Freezing and Sleeping:
Tracking Experts that Learn by Evolving Past Posteriors

Tim van Erven* Wouter M. Koolen
Centrum Wiskunde & Informatica (CWI)
Science Park 123, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
{Tim.van.Erven,Wouter.Koolen}@cwi.nl

A problem posed by Freund is how to efficiently track a small pool of experts out of a much larger set. This problem was solved when Bousquet and Warmuth introduced their mixing past posteriors (MPP) algorithm in 2001.

In Freund's problem the experts would normally be considered black boxes. However, in this paper we re-examine Freund's problem in case the experts have internal structure that enables them to learn. In this case the problem has two possible interpretations: should the experts learn from all data or only from the subsequence on which they are being tracked? The MPP algorithm solves the first case. We generalise MPP to address the second option. Our results apply to any expert structure that can be formalised using (expert) hidden Markov models. Curiously enough, for our interpretation there are *two* natural reference schemes: freezing and sleeping. For each scheme, we provide an efficient prediction strategy and prove the relevant loss bound.

Introduction Freund's problem arises in the context of prediction with expert advice [2]. In this setting a sequence of outcomes $x_{1:T} = x_1, \ldots, x_T$ needs to be predicted, one outcome at a time. Thus, prediction proceeds in rounds: in each round t we first consult a set of experts, who give us their predictions. Then we make our own prediction p_t and incur some loss $\ell(\mathrm{p}_t, x_t)$ based on the discrepancy between this prediction and the actual outcome. In this abstract predictions are probability distributions on a single outcome, and we restrict attention to the log loss $\ell(\mathrm{p}_t, x_t) = -\log \mathrm{p}_t(x_t)$. In the full paper we show how to turn any prediction strategy for log loss that satisfies certain weak requirements, into a strategy for arbitrary mixable loss.

The goal is to minimise the difference between our cumulative loss and some reference scheme. For this reference there are several options; we may, for example, compare ourselves to the cumulative loss of the best expert in hindsight. A more ambitious reference scheme was proposed by Yoav Freund in 2000.

Freund's Problem Freund asked for an efficient prediction strategy that suffers low additional loss compared to the following reference scheme:

(a) Partition the data into several subsequences.
(b) Select an expert for each subsequence.
(c) Sum the loss of the selected experts on their subsequences.

In 2001, Freund's problem was solved by Bousquet and Warmuth, who developed the efficient mixing past posteriors (MPP) algorithm [1]. To state its loss bound, we need the following notation. If members of a family $\mathbb{C}$ are pairwise disjoint and together cover $\{1, \ldots, T\}$, then we call $\mathbb{C}$ a *partition*. For any cell $\mathcal{C} = \{i_1, \ldots, i_k\} \in \mathbb{C}$ we write $x_{\mathcal{C}}$ for the subsequence $x_{i_1}, \ldots, x_{i_k}$.

Theorem 1 (Bousquet and Warmuth 2002, Theorem 7)**.** *For any mixing scheme β, data $x_{1:T}$, expert predictions and partition $\mathbb{C}$*

$$\ell(\mathrm{MPP}, x_{1:T}) \;\leq\; \sum_{\mathcal{C} \in \mathbb{C}} \ell(\textsc{Bayes}, x_{\mathcal{C}}) - \ln \beta(\mathbb{C}).$$

For each cell $\mathcal{C}$, $\ell(\textsc{Bayes}, x_{\mathcal{C}})$ denotes the loss of the Bayesian mixture of experts on the subsequence $x_{\mathcal{C}}$ in isolation, which is close to the loss of the best expert for that subsequence. The additional overhead $-\ln \beta(\mathbb{C})$ is related to the number of bits required to encode the reference partition $\mathbb{C}$. Multiple mixing schemes are possible. For an extensive discussion see [1]. This seems to settle Freund's problem, but does it really?

* Presenting author

Topic: estimation, prediction, and sequence modeling
Preference: oral presentation

The Loss of an Expert on a Subsequence In our view Freund's problem has two possible interpretations, which differ most clearly for learning experts. Namely, to measure the predictive performance of an expert on a subsequence, do we show her the data *outside* her subsequence or not? An expert that sees all outcomes will track the *global* properties of the data. This is (implicitly) the case for mixing past posteriors. But an expert that only observes the subsequence that she has to predict might see and thus exploit its *local* structure, resulting in decreased loss. The more the characteristics of the subsequences differ, the greater the gain. Let us illustrate this by an example.

The data consist of a block of ones, followed by a block of zeros, again followed by a block of ones. In Figure 1 we compare two partitions. Either we put all data into a single cell, or we split the data into the subsequence of ones and the subsequence of zeroes. Our expert predicts the probability of a one using Laplace's rule of succession, i.e. she *learns* the frequency of ones in the data that she observes [2]. Note that one learning expert suffices, as we can select (a separate copy of) her for two subsequences.

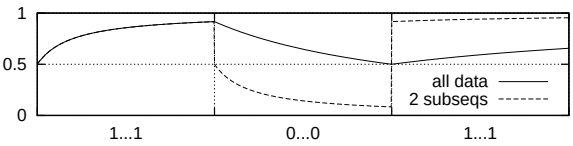

Figure 1: Estimated probability of a one

When learning on all data, the subsequences interfere with each other and the expert's predictions on block two and three track the global frequency. If the subsequences are separated, the expert's predictions converge quickly to the local frequencies (one and zero). This shows that the predictive performance of a learning expert on a subsequence in isolation can be dramatically higher than that on the same sequence in the context of all data. This behaviour is typical: on all data a learning expert will learn the average, global pattern, while on a well-chosen subsequence she can zoom in on local structure.

Sleeping or Freezing We solve Freund's problem under the interpretation that experts only observe the subsequence on which they are evaluated. This requires knowledge about the experts' internal structure. We therefore represent a learning expert as an *expert hidden Markov model* (EHMM), which is an HMM in which the production probabilities are determined by the advice of simpler experts. Many adaptive prediction strategies (i.e. learning experts) from the literature can be represented as efficient EHMMs [3].

There are two ways to evaluate the performance of a learning expert $\mathfrak{H}$ on a subsequence $x_{\mathcal{C}}$ in isolation: *freezing* ($\mathfrak{H}^{\mathsf{fr}}_{\mathcal{C}}$) and *sleeping* ($\mathfrak{H}^{\mathsf{sl}}_{\mathcal{C}}$). To illustrate the difference, imagine a sequence $x_{1:T}$ of images shown on a television screen. Suppose we ask $\mathfrak{H}$ to predict the subsequence $x_{\mathcal{C}}$ of images belonging to our favourite show. We want to *freeze* $\mathfrak{H}$ during commercial breaks: $\mathfrak{H}^{\mathsf{fr}}_{\mathcal{C}}$ simply ignores them and continues predicting the show where it left off. We want to put $\mathfrak{H}$ to *sleep* when we zap to another channel: $\mathfrak{H}^{\mathsf{sl}}_{\mathcal{C}}$ knows the time and, after we zap back, predicts the show as it has advanced.

EPP We introduce an efficient prediction strategy, called *evolving past posteriors*, that generalises MPP. Its two variants $\mathrm{EPP}^{\mathsf{fr}}$ and $\mathrm{EPP}^{\mathsf{sl}}$ achieve small additional loss compared to Freund's scheme for freezing and sleeping:

Theorem 2. *Let* f/s *denote either* fr *or* sl. *For any learning expert $\mathfrak{H}$ in EHMM form, mixing scheme β, data $x_{1:T}$, expert predictions and partition $\mathbb{C}$*

$$\ell(\mathrm{EPP}^{\mathsf{f/s}}, x_{1:T}) \;\leq\; \sum_{\mathcal{C}\in\mathbb{C}} \ell(\mathfrak{H}^{\mathsf{f/s}}_{\mathcal{C}}, x_{\mathcal{C}}) - \ln \beta(\mathbb{C}).$$

Acknowledgements We would like to thank Manfred Warmuth for raising our interest in this subject during COLT 2008. We also thank Steven de Rooij and Peter Grünwald for fruitful discussions and suggestions. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

[1] O. Bousquet and M. K. Warmuth. Tracking a small set of experts by mixing past posteriors. *JMLR*, 3:363–396, 2002.

[2] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[3] W. M. Koolen and S. de Rooij. Combining expert advice efficiently. In *Proc. of the 21st Ann. Conf. on Learning Theory*, pages 275–286, 2008.