

---

# Tracking the Best Predicting Model

---

Steven de Rooij\*

## Abstract

According to standard MDL and Bayesian model selection, we should (roughly) prefer the model that minimises overall prediction error. But if the goal is to predict well, it may well depend on the sample size which model is most useful to predict the next outcome. By re-interpreting the Bayesian prediction strategies associated with the models as “experts”, we can use the various algorithms for “expert tracking” to improve model selection for prediction without introducing a substantial computational overhead.

## 1 Model Selection Preliminaries

A *model*  $\mathcal{M} = \{P_\theta | \theta \in \Theta\}$  is a set of probability distributions. *Model selection* is choosing the “most useful” model based on the available observations  $x^n := x_1, \dots, x_n \in \mathcal{X}^n$ . For simplicity, we consider only model selection criteria that satisfy Dawids *weak prequential principle* [1, 2]. That is, models are considered “useful” if we can use them to construct *prediction strategies* that give high probability to the data, or, equivalently, achieve low accumulated prediction error, where prediction error is measured using logarithmic loss. More discussion about how our results relate to model selection for other applications, such as truth finding, can be found in [4]. To further simplify the presentation, we assume that  $\mathcal{X}$  is countable, we identify probability distributions with their defining mass functions, and we treat the sample size  $n$  as a given rather than considering random processes.

As the most important special case, we consider Bayes factor model selection, where prior distributions  $w_1, \dots, w_k$  are defined on the parameter spaces  $\Theta_1, \dots, \Theta_K$  of each of the models. By “integrating out” the parameter we obtain, for each model  $\mathcal{M}_k$ , an associated marginal distribution:

$$P_k(x^n) := \int_{\theta \in \Theta_k} P_\theta(x^n) w_k(\theta) d\theta. \quad (1)$$

By subsequently defining a prior distribution  $W$  on the models, we can then use Bayes’ rule to compute the posterior odds

$$\frac{P(\mathcal{M}_i | x^n)}{P(\mathcal{M}_j | x^n)} = \frac{W(i)}{W(j)} \cdot \frac{P_i(x^n)}{P_j(x^n)},$$

in other words, the posterior odds are the prior odds multiplied by the probability ratio of the data, which is called the “Bayes factor”.

We now take a step back and use the chain rule for conditional probability to rewrite (1) as

$$P_k(x^n) = P_k(x_1) \cdot P_k(x_2 | x^1) \cdot \dots \cdot P_k(x_n | x^{n-1}),$$

to obtain a prediction strategy. Thus, Bayes factor model selection satisfies the weak prequential principle, and it is an example of the model selection criteria we consider.

---

\*Based on joint work with Tim van Erven, Wouter Koolen and Peter Grünwald

## 2 Example: First vs Second Order Markov Chains

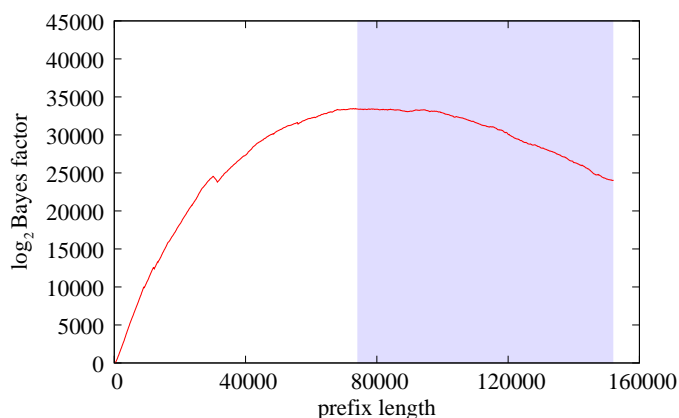
We give a concrete, simple example of Bayes factor model selection. Let  $\mathcal{M}_1$  and  $\mathcal{M}_2$  be the sets of all first and second order Markov chains on the 8-bit ASCII set  $\mathcal{X}$ ,  $|\mathcal{X}| = 256$ . The models are parameterised by their transition probabilities. Now let  $P_1$  and  $P_2$  be corresponding Bayesian prediction strategies based on uniform priors  $w_1(\theta) = 1$  and  $w_2(\theta) = 1$ . We also use a uniform prior  $W(1) = W(2) = \frac{1}{2}$  on the models. Finally let  $x^n$  be the sequence of ASCII symbols that constitute Alice in Wonderland, which has  $n = 152089$ . We can now calculate

$$\frac{P(\mathcal{M}_1|x^n)}{P(\mathcal{M}_2|x^n)} = \frac{P_1(x^n)}{P_2(x^n)} = \frac{2^{-569147}}{2^{-593132}} = 2^{23985}.$$

Thus, Bayes factor model selection tells us that the odds are overwhelmingly in favour of the first order Markov model. This suggests that we should also expect  $P_1$  to issue better predictions, i.e. if Carroll were to rise from the grave and write an additional chapter to his beloved story, we might expect that  $P_1$  assigns higher probability to, and accumulates less loss on, that new chapter.

This assumption turns out to be false, certainly in this example. The reason is that the incurred loss for  $P_1$  and  $P_2$  is not evenly distributed over the entire sample, which means that even though  $P_1$  may have accumulated less loss overall, it may still be the case that  $P_2$  is making better predictions *at the current sample size*. This becomes very clear if we look at the  $\log_2$  of the Bayes factor as a function of the length of the prefix of the novel.

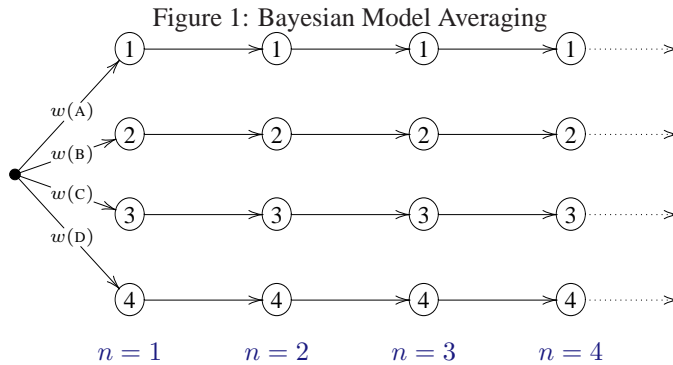
From the graph it is clear that around prefix length 78,000, the two prediction strategies perform more or less equally well, since the Bayes factor hardly changes there. Beyond sample size 78,000, the strategy based on the second order Markov chain model accumulates less loss, causing the Bayes factor to decrease. However  $P_1$  has acquired so much evidence in its favour, that it will take many outcomes before  $P_2$  can finally catch up in terms of accumulated prediction error. Only then will  $P_2$  be preferred by Bayes factor model selection. We call this the *catch-up phenomenon*.



To put this example in perspective, note that we are *not* trying to suggest that either the models or the priors we used are reasonable. We used this extremely naive example only for simplicity and because it *illustrates* the phenomenon we are interested in so well. One may ask if the phenomenon would still occur if we had used better models. The answer is yes: even if the models are chosen carefully so that one of the considered models is “true”, i.e. it contains the distribution from which the data were sampled, then it may still be the case that, at lower sample sizes, the prediction strategies associated with other, simpler models may be much more effective, so that the catch-up phenomenon will still occur. Furthermore, the processes we encounter in practice are often so complex that even the best models we can come up with are naive, and we are forced to use uninformative priors. One example is the nonparametric setting, see [4]. We would still like to make the best predictions we can under those circumstances!

## 3 Expert Tracking

To improve predictive performance when the catch-up phenomenon occurs, we would like to figure out which prediction strategy issues the best predictions, not just overall, but *at each sample size*. For in-



stance, in the Alice example we would want to switch from prediction according to  $P_1$  to prediction according to  $P_2$  around sample size 78,000 rather than never.

It turns out that algorithms to do just this already exist. Namely, we may think of the prediction strategies  $P_1, \dots, P_K$  associated with the models as “experts”, which is just another word for an algorithm that issues predictions given a sequence of past observations. We will now describe some known algorithms for prediction advice that are useful in this context.

As explained in [5], many algorithms for prediction with expert advice can be implemented by forward propagation on hidden Markov models (HMMs). Bayesian Model Averaging (BMA) is one of the simplest. It mixes the expert predictions according to their posterior weights as follows:

$$P(x_{n+1}|x^n) = \sum_{k=1}^K P_k(x_{n+1}|x^n)w(k|x^n). \quad (2)$$

Figure 1 shows the corresponding HMM for four experts labeled  $1, \dots, 4$ . It can be interpreted as a description of a prior distribution, not on the experts, but on *sequences* of experts. Namely, the prior probability that expert  $k$  is used at sample size  $n$  is the sum of the weights of all paths from the starting state to the state(s) associated with expert  $k$  at sample size  $n$ . The weight of a path is the product of its transition probabilities. The HMM in Figure 1 contains only one path to each expert and that path visits no other experts. Thus, only expert sequences that contain exactly one expert receive nonzero prior probability: *switching* between experts is not catered for. This inability to switch between experts needs to be addressed in order to alleviate the catch-up problem.

A second simple algorithm for prediction with expert advice goes to the other extreme: here it is not only possible to use different experts at different sample sizes, but which expert is used at sample size  $n + 1$  does not even *depend* on which expert is used at sample size  $n$ ! The corresponding prediction strategy is

$$P(x_{n+1}|x^n) = \sum_{k=1}^K P_k(x_{n+1}|x^n)w(k). \quad (3)$$

In their groundbreaking paper “Tracking the Best Expert” [3], Herbster and Warmuth interpolate between these two extreme approaches: rather than never or always, they switch to a different expert with fixed probability  $\alpha$ , as in Figure 3.<sup>1</sup> Note that, as before, forward propagation on this HMM only needs to maintain  $K$  weights, and requires total running time proportional to the number of experts  $K$  and the

<sup>1</sup>Actually, unlike FixedShare, the HMM in Figure 3 allows switching to the same expert. However, it can be made to simulate FixedShare by using a slightly lower value for  $\alpha$ .

Figure 2: Elementwise Mixture

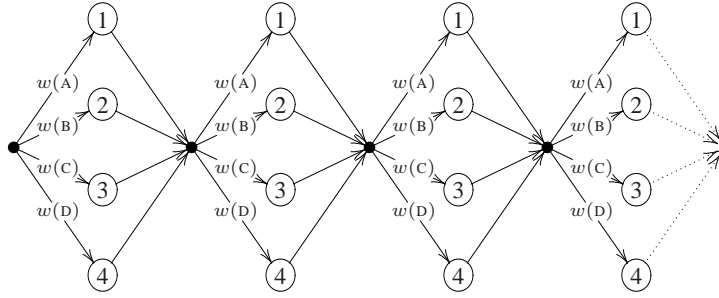
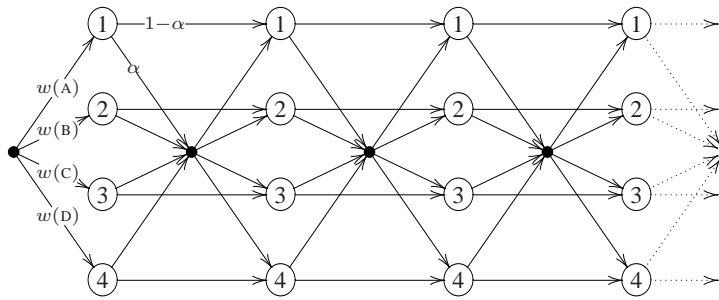


Figure 3: Fixed-Share



sample size  $n$ . The corresponding prediction strategy is in between (2) and (3):

$$P(x_{n+1}|x^n) = \sum_{k=1}^K P_k(x_{n+1}|x^n) ((1 - \alpha)P(K_n = k|x^n) + \alpha w(k)). \quad (4)$$

Herbster and Warmuth compare the logarithmic loss incurred by Fixed-Share to that incurred by any partition of the data into  $m$  blocks, where any expert is used within each block. They show that

$$\log \frac{P_{m\text{-part}}(x^n)}{P_{\text{fs}}(x^n|\alpha)} \leq (n - 1)H(\alpha) + \log K + (m - 1) \log(K - 1),$$

provided that the parameter  $\alpha$  is optimally tuned to  $(m - 1)/(n - 1)$ . This result can be applied directly to our Alice in Wonderland example: ideally we would partition the data into  $m = 2$  blocks, with the split appearing somewhere around sample size 78,000. The Fixed-Share bound tells us that, compared to this optimal partition, the logarithmic loss using the Fixed-Share algorithm is *at most*  $(n - 1)H(\frac{1}{n - 1}) + 1 \leq 17$  bits higher. This overhead is negligible compared to the gain, which is of the order of 9,000 bits, as can be read from the log Bayes factor graph. Namely, compared to using  $P_1$  for the entire book, we gain the difference between the height of the graph at index 78,000 (around 33,000 bits) and at full sample size (around 24,000 bits).

The Fixed-Share algorithm does require tuning of the switching rate alpha. However, by letting the probability of the switching transitions decrease as a function of the sample size, it is possible to do away with the parameter  $\alpha$  at only a moderate cost in terms of the performance guarantee. More information about other expert tracking algorithms in HMM format is given in [5].

## References

- [1] A.P. Dawid. Present position and potential developments: Some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society, Series A*, 147(2):278–292, 1984.
- [2] A.P. Dawid. Prequential data analysis. In M. Ghosh and P.K. Pathak, editors, *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, Lecture Notes-Monograph Series, pages 113–126. Institute of Mathematical Statistics, 1992.
- [3] M. Herbster and M.K. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.
- [4] T. van Erven, P.D. Grünwald, and S. de Rooij. Catching up faster in Bayesian model selection and model averaging. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, 2008.
- [5] W.Koolen and S. de Rooij. Expert automata for efficient tracking. In *Proceedings of the 21st Annual Conference on Computational Learning Theory (COLT)*, 2008.