Tracking the Best Predicting Model

Steven de Rooij*

July 9, 2008

*Joint work with Peter Grünwald, Tim van Erven and Wouter Koolen, CWI institute, Amsterdam

Model Selection for Prediction

Bayesian Model Selection

- A *model* is a set of probability distributions.
- *Model selection* is using the available data $x^n = x_1, \ldots, x_n$ to determine which of a set of models $\mathcal{M}_1 = \{P_\theta | \theta \in \Theta_1\}, \mathcal{M}_2 = \{P_\theta | \theta \in \Theta_2\}, \ldots$ is "best".

Bayesian approach: impose prior distributions $w_1, w_2, \ldots$ on the model parameters. We are only concerned with the marginal likelihood:

$$P_k(x^n) = \int_{\theta \in \Theta_k} P_\theta(x^n) w_k(\theta) \mathrm{d}\theta$$

Example: Wonderland

$\mathcal{M}_1$ and $\mathcal{M}_2$ are first and second order Markov chain models on byte sequences.
x^n is the text of "Alice in Wonderland".

- We use uniform priors.
- Simple models, I know!

$$\text{Bayes factor:} \qquad \frac{P_1(x^n)}{P_2(x^n)} = \frac{2^{-569147}}{2^{-593132}} = 2^{23985}.$$

Conclusion: the first order Markov chain model is "best".

Example: Wonderland

$\mathcal{M}_1$ and $\mathcal{M}_2$ are first and second order Markov chain models on byte sequences.
x^n is the text of "Alice in Wonderland".

- We use uniform priors.
- Simple models, I know!

$$\text{Bayes factor:} \qquad \frac{P_1(x^n)}{P_2(x^n)} = \frac{2^{-569147}}{2^{-593132}} = 2^{23985}.$$

Conclusion: the first order Markov chain model is "best".

But is it?

Continued Example: Sequential Prediction

We reinterpret P_k as a sequential prediction strategy.
Using the chain rule, we can rewrite:

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

We score P_k using log loss:

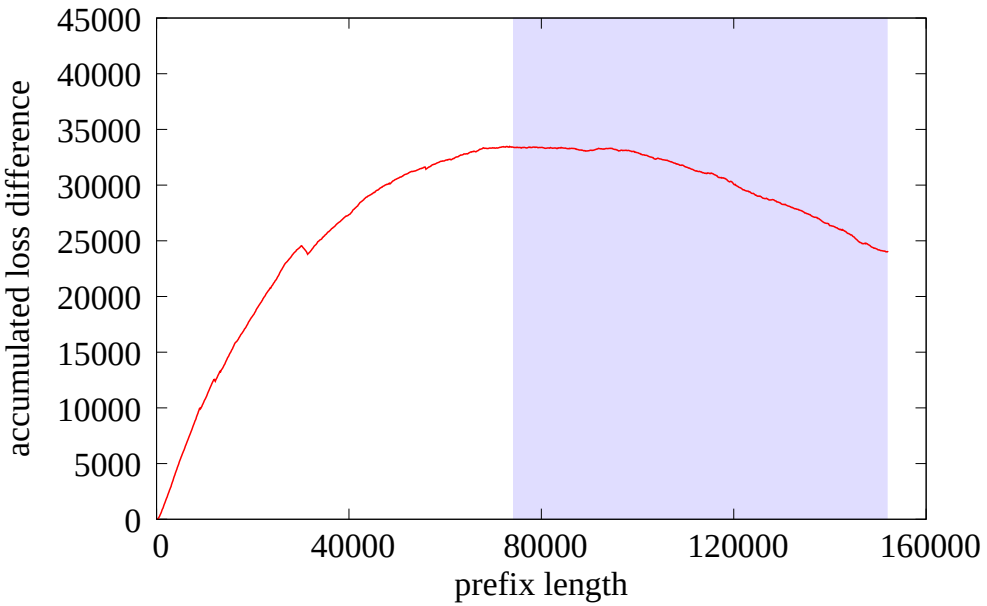
$$-\log P_k(x^n) = \sum_{i=1}^{n} -\log P_k(x_i|x^{i-1}).$$

Thus,

$$-\log \text{ probability} = \text{accumulated log loss.}$$

Observation

Plot $-\log P_2(x^i) - (-\log P_1(x^i))$ for prefixes of *Alice*:



- From the start, odds overwhelmingly in favour of P_1
- But: P_1 only predicts best during the first 78,000 outcomes!
- We would like to *switch* from P_1 to P_2 at $i \approx 78,000$.
- Idea: interpret P_k as *experts* and use *expert tracking*.

Tracking the Best Model

Experts

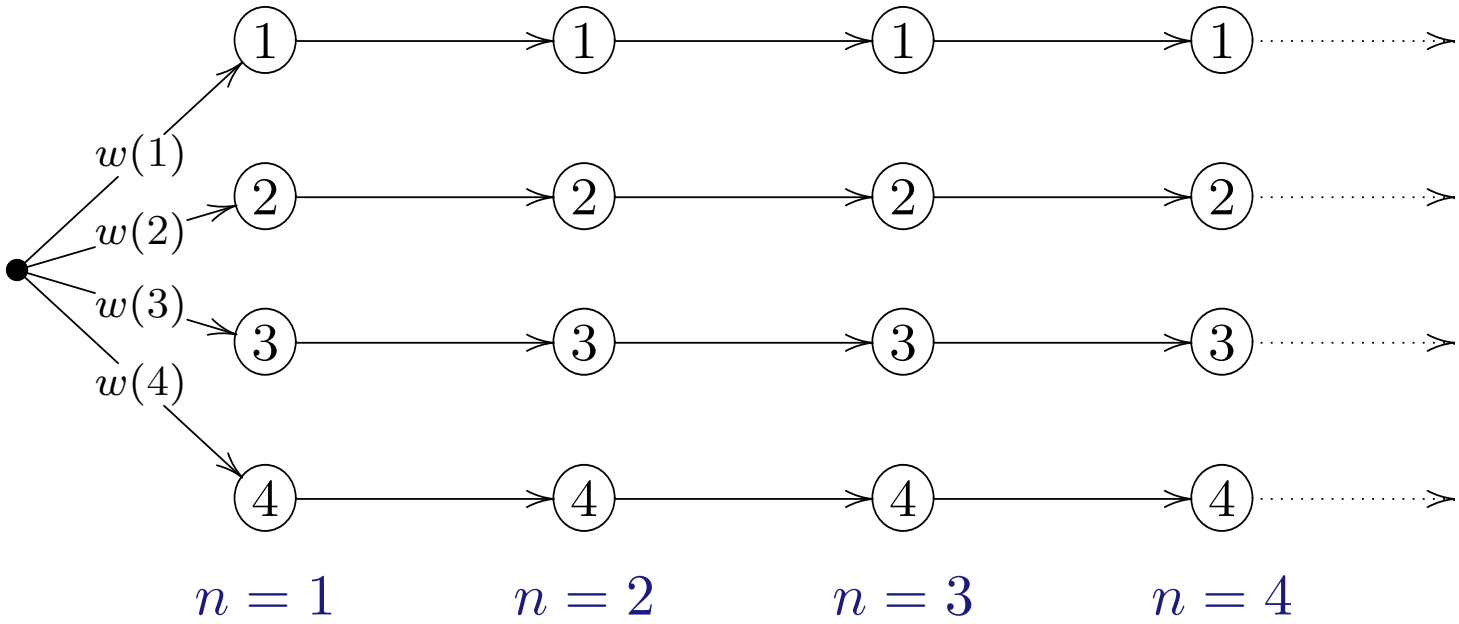
Setup:

- An *expert* P issues a prediction of X_{n+1} on input x^n.
- Our setting: probabilistic predictions, logarithmic loss.
- Similar to Dawid's *prequential forecasting systems* (see [1]).
- The predictive distribution $P_k(X_{n+1}|x^n)$ is an expert!

Our contribution: we describe algorithms using HMMS [2].

- This allows us to use the standard HMM algorithms.
- State transition diagrams provide intuitive language.
- New expert tracking algorithm: "Switch Distribution".

A first HMM example



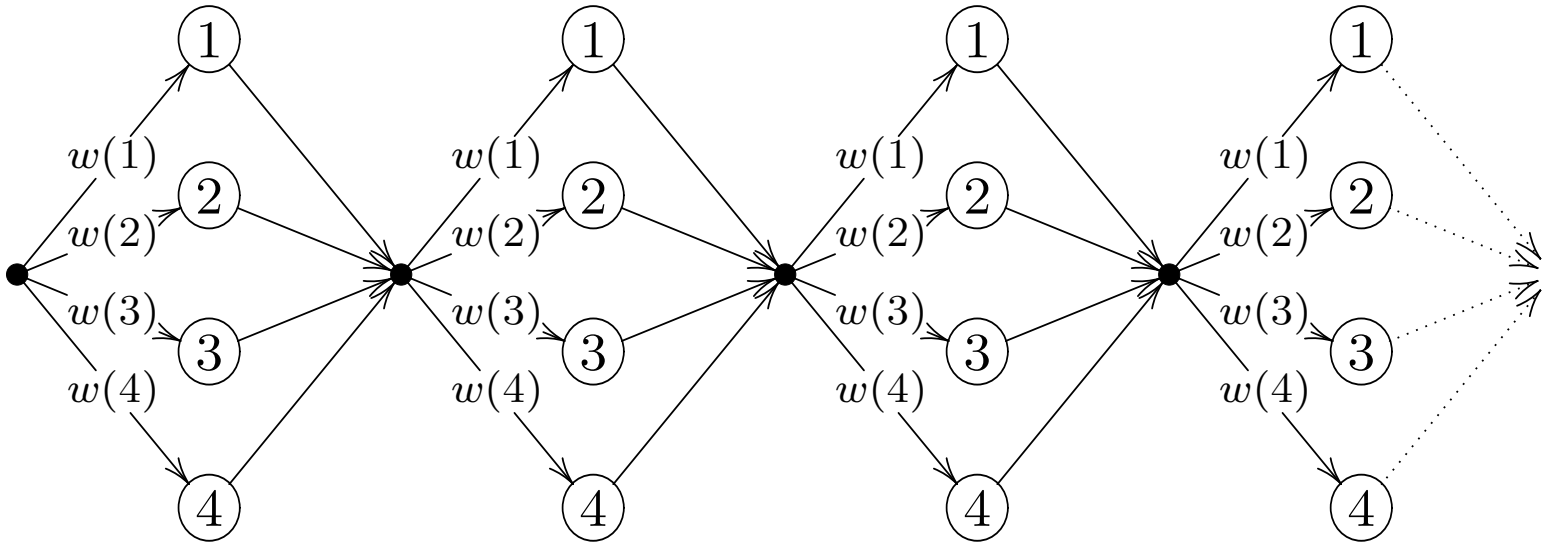
- Interpretation: defines prior on expert *sequences*.
- This example defines a standard Bayesian mixture:

$$P(X_{i+1}|x^i) = \sum_{k\in\{1,2,3,4\}} P_k(X_{i+1}|x^i)w(k|x^i)$$

- *Switching* between experts has prior probability 0.

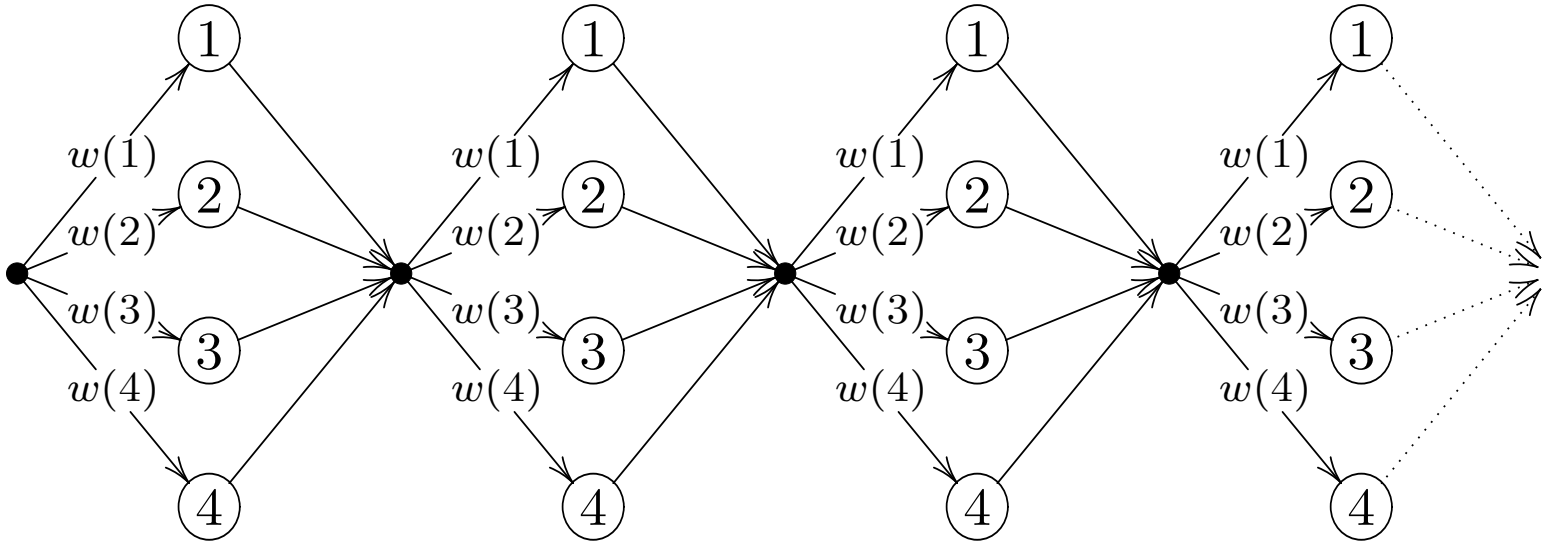
A second HMM

Q: What does this funny HMM do?



A second HMM

Q: What does this funny HMM do?



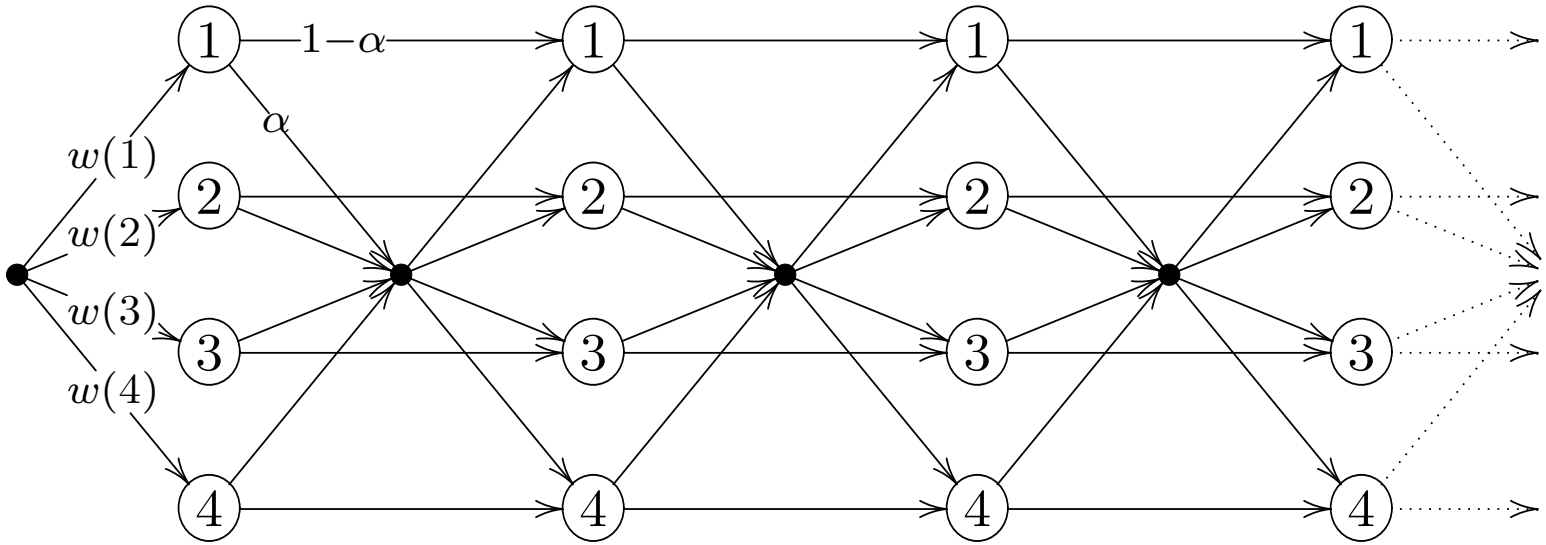
A: It is a weighted mixture of individual expert predictions:

$$P(X_{i+1} \mid x^i) = \sum_{k \in \{1,2,3,4\}} P_k(X_{i+1} \mid x^i) w(k)$$

- Other extreme: consecutive experts are *independent*.

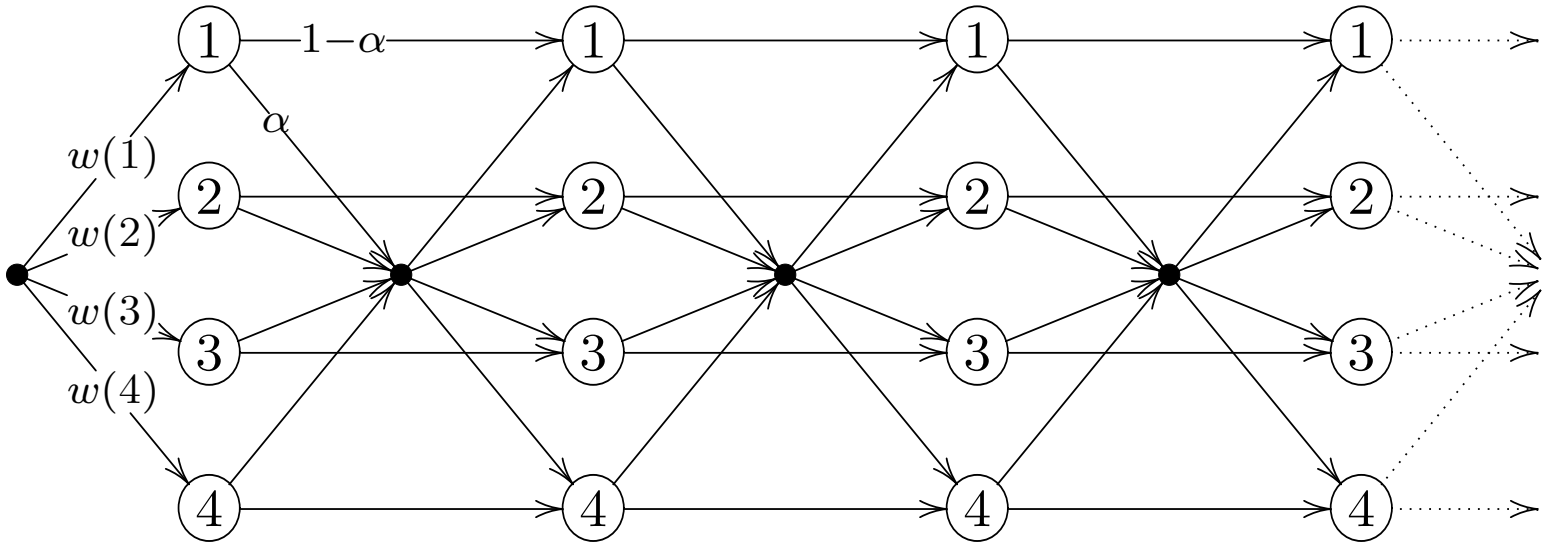
Fixed-Share

Herbster and Warmuth's Fixed-Share [3] mixes both ways:



Fixed-Share

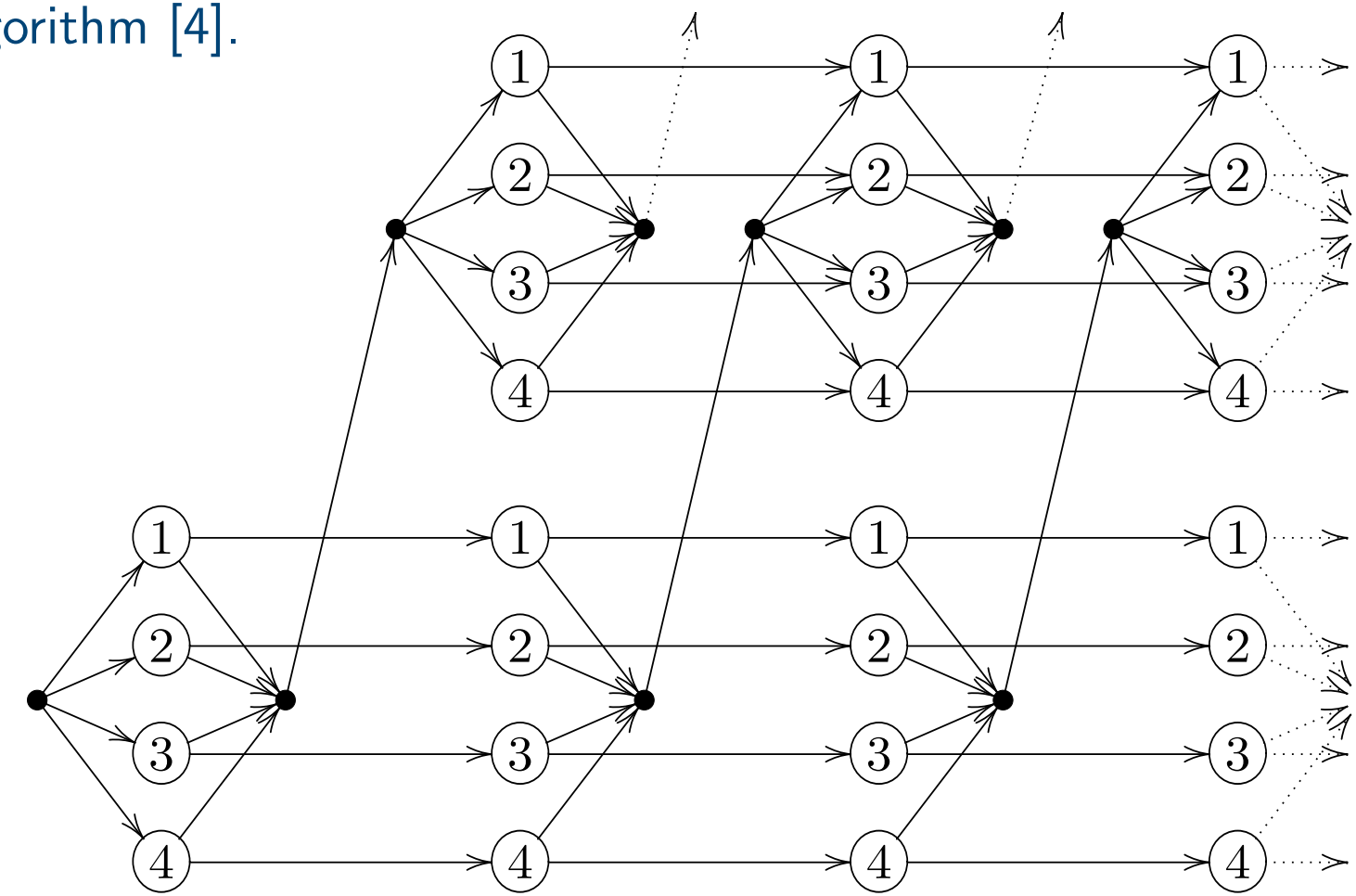
Herbster and Warmuth's Fixed-Share [3] mixes both ways:



- Predicts well if the best expert *changes occasionally*
- Difficult to tune α

Universal Share

Volf and Willems "integrate out" α and obtain an $O(n^2)$ algorithm [4].



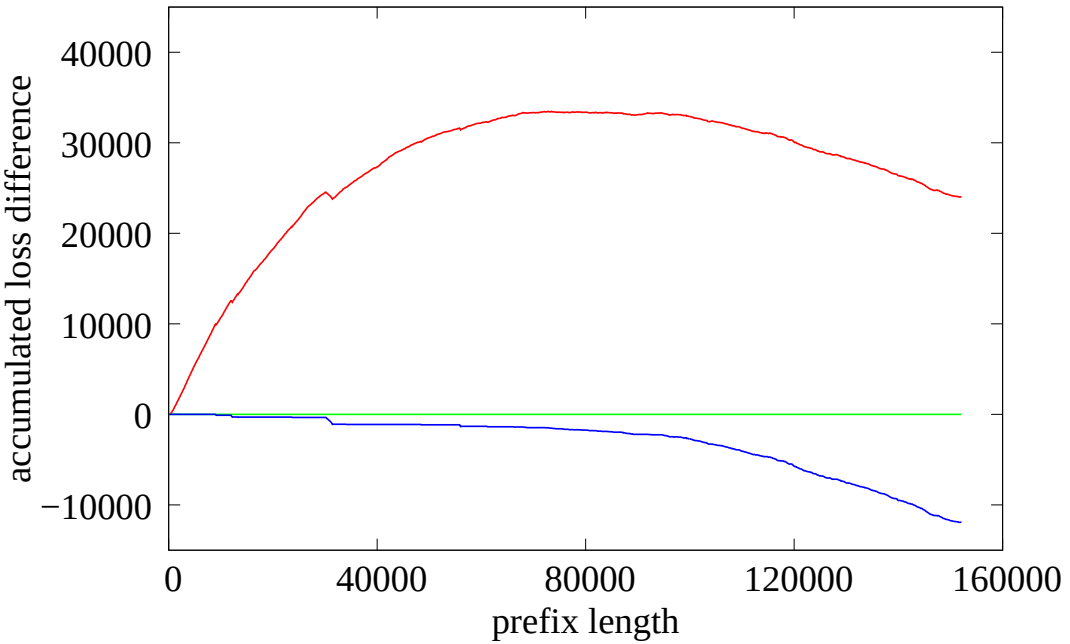
Switch Distribution

New alternative: set $\alpha_i = 1/i$ in Fixed-Share (see [5]).
Compare loss overhead to best partition into m blocks:

- Like Universal Share when m is small (say, $O(\log n)$).
- *Bounded* when m is bounded in n (using trick)

However, the running time stays linear.

Switching in the Alice example



- Switches to P_2 during "The Mouse's Tale" and at $n \approx 78,000$
- Wins significantly over both P_1 and P_2.

Conclusion

- Expert tracking can improve model selection for prediction.
- Especially in nonparametric problems with unbounded m.
- Algorithms are conveniently expressed with HMMs.

Also see:

Peter Grünwald	*The Catch-Up Phenomenon in Bayesian Inference*	invited talk thursday 10.40
Wouter Koolen	*Combining Expert Advice Efficiently*	COLT friday, 17.15

References

[1] A.P. Dawid, *Statistical Theory: The Prequential Approach*. In: Journal of the Royal Statistical Society B 147-2, pp. 278-292, 1984.

[2] W. Koolen and S. de Rooij, *Combining Expert Advice Efficiently*. In: proceedings of COLT 2008.

[3] M. Herbster and M.K. Warmuth, *Tracking the Best Expert*. In: Machine Learning 32, pp. 151-178, 1998.

[4] P.A.J. Volf and F.M.J. Willems, *Switching between Two Universal Source Coding Algorithms*. In: proceedings of DCC pp.491-500, 1998.

[5] T. van Erven, P.D. Grünwald and S. de Rooij, *Catching Up Faster in Bayesian Model Selection and Model Averaging*. In: proceedings of NIPS 2008.