Convergence Guarantees for Bayesian Partial Models

Tyra Burgess
MATS Program

Michael K. Cohen
University of California, Berkeley

1 Bayesian Sequence Prediction

Suppose a sequence unfolds z_1, z_2, and so on, for $z_t \in \mathcal{Z}$, and we would like to predict the next element given the past. We could start with a set of possible theories τ, where each τ determines a function for computing the conditional probabilities $\mathbb{P}_\tau(z_t \mid z_{<t})$. We could then start with prior weights $\mathbb{P}_w(\tau)$ for each τ, and then update those weights with Bayes' rule:

$$\mathbb{P}_w(\tau \mid z_{<t}z_t) = \mathbb{P}_w(\tau \mid z_{<t})\mathbb{P}_\tau \frac{z_t \mid z_{<t}}{\sum_{\tau'} \mathbb{P}_w(\tau' \mid z_{<t})\mathbb{P}_{\tau'}(z_t \mid z_{<t})}$$

And then one can prove that if the sequence $z_{<\infty}$ in fact unfolds according to a true theory τ^*, and $\mathbb{P}_w(\tau^*) > 0$, then with probability 1, for any $z \in \mathcal{Z}$,

$$\lim_{t\to\infty} \sum_{\tau} \mathbb{P}_w(\tau \mid z_{<t})\mathbb{P}_\tau(z \mid z_{<t}) - \mathbb{P}_{\tau^*}(z \mid z_{<t}) = 0$$

(One can also show the rate of convergence: the sum of squares of these errors is finite).

2 Partial Theories

In the standard formalism above, we considered a set of theories τ which always give a probability distribution over what data would be observed next $\mathbb{P}_\tau(z_t \mid z_{<t})$. Now we consider the case where we have a set of partial theories $\pi \in \mathbb{M}$, which are like theories, except that they only sometimes give a probability distribution over what comes next; otherwise, they are inactive, outputting $\emptyset$. Real theories are like this, making predictions only about certain parts of the world. How should we update our credence in partial theories and use them for prediction? A natural proposal is as follows.

We should start with relative credences in the partial theories in the absence of any data. Let $\mathbb{P}_w$ be a prior/posterior weight function, where $\sum_{\pi\in\mathbb{M}} \mathbb{P}_w(\pi) = 1$. Let $\mathbb{M}_{\{z<t\}}$ be the subset of partial theories which are active in the context $z_{<t}$. We let the prior $\mathbb{P}_w(\pi) > 0$ be defined arbitrarily. And we define $\mathbb{P}_w(\pi \mid z_{<t}z_t)$ inductively as follows:

$$\mathbb{P}_w(\pi \mid z_{<t}z_t) = \begin{cases} \mathbb{P}_w(\pi \mid z_{<t}) & \text{if } \mathbb{P}_\pi(z_t \mid z_{<t}) = \emptyset \\ \mathbb{P}_w(\pi \mid z_{<t})\mathbb{P}_\pi(z_t \mid z_{<t}) \frac{\sum_{\pi'\in\mathbb{M}_{z_{<t}}} \mathbb{P}_w(\pi' \mid z_{<t})}{\sum_{\pi'\in\mathbb{M}_{z_{<t}}} \mathbb{P}_w(\pi' \mid z_{<t})\mathbb{P}_{\pi'}(z_t \mid z_{<t})} & \text{otherwise} \end{cases}$$

So when we condition on an ordered dataset $z_{<t}z_t$, if a particular theory doesn't give a prediction about a particular data point, its posterior weight is not updated. If it does, then its weight is updated proportionally to the probability it assigned to that data point. But the total weight on partial theories that make a prediction doesn't change. Otherwise, they would be penalised for making predictions in hard-to-predict contexts.

To make a prediction using an accuracy-weighted mixture over these partial theories, we take an essentially Bayesian approach:

$$\mathbb{P}_w(z_t \mid z_{<t}) = \sum_{\pi\in\mathbb{M}_{z_{<t}}} \frac{\mathbb{P}_w(\pi \mid z_{<t})}{\sum_{\pi'\in\mathbb{M}_{z_{<t}}} \mathbb{P}_w(\pi' \mid z_{<t})} \mathbb{P}_\pi(z_t \mid z_{<t})$$

This just mixes over all the partial theories that apply in the context in question, proportional to their posterior weight. For this to make sense, we have to assume that $\mathbb{M}_{z_{<t}}$ is always nonempty (though this is a fairly trivial assumption that can be satisfied by e.g. including the uniform predictor in the model class.).

As a caveat, while the posterior weights always sum to 1, the posterior weights should not be considered to have the semantic meaning of "the probability that this partial theory is true". Many partial theories could be true and non-identical to each other, provided that they make predictions in different contexts.

If we can show that those predictions are equivalent to the predictions of a Bayesian predictor with a different set of (full) theories, then we can more easily apply theorems relating to the standard Bayesian setting. It wouldn't make sense to actually construct such a Bayesian predictor, since it would behave in the same way, but less efficiently.

For any partial theory π, we define a full theory τ as follows:

$$\mathbb{P}_{\tau_\pi}(z_t \mid z_{<t}) = \begin{cases} \mathbb{P}_w(z_t \mid z_{<t}) \text{ if } \mathbb{P}_\pi(z_t \mid z_{<t}) = \emptyset \\ \mathbb{P}_\pi(z_t \mid z_{<t}) \text{ otherwise} \end{cases}$$

Proposition 1: *Baysian predictions using the model class* $\mathcal{M}_\tau = \{\tau_\pi : \pi \in \mathcal{M}\}$ *and the prior* $\mathbb{P}_{w(\tau_\pi)} = \mathbb{P}_{w(\pi)}$ *are identical to predictions calculated according to the scheme described above.*

Before getting into the proof, a brief intuitive explanation of why this should be true. Firstly,

This involves both proving that $\mathbb{P}_w^{\mathcal{M}}(z_{t+1} \mid z_{\leq t}) = \mathbb{P}_w^{\mathcal{M}_\tau}(z_{t+1} \mid z_{\leq t})$ for all $z_{\leq t+1}$ and that $\mathbb{P}_w(\tau_\pi \mid z_{\leq t}) = \mathbb{P}_w(\pi \mid z_{\leq t})$, for all π and all $z_{\leq t}$.

Before getting into the proof, I will state some definitions for convenience. Define $\mathcal{A} = \{\tau_\pi : \pi \in \mathcal{M}, \mathbb{P}_\pi(z_{t+1} \mid z_{\leq t}) = \emptyset\}$, i.e. the set of all total models corresponding to the partial models that return $\emptyset$. Let $\mathcal{B} = \mathcal{M}_\tau \setminus \mathcal{A}$, and $\mathcal{C} = \{\pi : \tau_\pi \in \mathcal{B}\}$, i.e. the set of partial models that don't return $\emptyset$. The specific models that these sets refer to will depend on $z_{\leq t+1}$, which will be variable on t throughout the proof, I'm simply skipping the $\mathcal{A}_{\leq t+1}$ style of notation for conciseness.

Since the definitions of these two are intertwined, I will need to prove both together by induction on t. Firstly, proving for the case of $t = 0$; The second property for the case of $t = 0$ is simply $\mathbb{P}_w(\tau_\pi) = \mathbb{P}_w(\pi)$, which is true by the construction of the prior on $\mathcal{M}_\tau$. For the second property, it can be seen as follows:

$$
\begin{aligned}
\mathbb{P}_w^{\mathcal{M}_\tau}(z_1) &= \sum_{\tau_\pi \in \mathcal{M}_\tau} P_w(\tau_\pi) P_{\tau_\pi}(z_1) \\
&\underset{(1)}{=} \sum_{\tau_\pi \in \mathcal{M}_\tau} P_w(\pi) P_{\tau_\pi}(z_1) \\
&\underset{(2)}{=} \sum_{\tau_\pi \in \mathcal{A}} P_w(\pi) P_{\tau_\pi}(z_1) + \sum_{\tau_\pi \in \mathcal{B}} P_w(\pi) P_{\tau_\pi}(z_1) \\
&\underset{(3)}{=} \sum_{\tau_\pi \in \mathcal{A}} P_w(\pi) P_w^{\mathcal{M}}(z_1) + \sum_{\tau_\pi \in \mathcal{B}} P_w(\pi) P_\pi(z_1) \\
&\underset{(4)}{=} P_w^{\mathcal{M}}(z_1) \left(1 - \sum_{\tau_\pi \in \mathcal{B}} P_w(\pi)\right) + \sum_{\tau_\pi \in \mathcal{B}} P_w(\pi) P_\pi(z_1) \\
&\underset{(5)}{=} \frac{\sum_{\pi \in \mathcal{C}} P_w(\pi) P_\pi(z_1)}{\sum_{\pi \in \mathcal{C}} P_w(\pi)} \left(1 - \sum_{\pi \in \mathcal{C}} P_w(\pi)\right) + \sum_{\pi \in \mathcal{C}} P_w(\pi) P_\pi(z_1) \\
&\underset{(6)}{=} \frac{\sum_{\pi \in \mathcal{C}} P_w(\pi) P_\pi(z_1)}{\sum_{\pi \in \mathcal{C}} P_w(\pi)} - \sum_{\pi \in \mathcal{C}} P_w(\pi) P_\pi(z_1)) + \sum_{\pi \in \mathcal{C}} P_w(\pi) P_\pi(z_1) \\
&\underset{(7)}{=} \frac{\sum_{\pi \in \mathcal{C}} P_w(\pi) P_\pi(z_1)}{\sum_{\pi \in \mathcal{C}} P_w(\pi)} \\
&\underset{(8)}{=} \mathbb{P}_w^{\mathcal{M}}(z_1)
\end{aligned}
$$

(1) is by the definition of the prior on $\mathcal{M}_\tau$. (2) is by the fact that $\mathcal{A} \cup \mathcal{B} = \mathcal{M}_\tau$. (3) is by the piecewise definition of $P_\tau(z_t \mid z_{<t})$, and the fact that $\mathcal{A}$ and $\mathcal{B}$ split $\mathcal{M}_\tau$ by the cases of that piecewise definition. (4) is pulling out from the $\mathcal{A}$ sum the term that doesn't depend on τ_π, and substituting $P_w(\mathcal{A})$ for $(1 - P_w(\mathcal{B}))$ since they partition $\mathcal{M}_\tau$. (5) replaces $P_w^{\mathcal{M}}(z_1)$ with its definition, and replaces the sums over $\mathcal{B}$ with sums over $\mathcal{C}$ since at this point I was only using the τ_π models to refer to the π models backing them, which is exactly the difference between $\mathcal{B}$ and $\mathcal{C}$. The remainder is simple algebra.

Thus, the claim is proven for $t = 0$. Now, I prove that it is true for any $t > 0$, given that its true for $t - 1$. That is, we assume that $\mathbb{P}_w^{\mathcal{M}}(z_t \mid z_{<t}) = \mathbb{P}_w^{\mathcal{M}_\tau}(z_t \mid z_{<t})$, and that $\mathbb{P}_w(\tau_\pi \mid z_{<t}) = \mathbb{P}_w(\pi \mid z_{<t})$.

I will firstly show that this implies that $\mathbb{P}_w(\tau_\pi \mid z_{\leq t}) = \mathbb{P}_w(\pi \mid z_{\leq t})$.

There are two cases to consider: if $\mathbb{P}_\pi(z_t \mid z_{<t}) = \emptyset$, and if it is not equal to $\emptyset$.

if $\mathbb{P}_\pi(z_t \mid z_{<t}) = \emptyset$, then

$$
\begin{aligned}
\mathbb{P}_w(\tau_\pi \mid z_{\leq t}) &\underset{(1)}{=} \frac{P_w(\tau_\pi \mid z_{<t}) P_{\tau_\pi}(z_t \mid z_{<t})}{\sum_{\tau \in \mathcal{M}_\tau} P_w(\tau \mid z_{<t}) P_\tau(z_t \mid z_{<t})} \\
&\underset{(2)}{=} \frac{P_w(\tau_\pi \mid z_{<t}) P_{\tau_\pi}(z_t \mid z_{<t})}{P_w^{\mathcal{M}_\tau}(z_t \mid z_{<t})} \\
&\underset{(3)}{=} \frac{P_w(\pi \mid z_{<t}) P_w^{\mathcal{M}}(z_t \mid z_{<t})}{P_w^{\mathcal{M}}(z_t \mid z_{<t})} \\
&\underset{(4)}{=} P_w(\pi \mid z_{<t}) \\
&\underset{(5)}{=} P_w(\pi \mid z_{\leq t})
\end{aligned}
$$

(1) is by Bayes rule, while (2) is by the definition of $P_w^{\mathcal{M}_\tau}$. (3) makes use of the induction assumptions, and the definition of $\mathbb{P}_\tau$ when $\mathbb{P}_\pi$ returns $\emptyset$. (5) is by the definition of $\mathbb{P}_w$ when $\mathbb{P}_\pi$ returns $\emptyset$.

Alternatively, $\mathbb{P}_\pi(z_t \mid z_{<t}) \neq \emptyset$.

$$\begin{aligned}
\mathbb{P}_w(\tau_\pi \mid z_{\leq t}) &\underset{(1)}{=} \frac{\mathbb{P}_w(\tau_\pi \mid z_{<t})\mathbb{P}_{\tau_\pi}(z_t \mid z_{<t})}{\sum_{\tau \in \mathcal{M}_\tau} \mathbb{P}_w(\tau \mid z_{<t})\mathbb{P}_\tau(z_t \mid z_{<t})} \\
&\underset{(2)}{=} \frac{\mathbb{P}_w(\tau_\pi \mid z_{<t})\mathbb{P}_{\tau_\pi}(z_t \mid z_{<t})}{\mathbb{P}_w^{\mathcal{M}_\tau}(z_t \mid z_{<t})} \\
&\underset{(3)}{=} \frac{\mathbb{P}_w(\pi \mid z_{<t})\mathbb{P}_\pi(z_t \mid z_{<t})}{\mathbb{P}_w^{\mathcal{M}}(z_t \mid z_{<t})} \\
&\underset{(4)}{=} \mathbb{P}_w(\pi \mid z_{\leq t})
\end{aligned}$$

(1) is by definition of $\mathbb{P}_w$ for τ_π. (2) is by the definition of $\mathbb{P}_w^{\mathcal{M}_\tau}$. (3) is by the definition of $\mathbb{P}_{\tau_\pi}$ when $\mathbb{P}_\pi$ doesn't return $\emptyset$, and by the induction assumptions. (4) is by the definition of $\mathbb{P}_w(\pi \mid z_{\leq t})$.

Next, I will show that our assumptions imply that $\mathbb{P}_w^{\mathcal{M}}(z_{t+1} \mid z_{\leq t}) = \mathbb{P}_w^{\mathcal{M}_\tau}(z_{t+1} \mid z_{\leq t})$.

$$\begin{aligned}
\mathbb{P}_w^{\mathcal{M}_\tau}(z_{t+1} \mid z_{\leq t}) &\underset{(1)}{=} \sum_{\tau_\pi \in \mathcal{M}_\tau} \mathbb{P}_w(\tau_\pi \mid z_{\leq t})\mathbb{P}_{\tau_\pi}(z_{t+1} \mid z_{\leq t}) \\
&\underset{(2)}{=} \sum_{\tau_\pi \in \mathcal{A}} \mathbb{P}_w(\tau_\pi \mid z_{\leq t})\mathbb{P}_{\tau_\pi}(z_{t+1} \mid z_{\leq t}) + \sum_{\tau_\pi \in \mathcal{B}} \mathbb{P}_w(\tau_\pi \mid z_{\leq t})\mathbb{P}_{\tau_\pi}(z_{t+1} \mid z_{\leq t}) \\
&\underset{(3)}{=} \sum_{\tau_\pi \in \mathcal{A}} \mathbb{P}_w(\pi \mid z_{\leq t})\mathbb{P}_w^{\mathcal{M}}(z_{t+1} \mid z_{\leq t}) + \sum_{\tau_\pi \in \mathcal{B}} \mathbb{P}_w(\pi \mid z_{\leq t})\mathbb{P}_\pi(z_{t+1} \mid z_{\leq t}) \\
&\underset{(4)}{=} \mathbb{P}_w^{\mathcal{M}}(z_{t+1} \mid z_{\leq t})\left(1 - \sum_{\tau_\pi \in \mathcal{B}} \mathbb{P}_w(\pi \mid z_{\leq t})\right) + \sum_{\tau_\pi \in \mathcal{B}} \mathbb{P}_w(\pi \mid z_{\leq t})\mathbb{P}_\pi(z_{t+1} \mid z_{\leq t}) \\
&\underset{(5)}{=} \frac{\sum_{\pi \in \mathcal{C}} \mathbb{P}_w(\pi \mid z_{\leq t})\mathbb{P}_\pi(z_{t+1} \mid z_{\leq t})}{\sum_{\pi \in \mathcal{C}} \mathbb{P}_w(\pi \mid z_{\leq t})}\left(1 - \sum_{\pi \in \mathcal{C}} \mathbb{P}_w(\pi \mid z_{\leq t})\right) + \sum_{\pi \in \mathcal{C}} \mathbb{P}_w(\pi \mid z_{\leq t})\mathbb{P}_\pi(z_{t+1} \mid z_{\leq t}) \\
&\underset{(6)}{=} \frac{\sum_{\pi \in \mathcal{C}} \mathbb{P}_w(\pi \mid z_{\leq t})\mathbb{P}_\pi(z_{t+1} \mid z_{\leq t})}{\sum_{\pi \in \mathcal{C}} \mathbb{P}_w(\pi \mid z_{\leq t})} - \sum_{\pi \in \mathcal{C}} \mathbb{P}_w(\pi \mid z_{\leq t})\mathbb{P}_\pi(z_{t+1} \mid z_{\leq t}) + \sum_{\pi \in \mathcal{C}} \mathbb{P}_w(\pi \mid z_{\leq t})\mathbb{P}_\pi(z_{t+1} \mid z_{\leq t}) \\
&\underset{(7)}{=} \frac{\sum_{\pi \in \mathcal{C}} \mathbb{P}_w(\pi \mid z_{\leq t})\mathbb{P}_\pi(z_{t+1} \mid z_{\leq t})}{\sum_{\pi \in \mathcal{C}} \mathbb{P}_w(\pi \mid z_{\leq t})} \\
&\underset{(8)}{=} \mathbb{P}_w^{\mathcal{M}}(z_{t+1} \mid z_{\leq t})
\end{aligned}$$

(1) is the definition of $\mathbb{P}_w^{\mathcal{M}_\tau}$. (2) is splitting the sum across the $\mathcal{A}, \mathcal{B}$ partition. (3) is applying definitions and the induction assumptions. (4) is pulling out the term that doesn't depend on τ_π from the sum, and using $\mathbb{P}_w(\mathcal{A} \mid z_{\leq t}) = 1 - \mathbb{P}_{w(\mathcal{B} \mid z_{\leq t})}$. (5) is replacing $\mathbb{P}_w^{\mathcal{M}}$ with its definition, and replacing the enumeration over $\mathcal{B}$ with enumeration over $\mathcal{C}$. (6) and (7) are simple algebra, and (8) is by definition of $\mathbb{P}_w^{\mathcal{M}}$.

Thus, given Baysian predictions are identical to the partial theories predictions for π and any t, it follows that they are identical for that π and $t+1$, for any z_{t+1}. For $t = 0$ they give identical predictions, and therefore for all $z_{\leq t+1}$, the two schemes give identical predictions for all models.

3 Research Question

Typically, Bayesian error bounds depend on $\mathbb{P}_w(\tau^*) > 0$. where τ^* is the true sampling distribution of the observed sequence. But in the partial model scenario, we don't want to assume that any τ_π will match the true sampling distribution. I show here that if our countable set of martial models $\mathcal{M}$ contains a model π^* with $\mathbb{P}_w(\pi^*) > 0$ and whenever π^* is active it matches the true sampling distribution, then the total expected error over timesteps where π^* is active is finite, and specifically bounded above by $\ln\left(\mathbb{P}_w(\pi^*)^{-1}\right)$. The particular inequality that I'm proving is

Theorem 1:

$$\forall z \in \mathcal{Z} : \mathbb{E}_{\tau^*}\left[\sum_{t=1}^{N} \mathbb{1}(\pi^*, z_{<t})(\mathbb{P}_{\tau^*}(z \mid z_{<t}) - \mathbb{P}_w(z \mid z_{<t}))^2\right] \leq \ln\left(\mathbb{P}_w(\pi^*)^{-1}\right)$$

Some notational explanations. $\mathbb{1}(\pi^*, z_{<t})$ is an indicator variable that is 1 if π^* is active conditioned on $z_{<t}$, and 0 otherwise. Since it will be convenient later I will also define $\mathbb{1}(\pi^*, z_{<t}, a, b)$ as returning a if π^* is active and b otherwise.

It will also be worth noting that while the definition given for $\mathbb{P}_w(\pi \mid z_{<t})$ is the piecewise recursive function below,

$$\mathbb{P}_w(\pi \mid z_{\leq t}) = \begin{cases} \mathbb{P}_w(\pi \mid z_{<t}) & \text{if } \mathbb{P}_\pi(z_t \mid z_{<t}) = \emptyset \\ \frac{\mathbb{P}_w(\pi \mid z_{<t})\mathbb{P}_\pi(z_t \mid z_{<t})}{\mathbb{P}_w(z_t \mid z_{<t})} & \text{else} \end{cases}$$

I will make use of an equivalent statement during the proof that

$$\mathbb{P}_w(\pi \mid z_{\leq t}) = \mathbb{P}_w(\pi \mid z_{<t})\mathbb{U}(\pi, z_{\leq t})$$

$$\mathbb{U}(\pi, z_{\leq t}) = \begin{cases} 1 & \text{if } \mathbb{P}_\pi(z_t \mid z_{<t}) = \emptyset \\ \frac{\mathbb{P}_\pi(z_t \mid z_{<t})}{\mathbb{P}_w(z_t \mid z_{<t})} & \text{else} \end{cases}$$

Lemma 1:

$$\prod_{t=1}^{N} \mathbb{1}\left(\pi^*, z_{<t}, \frac{\mathbb{P}_{\tau^*}(z_t \mid z_{<t})}{\mathbb{P}_w(z_t \mid z_{<t})}, 1\right) \leq \frac{1}{\mathbb{P}_w(\pi^*)}$$

I.e. the product of all the $\frac{\mathbb{P}_{\tau^*}(z_t \mid z_{<t})}{\mathbb{P}_w(z_t \mid z_{<t})}$ on timesteps where π^* is active is bounded above by $\frac{1}{\mathbb{P}_w(\pi^*)}$. This will be the key part of the proof, which allows us to bound the error by something which doesn't depend on $z_{\leq t}$, and is sufficiently complex to be worth showing separately. My working to prove this lemma is as follows:

$$\begin{aligned}
\mathbb{P}_w(\pi^* \mid z_{\leq N}) &\underset{(1)}{=} \mathbb{P}_w(\pi^* \mid z_{1:N-1})\mathbb{U}(\pi, z_{\leq N}) \\
&\underset{(2)}{=} \mathbb{P}_w(\pi^*) \prod_{t=1}^{N} \mathbb{U}(\pi, z_{\leq t}) \\
&\underset{(3)}{=} \mathbb{P}_w(\pi^*) \prod_{t=1}^{N} \mathbb{1}\left(\pi, z_{<t}, \frac{\mathbb{P}_\pi(z_t \mid z_{<t})}{\mathbb{P}_w(z_t \mid z_{<t})}, 1\right) \\
\frac{\mathbb{P}_w(\pi^* \mid z_{\leq N})}{\mathbb{P}_w(\pi^*)} &\underset{(4)}{=} \prod_{t=1}^{N} \mathbb{1}\left(\pi, z_{<t}, \frac{\mathbb{P}_\pi(z_t \mid z_{<t})}{\mathbb{P}_w(z_t \mid z_{<t})}, 1\right) \\
\frac{1}{\mathbb{P}_w(\pi^*)} &\underset{(5)}{\geq} \prod_{t=1}^{N} \mathbb{1}\left(\pi, z_{<t}, \frac{\mathbb{P}_\pi(z_t \mid z_{<t})}{\mathbb{P}_w(z_t \mid z_{<t})}, 1\right)
\end{aligned}$$

(1) is by the alternate statement of $\mathbb{P}_w(\pi^* \mid z_{\leq t})$ given above, and (2) is by repeatedly applying that same expansion until the recursive definition is fully unwrapped to its base case. (3) is by expressing the definition of $\mathbb{U}(\pi, z_{\leq t})$ in the $\mathbb{1}$ notation, (3) to (4) is by simple algebra, and (4) to (5) is by the fact that $\mathbb{P}_w(\pi^* \mid z_{\leq N}) \leq 1$, and this is equivalent to the lemma since by assumption $\mathbb{P}_\pi(z_t \mid z_{<t}) = \mathbb{P}_{\tau^*}(z_t \mid z_{<t})$ if $\mathbb{P}_\pi(z_t \mid z_{<t}) \neq \emptyset$.

And with that out of the way, onto the actual proof!

$$\begin{aligned}
\mathbb{E}_{\tau^*}\left[\sum_{t=1}^{N} \mathbb{1}(\pi^*, z_{<t})(\mathbb{P}_{\tau^*}(z \mid z_{<t}) - \mathbb{P}_w(z \mid z_{<t}))^2\right] &\underset{(1)}{\leq} \mathbb{E}_{\tau^*}\left[\sum_{t=1}^{N} \mathbb{1}(\pi^*, z_{<t}) \sum_{z_t \in \mathcal{Z}} \mathbb{P}_{\tau^*}(z_t \mid z_{<t}) \ln\left(\frac{\mathbb{P}_{\tau^*}(z_t \mid z_{<t})}{\mathbb{P}_w(z_t \mid z_{<t})}\right)\right] \\
&\underset{(2)}{=} \sum_{t=1}^{N} \sum_{z_{<t} \in \mathcal{Z}^{t-1}} \sum_{z_t \in \mathcal{Z}} \mathbb{P}_{\tau^*}(z_{<t}) \mathbb{P}_{\tau^*}(z_t \mid z_{<t}) \mathbb{1}(\pi^*, z_{<t}) \ln\left(\frac{\mathbb{P}_{\tau^*}(z_t \mid z_{<t})}{\mathbb{P}_w(z_t \mid z_{<t})}\right) \\
&\underset{(3)}{=} \sum_{t=1}^{N} \sum_{z_{\leq t} \in \mathcal{Z}^t} \mathbb{P}_{\tau^*}(z_{\leq t}) \mathbb{1}(\pi^*, z_{<t}) \ln\left(\frac{\mathbb{P}_{\tau^*}(z_t \mid z_{<t})}{\mathbb{P}_w(z_t \mid z_{<t})}\right) \\
&\underset{(4)}{=} \sum_{z_{\leq N} \in \mathcal{Z}^N} \mathbb{P}_{\tau^*}(z_{\leq N}) \sum_{t=1}^{N} \mathbb{1}(\pi^*, z_{<t}) \ln\left(\frac{\mathbb{P}_{\tau^*}(z_t \mid z_{<t})}{\mathbb{P}_w(z_t \mid z_{<t})}\right) \\
&\underset{(5)}{=} \sum_{z_{\leq N} \in \mathcal{Z}^N} \mathbb{P}_{\tau^*}(z_{\leq N}) \ln\left(\prod_{t=1}^{N} \mathbb{1}\left(\pi^*, z_{<t}, \frac{\mathbb{P}_{\tau^*}(z_t \mid z_{<t})}{\mathbb{P}_w(z_t \mid z_{<t})}, 1\right)\right) \\
&\underset{(6)}{\leq} \sum_{z_{\leq N} \in \mathcal{Z}^N} \mathbb{P}_{\tau^*}(z_{\leq N}) \ln\left(\frac{1}{\mathbb{P}_w(\pi^*)}\right) \\
&\underset{(7)}{=} \ln\left(\frac{1}{\mathbb{P}_w(\pi^*)}\right)
\end{aligned}$$

(1) holds by Pinsker's Inequality. (2) through (4) hold by reordering the sums and by simple properties of $\mathbb{P}_{\tau^*}$. (5) holds by bringing the sum up to N inside the logarithm, while making sure that the $\mathbb{1}$ function still results in terms where $\mathbb{P}_\pi(z_t \mid z_{<t}) = \emptyset$ not affecting the product. (6) is by lemma 1, and (7) is by pulling out the constant and noting that $\mathbb{P}_{\tau^*}$ sums to 1 over all length N strings.

Theorem 2: From this, a trivial additional result is that if there is some finite set $\{\pi_1^*, ..., \pi_K^*\}$ which have positive initial weight, agree with τ^* when active, and for all $z_{<t}$ at least one is active, then we can put a bound on the expected prediction error over all timesteps.

$$\begin{aligned}
\mathbb{E}_{\tau^*}\left[\sum_{t=1}^{N} (\mathbb{P}_{\tau^*}(z \mid z_{<t}) - \mathbb{P}_w(z \mid z_{<t}))^2\right] &\underset{(1)}{\leq} \mathbb{E}_{\tau^*}\left[\sum_{t=1}^{N} \sum_{k=1}^{K} \mathbb{1}(\pi_k^*, z_{<t})(\mathbb{P}_{\tau^*}(z \mid z_{<t}) - \mathbb{P}_w(z \mid z_{<t}))^2\right] \\
&\underset{(2)}{=} \sum_{k=1}^{K} \mathbb{E}_{\tau^*}\left[\sum_{t=1}^{N} \mathbb{1}(\pi_k^*, z_{<t})(\mathbb{P}_{\tau^*}(z \mid z_{<t}) - \mathbb{P}_w(z \mid z_{<t}))^2\right] \\
&\underset{(3)}{\leq} \sum_{k=1}^{K} \ln\left(\frac{1}{\mathbb{P}_w(\pi_k^*)}\right)
\end{aligned}$$

(1) holds by the fact that for every t at least one of the π_k^* will have $\mathbb{1}(\pi_k^*, z_{<t}) = 1$. (2) holds simply by rearranging the sum and by the linearity of expectation, and (3) holds by theorem 1.

Tightness of Theorem 2: Theorem 2 assumes that we have a finite set of partial theories which satisfy our accuracy requirements, and here we will show by counterexample why that set must be finite.

Counterexample: Take the problem of predicting whether a binary variable is 0 or 1. The true model will be $\mathbb{P}_{\tau^*}(0 \mid z_{1:t-1}) = 1$, i.e. the variable will always be 0.

Our model class for predictions will consist of $\{\pi_k \mid k \in \mathbb{Z}^+\} \cup \{\pi_k^* \mid k \in \mathbb{Z}^+\}$, where the π_k models predict $\mathbb{P}_{\pi_k}(0 \mid z_{1:k-1}) = 0$ and the π_k^* models predict $\mathbb{P}_{\tau^*}(0 \mid z_{1:k-1}) = 1$, with both types returning $\emptyset$ when $t \neq k$. Thus, at each time step t only two models will be active, one of which assigns 100% to the binary variable being 0 and the other which is the opposite.

P_w will then be constructed as follows:

$$P_w(\pi_k) = \frac{1}{2^{k+1}}, \qquad P_w(\pi_k^*) = \frac{1}{2^{k+1}} \qquad \forall k \in \mathbb{Z}^+$$

From this, we can see that we have $\{\pi_k^* \mid k \in \mathbb{Z}^+\}$, a set of partial theories each with positive prior weight, such that each one in the set never makes a prediction that differs from τ^*, and tha for every timestep at least one makes a prediction. This satisfies all requirements of the research question except for the set of accurate models being finite.

Now I prove that this is not sufficient to have a prediction error that converges to 0, and therefore that a finite set is required.

$$\begin{aligned}
\mathbb{P}_{\tau^*}(0 \mid z_{1:t-1}) - \mathbb{P}_w(0 \mid z_{1:t-1}) &\underset{(1)}{=} 1 - \frac{\sum_{\{\pi \in \mathcal{M}_{1:t-1}\}} \mathbb{P}_w(\pi \mid z_{1:t-1}) \mathbb{P}_\pi(0 \mid z_{1:t-1})}{\sum_{\{\pi \in \mathcal{M}_{1:t-1}\}} \mathbb{P}_w(\pi \mid z_{1:t-1})} \\
&\underset{(2)}{=} 1 - \frac{\mathbb{P}_w(\pi_t \mid z_{1:t-1}) \mathbb{P}_{\pi_t}(0 \mid z_{1:t-1}) + \mathbb{P}_w(\pi_t^* \mid z_{1:t-1}) \mathbb{P}_{\pi_t^*}(0 \mid z_{1:t-1})}{\mathbb{P}_w(\pi_t \mid z_{1:t-1}) + \mathbb{P}_w(\pi_t^* \mid z_{1:t-1})} \\
&\underset{(3)}{=} 1 - \frac{\frac{1}{2^{t+1}} * 0 + \frac{1}{2^{t+1}} * 1}{\frac{1}{2^t}} \\
&\underset{(4)}{=} \frac{1}{2}
\end{aligned}$$

(1) is by definitions, while 2 is by the construction of the models in $\mathcal{M}$. (3) follows by the fact that $\mathbb{P}_w(\pi_t \mid z_{1:t-1}) = \mathbb{P}_w(\pi_t)$ since before timestep t model π_t has only returned $\emptyset$ and thus its prior wasn't updated. The same logic holds for π_t^*.

A similar calculation results for $z_t = 1$, showing that the prediction error of this mixture model doesn't converge to 0 as $t \to \infty$ even though it contains a set of partial theories that satisfy all of the requiremints of the research question except for being finite.

Further, this can be simply extended to show that even all models in the accurate model set triggering infinitely often isn't enough to show prediction error that converges to 0, if the set is infinite. Take the same setup as described above, except on every odd timestep introduce a new pair of models as described above, while on every odd timestep have every model that has triggered at least once trigger. Thus, every accurate model is triggering infinitely often, but there are still infinitely many timesteps on which the prediction error is $\frac{1}{2}$.