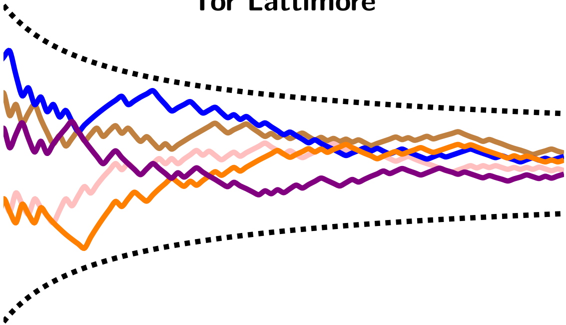
# Confident Bayesian Sequence Prediction

Tor Lattimore



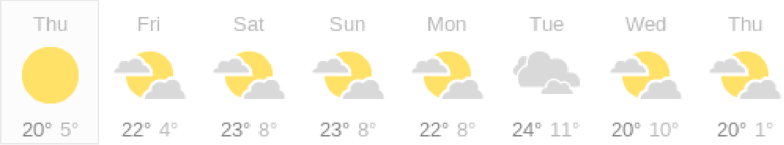
# Sequence Prediction

Can you guess the next number?

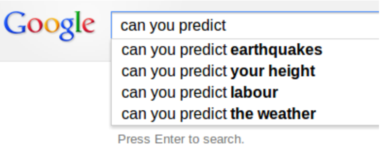
$1, 2, 3, 4, 5, \ldots$

$3, 1, 4, 1, 5, \ldots$

$1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, \ldots$







# Sequence Prediction

- $x$ is a sequence over countable alphabet $X$
- $\mu(x)$ is the $\mu$-probability of observing $x$
- $\mu(a|x)$ is the $\mu$-probability of observing $a \in X$ given $x$
- $\mathcal{M}$ is a countable set of measures (sequence generators)

$X = \{heads, tails\}$
$X = \{rain, shine\}$
$X = \{plague, smallpox, flu\}$
$X = \{1, 2, 3, \cdots\}$

Hellinger$^2$ Distance $\quad h_x(\rho, \mu) := \sum_{a \in X} \left(\sqrt{\rho(a|x)} - \sqrt{\mu(a|x)}\right)^2$

Total Variation Distance $\quad \delta_x(\rho, \mu) := \frac{1}{2} \sum_{a \in X} |\rho(a|x) - \mu(a|x)|$

$$\sqrt{h_x(\rho, \mu)} \approx \delta_x(\rho, \mu)$$

Goal: Construct predictor $\rho$ such that for all $\mu \in \mathcal{M}$

$$\rho(\cdot|x_{<t}) \xrightarrow{fast} \mu(\cdot|x_{<t})$$

when $x$ is sampled from $\mu$

# Example

- $X = \{\text{heads}, \text{tails}\}$
- $\mathcal{M} = \{\mu_\theta\}$ is a countable set of Bernoulli measures (coins)
- $\rho(\text{heads}|x_{<t}) = \frac{\text{number of heads in } x_{<t}}{\text{number of observations} = t-1}$

**Theorem (Law of Large Numbers)**

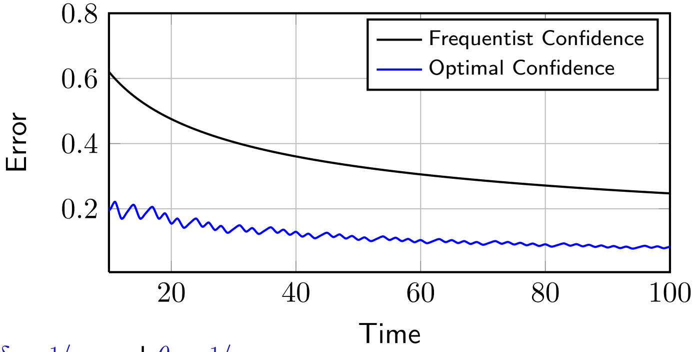
*If $\mu \in \mathcal{M}$, then $\lim_{t\to\infty} h_{x_{<t}}(\rho, \mu) = 0$ with $\mu$-probability 1*

# Example (continued)

**Theorem (Corollary of Hoeffding Bound)**

*If $\mu \in \mathcal{M}$, then with $\mu$-probability at least $1 - \delta$*

$$(\forall t) \quad |\rho(\textit{heads}|x_{<t}) - \mu(\textit{heads}|x_{<t})| \leq \sqrt{\frac{1}{2t} \log \frac{2t(t+1)}{\delta}}$$




$\delta = {}^1\!/_{10}$ and $\theta = {}^1\!/_2$

# Bayesian Predictors

- No assumption on $\mathcal{M}$ except that it is countable
- $w : \mathcal{M} \to (0, 1]$ is a prior on $\mathcal{M}$
- $\xi(x) := \sum_{\nu \in \mathcal{M}} w(\nu)\nu(x)$ is the Bayes measure

**Theorem (Solomonoff & Hutter)**

*If $\mu \in \mathcal{M}$, then*

- $\lim_{t \to \infty} h_{x_{<t}}(\mu, \xi) = 0$ *with $\mu$-probability 1*
- $\mathbf{E}_\mu \sum_{t=1}^{\infty} h_{x_{<t}}(\mu, \xi) \leq \ln \frac{1}{w(\mu)}$

Done? Not quite, $h_{x_{<t}}(\mu, \xi)$ is unknown. Can you construct a confidence bound on the error like in the Bernoulli case?

# Posterior Convergence

**Theorem (Bayes Law)**

$$P(H|D) = P(H)\frac{P(D|H)}{P(D)}$$

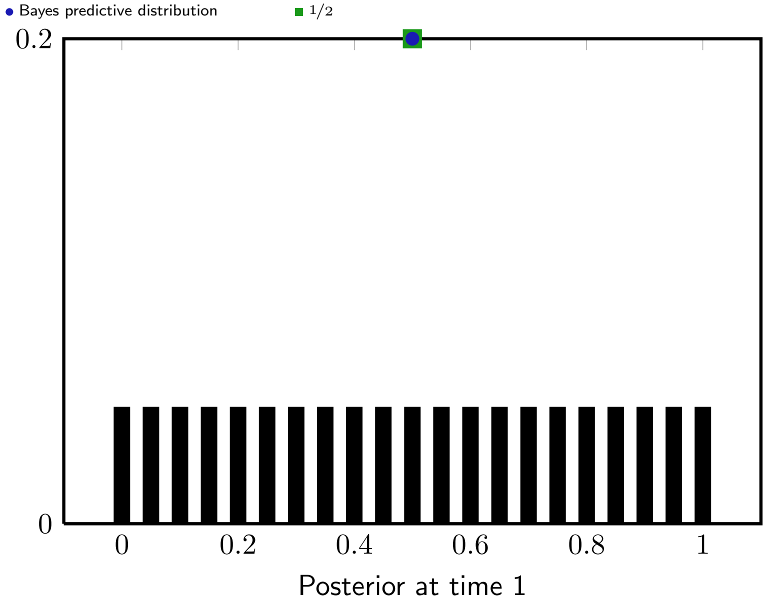
Posterior belief in hypothesis $\nu \in \mathcal{M}$ having observed $x$ is

$$w(\nu|x) := w(\nu)\frac{\nu(x)}{\xi(x)}$$

**Conjecture**

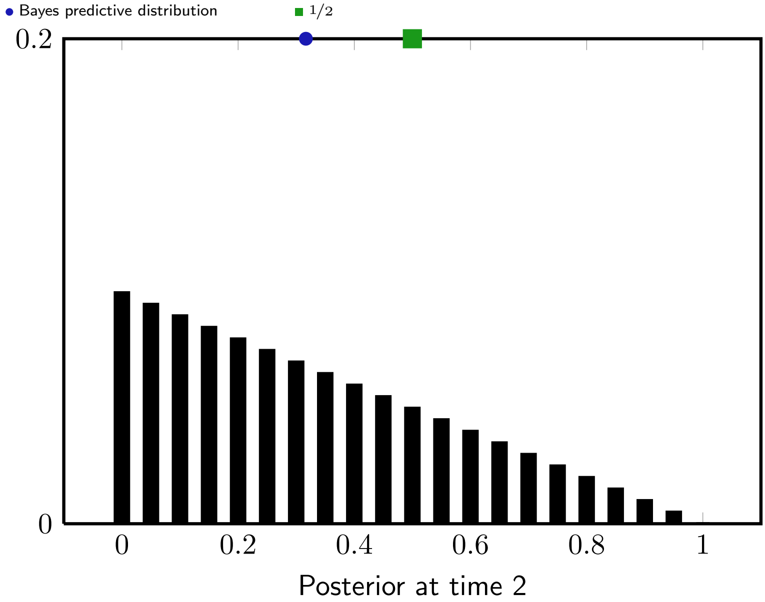
*The posterior $w(\cdot|x)$ concentrates about the truth as data is observed*

# Posterior Convergence



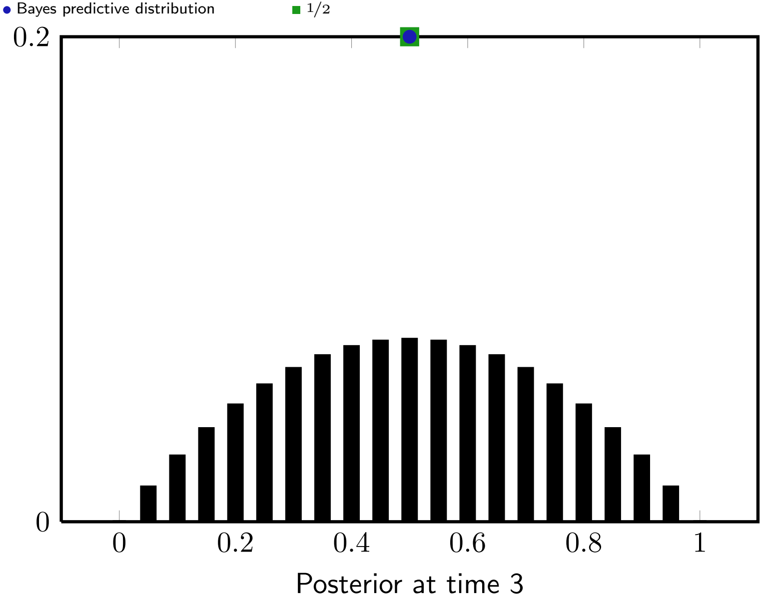

21 Bernoulli measures with true $\theta = 1/2$ and $w(\nu) = 1/21$

# Posterior Convergence



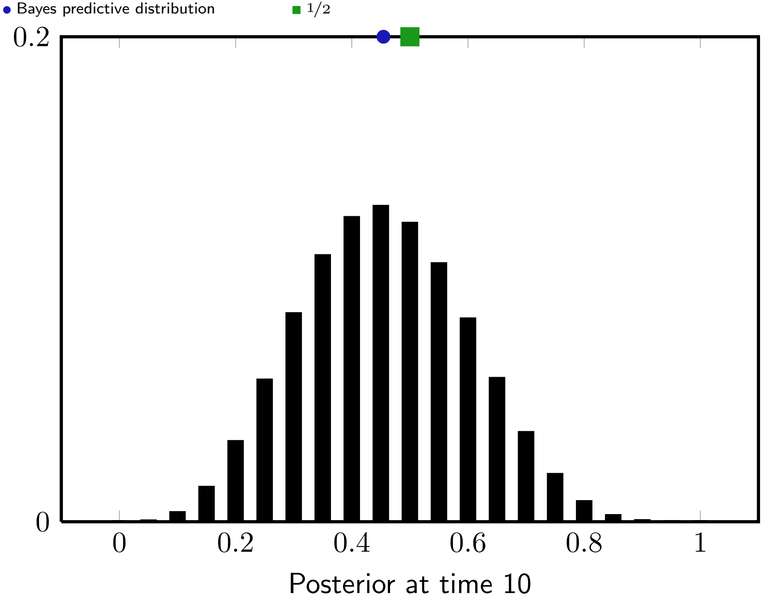

21 Bernoulli measures with true $\theta = 1/2$ and $w(\nu) = 1/21$

# Posterior Convergence



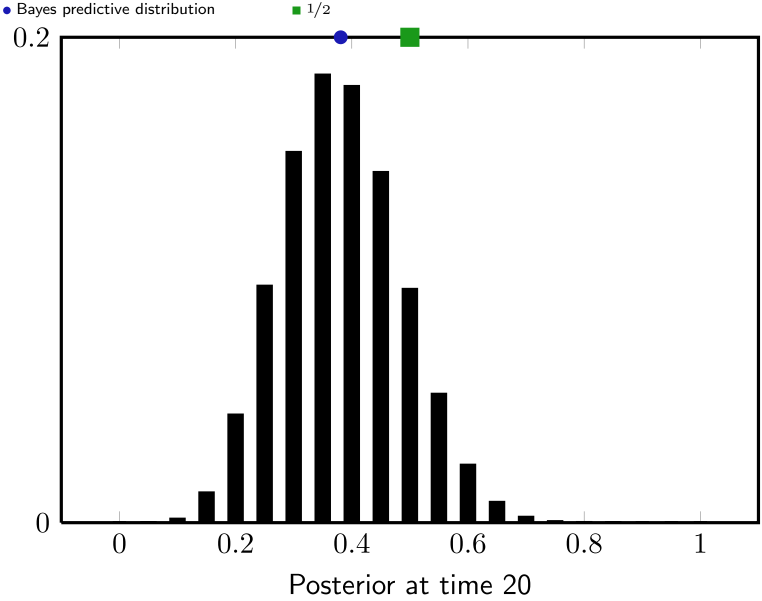

21 Bernoulli measures with true $\theta = 1/2$ and $w(\nu) = 1/21$

# Posterior Convergence




$21$ Bernoulli measures with true $\theta = 1/2$ and $w(\nu) = 1/21$

# Posterior Convergence




21 Bernoulli measures with true $\theta = 1/2$ and $w(\nu) = 1/21$

# Bayesian Confidence

**Theorem (Ville)**

*With $\mu$-probability at least $1 - \delta$*

$$(\forall n), \quad w(\mu | x_{<n}) \geq \delta w(\mu)$$

"With high probability the posterior belief in true hypothesis $\mu$ never falls below $\delta w(\mu)$"

Define set of plausible environments

$$\mathcal{M}(x_{<n}) := \{\nu \in \mathcal{M} : \forall \eta \leq n,\ w(\nu | x_{<\eta}) \geq \delta w(\nu)\}$$

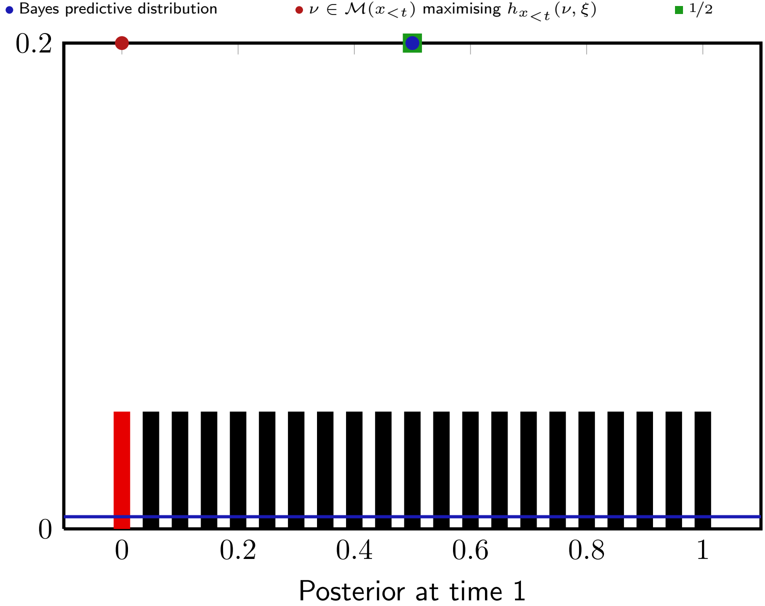
and confidence bound on error

$$\hat{h}(x_{<n}) := \max \{h_{x_{<n}}(\nu, \xi) : \nu \in \mathcal{M}(x_{<n})\}$$

**Theorem**

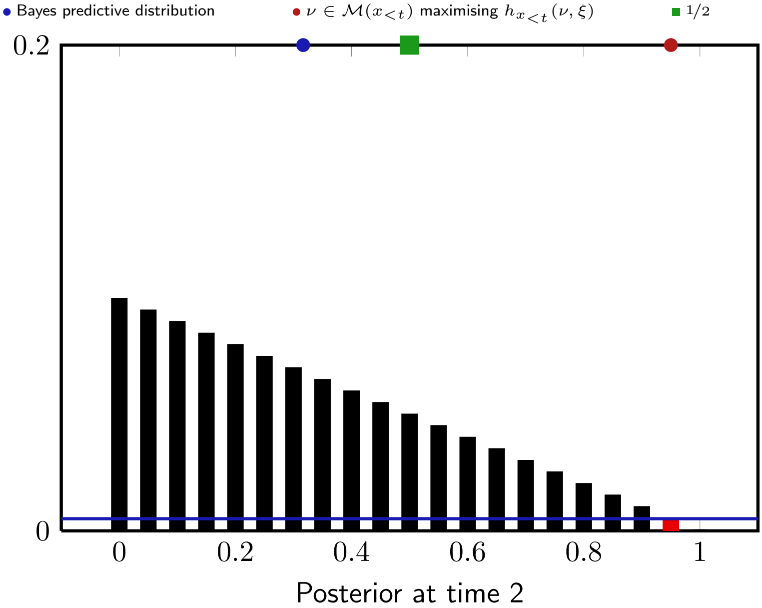
*With $\mu$-probablity at least $1 - \delta$ it holds that $\hat{h}(x_{<n}) \geq h_{x_{<n}}(\mu, \xi)$ for all $n$*

# Posterior Convergence



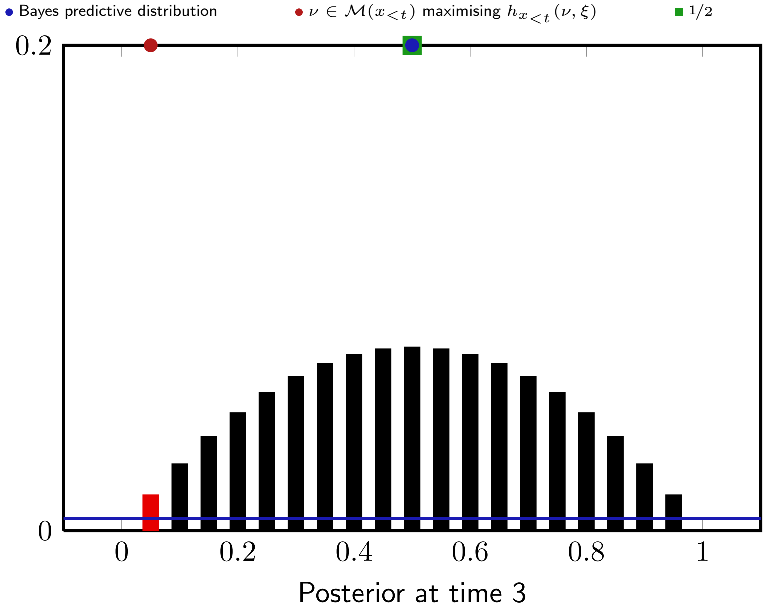

$21$ Bernoulli measures with true $\theta = 1/2$ and $w(\nu) = 1/21$ and $\delta = 1/10$

# Posterior Convergence



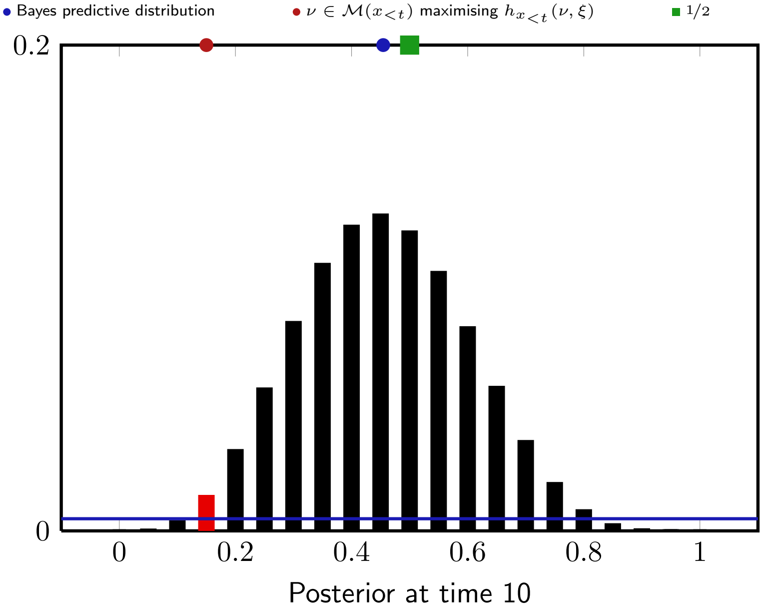

$21$ Bernoulli measures with true $\theta = 1/2$ and $w(\nu) = 1/21$ and $\delta = 1/10$

# Posterior Convergence



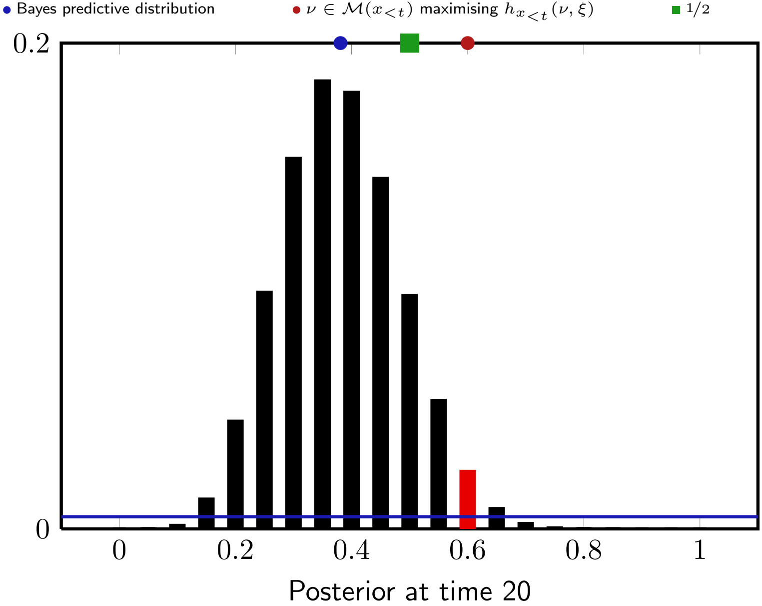

21 Bernoulli measures with true $\theta = 1/2$ and $w(\nu) = 1/21$ and $\delta = 1/10$

# Posterior Convergence




21 Bernoulli measures with true $\theta = 1/2$ and $w(\nu) = 1/21$ and $\delta = 1/10$

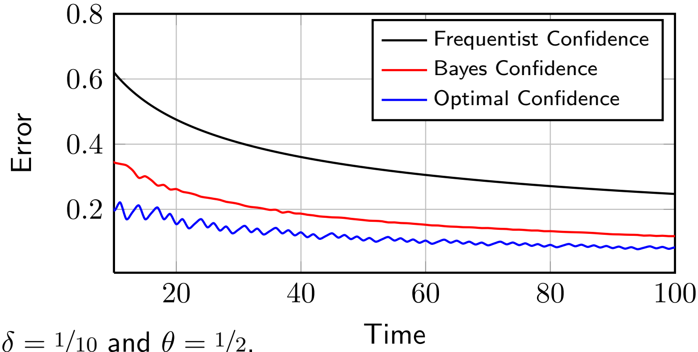
# Posterior Convergence




21 Bernoulli measures with true $\theta = 1/2$ and $w(\nu) = 1/21$ and $\delta = 1/10$

# Bayesian Confidence

$$\mathcal{M}(x_{<n}) := \{\nu \in \mathcal{M} : \forall \eta \leq n,\ w(\nu|x_{<\eta}) \geq \delta w(\nu)\}$$
$$\hat{h}(x_{<n}) := \max\{h_{x_{<n}}(\nu, \xi) : \nu \in \mathcal{M}(x_{<n})\}$$




$\delta = 1/10$ and $\theta = 1/2$.

# Bayesian Confidence

$$\mathcal{M}(x_{<n}) := \{\nu \in \mathcal{M} : \forall \eta \leq n,\ w(\nu | x_{<\eta}) \geq \delta w(\nu)\}$$
$$\hat{h}(x_{<n}) := \max\{h_{x_{<n}}(\nu, \xi) : \nu \in \mathcal{M}(x_{<n})\}$$

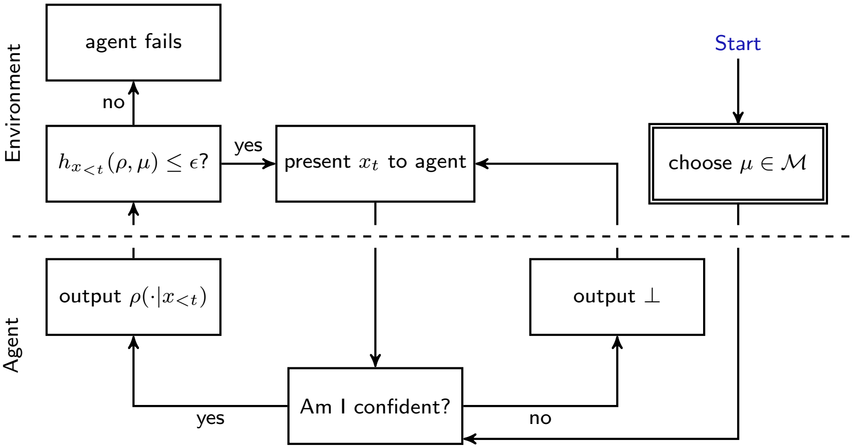
**Theorem**

*If $w$ is uniform, then with $\mu$-probability at least $1 - \delta$*

$$\sum_{n=1}^{\infty} \hat{h}(x_{<n}) \lesssim |\mathcal{M}| \left(\ln |\mathcal{M}| + \ln \frac{|\mathcal{M}|}{\delta}\right)$$

Therefore $\hat{h}(x_{<n})$ converges fast to zero

# Knows What It Knows Framework



# Knows What It Knows Algorithm

**Inputs:** $\varepsilon$, $\delta$ and $\mathcal{M} := \left\{\nu_1, \nu_2, \cdots, \nu_{|\mathcal{M}|}\right\}$.
$t \leftarrow 1$ and $x_{<t} \leftarrow \epsilon$ and $w : \mathcal{M} \to [0, 1]$ is uniform
**loop**
    **if** $\hat{h}_t(x_{<t}) \le \varepsilon$ **then**
        output $\xi(\cdot|x_{<t})$
    **else**
        output $\perp$
    observe $x_t$ and $t \leftarrow t + 1$

**Theorem**

*The following hold:*

1. *The agent fails with probability at most $\delta$*
2. *The number of times action $\perp$ is taken is at most*

$$O\left(\frac{|\mathcal{M}|}{\epsilon} \log \frac{|\mathcal{M}|}{\delta}\right)$$

*with probability at last $1 - \delta$*

# Summary

- Constructed frequentist-style confidence intervals for discrete non-i.i.d. Bayes
- Works well in theory and in practise
- Leads to state-of-the-art bounds for KWIK learning
- Generic and applicable elsewhere (Bandits/RL)
- Also have bounds for KL divergence
- Countable case also covered