no algorithm guaranteeing ECE below $O(T^{0.54389})$ [Qiao and Valiant, 2021, Dagan et al., 2024]. Qiao and Zheng [2024] give an algorithm that is the solution to a game of size doubly-exponential in T that obtains expected distance to calibration $O(\sqrt{T})$. Here we give an elementary analysis of a simple efficient deterministic algorithm (Algorithm 1) that obtains distance to calibration $2\sqrt{T}+1$.

Theorem 1. *Algorithm 1 (Almost-One-Step-Ahead) guarantees that against any sequence of outcomes,* $\text{CalDist}(p^{1:T}, y^{1:T}) \le 2\sqrt{T}+1$.

3 Analysis of Algorithm 1

Before describing the algorithm, we introduce some notation. We will make predictions that belong to a grid. Let $B_m = \{0, 1/m, ..., 1\}$ denote a discretization of the prediction space with discretization parameter $m > 0$, and let $p_i = i/m$. For a sequence of predictions $\tilde{p}^1, ..., \tilde{p}^t$ and outcomes $y^1, ..., y^t$, we define the bias conditional on a prediction p as:

$$\alpha_{\tilde{p}^{1:t}}(p) = \sum_{s=1}^{t} \mathbb{1}[\tilde{p}^s = p](\tilde{p}^s - y^s)$$

To understand our algorithm, it will be helpful to first state and analyze a hypothetical "lookahead" algorithm that we call "One-Step-Ahead", which is closely related to the algorithm and analysis given by

Gupta and Ramdas [2022] in a different model. One-Step-Ahead produces predictions $\tilde{p}^1, ..., \tilde{p}^T$ as follows. At round t, before observing y^t, the algorithm fixes two predictions p_i, p_{i+1} satisfying $\alpha_{\tilde{p}^{1:t-1}}(p_i) \le 0$ and $\alpha_{\tilde{p}^{1:t-1}}(p_{i+1}) \ge 0$. Such a pair is guaranteed to exist, because by construction, it must be that for any history, $\alpha_{\tilde{p}^{1:t-1}}(0) \le 0$ and $\alpha_{\tilde{p}^{1:t-1}}(1) \ge 0$. Note that a well known randomized algorithm obtaining diminishing ECE (and smooth calibration error) uses the same observation to carefully *randomize* between two such adjacent predictions [Foster, 1999, Foster and Hart, 2018]. Upon observing the outcome y^t, the algorithm outputs prediction $\tilde{p}^t = \text{argmin}_{p \in \{p_i, p_{i+1}\}} |p - y^t|$. Naturally, we cannot implement this algorithm, as it chooses its prediction only after observing the outcome, but our analysis will rely on a key property this algorithm maintains—namely, that it always produces a sequence of predictions with ECE upper bounded by $m+1$, the number of elements in the discretized prediction space.

Theorem 2. *For any sequence of outcomes, One-Step-Ahead achieves* $\text{ECE}(\tilde{p}^{1:T}, y^{1:T}) \le m+1$.

Proof. We will show that for any $p_i \in B_m$, we have $|\alpha_{\tilde{p}^{1:T}}(p_i)| \le 1$, after which the bound on ECE will follow: $\text{ECE}(\tilde{p}^{1:T}, y^{1:T}) = \sum_{p_i \in B_m} |\alpha_{\tilde{p}^{1:T}}(p_i)| \le m+1$. We proceed via an inductive argument. Fix a prediction $p_i \in B_m$. At the first round t_1 in which p_i is output by the algorithm, we have that $|\alpha_{\tilde{p}^{1:t_1}}(p_i)| = |p^{t_1} - y^{t_1}| \le 1$. Now suppose after round $t-1$, we satisfy $|\alpha_{\tilde{p}^{1:t-1}}(p_i)| \le 1$. If p_i is the prediction made at round t, it must be that either: $\alpha_{\tilde{p}^{1:t-1}}(p_i) \le 0$ and $p_i - y^t \ge 0$; or $\alpha_{\tilde{p}^{1:t-1}}(p_i) \ge 0$ and $p_i - y^t \le 0$. Thus, since $\alpha_{\tilde{p}^{1:t-1}}(p_i)$ and $p_i - y^t$ either take value 0 or differ in sign, we can conclude that

$$|\alpha_{\tilde{p}^{1:t}}(p_i)| = |\alpha_{\tilde{p}^{1:t-1}}(p_i) + p_i - y^t| \le \max\{|\alpha_{\tilde{p}^{1:t-1}}(p_i)|, |p_i - y^t|\} \le 1$$

which proves the theorem. □

Algorithm 1 (Almost-One-Step-Ahead) maintains the same state $\alpha_{\tilde{p}^{1:t}}(p)$ as One-Step-Ahead (which it can compute at round t after observing the outcome y_{t-1}). In particular, it does not keep track of the bias of its own predictions, but rather keeps track of the bias of the predictions that One-Step-Ahead *would have made*. Thus it can determine the pair p_i, p_{i+1} that One-Step-Ahead would commit to predict at round t. It cannot make the same prediction as One-Step-Ahead (as it must fix its prediction before the label is observed) — so instead it deterministically predicts $p^t = p_i$ (or $p^t = p_{i+1}$ — the choice can be arbitrary and does not affect the analysis). Since we have that $|p_i - p_{i+1}| \le \frac{1}{m}$, it must be that for whichever choice One-Step-Ahead would have made, we have $|\tilde{p}^t - p^t| \le \frac{1}{m}$. In other words, although Almost-One-Step-Ahead does not make the same predictions as One-Step-Ahead, it makes predictions that are within ℓ_1 distance T/m after T rounds. The analysis then follows by the ECE bound of One-Step-Ahead, the triangle inequality, and choosing $m = \sqrt{T}$.

Proof of Theorem 1. Observe that internally, Algorithm 1 maintains the sequence $\tilde{p}^1, ..., \tilde{p}^t$ which corresponds exactly to predictions made by One-Step-Ahead. Thus, by Lemma 1 and Theorem 2, we have that $\text{CalDist}(\tilde{p}^{1:T}, y^{1:T}) \le \text{ECE}(\tilde{p}^{1:T}, y^{1:T}) \le m+1$. Then, we can compute the distance to calibration of the sequence $p^1, ..., p^T$:

$$\begin{aligned}
\text{CalDist}(p^{1:T}, y^{1:T}) &= \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1 \\
&= \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - \tilde{p}^{1:T} + \tilde{p}^{1:T} - q^{1:T}\|_1 \\
&\le \|p^{1:T} - \tilde{p}^{1:T}\|_1 + \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|\tilde{p}^{1:T} - q^{1:T}\|_1 \\
&\le \frac{T}{m} + m + 1
\end{aligned}$$

where in the last step we use the fact that $|p^t - \tilde{p}^t| \le 1/m$ for all t and thus $\|p^{1:T} - \tilde{p}^{1:T}\|_1 \le T/m$. The result then follows by setting $m = \sqrt{T}$. □

Acknowledgements

This work was supported in part by the Simons Collaboration on the Theory of Algorithmic Fairness, NSF grants FAI-2147212 and CCF-2217062, an AWS AI Gift for Research on Trustworthy AI, and the Hans Sigrist Prize.

An Elementary Predictor Obtaining $2\sqrt{T}+1$ Distance to Calibration

Eshwar Ram Arunachaleswaran, Natalie Collina, Aaron Roth, and Mirah Shi

1 Introduction

Probabilistic predictions of binary outcomes are said to be *calibrated*, if, informally, they are unbiased conditional on their own predictions. For predictors that are not perfectly calibrated, there are a variety of ways to measure calibration error. Perhaps the most popular measure is Expected Calibration Error (ECE), which measures the average bias of the predictions, weighted by the frequency of the predictions. ECE has a number of difficulties as a measure of calibration, not least of which is that it is discontinuous in the predictions. Motivated by this, Błasiok et al. [2023] propose a different measure: distance to calibration, which measures how far a predictor is in ℓ_1 distance from the nearest perfectly calibrated predictor. In the online adversarial setting, it has been known since Foster and Vohra [1998] how to make predictions with ECE growing at a rate of $O(T^{2/3})$. Qiao and Valiant [2021] show that obtaining $O(\sqrt{T})$ rates for ECE is impossible. Recently, in a COLT 2024 paper, Qiao and Zheng [2024] showed that it was possible to make sequential predictions against an adversary guaranteeing expected distance to calibration growing at a rate of $O(\sqrt{T})$. Their algorithm is the solution to a minimax problem of size doubly-exponential in T. They leave as an open problem finding an explicit, efficient, deterministic algorithm for this problem. In this paper we resolve this problem, by giving an extremely simple such algorithm with an elementary analysis.

Algorithm 1: Almost-One-Step-Ahead

Input: Sequence of outcomes $y^{1:T} \in \{0,1\}^T$
Output: Sequence of predictions $p^{1:T} \in \{0, \frac{1}{m}, ..., 1\}^T$ for some discretization parameter $m > 0$
for $t = 1$ **to** T **do**
Given look-ahead predictions $\tilde{p}^{1:t-1}$, define the look-ahead bias conditional on a prediction p as:

$$\alpha_{\tilde{p}^{1:t-1}}(p) := \sum_{s=1}^{t-1} \mathbb{1}[\tilde{p}^s = p](\tilde{p}^s - y^s)$$

Choose two adjacent points $p_i = \frac{i}{m}, p_{i+1} = \frac{i+1}{m}$ satisfying:

$$\alpha_{\tilde{p}^{1:t-1}}(p_i) \le 0 \text{ and } \alpha_{\tilde{p}^{1:t-1}}(p_{i+1}) \ge 0$$

Arbitrarily predict $p^t = p_i$ or $p^t = p_{i+1}$;
Upon observing the (adversarially chosen) outcome y^t, set look-ahead prediction

$$\tilde{p}^t = \text{argmin}_{p \in \{p_i, p_{i+1}\}} |p - y^t|$$

2 Setting

We study a sequential binary prediction setting: at every round t, a forecaster makes a prediction $p^t \in [0,1]$, after which an adversary reveals an outcome $y^t \in \{0,1\}$. Given a sequence of predictions $p^{1:T}$ and outcomes $y^{1:T}$, we measure expected calibration error (ECE) as follows:

$$\text{ECE}(p^{1:T}, y^{1:T}) = \sum_{p \in [0,1]} \left| \sum_{t=1}^{T} \mathbb{1}[p^t = p](p^t - y^t) \right|$$

Following Qiao and Zheng [2024], we define *distance to calibration* to be the minimum ℓ_1 distance between a sequence of predictions produced by a forecaster and any *perfectly calibrated* sequence of predictions:

$$\text{CalDist}(p^{1:T}, y^{1:T}) = \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1$$

where $\mathcal{C}(y^{1:T}) = \{q^{1:T} : \text{ECE}(q^{1:T}, y^{1:T}) = 0\}$ is the set of predictions that are perfectly calibrated against outcomes $y^{1:T}$. First we observe that distance to calibration is upper bounded by ECE.

Lemma 1 (Qiao and Zheng [2024])**.** *Fix a sequence of predictions $p^{1:T}$ and outcomes $y^{1:T}$. Then,* $\text{CalDist}(p^{1:T}, y^{1:T}) \le \text{ECE}(p^{1:T}, y^{1:T})$.

Proof. For any prediction $p \in [0,1]$, define

$$\bar{y}^T(p) = \sum_{t=1}^{T} \frac{\mathbb{1}[p^t = p]}{\sum_{t=1}^{T} \mathbb{1}[p^t = p]} y^t$$

to be the average outcome conditioned on the prediction p. Consider the sequence $q^{1:T}$ where $q^t = \bar{y}^T(p^t)$. Observe that $q^{1:T}$ is perfectly calibrated. Thus, we have that

$$\begin{aligned}
\text{CalDist}(p^{1:T}, y^{1:T}) &\le \|p^{1:T} - q^{1:T}\|_1 \\
&= \sum_{t=1}^{T} |p^t - q^t| \\
&= \sum_{p \in [0,1]} \sum_{t=1}^{T} \mathbb{1}[p^t = p] |p - \bar{y}^T(p)| \\
&= \sum_{p \in [0,1]} |p - \bar{y}^T(p)| \sum_{t=1}^{T} \mathbb{1}[p^t = p] \\
&= \sum_{p \in [0,1]} \left| p \sum_{t=1}^{T} \mathbb{1}[p^t = p] - \bar{y}^T(p) \sum_{t=1}^{T} \mathbb{1}[p^t = p] \right| \\
&= \sum_{p \in [0,1]} \left| \sum_{t=1}^{T} \mathbb{1}[p^t = p](p - y^t) \right| \\
&= \text{ECE}(p^{1:T}, y^{1:T})
\end{aligned}$$

□

The upper bound is not tight, however. The best known sequential prediction algorithm obtains ECE bounded by $O(T^{2/3})$ [Foster and Vohra, 1998], and it is known that there is no algorithm guaranteeing ECE below $O(T^{0.54389})$ [Qiao and Valiant, 2021, Dagan et al., 2024]. Qiao and Zheng [2024] give an algorithm that is the solution to a game of size doubly-exponential in T that obtains expected distance to calibration $O(\sqrt{T})$. Here we give an elementary analysis of a simple efficient deterministic algorithm (Algorithm 1) that obtains distance to calibration $2\sqrt{T}+1$.

Theorem 1. *Algorithm 1 (Almost-One-Step-Ahead) guarantees that against any sequence of outcomes,* $\text{CalDist}(p^{1:T}, y^{1:T}) \le 2\sqrt{T} + 1$.

3 Analysis of Algorithm 1

Before describing the algorithm, we introduce some notation. We will make predictions that belong to a grid. Let $B_m = \{0, 1/m, ..., 1\}$ denote a discretization of the prediction space with discretization parameter $m > 0$, and let $p_i = i/m$. For a sequence of predictions $\tilde{p}^1, ..., \tilde{p}^t$ and outcomes $y^1, ..., y^t$, we define the bias conditional on a prediction p as:

$$\alpha_{\tilde{p}^{1:t}}(p) = \sum_{s=1}^{t} \mathbb{1}[\tilde{p}^s = p](\tilde{p}^s - y^s)$$

To understand our algorithm, it will be helpful to first state and analyze a hypothetical "lookahead" algorithm that we call "One-Step-Ahead", which is closely related to the algorithm and analysis given by

Gupta and Ramdas [2022] in a different model. One-Step-Ahead produces predictions $\tilde{p}^1, ..., \tilde{p}^T$ as follows. At round t, before observing y^t, the algorithm fixes two predictions p_i, p_{i+1} satisfying $\alpha_{\tilde{p}^{1:t-1}}(p_i) \le 0$ and $\alpha_{\tilde{p}^{1:t-1}}(p_{i+1}) \ge 0$. Such a pair is guaranteed to exist, because by construction, it must be that for any history, $\alpha_{\tilde{p}^{1:t-1}}(0) \le 0$ and $\alpha_{\tilde{p}^{1:t-1}}(1) \ge 0$. Note that a well known randomized algorithm obtaining diminishing ECE (and smooth calibration error) uses the same observation to carefully *randomize* between two such adjacent predictions [Foster, 1999, Foster and Hart, 2018]. Upon observing the outcome y^t, the algorithm outputs prediction $\tilde{p}^t = \text{argmin}_{p \in \{p_i, p_{i+1}\}} |p - y^t|$. Naturally, we cannot implement this algorithm, as it chooses its prediction only after observing the outcome, but our analysis will rely on a key property this algorithm maintains—namely, that it always produces a sequence of predictions with ECE upper bounded by $m+1$, the number of elements in the discretized prediction space.

Theorem 2. *For any sequence of outcomes, One-Step-Ahead achieves* $\text{ECE}(\tilde{p}^{1:T}, y^{1:T}) \le m + 1$.

Proof. We will show that for any $p_i \in B_m$, we have $|\alpha_{\tilde{p}^{1:T}}(p_i)| \le 1$, after which the bound on ECE will follow: $\text{ECE}(\tilde{p}^{1:T}, y^{1:T}) = \sum_{p_i \in B_m} |\alpha_{\tilde{p}^{1:T}}(p_i)| \le m + 1$. We proceed via an inductive argument. Fix a prediction $p_i \in B_m$. At the first round t_1 in which p_i is output by the algorithm, we have that $|\alpha_{\tilde{p}^{1:t_1}}(p_i)| = |p^{t_1} - y^{t_1}| \le 1$. Now suppose after round $t-1$, we satisfy $|\alpha_{\tilde{p}^{1:t-1}}(p_i)| \le 1$. If p_i is the prediction made at round t, it must be that either: $\alpha_{\tilde{p}^{1:t-1}}(p_i) \le 0$ and $p_i - y^t \ge 0$; or $\alpha_{\tilde{p}^{1:t-1}}(p_i) \ge 0$ and $p_i - y^t \le 0$. Thus, since $\alpha_{\tilde{p}^{1:t-1}}(p_i)$ and $p_i - y^t$ either take value 0 or differ in sign, we can conclude that

$$|\alpha_{\tilde{p}^{1:t}}(p_i)| = |\alpha_{\tilde{p}^{1:t-1}}(p_i) + p_i - y^t| \le \max\{|\alpha_{\tilde{p}^{1:t-1}}(p_i)|, |p_i - y^t|\} \le 1$$

which proves the theorem. □

Algorithm 1 (Almost-One-Step-Ahead) maintains the same state $\alpha_{\tilde{p}^{1:t}}(p)$ as One-Step-Ahead (which it can compute at round t after observing the outcome y_{t-1}). In particular, it does not keep track of the bias of its own predictions, but rather keeps track of the bias of the predictions that One-Step-Ahead *would have made*. Thus it can determine the pair p_i, p_{i+1} that One-Step-Ahead would commit to predict at round t. It cannot make the same prediction as One-Step-Ahead (as it must fix its prediction before the label is observed) — so instead it deterministically predicts $p^t = p_i$ (or $p^t = p_{i+1}$ — the choice can be arbitrary and does not affect the analysis). Since we have that $|p_i - p_{i+1}| \le \frac{1}{m}$, it must be that for whichever choice One-Step-Ahead would have made, we have $|\tilde{p}^t - p^t| \le \frac{1}{m}$. In other words, although Almost-One-Step-Ahead does not make the same predictions as One-Step-Ahead, it makes predictions that are within ℓ_1 distance T/m after T rounds. The analysis then follows by the ECE bound of One-Step-Ahead, the triangle inequality, and choosing $m = \sqrt{T}$.
Proof of Theorem 1. Observe that internally, Algorithm 1 maintains the sequence $\tilde{p}^1, ..., \tilde{p}^t$ which corresponds exactly to predictions made by One-Step-Ahead. Thus, by Lemma 1 and Theorem 2, we have that $\text{CalDist}(\tilde{p}^{1:T}, y^{1:T}) \le \text{ECE}(\tilde{p}^{1:T}, y^{1:T}) \le m + 1$. Then, we can compute the distance to calibration of the sequence $p^1, ..., p^T$:

$$\begin{aligned}
\text{CalDist}(p^{1:T}, y^{1:T}) &= \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - q^{1:T}\|_1 \\
&= \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|p^{1:T} - \tilde{p}^{1:T} + \tilde{p}^{1:T} - q^{1:T}\|_1 \\
&\le \|p^{1:T} - \tilde{p}^{1:T}\|_1 + \min_{q^{1:T} \in \mathcal{C}(y^{1:T})} \|\tilde{p}^{1:T} - q^{1:T}\|_1 \\
&\le \frac{T}{m} + m + 1
\end{aligned}$$

where in the last step we use the fact that $|p^t - \tilde{p}^t| \le 1/m$ for all t and thus $\|p^{1:T} - \tilde{p}^{1:T}\|_1 \le T/m$. The result then follows by setting $m = \sqrt{T}$. □

Acknowledgements

This work was supported in part by the Simons Collaboration on the Theory of Algorithmic Fairness, NSF grants FAI-2147212 and CCF-2217062, an AWS AI Gift for Research on Trustworthy AI, and the Hans Sigrist Prize.

Thanks!