Concentration and Confidence for Discrete Bayesian Sequence Predictors

Tor Lattimore and **Marcus Hutter** and **Peter Sunehag**

Research School of Computer Science
Australian National University
{tor.lattimore,marcus.hutter,peter.sunehag}@anu.edu.au

July 2, 2013

Abstract

Bayesian sequence prediction is a simple technique for predicting future symbols sampled from an unknown measure on infinite sequences over a countable alphabet. While strong bounds on the expected cumulative error are known, there are only limited results on the distribution of this error. We prove tight high-probability bounds on the cumulative error, which is measured in terms of the Kullback-Leibler (KL) divergence. We also consider the problem of constructing upper confidence bounds on the KL and Hellinger errors similar to those constructed from Hoeffding-like bounds in the i.i.d. case. The new results are applied to show that Bayesian sequence prediction can be used in the Knows What It Knows (KWIK) framework with bounds that match the state-of-the-art.

Contents

1 Introduction

Sequence prediction is the task of predicting symbol ω_t having observed $\omega_{<t} = \omega_1\omega_2\omega_3\cdots\omega_{t-1}$ where the underlying distribution from which the sequence is sampled is unknown and may be non-stationary. We assume sequences are sampled from an unknown measure μ known to be contained in a countable model class $\mathcal{M}$. At time-step t having observed $\omega_{<t}$ a predictor ρ should output a distribution ρ_t over the next symbol ω_t. A predictor may be considered good if for all $\mu \in \mathcal{M}$ the predictive distribution of ρ converges fast to that of μ

$$\Delta(\rho_t, \mu_t) \stackrel{fast}{\longrightarrow} 0$$

where $\Delta(\rho_t, \mu_t)$ is some measure of the distance between ρ_t and μ_t, typically either the Kullback-Leibler (KL) divergence d_t or the squared Hellinger distance h_t. One such predictor is the Bayesian mixture ξ over all $\nu \in \mathcal{M}$ with strictly positive prior. A great deal is already known about ξ. In particular the predictive distribution ξ_t converges to μ_t with μ-probability one and does so with finite expected cumulative error with respect to both the KL divergence and the squared Hellinger distance [BD62, Sol78, Hut01, Hut03, Hut05].

The paper is divided into three sections. In the first we review the main results bounding the expected cumulative error between μ_t and ξ_t and prove high-probability bounds on this quantity. Such bounds are already known for the squared Hellinger distance, but not the KL divergence until now [HM07]. We also bound the cumulative ξ-expected information gain. The second section relates to the confidence of the Bayes predictor. Even though h_t and d_t converge fast to zero, these quantities cannot be computed without knowing μ. We construct confidence bounds $\hat{h}_t$ and $\hat{d}_t$ that are computable from the observations and upper bound h_t and d_t with high probability respectively. Furthermore we show that $\hat{h}_t$ and $\hat{d}_t$ also converge fast to zero and so can be used in the place of the unknown h_t and d_t. The results serve a similar purpose to upper confidence bounds obtained from Hoeffding-like bounds in the i.i.d. case to which our bounds are roughly comparable (Appendix E). Finally we present a simple application of the new results by showing that Bayesian sequence prediction can be applied to the Knows What It Knows (KWIK) framework [LLWS11] where we achieve a state-of-the-art bound using a simple, efficient and principled algorithm.

2 Notation

A table summarising the notation presented in this section may be found in Appendix F. The natural numbers are denoted by $\mathbb{N}$. Logarithms are taken with respect to base e. The indicator function is $[\![expr]\!]$, which is equal to 1 if $expr$ is true and 0 otherwise. The alphabet X is a finite or countable set of symbols. A finite string x over alphabet X is a sequence $x_1x_2x_3\cdots x_n$ where $x_k \in X$. An infinite string is a sequence $\omega_1\omega_2\omega_3\cdots$. We denote the set of all finite strings by X^* and the set of infinite strings by X^∞. The length of finite string $x \in X^*$ is denoted by $\ell(x)$. Strings can be concatenated. If $x \in X^*$ and $y \in X^* \cup X^\infty$, then xy is the concatenation of x and y. For string $x \in X^* \cup X^\infty$, substrings are denoted by $x_{1:t} = x_1x_2\cdots x_t$ and $x_{<t} = x_{1:t-1}$. The empty string of length zero is denoted by ϵ.

Measures. The cylinder set of finite string x is $\Gamma_x := \{x\omega : \omega \in X^\infty\}$. Define σ-algebra $\mathcal{F}_{<t} := \sigma(\{\Gamma_x : x \in X^{t-1}\})$ and $\mathcal{F} := \sigma(\{\Gamma_x : x \in X^*\})$. Then $(X^\infty, \{\mathcal{F}_{<t}\}, \mathcal{F})$ is a filtered probability space. Let μ be a probability measure on this space. We abuse notation by using the shorthands $\mu(x) := \mu(\Gamma_x)$ and $\mu(y|x) := \mu(xy)/\mu(x)$. The intuition is that $\mu(x)$ represents the μ-probability that an infinite sequence sampled from μ starts with x and $\mu(y|x)$ is the μ-probability that an infinite sequence sampled from μ starts with xy given that it starts with x. We write $\mu \ll \xi$ if μ is absolutely continuous with respect to ξ. From now on, unless otherwise specified, all measures will be probability measures on filtered probability space $(X^\infty, \{\mathcal{F}_{<t}\}, \mathcal{F})$.

Bayes mixture. Let $\mathcal{M}$ be a countable set of measures and $w : \mathcal{M} \to (0, 1]$ be a probability distribution on $\mathcal{M}$. The Bayes mixture measure $\xi : \mathcal{F} \to [0, 1]$ is defined by $\xi(A) := \sum_{\nu\in\mathcal{M}} w_\nu\nu(A)$. By the definition $\xi(A) \geq w_\nu\nu(A)$ for all $A \in \mathcal{F}$ and $\nu \in \mathcal{M}$, which implies that $\nu \ll \xi$. Having observed data $x \in X^*$ the prior w is updated using Bayes rule to be $w_\nu(x) := w_\nu\nu(x)/\xi(x)$. Then $\xi(y|x)$ can

be written $\xi(y|x) = \sum_{\nu\in\mathcal{M}} w_\nu(x)\nu(y|x)$. The entropy of the prior is $\mathrm{Ent}(w) := -\sum_{\nu\in\mathcal{M}} w_\nu \ln w_\nu$.

Distances between measures. Let μ and ξ be measures. The squared Hellinger distance between the predictive distributions of μ and ξ given $x \in X^*$ is defined by $h_x(\mu,\xi) := \sum_{a\in X}(\sqrt{\mu(a|x)} - \sqrt{\xi(a|x)})^2$. If $\mu \ll \xi$, then the Kullback-Leibler (KL) divergence is defined by $d_x(\mu\|\xi) := \sum_{a\in X}\mu(a|x)\ln\frac{\mu(a|x)}{\xi(a|x)}$. The KL divergence is not a metric because it satisfies neither the symmetry nor the triangle inequality properties. Nevertheless, it is a useful measure of the difference between measures and is occasionally more convenient than the Hellinger distance. Let ξ be the Bayes mixture over $\nu \in \mathcal{M}$ with prior $w : \mathcal{M} \to (0,1]$. If $\rho \in \mathcal{M}$, then define random variables on X^∞ by

$$\rho_{1:t}(\omega) := \rho(\omega_{1:t}) \qquad \rho_{<t}(\omega) := \rho(\omega_{<t}) \qquad \rho_t(\omega) := \rho(\omega_t|\omega_{<t})$$

$$h_t(\rho,\xi)\,(\omega) := h_{\omega_{<t}}(\rho,\xi) \qquad d_t(\rho\|\xi)\,(\omega) := d_{\omega_{<t}}(\rho\|\xi)$$

The latter term can be rewritten as

$$d_t(\rho\|\xi) = \mathbb{E}_\rho\left[\ln\frac{\rho_{1:t}}{\rho_{<t}}\cdot\frac{\xi_{<t}}{\xi_{1:t}}\middle|\mathcal{F}_{<t}\right] = \mathbb{E}_\rho\left[\ln\frac{\rho_{1:t}}{\xi_{1:t}}\middle|\mathcal{F}_{<t}\right] + \ln\frac{\xi_{<t}}{\rho_{<t}}. \tag{1}$$

Now fix an unknown $\mu \in \mathcal{M}$ and define random variables (also on X^∞).

$$d_t := d_t(\mu\|\xi) \qquad h_t := h_t(\mu,\xi) \qquad c_t(\omega) := \sum_{\nu\in\mathcal{M}} w_\nu(\omega_{<t})d_{\omega_{<t}}(\nu\|\xi)$$

$$D_\infty := \sum_{t=1}^\infty d_t \qquad H_\infty := \sum_{t=1}^\infty h_t \qquad C_\infty := \sum_{t=1}^\infty c_t.$$

Both h_t and d_t are well-known "distances" between the predictive distributions of ξ and μ at time t. The other quantity c_t is the ξ-expected information gain of the posterior between times t and $t+1$ given the observed sequence at time t.

$$c_t = \sum_{\nu\in\mathcal{M}} w_\nu\frac{\nu_{<t}}{\xi_{<t}}d_t(\nu\|\xi) = \mathbb{E}_\xi\Bigg[\underbrace{\sum_{\nu\in\mathcal{M}} w_\nu\frac{\nu_{1:t}}{\xi_{1:t}}\ln\frac{\nu_t}{\xi_t}}_{\text{information gain}}\Bigg|\mathcal{F}_{<t}\Bigg]$$

An important observation is that c_t is independent of the unknown μ.

3 Convergence

In this section we consider the convergence of $\xi_t - \mu_t \to 0$ for all $\mu \in \mathcal{M}$ where convergence holds with μ-probability 1, in mean sum or with high μ-probability of a small cumulative error. The first theorem is a version of the celebrated result of Solomonoff that the predictive distribution of the Bayes mixture ξ converges fast to the truth in expectation [Sol78, Hut05]. The only modification is the alphabet is now permitted to be countable rather than finite.

Theorem 1 (Sol78, Hut05)**.** *The following hold:*

$$\mathbb{E}_\mu H_\infty \le \mathbb{E}_\mu D_\infty \le \ln\frac{1}{w_\mu} \qquad \lim_{t\to\infty} d_t = \lim_{t\to\infty} h_t = 0, \quad w.\mu.p.1.$$

The proof can be found in Appendix B. Theorem 1 shows that the predictive distribution of ξ converges to μ asymptotically and that it does so fast (with finite cumulative squared Hellinger/KL error) in expectation. We now move on to the question of high-probability bounds on D_∞ and H_∞. The following theorem is already known and essentially unimprovable.

Theorem 2 (HM07)**.** *For all $\delta \in (0,1)$ it holds with μ-probability at least $1-\delta$ that $H_\infty \leq \ln \frac{1}{w_\mu} + 2 \ln \frac{1}{\delta}$.*

We contribute a comparable concentration bound for D_∞. A weak bound can be obtained by applying Markov's inequality to show that $D_\infty \leq \frac{1}{\delta} \cdot \ln(\frac{1}{w_\mu})$ with μ-probability at least $1-\delta$, but a stronger result is possible.

Theorem 3. *For all $\delta \in (0,1)$ it holds with μ-probability at least $1-\delta$ that $D_\infty \leq e \cdot (\ln \frac{6}{\delta}) \cdot (\ln \frac{2}{\delta} + \ln \frac{1}{w_\mu})$.*

Proof. A stopping time is a random variable $t : X^\infty \to \mathbb{N} \cup \{\infty\}$ such that $t^{-1}(n)$ is $\mathcal{F}_{<n}$ measurable for all n. For stopping time t let $X(t) \subset X^*$ be the set of finite sequences where t becomes known

$$X(t) := \{x : t(x\omega) = \ell(x) + 1, \forall \omega\}\,.$$

Define random variable $z_{<t} := \xi_{<t}/\mu_{<t}$ and $L := \lceil \ln(2/\delta) \rceil \leq \ln(6/\delta)$ and stopping times $\{t_k\}$ inductively by

$$t_1 := 1 \qquad\qquad t_{k+1} := \min\left\{ s : \sum_{t=t_k}^{s} d_t > e \cdot \left(\ln z_{<t_k} + \ln \frac{1}{w_\mu} \right) \right\}.$$

The result follows from two claims, which are proven later.

$$\boxed{\mu\left(\sup_t \ln z_{<t} \geq \ln \frac{2}{\delta} \right) \leq \delta/2 \qquad (\star)} \qquad \boxed{\mu\Big(t_{L+1} < \infty \Big) \leq \delta/2 \qquad (\star\star)}$$

By the union bound we obtain that if A is the event that $t_{L+1} = \infty$ and $\sup_t \ln z_{<t} \leq \ln \frac{2}{\delta}$, then $\mu(A) \geq 1-\delta$ and for $\omega \in A$

$$\begin{aligned} D_\infty(\omega) = \sum_{t=1}^{\infty} d_t(\omega) &\overset{\text{(a)}}{=} \sum_{k=1}^{L} \sum_{t=t_k(\omega)}^{t_{k+1}(\omega)-1} d_t(\omega) \overset{\text{(b)}}{\leq} \sum_{k=1}^{L} e \cdot \left(\ln z_{<t_k}(\omega) + \ln \frac{1}{w_\mu} \right) \\ &\overset{\text{(c)}}{\leq} e \cdot L \left(\ln \frac{2}{\delta} + \ln \frac{1}{w_\mu} \right) \overset{\text{(d)}}{\leq} e \cdot \ln\left(\frac{6}{\delta}\right) \cdot \left(\ln \frac{2}{\delta} + \ln \frac{1}{w_\mu} \right) \end{aligned}$$

where (a) follows from the definition of t_k and because $t_{L+1}(\omega) = \infty$. (b) follows from the definition of t_k. (c) because $\sup_t \ln z_{<t} \leq \ln \frac{2}{\delta}$. (d) by the definition of L. The theorem is completed by proving $(\star)$ and $(\star\star)$. The first follows immediately from Lemma 14. For the second we use induction and Theorem 1. After observing $x \in X(t_n)$, $\xi(\cdot|x)$ is a Bayes mixture over $\nu(\cdot|x)$ where $\nu \in \mathcal{M}$ with prior weight $w(\nu(\cdot|x)) = w_\nu \nu(x)/\xi(x)$. Therefore by Theorem 1

$$\mathbb{E}_\mu \left[\sum_{t=\ell(x)+1}^{\infty} d_t \middle| x \right] \leq \ln \frac{1}{w(\mu(\cdot|x))} = \ln \frac{\xi(x)}{\mu(x)} + \ln \frac{1}{w_\mu}.$$

Therefore by Markov's inequality

$$\mu\left(\sum_{t=\ell(x)+1}^{\infty} d_t > e\cdot\left(\ln\frac{\xi(x)}{\mu(x)}+\ln\frac{1}{w_\mu}\right)\middle| x\right)\le\frac{1}{e}.$$

Let $n\in\mathbb{N}$ and assume $\mu(t_n<\infty)\le e^{1-n}$. By the definition $t_{n+1}\ge t_n$ we have

$$\begin{aligned}\mu\Big(t_{n+1}<\infty\Big)&=\sum_{x\in X(t_n)}\mu(x)\cdot\mu\left(\sum_{t=\ell(x)+1}^{\infty} d_t > e\cdot\left(\ln\frac{\xi(x)}{\mu(x)}+\ln\frac{1}{w_\mu}\right)\middle| x\right)\\&\le\frac{1}{e}\sum_{x\in X(t_n)}\mu(x)=\frac{1}{e}\mu(t_n<\infty)\le e^{-n}.\end{aligned}$$

Therefore $\mu(t_n<\infty)\le e^{1-n}$ for all n and so $\mu(t_{L+1}<\infty)\le e^{-L}\le\delta/2$, which completes the proof of $(\star\star)$ and so also the theorem. ■

Theorem 3 is close to unimprovable.

Proposition 4. *There exists an $\mathcal{M}=\{\mu,\nu\}$ such that with μ-probability at least δ it holds that $D_\infty > \frac{1}{4\ln 2}\ln\frac{1}{\delta}\left(\ln\frac{1}{\delta}+2\ln\frac{1-w}{w}-3\ln 2\right)$.*

Proof. Let $X=\{0,1\}$ and $\mathcal{M}:=\{\mu,\nu\}$ where the true measure μ is the Lebesgue measure and ν is the measure deterministically producing an infinite sequence of ones, which are defined by $\mu(x):=2^{-\ell(x)}$ and $\nu(x):=[\![x=1^{\ell(x)}]\!]$ where 1^n is the sequence of n ones.. Let $w=w_\mu$ and $w_\nu=1-w$. If $n=\left\lfloor\frac{1}{\ln 2}\ln\frac{1}{\delta}\right\rfloor\in\mathbb{N}$, then $\mu(\Gamma_{1^n})\ge\delta$ and for $\omega\in\Gamma_{1^n}$

$$\begin{aligned}D_\infty(\omega)&\overset{(a)}{\ge}\sum_{t=1}^{n+1}d_{1^{t-1}}(\mu\|\xi)\overset{(b)}{=}\sum_{t=1}^{n+1}\left(\frac{1}{2}\cdot\ln\frac{\frac{1}{2}}{\xi(1|1^{t-1})}+\frac{1}{2}\cdot\ln\frac{\frac{1}{2}}{\xi(0|1^{t-1})}\right)\\&\overset{(c)}{>}\frac{1}{2}\sum_{t=1}^{n+1}\ln\left(\frac{1}{4\xi(0|1^{t-1})}\right)\overset{(d)}{=}\frac{1}{2}\sum_{t=1}^{n+1}\ln\left(\frac{w\cdot 2^{1-t}+(1-w)}{4w\cdot 2^{-t}}\right)\\&\overset{(e)}{\ge}\frac{1}{2}\sum_{t=1}^{n+1}\left((t-2)\ln 2+\ln\frac{1-w}{w}\right)\overset{(f)}{=}\frac{(n+1)\left(2\ln\frac{1-w}{w}+(n-2)\ln 2\right)}{4}\end{aligned}$$

(a) follows from the definition of $D_\infty(\omega)$ and the positivity of the KL divergence, which allows the sum to be truncated. (b) follows by inserting the definitions of μ and the KL divergence. (c) by basic algebra and the fact that $\xi(1|1^{t-1})<1$. (d) follows from the definition of ξ while (e) and (f) are basic algebra. Finally substitute $n+1\ge\frac{1}{\ln 2}\ln\frac{1}{\delta}$. ■

In the next section we will bound d_t by a function of c_t, which can be computed without knowing μ. For this result to be useful we need to show that c_t converges to zero, which is established by the following theorems.

Theorem 5. *If $\mathrm{Ent}(w)<\infty$, then $\mathbb{E}_\mu C_\infty\le\mathrm{Ent}(w)/w_\mu$ and $\lim_{t\to\infty}c_t=0$ with μ-probability 1.*

Proof. We make use of the dominance $\xi(x)\ge w_\mu\mu(x)$, properties of expectation and Theorem 1.

$$\begin{aligned}\mathbb{E}_\mu C_\infty:=\mathbb{E}_\mu\sum_{t=1}^{\infty}c_t&\overset{(a)}{\le}\frac{1}{w_\mu}\mathbb{E}_\xi\sum_{t=1}^{\infty}c_t\overset{(b)}{=}\frac{1}{w_\mu}\mathbb{E}_\xi\sum_{t=1}^{\infty}\sum_{\nu\in\mathcal{M}}w_\nu\frac{\nu_{<t}}{\xi_{<t}}d_t(\nu\|\xi)\\&\overset{(c)}{=}\frac{1}{w_\mu}\sum_{\nu\in\mathcal{M}}w_\nu\mathbb{E}_\nu\sum_{k=1}^{\infty}d_t(\nu\|\xi)\overset{(d)}{\le}\frac{1}{w_\mu}\sum_{\nu\in\mathcal{M}}w_\nu\ln\frac{1}{w_\nu}\overset{(e)}{=}\frac{\mathrm{Ent}(w)}{w_\mu}\end{aligned}$$

(a) follows by dominance $\mu(A) \leq \xi(A)/w_\mu$ and linearity of expectation. (b) is the definition of c_t. (c) by exchanging sums and the definition of expectation. (d) is true by substituting the result in Theorem 1. Finally (e) follows from the definition of the entropy $\mathrm{Ent}(w)$. That $\lim_{t\to\infty} c_t = 0$ with μ-probability 1 follows from the first result by applying Markov's inequality to bound $C_\infty < \infty$ with probability 1. ■

In the finite case a stronger result is possible.

Theorem 6. *If $|\mathcal{M}| = K < \infty$ and w is the uniform prior, then $\mathbb{E}_\mu C_\infty \leq 6\ln^2 K + 14\ln K + 8$.*

Theorem 6 is tight in the following sense.

Proposition 7. *For each $K \in \mathbb{N}$ there exists an $\mathcal{M}$ of size K and $\mu \in \mathcal{M}$ such that if w is the uniform prior on $\mathcal{M}$, then $\mathbb{E}_\mu C_\infty > \frac{1}{2}\ln^2 K - 1$.*

See the appendix for the proofs of Theorem 6 and Proposition 7.

4 Confidence

In the previous section we showed that ξ_t converges fast to μ_t. One disadvantage of these results is that errors d_t and h_t cannot be determined without knowing μ. In this section we define $\hat{d}_t$ and $\hat{h}_t$ that upper bound d_t and h_t respectively with high probability and may be computed without knowing μ. Let $\mathcal{M} \supseteq \mathcal{M}_1 \supseteq \mathcal{M}_2 \cdots$ be a narrowing sequence of hypothesis classes where $\mathcal{M}_t$ contains the set of plausible models at time-step t and is defined by

$$\mathcal{M}_t := \left\{\nu \in \mathcal{M} : \forall \tau \leq t, \frac{\nu_{<\tau}}{\xi_{<\tau}} \geq \delta \frac{w_\mu}{w_\nu}\right\}$$

Then $\hat{h}_t$ is defined as the value maximising the weighted squared Hellinger distance between ν and ξ for all plausible $\nu \in \mathcal{M}_t$ and $\hat{d}_t$ is defined in terms of the expected information gain.

$$\boxed{\hat{d}_t := \frac{c_t}{w_\mu \delta}} \qquad \boxed{\hat{h}_t := \sup_{\nu \in \mathcal{M}_t}\left\{\frac{w_\nu}{w_\mu} h_t(\nu, \xi)\right\}}$$

Both d_t and h_t depend on w_μ, which is also typically unknown. If $\mathcal{M}$ is finite, then the problem is easily side-stepped by choosing w to be uniform. The countable case is discussed briefly in the conclusion. First we prove that $h_t \leq \hat{h}_t$ and $d_t \leq \hat{d}_t$ with high probability after which we demonstrate that they are non-vacuous by proving that $\hat{h}_t$ and $\hat{d}_t$ converge fast to zero with high probability. Now is a good time to remark that hypothesis testing using the factor $\nu_{<t}/\xi_{<t}$ is not exactly a new idea. For discussion, results, history and references see [SSVV11].

Theorem 8. *For all $\delta \in [0,1]$ it holds that:*

$$\mu(\forall t : d_t \leq \hat{d}_t) \geq 1 - \delta \qquad (\star) \qquad\qquad \mu(\forall t : h_t \leq \hat{h}_t) \geq 1 - \delta \qquad (\star\star)$$

Proof. To prove ($\star$) define event $A := \left\{\omega : \sup_t \xi(\omega_{<t})/\mu(\omega_{<t}) < \frac{1}{\delta}\right\}$. By Lemma 14 in the appendix we have that $\mu(A) \geq 1-\delta$. If $\omega \in A$, then $\mu(\omega_{<t})/\xi(\omega_{<t}) > \delta$ for all t and

$$c_t(\omega) \overset{\text{(a)}}{=} \sum_{\nu\in\mathcal{M}} w_\nu \frac{\nu(\omega_{<t})}{\xi(\omega_{<t})} d_{\omega_{<t}}(\nu\|\xi) \overset{\text{(b)}}{\geq} w_\mu \frac{\mu(\omega_{<t})}{\xi(\omega_{<t})} d_{\omega_{<t}}(\mu\|\xi)$$
$$\overset{\text{(c)}}{>} w_\mu \cdot \delta \cdot d_{\omega_{<t}}(\mu\|\xi) \overset{\text{(d)}}{=} w_\mu \cdot \delta \cdot d_t.$$

(a) is the definition of c_t. (b) follows by dropping all elements of the sum except μ. (c) by substituting the bound on μ/ξ. (d) is the definition of d_t. Therefore $d_t \cdot w_\mu \cdot \delta \leq c_t$ with μ-probability at least $1-\delta$ as required. For ($\star\star$) we note that by the definition of $\hat{h}_t$, if $\mu \in \mathcal{M}_t$, then $h_t \leq \hat{h}_t$. The result is completed by applying Lemma 14 in the appendix to show that $\mu \in \mathcal{M}_t$ for all t with probability at least $1-\delta$. ■

Theorem 9. *The following hold:*

1. $\mathbb{E}_\mu \sum_{t=1}^\infty \hat{d}_t \leq \frac{\mathrm{Ent}(w)}{\delta w_\mu^2}$.
2. *w.μ.p. at least* $1-\delta$ *it holds that* $\sum_{t=1}^\infty \hat{d}_t \leq \frac{\mathrm{Ent}(w)}{\delta^2 w_\mu^2}$.

Theorem 10. *The following hold:*

1. $\mathbb{E}_\mu \sum_{t=1}^\infty \hat{h}_t \leq \frac{2}{w_\mu}\left(\ln\frac{1}{w_\mu} + \ln\frac{1}{\delta} + \mathrm{Ent}(w)\right)$
2. *w.μ.p. at least* $1-\delta$, $\sum_{t=1}^\infty \hat{h}_t \leq \frac{2}{w_\mu}\left(2\ln\frac{1}{w_\mu} + 5\ln\frac{1}{\delta} + 3\,\mathrm{Ent}(w)\right)$.

The consequences of Thereoms 6, 9 and 10 are summarised in Figure 1 for both countable and finite hypothesis classes. The proof of Theorem 9 follows immediately from Theorem 5 and Markov's inequality. If $\mathcal{M}$ is finite and w uniform, then one can use Theorem 6 instead to improve dependence on $1/w_\mu$. For Theorem 10 we use Theorem 2 and the following lemma, which is a generalization of Lemma 4 in [HM07].

Lemma 11. *Let* $\kappa > 0$ *and stopping time* $\tau := \min_t\{t : \nu_{<t}/\mu_{<t} < \kappa\}$. *Then* $\mathbb{E}_\mu \sum_{t=1}^{\tau-1} h_t(\nu,\mu) \leq 2\ln \mathbb{E}_\mu \exp\left(\frac{1}{2}\sum_{t=1}^{\tau-1} h_t(\nu,\mu)\right) \leq \ln\frac{1}{\kappa}$.

Proof. We borrow some tricks from [Vov87] and [HM07, Lem 4]. Define ρ inductively by $\rho(a|x) := \sqrt{\nu(a|x)\mu(a|x)}/\sum_{b\in X}\sqrt{\nu(b|x)\mu(b|x)}$.

$$\rho_{<\tau} \overset{\text{(a)}}{=} \prod_{t=1}^{\tau-1}\rho_t \overset{\text{(b)}}{\geq} \prod_{t=1}^{\tau-1}\sqrt{\nu_t\mu_t}\exp\left(\frac{1}{2}h_t(\nu,\mu)\right)$$
$$\overset{\text{(c)}}{=} \prod_{t=1}^{\tau-1}\mu_t\sqrt{\frac{\nu_t}{\mu_t}}\exp\left(\frac{1}{2}h_t(\nu,\mu)\right) \overset{\text{(d)}}{=} \mu_{<\tau}\sqrt{\frac{\nu_{<\tau}}{\mu_{<\tau}}}\prod_{t=1}^{\tau-1}\exp\left(\frac{1}{2}h_t(\nu,\mu)\right)$$
$$\overset{\text{(e)}}{\geq} \mu_{<\tau}\sqrt{\kappa}\prod_{t=1}^{\tau-1}\exp\left(\frac{1}{2}h_t(\nu,\mu)\right) \overset{\text{(f)}}{=} \mu_{<\tau}\sqrt{\kappa}\exp\left(\frac{1}{2}\sum_{t=1}^{\tau-1}h_t(\nu,\mu)\right)$$

where (a) and (d) follow from the definition of conditional probability. (b) by inserting the definition of ρ_t and applying Lemma 13. (c) by factoring. (e) by noting that $\nu_{<\tau}/\mu_{<\tau} \geq \kappa$ by the

definition of τ. (f) by exchanging $\prod \exp = \exp \sum$. Therefore $\sqrt{\kappa} \exp(\frac{1}{2} \sum_{t=1}^{\tau-1} h_t(\nu, \mu)) \leq \frac{\rho_{<\tau}}{\mu_{<\tau}}$. Taking the expectation with respect to μ

$$\mathbb{E}_\mu \exp\left(\frac{1}{2} \sum_{t=1}^{\tau-1} h_t(\nu, \mu)\right) \leq \mathbb{E}_\mu \frac{1}{\sqrt{\kappa}} \frac{\rho_{<\tau}}{\mu_{<\tau}} \leq \mathbb{E}_\rho \frac{1}{\sqrt{\kappa}} = \frac{1}{\sqrt{\kappa}}.$$

By Jensen's inequality $\mathbb{E}X = 2 \ln \exp \mathbb{E} \frac{1}{2} X \leq 2 \ln \mathbb{E} \exp \frac{1}{2} X$ and so

$$\mathbb{E}_\mu \sum_{t=1}^{\tau} h_t(\nu, \mu) \leq 2 \ln \mathbb{E}_\mu \exp\left(\frac{1}{2} \sum_{t=1}^{\tau} h_t(\nu, \mu)\right) \leq 2 \ln \frac{1}{\sqrt{\kappa}} = \ln \frac{1}{\kappa}$$

as required. ■

Proof of Theorem 10. Define stopping times τ_ν and $\bar{\tau}_\nu$ by

$$\tau_\nu := \min_t \{t : \nu_{<t}/\xi_{<t} < w_\nu / w_\mu \delta\} \qquad \bar{\tau}_\nu := \min_t \{t : \nu_{<t}/\mu_{<t} < w_\nu \delta\}$$

First we show that $\bar{\tau}_\nu \geq \tau_\nu$. By dominance $\xi_{<t} \geq w_\mu \mu_{<t}$ we have that

$$\frac{\nu_{<t}}{\mu_{<t}} < w_\nu \delta \implies \frac{\nu_{<t} \xi_{<t}}{\mu_{<t} \xi_{<t}} < w_\nu \delta \implies \frac{\nu_{<t}}{\xi_{<t}} < \frac{w_\nu}{w_\mu} \delta$$

Therefore $\bar{\tau}_\nu \geq \tau_\nu$. Let $\nu_t := \arg\max_{\nu \in \mathcal{M}_t} h_t(\nu, \xi)$ and bound

$$\begin{aligned}
\sum_{t=1}^{\infty} \hat{h}_t &\overset{\text{(a)}}{=} \sum_{t=1}^{\infty} \frac{w_{\nu_t}}{w_\mu} h_t(\nu_t, \xi) \overset{\text{(b)}}{\leq} \frac{2}{w_\mu} \sum_{t=1}^{\infty} (w_{\nu_t} h_t(\mu, \xi) + w_{\nu_t} h_t(\mu, \nu_t)) \\
&\overset{\text{(c)}}{\leq} \frac{2}{w_\mu} \sum_{t=1}^{\infty} \left(h_t(\mu, \xi) + \sum_{\nu \in \mathcal{M}_t} w_\nu h_t(\nu, \mu)\right) \\
&\overset{\text{(d)}}{=} \frac{2}{w_\mu} \left(H_\infty + \sum_{\nu \in \mathcal{M}} \sum_{t=1}^{\tau_\nu - 1} w_\nu h_t(\nu, \mu)\right) \overset{\text{(e)}}{\leq} \frac{2}{w_\mu} \left(H_\infty + \sum_{\nu \in \mathcal{M}} w_\nu \sum_{t=1}^{\bar{\tau}_\nu - 1} h_t(\nu, \mu)\right)
\end{aligned} \tag{2}$$

where (a) is the definition of $\hat{h}_t$. (b) follows by the inequality $h_t(\nu_t, \xi) < 2(h_t(\nu_t, \mu) + h_t(\mu, \xi))$. (c) by dropping $w_{\nu_t} \leq 1$ and bounding the single term $w_{\nu_t} h_t(\nu_t, \mu)$ by the sum over all ν in the plausible class $\mathcal{M}_t$. (d) by the definitions of H_∞, $\mathcal{M}_t$ and τ_ν. (e) by the fact that $\bar{\tau}_\nu \geq \tau_\nu$ as previously shown. Let $\Delta_\nu := \sum_{t=1}^{\bar{\tau}_\nu - 1} h_t(\nu, \mu)$. The first claim is proven by taking the expectation with respect to μ and substituting Theorem 1 to bound $\mathbb{E}_\mu H_\infty \leq \ln \frac{1}{w_\mu}$ and Lemma 11 with $\tau = \bar{\tau}_\nu$ and $\kappa = w_\nu \delta$ to bound $\mathbb{E}_\mu \Delta_\nu \leq \ln \frac{1}{w_\nu} + \ln \frac{1}{\delta}$. For the high probability bound let $\lambda_\nu := 3 \ln \frac{1}{\delta w_\nu} + \ln \frac{1}{w_\mu}$ and apply Lemma 11 and Markov's inequality.

$$\mu\left(\Delta_\nu \geq \lambda_\nu\right) = \mu\left(e^{\Delta_\nu/2} \geq e^{\lambda_\nu/2}\right) \leq e^{-\lambda_\nu/2} \mathbb{E}_\mu[e^{\Delta_\nu/2}] \leq \frac{e^{-\lambda_\nu/2}}{\sqrt{w_\nu \delta}} = w_\nu \delta.$$

By Theorem 2 we have that $H_\infty \leq \ln 1/w_\mu + 2 \ln 1/\delta w_\mu$ with μ-probability at least $1 - w_\mu \delta$ and by the union bound and the fact that $\sum_\nu w_\nu = 1$ we obtain with probability at least $1 - \delta$ that $\Delta_\nu \leq \lambda_\nu$ for all ν and $H_\infty \leq \ln \frac{1}{w_\mu} + 2 \ln \frac{1}{\delta}$, which when substituted into Equation (2) leads to $\sum_{t=1}^{\infty} \hat{h}_t \leq \frac{2}{w_\mu}(2 \ln \frac{1}{w_\mu} + 5 \ln \frac{1}{\delta} + 3 \operatorname{Ent}(w))$ as required. ■

$\lvert\mathcal{M}\rvert$	Expectation	High Probability
∞	$\mathbb{E}_\mu \sum_{t=1}^\infty \hat{d}_t \lesssim \frac{\mathrm{Ent}(w)}{\delta w_\mu^2}$	$\sum_{t=1}^\infty \hat{d}_t \lesssim \frac{\mathrm{Ent}(w)}{\delta^2 w_\mu^2}$
	$\mathbb{E}_\mu \sum_{t=1}^\infty \hat{h}_t \lesssim \frac{1}{w_\mu}\left(\mathrm{Ent}(w) + \ln \frac{1}{w_\mu \delta}\right)$	$\sum_{t=1}^\infty \hat{h}_t \lesssim \frac{1}{w_\mu}\left(\mathrm{Ent}(w) + \ln \frac{1}{w_\mu \delta}\right)$
K	$\mathbb{E}_\mu \sum_{t=1}^\infty \hat{d}_t \lesssim \frac{K}{\delta} \ln^2 K$	$\sum_{t=1}^\infty \hat{d}_t \lesssim \frac{K}{\delta^2} \ln^2 K$
$w_\nu = \frac{1}{K}$	$\mathbb{E}_\mu \sum_{t=1}^\infty \hat{h}_t \lesssim K\left(\ln K + \ln \frac{1}{\delta}\right)$	$\sum_{t=1}^\infty \hat{h}_t \lesssim K\left(\ln K + \ln \frac{1}{\delta}\right)$

$\lesssim$ ignores constant multiplicative factors

Figure 1: Confidence bounds

5 KWIK Learning

The KWIK learning framework involves an environment and agent interacting sequentially as depicted below. Suppose $|\mathcal{M}| = K < \infty$ and $\varepsilon, \delta > 0$ are known to both parties. A run starts with the environment choosing an unknown $\mu \in \mathcal{M}$. At each time-step t thereafter the agent chooses between outputting a predictive distribution $\rho(\cdot|\omega_{<t})$ and special symbol $\perp$. The run is failed if the agent outputs ρ and $h_{\omega_{<t}}(\rho, \mu) > \varepsilon$, otherwise ω_t is observed and the run continues. An agent is said to be KWIK if it fails the run with probability at most δ and chooses $\perp$ at most $B(\varepsilon, \delta)$ times with probability at least $1 - \delta$. Ideally, $B(\varepsilon, \delta)$ should be polynomial in $1/\varepsilon$ and $1/\delta$ [LLWS11].

Algorithm 1 KWIK Learner

Inputs: ε, δ and $\mathcal{M} := \{\nu_1, \nu_2, \cdots, \nu_K\}$.
$t \leftarrow 1$ and $\omega_{<t} \leftarrow \epsilon$ and $w_\nu = \frac{1}{K}$
loop
  if $\hat{h}_t(\omega_{<t}) \leq \varepsilon$ **then**
    output $\xi(\cdot|\omega_{<t})$
  else
    output $\perp$
  observe ω_t and $t \leftarrow t + 1$

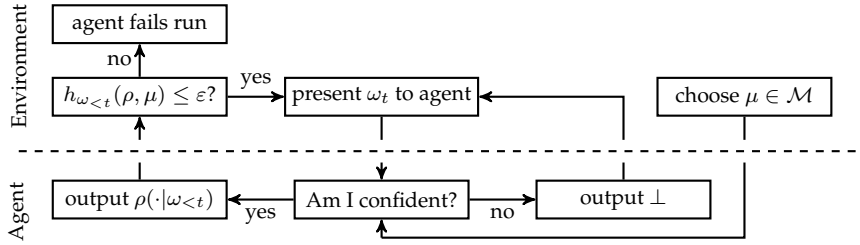


Figure 2: KWIK learning framework

Theorem 12. *Algorithm 1 is KWIK.*

Proof. By Theorem 8, Algorithm 1 fails a run with probability at most δ. Using $\gtrsim$ to ignore constant multiplicative factors, by Theorem 10 we have that

$$\mu\left(\left|\left\{t : \hat{h}_t \geq \varepsilon\right\}\right| \gtrsim \frac{K}{\varepsilon} \ln \frac{K}{\delta}\right) \leq \mu\left(\sum_{t=1}^\infty \hat{h}_t \gtrsim K \ln \frac{K}{\delta}\right) \leq \delta.$$

Therefore the agent will choose $\perp$ at most $O\left(\frac{K}{\varepsilon} \ln \frac{K}{\delta}\right)$ times with probability at least $1 - \delta$. ■

The Hellinger distance upper bounds the total variation distance. $\delta_x(\mu, \xi) = \frac{1}{2} \sum_{a \in X} |\mu(a|x) - \xi(a|x)| \leq \sqrt{h_x(\mu, \xi)}$. Therefore if Algorithm 1 is run with $\varepsilon = \varepsilon_1^2$, then with high probability when predicting it will be ε_1-optimal with respect to the total variation distance and it will output $\perp$ at most $O\left(\frac{K}{\varepsilon_1^2} \ln \frac{K}{\delta}\right)$ times, which is the same bound achieved by the k-meteorologist algorithm [DLL09].

6 Conclusions

The bound on the squared Hellinger distance $\hat{h}_t$ is especially nice because the results are rather clean. While the super-linear dependence on the size of the model class in Figure 1 is unfortunate, it is a worst-case bound that is only achieved when at each time-step only one model differs from ξ (see the proof of Proposition 7 for an example environment class when this occurs). For Bernoulli classes the estimator performs comparably with the Hoeffding bound (Appendix E). In the case when $\mathcal{M}$ is countable $\hat{h}_t$ is independent of μ, but not w_μ, which is also typically unknown. Either choose a conservatively small w and pay the $\frac{1}{w}\ln\frac{1}{w}$ price, or decrease w with t at some slow rate, say $w = \sqrt{1/t}$. Analyzing this situation is interesting future work.

There is opportunity for some improvement on the bound $\hat{d}$. Intuitively we expect the real dependence on $1/\delta$ ought to be logarithmic, not linear. The unimprovable result of Theorem 3 is interesting when compared to Theorem 2. Researchers frequently bound the total variation distance via the KL divergence. These results show that this is sometimes weaker than using the Hellinger distance when high-probability bounds are required.

KWIK learning for sequence prediction was chosen because our new results can easily be applied to prove a state-of-the-art bound in that setting. Although we have the same theoretical guarantee as the k-meteorologist algorithm [DLL09], our simple algorithm eliminates environments smoothly as they become unlikely while in that work no model (expert) is discarded before at least $m = O(\frac{1}{\varepsilon^2}\ln\frac{1}{\delta})$ differentiating samples have been observed. This distinction makes us suspect that Algorithm 1 may perform more efficiently in practice. Additionally, assuming $\nu(\cdot|x)$ can be computed in constant time, then Algorithm 1 runs in $O(K)$ time per time-step, while a naive implementation of the k-meteorologist algorithm appears to have $O(K^2)$ running time per time-step.

Finally, we want to emphasize the generality of the results, especially Theorem 10, which although tight in a minimax sense, can likely be improved in easier cases without changing the definition of $\hat{h}_t$. An interesting continuation is the parametric case that is intuitively straight-forward, but technically challenging (see [CB90] and [Hut05, §3] for some of the required techniques).

References

[BD62] David Blackwell and Lester Dubins. Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, 33(3):882–886, 1962.

[CB90] Bertrand Clarke and Andrew Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36:453–471, 1990.

[DLL09] Carlos Diuk, Lihong Li, and Bethany Leffler. The adaptive k-meteorologists problem and its application to structure learning and feature selection in reinforcement learning. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, *Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009)*, pages 249–256. ACM, 2009.

[HM07] Marcus Hutter and Andrei Muchnik. On semimeasures predicting Martin-Löf random sequences. *Theoretical Computer Science*, 382(3):247–261, 2007.

[Hut01] Marcus Hutter. Convergence and error bounds for universal prediction of nonbinary sequences. In *Proc. 12th European Conf. on Machine Learning (ECML-2001)*, volume 2167 of *LNAI*, Freiburg, 2001. Springer, Berlin.

[Hut03] Marcus Hutter. Optimality of universal Bayesian prediction for general loss and alphabet. *Journal of Machine Learning Research*, 4:971–997, 2003.

[Hut05] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

[LLWS11] Lihong Li, Michael Littman, Thomas Walsh, and Alexander Strehl. Knows what it knows: a framework for self-aware learning. *Machine Learning*, 82(3):399–443, 2011.

[Sol78] Ray Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, 24(4):422–432, 1978.

[SSVV11] Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

[Vil39] Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

[Vov87] Vladimir Vovk. On a randomness criterion. *Soviet Mathematics Doklady*, 35:656–660, 1987.

A Technical Lemmas

Lemma 13 (Vov87)**.** *Let p and q be distribution on X, then*

$$\sum_{a\in X}\sqrt{p(a)q(a)} \leq \exp\left(-\frac{1}{2}\sum_{a\in X}\left(\sqrt{p(a)}-\sqrt{q(a)}\right)^2\right).$$

Lemma 14 (Vil39)**.** *If $z_{<t} := \xi_{<t}/\mu_{<t}$, then $z_{<t}$ is a μ-super-martingale, $\mu\left(\lim_{t\to\infty} z_{<t} < \infty\right) = 1$ and $\mu\left(\sup_t z_{<t} \geq \frac{1}{\delta}\right) \leq \delta$.*

Proof. The proof is straight-forward and is included for the sake of completeness. Define $A \subset X^*$ to be the set of finite strings defined by

$$A := \{x \in X^* : \xi(x)/\mu(x) \geq 1/\delta \wedge \forall t \leq \ell(x), \xi(x_{<t})/\mu(x_{<t}) < 1/\delta\}$$

So A is the set of finite strings where $\xi(x)/\mu(x)$ first drops below δ. Let $\omega \in X^\infty$ and $z_t(\omega) \geq 1/\delta$, then there exists a t such that $\omega_{1:t} \in A$. Therefore if $\bar{A} = \{x \in A : \mu(x) > 0\}$, then

$$\mu(\lim_{t\to\infty} z_t \geq 1/\delta) \overset{\text{(a)}}{=} \sum_{x\in A}\mu(x) \overset{\text{(b)}}{=} \sum_{x\in\bar{A}}\mu(x) \overset{\text{(c)}}{=} \sum_{x\in\bar{A}}\mu(x)\xi(x)/\xi(x) \overset{\text{(d)}}{\leq} \delta\sum_{x\in\bar{A}}\xi(x) \overset{\text{(e)}}{\leq} \delta$$

where (a) follows from the definition of A and z_t. (b) since $\mu(x) = 0$ when $x \in A - \bar{A}$. (c) since $\xi(x) \geq w_\mu\mu(x) > 0$ for $x \in \bar{A}$. (d) by the bound $\mu(x)/\xi(x) \leq \delta$ for $x \in A$. (e) since ξ is a measure. ■

Lemma 15. *Both $\mathbb{E}_\mu \ln(\mu_{<n}/\xi_{<n})$ and $\mathbb{E}_\mu d_n$ exist and are finite.*

Proof. Let $A := \{x \in X^{n-1} : \mu(x) > 0\}$ and $B = \{x \in A : \mu(x)/\xi(x) \geq 1\}$ and $C = \{x \in A : \mu(x)/\xi(x) < 1\}$. If $\mathbb{1}_B$ is the indicator event $\mathbb{1}_B(\omega) := [\![\omega \in \Gamma_x : x \in B]\!]$ and $\mathbb{1}_C$ is defined similarly, then

$$\begin{aligned}\mathbb{E}_\mu\left[|\ln \mu_{<n}/\xi_{<n}|\right] &= \mathbb{E}_\mu\left[\mathbb{1}_B \ln \mu_{<n}/\xi_{<n}\right] + \mathbb{E}_\mu\left[\mathbb{1}_C \ln \xi_{<n}/\mu_{<n}\right] \\ &\leq \mathbb{E}_\mu\left[-\mathbb{1}_B \ln w_\mu\right] + \mathbb{E}_\mu\left[\mathbb{1}_C \xi_{<n}/\mu_{<n}\right] \leq -\ln w_\mu + 1\end{aligned}$$

where the first inequality is due to the dominance $\xi_{<n} \geq w_\mu \mu_{<n}$ and the inequality $\ln x \leq x$ for all $x \geq 1$. The second inequality follows from basic properties of expectation. For the second part note that d_{n+1} is positive and by Equation (1) that $d_n \leq \ln \xi_{<n}/\mu_{<n} - \ln w_\mu \leq |\ln \mu_{<n}/\xi_{<n}| - \ln w_\mu$. Then proceed as in the first part. ■

B Proof of Theorem 1

First we note that the squared Hellinger distances is bounded by the KL divergence, so $H_\infty \leq D_\infty$. We now bound $\mathbb{E}_\mu D_\infty$, which follows from the chain rule for the conditional relative entropy. Fix $n \in \mathbb{N}$ and assume that

$$\Delta_{n-1} := \mathbb{E}_\mu \sum_{t=1}^{n} d_t = \mathbb{E}_\mu \ln \frac{\mu_{<n}}{\xi_{<n}}, \tag{$\star$}$$

which is easily verified when $n = 1$. Therefore

$$\Delta_n \overset{\text{(a)}}{=} \mathbb{E}_\mu \ln \frac{\mu_{<n}}{\xi_{<n}} + \mathbb{E}_\mu d_n \overset{\text{(b)}}{=} \mathbb{E}_\mu \left[\mathbb{E}_\mu \left[\ln \frac{\mu_{1:n}}{\xi_{1:n}} \middle| \mathcal{F}_{<n}\right]\right] \overset{\text{(c)}}{=} \mathbb{E}_\mu \ln \frac{\mu_{1:n}}{\xi_{1:n}}.$$

(a) holds by Lemma 15. (b) by Equation (1) and the definition of expectation. (c) by the definition of (conditional) expectation. Therefore $(\star)$ holds for all n by induction. By substituting dominance $\xi_n \geq w_\mu \mu_n$ into $(\star)$ one obtains that $\Delta_n \leq -\mathbb{E}_\mu \ln w_\mu = -\ln w_\mu$. The proof is completed by taking the limit as $n \to \infty$ and applying the Lebesgue monotone convergence theorem to show that $\mathbb{E}_\mu D_\infty = \lim_{n\to\infty} \Delta_n \leq -\ln w_\mu$. That d_t and h_t converge to 0 with μ-probability 1 follows from Markov's inequality applied to D_∞ and H_∞ respectively.

C Proof of Theorem 6

If $t \leq t'$ are stopping times, then $I(\omega) = [t(\omega), t'(\omega))$ is called a stopping interval and $X(I) := X(t)$ is the set of finite sequences when the start of I becomes known. If ρ is a measure, then $\rho(I) := \sum_{x \in X(I)} \rho(x)$ is the ρ-probability of encountering interval I at some point.

Lemma 16. *Let $\nu \in \mathcal{M}$ and I be a stopping interval. Then*

$$\mathbb{E}_\nu \sum_{t \in I} d_t(\nu\|\xi) \leq \sum_{x \in X(I)} \nu(x) \left(\ln \frac{1}{w_\nu} + \ln \frac{\xi(x)}{\nu(x)}\right).$$

Proof. The result follows from Theorem 1 and definitions. Let t be the stopping time governing the start of interval I. Then

$$\begin{aligned}\mathbb{E}_\nu \sum_{t\in I} d_t(\nu\|\xi) &\overset{\text{(a)}}{=} \sum_{x\in X(I)} \nu(x)\mathbb{E}_\nu \left[\sum_{t\in I} d_t(\nu\|\xi) \,\middle|\, x\right] \\ &\overset{\text{(b)}}{\le} \sum_{x\in X(I)} \nu(x)\mathbb{E}_\nu \left[\sum_{t=\ell(x)+1}^{\infty} d_t(\nu\|\xi) \,\middle|\, x\right] \\ &\overset{\text{(c)}}{\le} \sum_{x\in X(I)} \nu(x) \ln \frac{1}{w_\nu(x)} \overset{\text{(d)}}{=} \sum_{x\in X(I)} \nu(x) \left(\ln \frac{1}{w_\nu} + \ln \frac{\xi(x)}{\nu(x)}\right).\end{aligned}$$

(a) follows by by the definition of expectation. (b) by increasing the size of the interval. (c) follows from Theorem 1 by noting that $\xi(\cdot|x)$ is a mixture over $\{\nu(\cdot|x) : \nu \in \mathcal{M}\}$ with prior $w(\cdot|x)$. (d) because $w_\nu(x) = w_\nu\nu(x)/\xi(x)$ and by expanding the logarithm. ■

Proof of Theorem 6. First, the quantity to be bounded can be rewritten as an average of ν-expectations of a certain random variable.

$$\begin{aligned}\Delta &:= \mathbb{E}_\mu \sum_{t=1}^{\infty} c_t \overset{\text{(a)}}{=} \sum_{t=1}^{\infty} \mathbb{E}_\mu c_t \overset{\text{(b)}}{=} \sum_{t=1}^{\infty} \sum_{x\in X^{t-1}} \mu(x) \sum_{\nu\in\mathcal{M}} \frac{1}{K} \cdot \frac{\nu(x)}{\xi(x)} d_x(\nu\|\xi) \\ &\overset{\text{(c)}}{=} \frac{1}{K} \sum_{\nu\in\mathcal{M}} \sum_{t=1}^{\infty} \sum_{x\in X^{t-1}} \nu(x) \frac{\mu(x)}{\xi(x)} d_x(\nu\|\xi) \overset{\text{(d)}}{=} \frac{1}{K} \sum_{\nu\in\mathcal{M}} \sum_{t=1}^{\infty} \mathbb{E}_\nu \frac{\mu_{<t}}{\xi_{<t}} d_t(\nu\|\xi) \\ &\overset{\text{(e)}}{=} \frac{1}{K} \sum_{\nu\in\mathcal{M}} \underbrace{\mathbb{E}_\nu \sum_{t=1}^{\infty} \frac{\mu_{<t}}{\xi_{<t}} d_t(\nu\|\xi)}_{\Delta(\nu)}.\end{aligned}$$

(a) follows by the linearity of expectation and positivity of c_t. (b) by writing out the definition of the expectation. (c), (d) and (e) exchanging sums and the definition of expectation. Define $a_t, b_t : X^\infty \to \mathbb{N}$ by

$$a_t(\omega) := \sup_{t'\le t} \lfloor \ln \xi(\omega_{<t'})/\nu(\omega_{<t'}) \rfloor \qquad b_t(\omega) := \sup_{t'\le t} \lfloor \ln \mu(\omega_{<t'})/\xi(\omega_{<t'}) \rfloor,$$

which are monotone non-decreasing. By the definition of ξ as a uniform mixture over $\mathcal{M}$, $\mu(x)/\xi(x) \le K$, so $b_t(\omega) \le \ln K =: L$. Furthermore, $\mu(\epsilon) = \nu(\epsilon) = \xi(\epsilon) = 1$ implies that $a_t(\omega), b_t(\omega) \ge 0$. Define intervals of the following form

$$I_\beta(\omega) := \{t : b_t = \beta \wedge a_t \le \beta\} \qquad I_{\alpha,\beta}(\omega) := \{t : a_t = \alpha \wedge b_t = \beta\}.$$

Then $\mathbb{N}$ can be divided into disjoint intervals of the form I_β and $I_{\alpha,\beta}$ where $\alpha > \beta$.

$$\forall(\omega \in X^\infty), \quad \mathbb{N} = \bigcup_{\beta=0}^{L} \left(I_\beta(\omega) \cup \bigcup_{\alpha>\beta\in\mathbb{N}} I_{\alpha,\beta}(\omega) \right) \tag{3}$$

See Figure 3 for an example of the definition of I_β and $I_{\alpha,\beta}$. Then $\Delta(\nu)$ can be decomposed as

follows

$$\begin{aligned}\Delta(\nu) &\equiv \mathbb{E}_\nu \sum_{t=1}^{\infty} \frac{\mu_{<t}}{\xi_{<t}} d_t(\nu\|\xi) \\ &= \underbrace{\sum_{\beta=0}^{L} \mathbb{E}_\nu \sum_{t\in I_\beta} \frac{\mu_{<t}}{\xi_{<t}} d_t(\nu\|\xi)}_{\Delta_1(\nu)} + \underbrace{\sum_{\beta=0}^{L} \sum_{\alpha=\beta+1}^{\infty} \mathbb{E}_\nu \sum_{t\in I_{\alpha,\beta}} \frac{\mu_{<t}}{\xi_{<t}} d_t(\nu\|\xi)}_{\Delta_2(\nu)}\end{aligned}$$

where the second equality follows from Equation (3) and by linearity of the expectation. We now bound $\Delta_1(\nu)$ and $\Delta_2(\nu)$.

$$\begin{aligned}\Delta_1(\nu) &\equiv \sum_{\beta=0}^{L} \mathbb{E}_\nu \sum_{t\in I_\beta} \frac{\mu_{<t}}{\xi_{<t}} d_t(\nu\|\xi) \overset{\text{(a)}}{\leq} \sum_{\beta=0}^{L} e^{\beta+1} \mathbb{E}_\nu \sum_{t\in I_\beta} d_t(\nu\|\xi) \\ &\overset{\text{(b)}}{\leq} \sum_{\beta=0}^{L} e^{\beta+1} \sum_{x\in X(I_\beta)} \nu(x) \left(L + \ln \frac{\xi(x)}{\nu(x)}\right) \overset{\text{(c)}}{\leq} \sum_{\beta=0}^{L} e^{\beta+1} \nu(I_\beta) \left(L+\beta+1\right).\end{aligned}$$

(a) follows since on the interval I_β the quantity $\mu_{<t}/\xi_{<t} < e^{\beta+1}$. (b) follows from Lemma 16 and by noting that $\ln 1/w_\mu = \ln K = L$. (c) by the definition of $\nu(I_\beta)$ and because $\xi_{<t}/\nu_{<t} < e^{\beta+1}$ on the interval I_β. $\Delta_2(\nu)$ is bounded in a similar fashion.

$$\begin{aligned}\Delta_2(\nu) &\equiv \sum_{\beta=0}^{L} \sum_{\alpha=\beta+1}^{\infty} \mathbb{E}_\nu \sum_{t\in I_{\alpha,\beta}} \frac{\mu_{<t}}{\xi_{<t}} d_t(\nu\|\xi) \overset{\text{(a)}}{\leq} \sum_{\beta=0}^{L} e^{\beta+1} \sum_{\alpha=\beta+1}^{\infty} \mathbb{E}_\nu \sum_{t\in I_{\alpha,\beta}} d_t(\nu\|\xi) \\ &\overset{\text{(b)}}{\leq} \sum_{\beta=0}^{L} e^{\beta+1} \sum_{\alpha=\beta+1}^{\infty} \sum_{x\in X(I_{\alpha,\beta})} \nu(x) \left(L + \ln \frac{\xi(x)}{\nu(x)}\right) \\ &\overset{\text{(c)}}{\leq} \sum_{\beta=0}^{L} e^{\beta+1} \sum_{\alpha=\beta+1}^{\infty} \nu(I_{\alpha,\beta}) \left(L+\alpha+1\right) \overset{\text{(d)}}{\leq} \sum_{\beta=0}^{L} e^{\beta+1} \sum_{\alpha=\beta+1}^{\infty} e^{-\alpha} \left(L+\alpha+1\right) \\ &\overset{\text{(e)}}{=} \sum_{\beta=0}^{L} e^{\beta+1} e^{-\beta} \left(L+\beta+3\right) \overset{\text{(f)}}{=} 3(L+1)(L+2).\end{aligned}$$

(a) follows because $\mu_{<t}/\xi_{<t} < e^{\beta+1}$ on the interval $I_{\alpha,\beta}$ and by expanding the interval. (b) by Lemma 16. (c) because $\nu(I_{\alpha,\beta}) = \sum_{x\in X(I_{\alpha,\beta})} \nu(x)$. By definition, if $x \in X(I_{\alpha,\beta})$, then a $\xi(x)/\nu(x) \geq e^{\alpha}$. By Lemma 14 the ν-probability of this ever occurring is at most $e^{-\alpha}$, which implies $\nu(I_{\alpha,\beta}) \leq e^{-\alpha}$ and so gives (d). (e) and (f) follow from simple algebra. Combining the bounds of $\Delta_1(\nu)$ and $\Delta_2(\nu)$ leads to

$$\begin{aligned}\sum_{\nu\in\mathcal{M}} w_\nu \Delta(\nu) &\equiv \sum_{\nu\in\mathcal{M}} w_\nu \left(\Delta_1(\nu) + \Delta_2(\nu)\right) \\ &\overset{\text{(a)}}{\leq} 3(L+1)(L+2) + \sum_{\nu\in\mathcal{M}} w_\nu \sum_{\beta=0}^{L} e^{\beta+1} \nu(I_\beta)(L+\beta+1) \\ &\overset{\text{(b)}}{=} 3(L+1)(L+2) + \sum_{\beta=0}^{L} e^{\beta+1} \xi(I_\beta)(L+\beta+1) \\ &\overset{\text{(c)}}{\leq} 3(L+1)(L+2) + \sum_{\beta=0}^{L} 2(L+\beta+1) \overset{\text{(d)}}{=} 6L^2 + 14L + 8\end{aligned}$$

(a) by substituting the bounds for $\Delta_1(\nu)$ and $\Delta_2(\nu)$. (b) by exchanging sums and recalling that $\sum_{\nu\in\mathcal{M}} w_\nu \nu(A) = \xi(A)$ for all measurable A. (c) from Lemma 14 applied to bound $\xi(I_\beta) \le e^{-\beta}$ in the same way as $\nu(I_{\alpha,\beta})$ was bounded. (d) by simple algebra. The theorem is completed by substituting $L := \ln K$. ■

time, t	1	2	3	4	5	6	7	8
$\xi(\omega_{<t})/\nu(\omega_{<t})$	1	2	5	5	5	3	2	3
$\mu(\omega_{<t})/\xi(\omega_{<t})$	1	2	2	1	5	2	1	9
a_t	0	1	2	2	2	2	2	2
b_t	0	1	1	1	2	2	2	3
	I_0	I_1	$I_{2,1}$		I_2			I_3

Figure 3: Example $I_{\alpha,\beta}$ and I_β

D Proof of Proposition 7

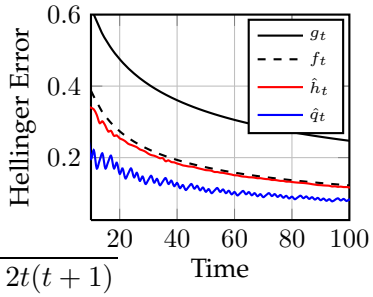
Let $X = \{0,1\}$ and define measure ν^k to be the deterministic measure producing k ones followed by zeros $\nu^k(1|x) := [\![\ell(x) < k]\!]$. Let $\mathcal{M} := \{\nu^k : 0 \le k \le K-1\}$ and the true measure be $\mu := \nu_{K-1}$. The Bayes mixture over $\mathcal{M}$ under the uniform prior becomes $\xi(x) := \frac{1}{K}\sum_{k=0}^{K-1} \nu^k(x)$. If $t < K$, then by substituting definitions one obtains $\xi(1^t) = (K-t)/K$ and $\xi(0|1^t) = 1/(K-t)$. Therefore

$$\mathbb{E}_\mu \sum_{t=1}^{\infty} c_t \overset{\text{(a)}}{\ge} \mathbb{E}_\mu \sum_{t=1}^{K} c_t \overset{\text{(b)}}{=} \sum_{t=0}^{K-1}\sum_{k=0}^{K-1} \frac{1}{K}\frac{\nu^k(1^t)}{\xi(1^t)} d_{1^t}\left(\nu^k \| \xi\right)$$
$$\overset{\text{(c)}}{\ge} \sum_{t=0}^{K-1} \frac{\nu^t(1^t)}{K\xi(1^t)} d_{1^t}\left(\nu^t\|\xi\right) \overset{\text{(d)}}{=} \sum_{t=1}^{K} \frac{\ln t}{t} \overset{\text{(e)}}{\ge} \frac{1}{2}\ln K - 1.$$

(a) follows by truncating the sum and positivity of c_t. (b) by the definition of c_t, the expectation and because $\mu(1^t) = 1$ for all $t \le K-1$. (c) by dropping all terms in the sum over k except for $k = t$ and positivity of all quantities. (d) and (e) follow by substituting definitions and simple calculus/algebra.

E Experiments

We set $\delta = 1/10$ and $\mathcal{M} = \{\nu_0, \cdots, \nu_{40}\}$ where ν_k is the Bernoulli measure with parameter $\theta_k := k/40$. We then sampled 20,000 sequences of length 100 from the Lebesgue measure $\mu = \nu_{20}$ and computed the average value of $\hat{h}_t$. For each t we computed the estimated quantile $\hat{q}_t$ as the value such that $h_t(\mu,\xi) < \hat{q}_t$ for 90% of the samples. We compare to



$$f_t := \sqrt{\frac{1}{2t}\ln\frac{2}{\delta}} \qquad g_t := \sqrt{\frac{1}{2t}\ln\frac{2t(t+1)}{\delta}}$$

which are obtained from the Hoeffding and union bounds and satisfy

$$\mu\left(\left|\hat{\theta}_t - 1/2\right| \le f_t\right) \ge 1-\delta \qquad \mu\left(\forall t : \left|\hat{\theta}_t - 1/2\right| \le g_t\right) \ge 1-\delta$$

where $\hat{\theta}_t$ is the empiric estimator of parameter θ. Some remarks:

- The frequentist estimator $\hat{\theta}_t(\omega_{<t}) \approx \xi(1|\omega_{<t})$ is very tight with high probability. Therefore comparing error between $\hat{\theta}_t$ and the true parameter $1/2$ is essentially the same as comparing $\xi(\cdot|\omega_{<t})$ and $\mu(\cdot|\omega_{<t})$.
- The comparison to f_t is not entirely fair to $\hat{h}_t$ for two reasons. First because $\hat{h}_t$ upper bounds h_t with high probability for all t while f_t does so only for each t and secondly because f_t was based on the total variation distance, which is smaller than the Hellinger distance.
- The comparison between $\hat{h}_t$ and g_t is not fair to g_t because the application of the union bound was rather weak.
- The comparison to the quantile is not entirely fair to $\hat{h}_t$, since $\hat{q}_t$ is computed for a single θ and individually for each t, while $\hat{h}_t$ must work for all models in $\mathcal{M}$ and all t.
- We also ran the experiment with 21 uniformly distributed environments with almost identical results showing that $\hat{h}_t$ is an excellent bound and strengthening our belief that at least in this simple setting the bound of Theorem 10 can be substantially improved in the i.i.d. case.
- The results indicate that $\hat{h}_t$ tracks close to f_t and $\hat{q}_t$, which essentially lower-bounds the optimum. We expect the definition of g_t can be improved to follow close to $\hat{h}_t$ and f_t without weakening the bound (holding for all t), but doubt that anything does much better than $\hat{h}_t$.

F Table of Notation

$\mathbb{N}$	natural numbers
$[\![expr]\!]$	indicator function of expression $expr$
$\ln$	natural logarithm
X	finite or countable alphabet
X^*	set of finite strings on X
X^∞	set of infinite strings on X
x, y	symbols or strings in X^*
$\ell(x)$	length of string x
ϵ	empty string of length 0
Γ_x	cylinder set of x, $\Gamma_x := \{x\omega : \omega \in X^\infty\}$
$\mathcal{F}_{<t}$	sigma algebra generated by cylinders on strings of length $t-1$
$\mathcal{F}$	sigma algebra generated by by cylinders on strings of all finite lengths
$\mathcal{M}$	environment class of hypothesis measures
μ	true measure from which sequences are sampled
ξ	bayes mixture over all $\nu \in \mathcal{M}$ with prior $w : \mathcal{M} \to (0,1]$
ν, ρ	measures in $\mathcal{M}$
ρ_t	random variable defined by $\rho_t(\omega) := \rho(\omega_t \vert \omega_{<t})$
$\rho_{<t}$	random variable defined by $\rho_{<t}(\omega) := \rho(\omega_{<t})$
$\mathbb{E}_\mu$	expectation w.r.t. μ
w	prior distribution $w : \mathcal{M} \to (0,1]$
w_ν	prior weight of measure $\nu \in \mathcal{M}$
$w_\nu(x)$	posterior of measure $\nu \in \mathcal{M}$ having observed x
Ent	shannons entropy function
$\mu \ll \xi$	μ is absolutely continuous w.r.t. ξ
$d_x(\mu\Vert\xi)$	KL divergence between predictive distributions of μ and ξ given finite sequence x
$h_x(\mu,\xi)$	squared Hellinger distance between predictive distributions of μ and ξ given finite sequence x
h_t	random variable $h_t(\omega) := h_{\omega_{<t}}(\mu,\xi)$
d_t	random variable $d_t(\omega) := d_{\omega_{<t}}(\mu\Vert\xi)$
c_t	random variable $c_t(\omega) := \sum_{\nu\in\mathcal{M}} w_\nu(\omega_{<t}) d_{\omega_{<t}}(\nu\Vert\xi)$
H_t, D_t, C_t	$\sum_{t=1}^\infty h_t$ and $\sum_{t=1}^\infty d_t$ and $\sum_{t=1}^\infty c_t$ respectively
K	number of models in $\mathcal{M}$
$\hat{h}_t$	upper confidence bound on h_t
$\hat{d}_t$	upper confidence bound on d_t
ε, δ	small positive reals with ε typically an accuracy parameter and δ a confidence (probability)