



UvA-DARE (Digital Academic Repository)

Combining strategies efficiently: high-quality decisions from conflicting advice

Koolen, W.M.

Publication date
2011

[Link to publication](#)

Citation for published version (APA):

Koolen, W. M. (2011). *Combining strategies efficiently: high-quality decisions from conflicting advice*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights

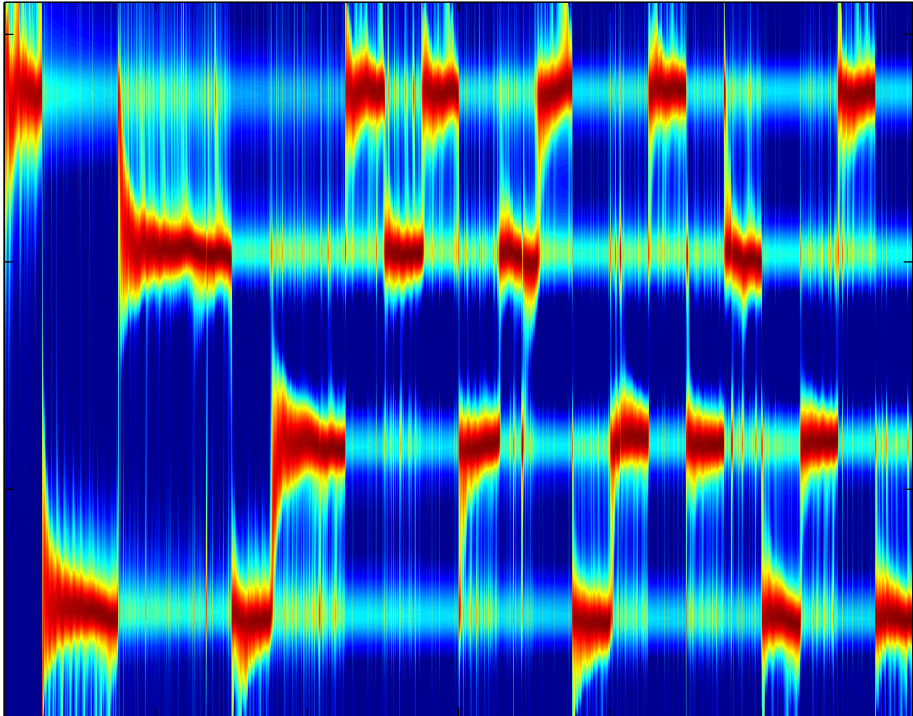
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

Chapter 4

Freezing & Sleeping



Abstract 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 chapter 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. Our contribution is to generalise MPP to address the second option. The results we obtain 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.

4.1 Introduction

Freund's problem arises in the context of prediction with expert advice [25]. In this setting a sequence of outcomes needs to be predicted, one outcome at a time. Thus, prediction proceeds in rounds: in each round we first consult a set of experts, who give us their predictions. Then we make our own prediction and incur some loss based on the discrepancy between this prediction and the actual outcome. 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 small 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.

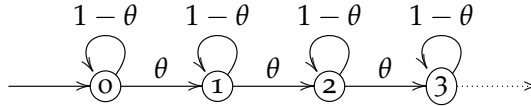
In 2001, Freund's problem was addressed by Bousquet and Warmuth, who developed the efficient algorithm called mixing past posteriors (MPP) [19]. MPP's loss is bounded by the loss of Freund's scheme plus some overhead that depends on the number of bits required to encode the partition of the data, and it has found successful application in [70]. Problem solved. Or is it?

4.1.1 Three Reference Schemes

In this paper we take another look at Freund's reference scheme for *learning experts* and ask: if an expert is selected for some segment, then should the expert learn from all data or only from the data in that segment?

We may assume that the experts do not know the segmentation chosen in step a of the reference scheme. (Otherwise, why wouldn't we just ask them?) Hence if we treat the experts as black boxes and only ask for their prediction at each time step as in [19], it is natural that they

Figure 4.1 Example learning expert $DM[\theta]$, which learns a drifting mean, specified by its state transition diagram.



learn from all data. We call this interpretation of Freund's problem the *full reference scheme*.

However, as the following example will illustrate, it may be beneficial if experts learn only from the segment for which they are selected, because they may get confused by data in other segments that follow a different pattern. As a slight complication, it will turn out that we have a further choice: whether to tell a learning expert the timing of its segment or not, which generally makes a difference. When segment timing is preserved, we obtain the *sleeping reference scheme*; when segment timing is *not* preserved we obtain the *freezing reference scheme*. The next intuitive example demonstrates that the full, freezing and sleeping reference schemes are fundamentally different, and that the latter two can be dramatically more appropriate for prediction with learning experts.

4.1.1.1 Motivating Example: Drifting Mean

In applications one would usually build up complicated prediction strategies from simpler ones in a hierarchical fashion. Following that fashion, we first define simple constant experts, parametrised by $\mu \in \mathbb{R}$, which predict according to a normal distribution with mean μ and unit variance in each round.

Learning Experts Now define a learning expert $DM[\theta]$, as displayed in Figure 4.1, that has a stochastic model for the (unobservable) drift of μ over time. This *drifting mean* learning expert predicts according to a hidden Markov model in which the hidden state at time t is μ_t and the production probability of an outcome given μ_t is determined by the simple expert with parameter μ_t . Initially, $\mu_1 = 0$ with probability one. Then $\mu_{t+1} = \mu_t + 1$ with probability θ and $\mu_{t+1} = \mu_t$ with probability $1 - \theta$ for some fixed parameter θ .

The expert $DM[\theta]$ may be said to be learning, because its posterior distribution of μ_t given outcomes x_1, \dots, x_{t-1} indicates how much credibility the expert assigns to each value of μ_t : high weight on, say, $\mu_t = 3$ indicates that $DM[\theta]$ considers it likely for $\mu_t = 3$ to give the best prediction for x_t .

Data Consider the two artificial data sets displayed in Figures 4.2a and 4.2b. These data sets were obtained as follows. First, we generated two straight-line data sets, with outcomes increasing at a rate of 0.1 and 0.3 per trial respectively. Then we divided both data sets in segments of 100 outcomes each. The data in Figure 4.2a were obtained by *interleaving* 10 segments from the 0.1 and 0.3 data sets, whereas the data in Figure 4.2b were obtained by *alternating* 10 segments from the 0.1 and 0.3 data sets. By construction, the freezing reference scheme is suited for the data in Figure 4.2a, while the sleeping reference scheme is appropriate for the data in Figure 4.2b.

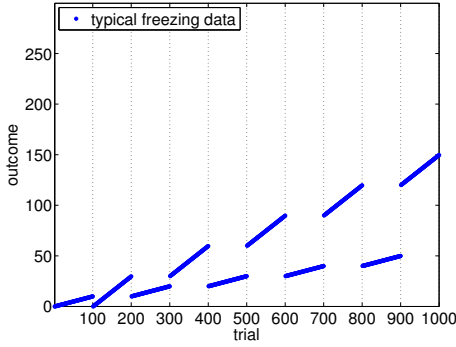
Prediction Task We now evaluate the performance of the three reference schemes on the two data sets. In each case we consider two experts: $DM[0.1]$ and $DM[0.3]$, and split the data into two subsequences (step (a), according to the true rate, either 0.1 or 0.3. We predict all outcomes for which the actual rate was $\theta \in \{0.1, 0.3\}$ using the expert $DM[\theta]$.

The difference between the three schemes lies in which data is used by both experts to learn from. In the full reference scheme $DM[0.1]$ and $DM[0.3]$ are shown all the data, even those samples they do not predict. In the two other reference schemes, on the other hand, $DM[0.1]$ only sees the data for which it is selected, that is, the data with true rate 0.1. Similarly, $DM[0.3]$ only sees the data with true rate 0.3. For freezing $DM[\theta]$ predicts as if the data it has observed are the only data, thus the original timing of the samples is lost. For sleeping the original timing of the samples is preserved, and $DM[\theta]$ has to predict with uncertainty about the intermediate unobserved samples.

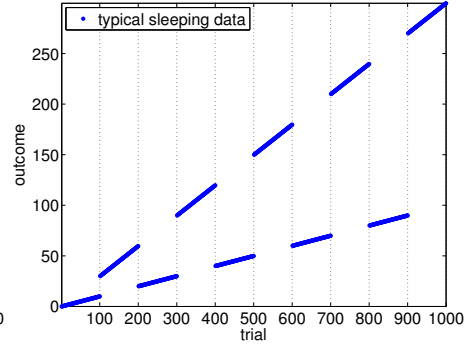
Posteriors Figures 4.2c and 4.2d show the posterior distribution of the expert $DM[0.1]$ on states after 200 trials for each reference scheme. These posterior distributions can be interpreted as the belief of the learning expert $DM[0.1]$ about the unobserved drifting mean after 200 trials.

Figure 4.2 The difference between the full, freezing and sleeping reference schemes. Note the logarithmic scale of the y-axis in (e) and (f)!

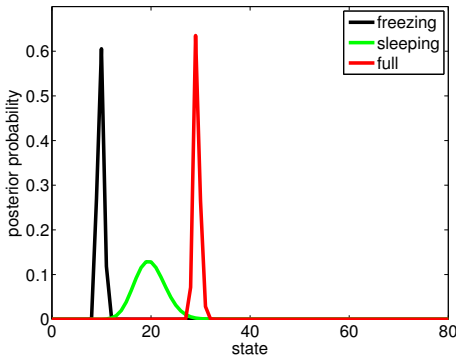
(a) Suitable freezing data



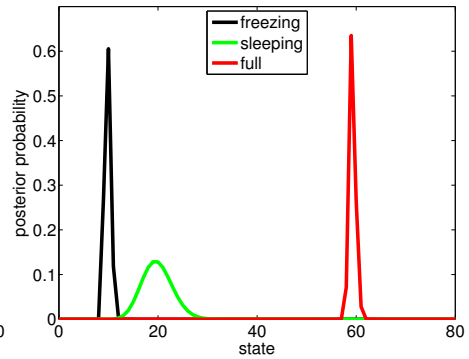
(b) Suitable sleeping data



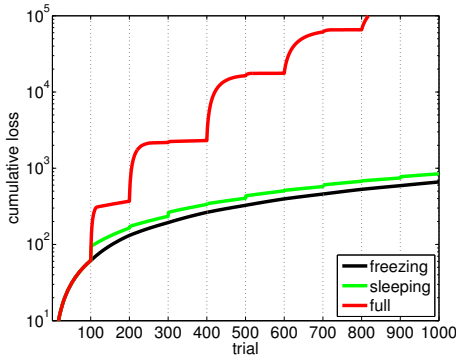
(c) Belief of DM[0.1] after 200 trials of (a)



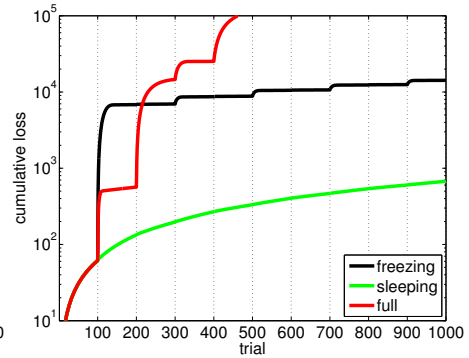
(d) Belief of DM[0.1] after 200 trials of (b)



(e) Cumulative loss on data (a)



(f) Cumulative loss on data (b)



We see in Figure 4.2c that, for the freezing data, the expert posterior obtained by the freezing reference scheme matches the 201st outcome (which is 10 in the freezing data set) best. Recall that this posterior is obtained by first showing DM[0.1] outcomes 1 through 100, and then asking it to predict outcome 201 as if it was the next outcome in the sequence.

We also see in Figure 4.2d that, for the sleeping data, the expert posterior obtained by the sleeping reference scheme matches the 201st outcome (which is 20 in the sleeping data set) best. Recall that this posterior is obtained by first showing DM[0.1] outcomes 1 through 100, and then asking it to predict outcome 201 with all intermediate outcomes unobserved.

Finally, we see that in both cases, the expert posterior obtained by the full reference scheme, which shows *all* outcomes to DM[0.1], overshoots: the expert is confused by observing the intermediate outcomes.

Loss These snapshots of the expert's posteriors provide an intuitive understanding of what the reference schemes do and which one is appropriate. We now quantify the predictive performance by looking at the resulting cumulative loss. Figures 4.2e and 4.2f show the cumulative log(arithmetic) loss for all three reference schemes. Note that the difference between the schemes is so large that their losses had to be plotted on a logarithmic scale.

We see in Figure 4.2f that for the sleeping data the sleeping reference scheme has much smaller loss than the other two schemes. And for the freezing data the freezing reference scheme has the smallest loss by far, as shown in Figure 4.2e. (Mind the logarithmic scale of the y-axis, which puts the loss of sleeping deceptively close to the loss of freezing in Figure 4.2e: a constant offset indicates a fixed multiplicative overhead.) In both cases the reason for the large differences between the reference schemes is that both experts DM[0.1] and DM[0.3] get confused if they learn from the wrong data.

Note that for this synthetic example, we knew which partitioning into subsequences to choose, since we constructed the data ourselves. For real data a partitioning is not readily available. The challenge addressed in this chapter is to *learn* the best partition of the data online.

4.1.1.2 Structured Experts

In this chapter, we solve Freund’s problem under the interpretation that experts only observe the subsequence on which they are evaluated. Of course, for *arbitrary* experts, this is impossible. For in the setting of prediction with expert advice (see [25]), the expert predictions that we receive each round are *always* in the context of all data. We have no access to the experts’ predictions in the context of any subsequence, and these predictions may differ drastically from those on the whole data.

Often however, experts have internal structure. For example, in [108, 80, 180, 181] adaptive prediction strategies (i.e. learning experts) are explicitly constructed from basic experts. To represent such structured experts, we use the general framework called *expert hidden Markov models* (EHMMs), that was introduced in Chapter 3. EHMMs are hidden Markov models in which the production probabilities are determined by expert advice. A structured expert in EHMM form provides sufficient information about its predictions on any isolated subsequence.

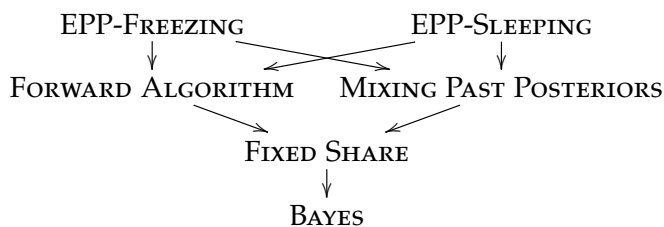
Many strategies for prediction with expert advice (i.e. learning experts) can be rendered as EHMMs. For example all adaptive strategies in the papers above (see Chapter 3). But there are also strategies that cannot be brought into EHMM form, like e.g. *follow the perturbed leader* [73] and *variable share* [80].

Our approach may also be of interest to machine learning with regular hidden Markov models (HMMs) [146]. Although existing approaches to shift between multiple HMMs [65, 66, 104] usually focus on change-point detection, prediction seems a highly related issue.

4.1.2 Overview

After preliminaries we start by reviewing the main existing loss bound for mixing past posteriors in Section 4.3. Then, in Section 4.4, we review EHMMs as a way to represent structured experts.

The next section, Section 4.5, contains our results for Freund’s problem when structured experts are evaluated on isolated subsequences. We formalise sleeping and freezing as two different ways of presenting a subsequence of the data to an EHMM, and present the *evolving past posteriors* (EPP) algorithm that takes an EHMM as input. The EPP algorithm has two variants, which both generalise the mixing past pos-

Figure 4.3 Generalisation relation among prediction strategies

teriors algorithm in a different way: EPP-SLEEPING for sleeping and EPP-FREEZING for freezing. The relation between EPP and other existing prediction strategies is shown in Figure 4.3. There $A \rightarrow B$ means that by carefully choosing prediction strategy A 's parameters it reduces to strategy B .

In order to understand EPP, we verify that it produces the same predictions for any two EHMMs that are equivalent in an appropriate sense, and analyse its running time. We then proceed to show our main result, which is that the losses of EPP-FREEZING and EPP-SLEEPING are bounded by the loss of their appropriate reference scheme plus a complexity penalty that depends on the number of bits required to encode the reference partition in the same way as for mixing past posteriors. In fact, our bounds (slightly) improve the known loss bound for mixing past posteriors. Thus we solve Freund's problem with learning experts presented as EHMMs, both for freezing and for sleeping.

We first derive our results only for logarithmic loss. This allows us to use familiar concepts and results from probability theory and refer to the interpretation of log loss as a codelength [25]. In Section 4.6 we conclude by proving that any algorithm that satisfies certain weak conditions, in particular EPP, directly generalises to an algorithm for arbitrary mixable losses with the appropriate loss bounds.

4.2 Preliminaries

Prediction With Expert Advice Each round t , we first receive advice from each expert $e \in \mathcal{E}$ in the form of an action $a_t^e \in \mathcal{A}$. Then we distill our own action $a_t^{\text{alg}} \in \mathcal{A}$ from the expert advice. Finally, the actual outcome $x_t \in \mathcal{X}$ is observed, and everybody suffers loss as specified

by a fixed loss function $\ell: \mathcal{A} \times \mathcal{X} \rightarrow [0, \infty]$. Thus, the performance of a sequence of actions $a_1 \cdots a_T$ upon data $x_1 \cdots x_T$ is measured by the cumulative loss $\sum_{t=1}^T \ell(a_t, x_t)$.

Log Loss For *log loss* the actions \mathcal{A} are probability distributions on \mathcal{X} and $\ell(p, x) = -\log p(x)$, where \log denotes the natural logarithm. It is important to notice that minimising log loss is equivalent to maximising the predicted probability of outcome x . We write p_t^e for the prediction of expert e at time t and denote these predictions jointly by $p_t^\mathcal{E}$.

Subsequences For $m \leq n$, we abbreviate $\{m, \dots, n\}$ to $m:n$. For completeness, we set $m:n = \emptyset$ for $m > n$. For any sequence y_1, y_2, \dots and any set of indices $\mathcal{C} = \{i_1, i_2, \dots\}$ we write $y_{\mathcal{C}}$ for the subsequence $\langle y_i \rangle_{i \in \mathcal{C}}$. For example, $x_{\mathcal{C}} = \langle x_i \rangle_{i \in \mathcal{C}}$ and $p_{1:T}^\mathcal{E} = p_1^\mathcal{E}, \dots, p_T^\mathcal{E}$. If members of a family $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots\}$ are pairwise disjoint and together cover $1:T$ ($\bigcup \mathcal{C} = 1:T$), then we call \mathcal{C} a *partition* of $1:T$, and its members *cells*.

4.3 Mixing Past Posteriors

Mixing past posteriors (MPP) is a strategy for prediction with expert advice. It operates by maintaining a table of so-called posterior distributions on the set of experts. Each round, we first compute the predictive distribution on experts by mixing all the posteriors in the table. Then the next outcome is predicted by mixing the expert predictions according to this distribution. Finally, the next outcome is observed. The predictive distribution on experts is conditioned on this outcome, and the posterior distribution thus obtained is appended to the table of posteriors. Note the recursive construction of the distributions in the table; they are not Bayesian posteriors, but conditioned mixtures of all earlier distributions from that same table.

We will not formally introduce MPP here, but recover it as a special case of both the freezing and sleeping algorithms in Section 4.5.4. Here we state the classical loss bound [19, Theorem 7], introducing our notation along the way. This loss bound relates the loss of MPP to Freund's full reference scheme, where we choose a partition of the data (step a) and select an expert for each partition cell (step b). We measure expert

performance (step c) using the predictions issued in the context of all data, i.e. the full interpretation of Freund's scheme.

4.3.1 Loss Bound

We bound the overhead of MPP over the full reference scheme in terms of the complexity of the reference partition. We first state the theorem, and then explain the ingredients. We write $P_w^{\text{MPP}}(x_{1:T})$ for the probability that MPP assigns to data $x_{1:T}$ (so $-\log(P_w^{\text{MPP}}(x_{1:T}))$ is MPP's cumulative log loss).

4.3.1. THEOREM ([19, Theorem 7]). *For any mixing scheme β , Bayesian joint distribution P^{B} with prior distribution w on experts, partition \mathbf{C} of $1:T$, data $x_{1:T}$ and expert predictions $p_{1:T}^{\mathcal{E}}$*

$$P_w^{\text{MPP}}(x_{1:T}) \geq \beta(\mathbf{C})P_{\mathbf{C}}^{\text{B}}(x_{1:T}). \quad (4.1)$$

A mixing scheme β is a sequence β_1, β_2, \dots of distributions, where β_{j+1} is a probability distribution on $0:j$. In [19] several mixing schemes are listed, e.g. *Uniform Past* and *Decaying Past*. A mixing scheme is turned into a distribution on partitions as follows. Let \mathbf{C} be a partition of $1:T$, and let $i \in 1:T$. The cell of i , denoted $\mathbf{C}(i)$, is the unique $\mathcal{C} \in \mathbf{C}$ such that $i \in \mathcal{C}$. We write $\text{prev}^{\mathbf{C}}(i)$ for the predecessor of i , defined as the largest element in $\mathbf{C}(i) \cup \{0\}$ that is smaller than i . Using this notation, the distribution on partitions is given by

$$\beta(\mathbf{C}) := \prod_{t \in 1:T} \beta_t(\text{prev}^{\mathbf{C}}(t)).$$

Note that this distribution is potentially *defective*; two elements $i < j$ cannot share the same nonzero predecessor, but β_i may assign nonzero probability to $\text{prev}^{\mathbf{C}}(j)$ nonetheless.

Now that we have seen how the loss bound encodes partition, we turn to $P_{\mathbf{C}}^{\text{B}}(x_{1:T})$, the probability of the data $x_{1:T}$ given a particular partition \mathbf{C} . To compute it, we treat the cells independently (4.2), and per cell we use the Bayesian mixture with prior w on experts (4.3), thus

mixing the predictions the experts issued in the context of all data (4.4).

$$P_{\mathbb{C}}^{\text{B}}(x_{1:T}) := \prod_{\mathcal{C} \in \mathbb{C}} P_{\mathcal{C}}^{\text{B}}(x_{\mathcal{C}}), \text{ where} \quad (4.2)$$

$$P_{\mathcal{C}}^{\text{B}}(x_{\mathcal{C}}) := \sum_{e \in \mathcal{E}} w(e) p_{\mathcal{C}}^e(x_{\mathcal{C}}) \text{ and} \quad (4.3)$$

$$p_{\mathcal{C}}^e(x_{\mathcal{C}}) := \prod_{i \in \mathcal{C}} p_i^e(x_i). \quad (4.4)$$

A second bounding step allows us to relate the performance of MPP directly to Freund's full scheme. Let w be the uniform prior over a finite set of experts \mathcal{E} , and select an expert $e^{\mathcal{C}}$ for each partition cell $\mathcal{C} \in \mathbb{C}$. Then bound each sum (4.3) from below by one of its terms to obtain

4.3.2. COROLLARY.
$$P_w^{\text{MPP}}(x_{1:T}) \geq \beta(\mathbb{C}) |\mathcal{E}|^{-|\mathbb{C}|} \prod_{\mathcal{C} \in \mathbb{C}} p_{\mathcal{C}}^{e^{\mathcal{C}}}(x_{\mathcal{C}}).$$

Thus the log-loss overhead of MPP over the full reference scheme is bounded by $-\log \beta(\mathbb{C}) + |\mathbb{C}| \log |\mathcal{E}|$, which can be related to the number of bits to encode the chosen partition and the selected experts for each cell [19].

Convex Combinations In [19], the authors make a point of selecting a *convex combination of experts* for each subsequence, where the loss of a convex combination of experts is the weighted average *loss* of the experts. The loss of such a convex combination is therefore *always* higher than the loss of its best expert. Uniform bounds in terms of arbitrary experts, like Corollary 4.3.2, apply in particular to the best expert, and hence to any convex combination. Therefore, without loss of generality, we do not discuss convex combinations any further.

Interpreting Freund's Problem The loss bound Theorem 4.3.1 shows that MPP solves the black-box-experts interpretation of Freund's problem. This can be seen clearly in (4.4). To predict the subsequence $x_{\mathcal{C}}$, it uses predictions $p_{\mathcal{C}}^e$ which were issued in the context of all data. This means that the experts observe the entire history $x_{1:i}$ before predicting the next outcome X_{i+1} .

Switching between *learning* experts that observe all data is useful when the data are homogeneous, and the experts learn its global pattern at different speeds. In such cases we want to train each expert on

all observations, for then by switching at the right time, we can predict each outcome using the expert that has learned most *until then*. This scenario is analysed in [178], where experts are parameter estimators for a series of statistical models of increasing complexity.

On the other hand, if the data have local patterns then our new interpretation of Freund's problem applies, and we want to train each expert on the subsequence on which it is evaluated, so that it can exploit its local patterns. To solve Freund's problem for such learning experts, we need to know about its internal structure.

4.4 Structured Experts

Assume there is only a single expert and fix a reference partition. Suppose we want to predict as if the expert is restarted on each cell of the partition, when in reality the expert just makes her predictions as if all the data were in a single cell. Then clearly this is impossible if we treat this expert completely as a black box: if we do not know what the expert's predictions would have been if a certain outcome were, say, the start of a new cell, then we cannot match these predictions.

The expert therefore needs to reveal to us some of her internal state. To this end, we will represent the parts of her internal state that will *not* be revealed to us by lower level experts that we will treat as black boxes, and assume our main expert combines the predictions of these base experts using an *expert hidden Markov model* (EHMM).

4.4.1 EHMMs

Expert Hidden Markov Models (EHMMs) were introduced in Chapter 3 as a language to specify strategies for prediction with expert advice. We briefly review them here. An *EHMM* \mathcal{A} is a probability distribution that is constructed according to the Bayesian network in Figure 4.4. It is used to sequentially predict outcomes X_1, X_2, \dots which take values in outcome space \mathcal{X} . At each time t , the distribution of X_t depends on a hidden state Q_t , which determines mixing weights for the experts' predictions. Formally, the *production function* p_{\downarrow} determines the interpretation of a state: it maps any state $q_t \in \mathcal{Q}$ to a distribution $p_{\downarrow}^{q_t}$ on the identity E_t of the expert that should be used to predict X_t . Then given $E_t = e$, the distribution of X_t is base expert e 's prediction p_t^e . It remains

Figure 4.4 Bayesian network specification of an EHMM

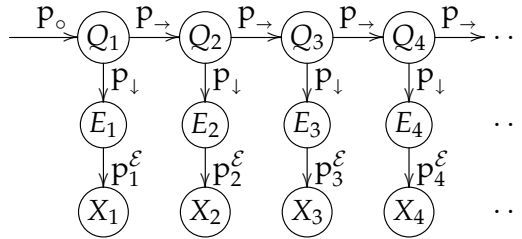
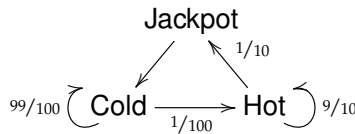


Figure 4.5 Hidden state transitions in slot machine HMM



to define the distribution of the hidden states. The starting state Q_1 has *initial distribution* p_o , and the state evolves according to the *transition function* $p_{->}$, which maps any state q_t to a distribution $p_{->}^{q_t}$ on states.

An EHMM \mathcal{A} defines a prediction strategy as follows; after observing $x_{1:t}$, predict outcome X_{t+1} using the marginal $\mathcal{A}(X_{t+1}|x_{1:t})$, which is a *mixture* of the expert's predictions $p_{t+1}^{\mathcal{E}}$.

4.4.1. EXAMPLE (Any Ordinary HMM). To illustrate how ordinary hidden Markov models are a special case of EHMMs, consider the following naive gambler's HMM model of an old-fashioned slot machine: in each round the gambler inserts one nickel into the slot machine and then the machine pays out a certain number of nickels depending on its hidden internal state: in state Cold it pays out nothing; in state Hot it pays out an amount between one and five nickels, uniformly at random; and then there's Jackpot in which it always pays out ten nickels. The machine always starts in state Cold and the state transitions are as in Figure 4.5.

To make an EHMM out of this HMM, we just identify experts with states: $\mathcal{Q} = \mathcal{E} = \{\text{Cold}, \text{Hot}, \text{Jackpot}\}$, $p_{\downarrow}^e(e) = 1$, and each expert predicts according to the corresponding payout scheme. The distributions on states follow the original HMM: $p_o(\text{Cold}) = 1$ and $p_{->}$ as in Figure 4.5. ◇

4.4.2. EXAMPLE (Bayes on base experts). We identify the Bayesian distribution with prior w on base experts \mathcal{E} and the EHMM with $\mathcal{Q} = \mathcal{E}$, $p_\circ = w$, and $p_{\rightarrow}^e(e) = p_{\downarrow}^e(e) = 1$, since their marginals coincide. Despite its deceptive simplicity, this EHMM *learns*: its marginal distribution on the next outcome is a mixture of the expert's predictions according to the Bayesian posterior. \diamond

4.4.3. EXAMPLE (Bayes on EHMMs). Fix EHMMs $\mathcal{A}^1, \dots, \mathcal{A}^n$ with disjoint state spaces and the same basic experts, and let w be a prior distribution on $1:n$. The Bayesian mixture EHMM has state space $\mathcal{Q} = \bigcup_i \mathcal{Q}^i$, and for any two states $q, q' \in \mathcal{Q}^i$ belonging to the same original EHMM, $p_\circ(q) = w(i) p_\circ^i(q)$, $p_{\rightarrow}^q(q') = p_{\rightarrow}^{i,q}(q')$ and $p_{\downarrow}^q(e) = p_{\downarrow}^{i,q}(e)$. Again, this EHMM *learns* which of the given EHMMs is the best predictor. \diamond

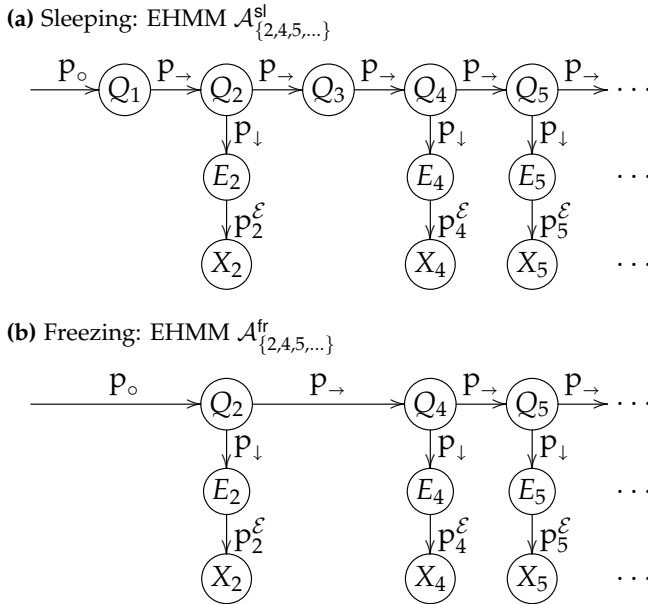
4.4.2 The Forward Algorithm

Sequential predictions for EHMMs can be computed efficiently using the *forward algorithm* (see Algorithm 3.1 on page 77), which maintains a posterior distribution over states, and predicts each outcome with a mixture of the experts' predictions. Given a posterior $\lambda_t(Q_t) = \mathcal{A}(Q_t | x_{1:t-1})$ for the hidden state at time t , the forward algorithm predicts x_t using the marginal of $\mathcal{A}(Q_t, E_t, X_t | x_{1:t-1})$. Then, after observing outcome x_t , it updates its posterior λ_t for Q_t to a posterior λ_{t+1} for Q_{t+1} .

For finite \mathcal{Q} , \mathcal{E} and \mathcal{X} , the running time of the algorithm is determined by this last posterior update step, which in general may require $O(|\mathcal{Q}|^2)$ computation steps for each round t . On T outcomes, this gives a total running time of $O(|\mathcal{Q}|^2 \cdot T)$. In Appendix 4.A we provide a more careful analysis.

4.5 Freezing & Sleeping

Let $x_{1:T} = x_1, \dots, x_T$ be a sequence of data and suppose that a reference partition \mathbb{C} of $1:T$ is given in advance. We are interested in the performance of a structured expert $\mathcal{A}_{\mathbb{C}}$, which for each cell $\mathcal{C} \in \mathbb{C}$ runs a separate instance of the structured expert \mathcal{A} on the subsequence $x_{\mathcal{C}}$. This leaves unspecified, however, whether the original timing of $x_{\mathcal{C}}$ should be preserved when $x_{\mathcal{C}}$ is presented to \mathcal{A} . This is a modelling choice,

Figure 4.6 Sleeping and Freezing EHMMs on outcomes $x_{\{2,4,5,\dots\}}$ 

which depends on the application at hand. We therefore treat both the case where the timing is preserved, which we call *sleeping*, and the case where the timing is not preserved, which we call *freezing*. (See also Figure 4.2 in the introduction.)

Sleeping We say that the instance of \mathcal{A} that is used to predict cell \mathcal{C} is sleeping if it does notice the passing of time during outcomes outside of \mathcal{C} , even though it does not observe them. We write $\mathcal{A}_{\mathcal{C}}^{\text{sl}}$ for the resulting EHMM, which is shown in Figure 4.6a for the example $\mathcal{C} = \{2, 4, 5, \dots\}$. Notice that $\mathcal{A}_{\mathcal{C}}^{\text{sl}}$ contains all five states $Q_{1:5}$, even though it does not observe x_1 or x_3 . This has the effect that state transitions from e.g. Q_2 to Q_4 are composed of two transition steps according to p_{\rightarrow} . The distributions on individual cells combine into the following distribution on all data $x_{1:T}$:

$$\mathcal{A}_{\mathcal{C}}^{\text{sl}}(x_{1:T}) := \prod_{\mathcal{C} \in \mathcal{C}} \mathcal{A}_{\mathcal{C}}^{\text{sl}}(x_{\mathcal{C}}).$$

To memorise the nature of sleeping, one may think of the way television channels get interleaved as you zap between them: a channel not being

watched is not paused, but instead continues broadcasting even when its content is not observed.

Freezing In freezing, the instance of \mathcal{A} that is used to predict cell $\mathcal{C} \in \mathbb{C}$ is frozen when outcomes outside of \mathcal{C} occur: its internal state should not change based on those outcomes. (Of course we have no control over the base experts on which \mathcal{A} is based, so they may do whatever they please with such data. We therefore do have to preserve the timing of the base experts' predictions.) The resulting EHMM $\mathcal{A}_{\mathcal{C}}^{\text{fr}}$ is shown for the example $\mathcal{C} = \{2, 4, 5, \dots\}$ in Figure 4.6b. Note that Q_2 , Q_4 and Q_5 are the first, second and third state of $\mathcal{A}_{\mathcal{C}}^{\text{fr}}$; state transitions between them consist of a single transition step according to p_{\cdot} . The resulting distribution on all data is defined by

$$\mathcal{A}_{\mathcal{C}}^{\text{fr}}(x_{1:T}) := \prod_{\mathcal{C} \in \mathbb{C}} \mathcal{A}_{\mathcal{C}}^{\text{fr}}(x_{\mathcal{C}}).$$

One might associate freezing with the way different e-mail conversations get interleaved in your inbox (if it is sorted by order of message arrival): a conversation about your latest research is paused (remains frozen) regardless of how much spam you receive in between.

4.5.1 An Infeasible Solution

The freezing or sleeping distributions can be computed if the reference partition \mathbb{C} is given in advance. The problem we are addressing, however, is that we do not assume \mathbb{C} to be known. An easy (but impractical) solution to this problem is to predict according to the Bayesian mixture of all possible partitions: let w be a prior on the set of all possible partitions and predict such that the joint distribution on all data is given by

$$\mathfrak{B}(x) := \sum_{\mathbb{C}} w(\mathbb{C}) \mathcal{A}_{\mathbb{C}}^{\text{f/s}}(x),$$

where f/s denotes either fr for freezing or sl for sleeping. Lower bounding the sum by the term for the reference partition \mathbb{C} directly gives an upper bound on the log loss:

$$-\log \mathfrak{B}(x) \leq -\log w(\mathbb{C}) - \log \mathcal{A}_{\mathbb{C}}^{\text{f/s}}(x).$$

To predict according to \mathfrak{B} in general would require an exponential amount of state to keep track of all possible partitions, which is completely impractical. In the following section we therefore present generalisations to both sleeping and freezing of the mixing past posteriors algorithm and show that their running time is comparable to that of the forward algorithm on \mathcal{A} itself. Then in section Section 4.5.3 we prove bounds that relate the additional loss to the encoding cost of the reference partition \mathcal{C} .

4.5.2 The EPP Algorithm

Here we present a generalisation of the mixing past posteriors (MPP) algorithm, which we call *evolving past posteriors* (EPP). It is based on the view that MPP internally uses the Bayesian mixture of base experts, which is a standard EHMM. Given this perspective and after making the distinction between sleeping and freezing, the generalisation to other EHMMs is straightforward. We will discuss the connections between MPP and EPP in more detail in Section 4.5.4.

The EPP algorithm has variants for sleeping and freezing, which are both given in Algorithm 4.1. It takes an EHMM \mathcal{A} and mixing scheme β (see Section 4.3.1) as input. Given a distribution λ_t on the hidden state Q_t at time t , the EPP algorithm predicts X_t exactly like the forward algorithm. It differs from the forward algorithm, however, in the way it computes λ_t . Whereas in the forward algorithm λ_t may be interpreted as the posterior distribution on Q_t , in the EPP algorithm λ_t is a β -mixture of *the algorithm's own past posteriors*. This recursive nature of EPP, which it inherits from the MPP algorithm, makes it hard to analyse.

We denote by $P_{\mathcal{A}}^{\text{fr}}$ and $P_{\mathcal{A}}^{\text{sl}}$ the probability distributions on random variables $\langle Q_t, E_t, X_t \rangle_{t \in \mathbb{N}}$ defined by EPP-FREEZING and EPP-SLEEPING on EHMM \mathcal{A} and mixing scheme β . For both $\text{f/s} \in \{\text{sl}, \text{fr}\}$

$$P_{\mathcal{A}}^{\text{f/s}}(q_{1:T}, e_{1:T}, x_{1:T}) = \prod_{t \in 1:T} p_t^{\text{alg}}(q_t, e_t, x_t).$$

4.5.2.1 Representation Invariance

Let \mathcal{A}^1 and \mathcal{A}^2 be EHMMs that are based on the same set of experts \mathcal{E} , but have different state spaces. We call \mathcal{A}^1 and \mathcal{A}^2 *equivalent* if

Algorithm 4.1 Evolving past posteriors (EPP)

Input:

- An EHMM \mathcal{A} with components p_{\circ} , p_{\rightarrow} and p_{\downarrow} (see Section 4.4)
- A mixing scheme β_1, β_2, \dots (see Section 4.3.1 and Section 4.5.2.2)
- Expert predictions $p_1^{\mathcal{E}}, p_2^{\mathcal{E}}, \dots$ and data x_1, x_2, \dots

Output: Predictions $p_1^{\text{alg}}, p_2^{\text{alg}}, \dots$ **Storage:** Past posteriors π_1, π_2, \dots on \mathcal{Q} , the states of \mathcal{A}

Algorithm

- 1: Set the first posterior to the initial distribution of
- \mathcal{A}

$$\pi_1(q_1) \leftarrow p_{\circ}(q_1)$$

- 2:
- for**
- $t = 1, 2, \dots$
- do**

- 3: Form
- λ_t
- , the current configuration, as the
- β_t
- mixture of past posteriors:

$$\lambda_t(q_t) \leftarrow \sum_{0 \leq j < t} \beta_t(j) \pi_{j+1}(q_t).$$

- 4: Compute
- p_t^{alg}
- , the joint distribution on states, experts and outcomes:

$$p_t^{\text{alg}}(q_t, e_t, x_t) \leftarrow \lambda_t(q_t) p_{\downarrow}^{q_t}(e_t) p_t^{e_t}(x_t).$$

- 5: Predict
- x_t
- using the marginal
- $p_t^{\text{alg}}(x_t)$
- ,

- 6: Observe
- x_t
- . Suffer log loss

$$\ell_t^{\text{alg}} \leftarrow -\log(p_t^{\text{alg}}(x_t)).$$

- 7: Perform loss update and state evolution to obtain the next posterior

$$\pi_{t+1}(q_{t+1}) \leftarrow \sum_{q_t \in \mathcal{Q}} p_t^{\text{alg}}(q_t | x_t) p_{\rightarrow}^{q_t}(q_{t+1}).$$

- 8: Only for sleeping: perform state evolution for all past posteriors (
- $1 \leq j \leq t$
-)

$$\pi_j(q_{t+1}) \leftarrow \sum_{q_t \in \mathcal{Q}} \pi_j(q_t) p_{\rightarrow}^{q_t}(q_{t+1}).$$

- 9:
- end for**
-

Table 4.1 Mixing schemes

Mixing scheme	$\beta_{t+1}(t)$	$\beta_{t+1}(j)$ for $0 \leq j < t$
Yesterday	1	0
Fixed Share(α)	$1 - \alpha$	α if $j = 0$ and 0 o.w.
Uniform past(α)	$1 - \alpha$	α/t
Decaying past(α, γ)	$1 - \alpha$	$\alpha(t-j)^{-\gamma}/Z_t$

$\mathcal{A}^1(e_{1:T}) = \mathcal{A}^2(e_{1:T})$ for all $e_{1:T}$. Consequently, equivalent EHMMs assign the same probability $\mathcal{A}^1(x_{1:T}) = \mathcal{A}^2(x_{1:T})$ to all data $x_{1:T}$, hence the difference between \mathcal{A}^1 and \mathcal{A}^2 is merely a matter of *representation*. As an important sanity check, we need to verify that EPP on either EHMM issues the same predictions.

4.5.1. THEOREM (Invariance). *Let f/s denote either fr or sl. Fix equivalent EHMMs \mathcal{A}^1 and \mathcal{A}^2 . Then for all data $x_{1:T}$*

$$P_{\mathcal{A}^1}^{f/s}(x_{1:T}) = P_{\mathcal{A}^2}^{f/s}(x_{1:T}).$$

Proof. Given in Appendix 4.C. □

Thus, from the perspective of predictive performance, the difference between \mathcal{A}^1 and \mathcal{A}^2 is irrelevant. Of course, it does matter for the computational cost of EPP, see Section 4.5.2.3.

4.5.2.2 Mixing Schemes

Bousquet and Warmuth [19] provide an extensive discussion of possible mixing schemes. Their loss bounds for various schemes carry over directly to our setting. It is interesting, however, to analyse the running times of the Fixed-Share to *uniform past* and to *decaying past* mixing schemes for EPP. For further information we refer the reader to [19].

Both schemes (see Table 4.1) depend on a *switching rate* $\alpha \in [0, 1]$, which determines whether to continue with yesterday's posterior or switch back to an earlier one: $\beta_{t+1}(t) = 1 - \alpha$ and $\sum_{0 \leq j < t} \beta_{t+1}(j) = \alpha$.

Uniform Past Given the choice to switch back, the uniform past mixing scheme gives equal weights to the entire past: $\beta_{t+1}(j) = \alpha/t$ for $0 \leq j < t$.

Decaying Past The decaying past scheme assigns larger weight to the recent past: $\beta_{t+1}(j) = \alpha(t-j)^{-\gamma}/Z_t$ for $0 \leq j < t$, where $Z_t = \sum_{0 \leq j < t} (t-j)^{-\gamma}$ is a normalising constant and $\gamma \geq 0$ is a parameter that determines the rate of decay.

4.5.2.3 Running Times

Appendix 4.A provides a detailed comparison of the running times and space requirements of EPP and the forward algorithm. The upshot is that for the uniform past mixing scheme the sleeping variant of EPP is as efficient as the forward algorithm, in terms of both running time and space requirements; the freezing variant is equally efficient if the set of hidden states \mathcal{Q} is finite, but may be a factor $O(T)$ less efficient on T outcomes for countably infinite \mathcal{Q} . The decaying past mixing scheme is a factor $O(T)$ less efficient (for both time and space) than uniform past in all cases, but may be approximated by a scheme described in [19] that reduces this factor to $O(\log T)$.

4.5.3 Loss Bound

We relate the performance of EPP-FREEZING and EPP-SLEEPING (defined in Algorithm 4.1) to that of $\mathcal{A}_C^{\text{fr}}$ and $\mathcal{A}_C^{\text{sl}}$ for all partitions \mathbf{C} jointly.

4.5.2. THEOREM (EPP Loss Bounds). *For both f/s $\in \{\text{fr}, \text{sl}\}$ and any mixing scheme β , data $x_{1:T}$ and expert predictions $\mathbf{p}_{1:T}^\mathcal{E}$*

$$P_{\mathcal{A}}^{\text{f/s}}(x_{1:T}) \geq \sum_{\mathbf{C}} \beta(\mathbf{C}) \mathcal{A}_C^{\text{f/s}}(x_{1:T}). \quad (4.5)$$

Proof. Given in Appendix 4.B. □

Using this bound, we can relate the predictive performance of EPP-SLEEPING and EPP-FREEZING to that of $\mathcal{A}_C^{\text{sl}}$ and $\mathcal{A}_C^{\text{fr}}$ for any reference partition \mathbf{C} .

4.5.3. COROLLARY. $P_{\mathcal{A}}^{\text{f/s}}(x_{1:T}) \geq \beta(\mathbf{C}) \mathcal{A}_C^{\text{f/s}}(x_{1:T})$.

From the brutal way in which Corollary 4.5.3 was obtained, we may expect to often do much better in practice; *many* partitions may contribute significantly to (4.5).

4.5.4 Recovering MPP

We now substantiate our claim that EPP generalises MPP by proving that MPP results from running EPP-FREEZING or EPP-SLEEPING on the Bayesian EHMM (Example 4.4.2).

4.5.4. THEOREM. *Let \mathcal{A} be the Bayesian EHMM with initial distribution w , and let P_w^{MPP} denote the probability distribution defined by MPP with prior w . Then for all data $x_{1:T}$*

$$P_{\mathcal{A}}^{\text{fr}}(x_{1:T}) = P_{\mathcal{A}}^{\text{sl}}(x_{1:T}) = P_w^{\text{MPP}}(x_{1:T}).$$

Proof. The difference between freezing and sleeping (line 8) evaporates since state evolution is the identity operation. By identifying states and experts the MPP algorithm [19, Figure 1] remains. \square

The theorem does not require the set of experts \mathcal{E} to be finite. If \mathcal{E} is infinite (or too large), MPP is intractable. Still, a small EHMM may exist that implements Bayes (say with the uniform prior) on \mathcal{E} , and we can use EPP-SLEEPING (which is faster than EPP-FREEZING) for sequential prediction. For example, we may implement MPP on the infinite set of Bernoulli experts efficiently, in time $O(T^2)$, using EPP-SLEEPING on the *universal element-wise mixture* EHMM of [100, §4.1].

4.5.4.1 Improved MPP Loss Bound

[19, Theorem 7] (our Theorem 4.3.1) bounds the overhead of MPP over Freund’s full scheme in terms of $\beta(\mathbb{C})$, the complexity of the reference partition \mathbb{C} according to the mixing scheme β . A more general bound follows directly from Theorems 4.5.2 and 4.5.4:

4.5.5. COROLLARY.
$$P_w^{\text{MPP}}(x_{1:T}) \geq \sum_{\mathbb{C}} \beta(\mathbb{C}) P_{\mathbb{C}}^{\text{B}}(x_{1:T}).$$

Even with a fixed reference partition \mathbb{C} in mind, we get a better bound by considering small modifications of \mathbb{C} , e.g. finer partitions or partitions that disagree about a single round.

Adversarial Experts For each number of rounds T one can construct a set of T base experts and data $x_{1:T}$ such that the loss of Freund’s full scheme is infinite for all partitions except the finest one. We simply

have expert t suffer infinite loss in all rounds other than t . In this pathological case the bounds in Theorem 4.3.1 for that partition and Corollary 4.5.5 are equal and tight.

4.5.4.2 Is EPP strictly more general than MPP?

A natural question is whether either EPP-SLEEPING or EPP-FREEZING can be implemented using MPP on a rich set of meta-experts. To preclude the trivial answer that regards either algorithm as a single meta-expert, we ask for a fixed construction that works for all mixing schemes.

Sleeping For any EHMM \mathcal{A} , EPP-SLEEPING can be reduced to MPP on meta-experts. Let the set of meta-experts be \mathcal{Q}^∞ , the set of paths through the hidden states of \mathcal{A} . Each meta-expert $q_{\mathbb{N}}$ predicts x_t using the $p_{\downarrow}^{q_t}$ -mixture of base expert predictions. We set the prior w in MPP equal to the marginal probability measure of \mathcal{A} on paths (as determined by p_{\circ} and p_{\rightarrow}). We omit the proof that the predictions made by MPP on these meta-experts with prior w are equal to those made by EPP on \mathcal{A} .

Freezing The next example shows that EPP-FREEZING really is more general than MPP. Fix two experts $\mathcal{E} = \{a, b\}$. Consider the EHMM \mathcal{A} that predicts the first outcome using expert a , and the second outcome using expert b , i.e. $\mathcal{Q} = \mathcal{E}$, and $p_{\circ}(a) = p_{\rightarrow}(b) = p_{\downarrow}(q) = 1$. Running EPP-FREEZING on \mathcal{A} results in $\pi_2(b) = \pi_1(a) = 1$, so that the first outcome is predicted using expert a , and the second outcome is predicted using the β_2 -mixture of experts. Thus any candidate meta-expert *must* predict the first outcome using base expert a . But that means that for MPP with prior w on meta-experts, the loss update has no effect, so that $w = \pi_1 = \pi_2 = \lambda_2$. Hence the second outcome will be predicted according to the prior mixture of experts. Since β_2 is arbitrary and w is fixed, there can be no general scheme to reduce EPP-FREEZING to MPP.

4.6 Other Loss Functions

We will now show how the EPP algorithm for logarithmic loss can be directly translated into an algorithm with corresponding loss bound for any other mixable loss function. The same construction works for any

logarithmic loss algorithm that predicts according to a mixture of the experts' predictions at each trial and whose predictions only depend on the experts' past losses on outcomes that actually occurred.

Mixability A loss function $\ell: \mathcal{A} \times \mathcal{X} \rightarrow [0, \infty]$ is called η -mixable for $\eta > 0$ if any distribution p on experts \mathcal{E} can be mapped to a single action $\text{Pred}(p) \in \mathcal{A}$ in a way that guarantees that

$$\ell(\text{Pred}(p), x) \leq -\frac{1}{\eta} \log \mathbb{E}_{e \sim p} \left[\exp(-\eta \ell(a^e, x)) \right] \quad (4.6)$$

for all outcomes $x \in \mathcal{X}$ and expert predictions a^e . It is called *mixable* if it is η -mixable for some $\eta > 0$ [25]. Mixability ensures that expert predictions for ℓ loss can be mixed in essentially the same way as for log loss.

For example, logarithmic loss itself is 1-mixable. And for $\mathcal{A} = [0, 1]$ and $\mathcal{X} = \{0, 1\}$ the *square loss* $\ell(a, x) := (a - x)^2$ is 2-mixable and the *Hellinger loss* $\ell(a, x) := ((\sqrt{1-x} - \sqrt{1-a})^2 + (\sqrt{x} - \sqrt{a})) / 2$ is $\sqrt{2}$ -mixable.[75, 25]

The Benefits of Lying Given data $x_{1:t}$ and expert predictions $a_{1:t}^e$, let $\ell_{1:t}^e := \ell(a_1^e, x_1), \dots, \ell(a_t^e, x_t)$ denote the sequence of losses of expert e , and let $\ell_{1:t}^{\mathcal{E}}$ denote these losses jointly for all experts. In the special case that ℓ is the logarithmic loss we write $\ell \ell_{1:t}^e$ and $\ell \ell_{1:t}^{\mathcal{E}}$, respectively.

Suppose ALG is an algorithm for log loss that predicts each outcome x_t by mixing the experts' predictions p_t^e according to the distribution $p_t^{\text{alg}}[x_{<t}, \ell \ell_{<t}^{\mathcal{E}}]$ on *experts*. The square-bracket expression indicates that p_t^{alg} may depend on the past outcomes $x_{1:t-1}$ and the losses of the experts on these outcomes, but not on the experts' past or current predictions in any other way. Following this convention, the algorithm predicts x_t using:

$$p_t^{\text{alg}}[x_{<t}, \ell \ell_{<t}^{\mathcal{E}}](x_t) := \sum_e p_t^{\text{alg}}[x_{<t}, \ell \ell_{<t}^{\mathcal{E}}](e) p_t^e(x_t).$$

Now for any game with η -mixable loss ℓ and the same set of experts \mathcal{E} , we can derive from ALG an algorithm ALG_ℓ^η that predicts x_t according to

$$a_t^{\text{alg}_\ell^\eta} := \text{Pred} (p_t^{\text{alg}}[x_{<t}, \eta \ell_{<t}^{\mathcal{E}}]).$$

Note that ALG_ℓ^η is lying to ALG: while ALG thinks it is playing a game for log loss in which experts have incurred log losses $\eta\ell_{<t}^\mathcal{E}$, in reality ALG_ℓ^η is playing a game for loss ℓ and is feeding ALG fake inputs and redirecting ALG's outputs. Let us now analyse the loss of the derived algorithm ALG_ℓ^η .

4.6.1. LEMMA (Other Loss Functions). *Suppose ALG is an algorithm for logarithmic loss that predicts according to*

$\mathbf{p}_t^{\text{alg}}[x_{<t}, \ell\ell_{<t}^\mathcal{E}]$ *at each time t , ℓ is an η -mixable loss function, and $f(x_{1:T}, \ell_{1:T}^\mathcal{E})$ is an arbitrary function that maps outcomes and expert losses to real numbers. Then any log loss bound for ALG of the form*

$$-\log \mathbf{P}^{\text{alg}}(x_{1:T}) \leq f(x_{1:T}, \ell\ell_{1:T}^\mathcal{E}) \quad \text{for all } \mathbf{p}_{1:T}^\mathcal{E}, \quad (4.7)$$

directly implies the ℓ loss bound for ALG_ℓ^η :

$$\ell(a_{1:T}^{\text{alg}_\ell^\eta}, x_{1:T}) \leq \frac{1}{\eta} f(x_{1:T}, \eta\ell_{1:T}^\mathcal{E}) \quad \text{for all } a_{1:T}^\mathcal{E}. \quad (4.8)$$

Proof. Construct a log loss game in which at any time t each expert e predicts according to a distribution \mathbf{p}_t^e such that $\mathbf{p}_t^e(x_t) = \exp(-\eta\ell_t^e)$ for the actual outcome x_t and \mathbf{p}_t^e is arbitrary on other outcomes such that $\sum_{x_t} \mathbf{p}_t^e(x_t) = 1$. In this game the log loss of ALG is

$$-\log \mathbf{P}^{\text{alg}}(x_{1:T}) = \sum_{t \in 1:T} -\log \mathbf{p}_t^{\text{alg}}[x_{<t}, \eta\ell_{<t}^\mathcal{E}](x_t).$$

By η -mixability of ℓ

$$\begin{aligned} \ell(a_{1:T}^{\text{alg}_\ell^\eta}, x_{1:T}) &= \sum_{t \in 1:T} \ell\left(\text{Pred}\left(\mathbf{p}_t^{\text{alg}}[x_{<t}, \eta\ell_{<t}^\mathcal{E}]\right), x_t\right) \\ &\leq \frac{1}{\eta} \sum_{t \in 1:T} -\log \mathbf{p}_t^{\text{alg}}[x_{<t}, \eta\ell_{<t}^\mathcal{E}](x_t). \end{aligned} \quad (4.9)$$

Combining with (4.7) and (4.9) completes the proof. \square

Algorithms that satisfy the requirements of the lemma include Bayes, follow the (perturbed) leader, the forward algorithm, MPP and EPP. An algorithm that does not satisfy them is the last-step minimax algorithm [172], because it takes into account the experts' predictions on outcomes that do not occur.

In the literature it is common to construct algorithms for arbitrary mixable losses and point out their probabilistic interpretation for the special case of log loss [75, 80, 19]. Instead, we have proceeded the other way around: first we derived results for log loss and then we showed that they generalise to other losses. This allowed us to draw on concepts and results from probability theory like conditional probabilities, HMMs and the forward algorithm, without reproving them in a more general setting.

Lemma 4.6.1 generalises results by Vovk [183], who shows that the most important loss bounds for Bayes with logarithmic loss can actually also be derived for arbitrary mixable losses. Our algorithm ALG plays a role similar to his APA algorithm.

4.7 Discussion

Relearning vs Continuing to Learn Corollary 4.5.3 bounds the regret of EPP with respect to a reference partition \mathbb{C} by $-\log \beta(\mathbb{C})$. Consider the asymptotic behaviour of this bound if \mathbb{C} has infinitely many shifts. (A shift occurs when $\text{prev}^{\mathbb{C}}(t+1) \neq t$.) For both decaying past with $\gamma \leq 1$ (e.g. following recommendations in [19]) and uniform past (see Table 4.1) $\max_{0 \leq j < t} \beta_{t+1}(j)$ goes to zero as a function of t . Thus, the cost per shift (be it to continue an earlier cell or to start a new one) grows without bound. On the other hand for fixed share $\beta_{t+1}(0) = \alpha$ for all t , hence fixed share can start a new cell at fixed cost. It depends on the structured expert whether continuing previously selected cells at increasing cost is advantageous over relearning from scratch after each shift at fixed cost. For EHMM experts with a finite state space \mathcal{Q} (including Bayes), relearning from scratch will cost at most a factor $|\mathcal{Q}|$ over learning on. This factor is constant, so that fixed share will eventually win.

4.8 Conclusion

We revisited Freund's problem, which asks for a strategy for prediction with expert advice that suffers small additional loss compared to Freund's reference scheme. We discussed the solution by Bousquet and Warmuth, which interprets the experts as black boxes. We proposed

a new interpretation of Freund's scheme which is natural for learning experts, namely to train experts on the subsequence on which they are evaluated. This allows the reference scheme to exploit local patterns in the data, and thus makes the problem harder.

We solved Freund's problem for structured experts that are represented as EHMMs, building on the work of Bousquet and Warmuth. We showed that our prediction strategies are efficient, and have desirable loss bounds that apply to all mixable losses.

4.A Running Times

We compare the running times on T outcomes of EPP and the forward algorithm, with respect to an arbitrary EHMM \mathcal{A} with a countable set of hidden states \mathcal{Q} . For simplicity we assume that the sets of experts \mathcal{E} and outcomes \mathcal{X} are finite.

Let Q_t denote the hidden state of \mathcal{A} at time t , and let p_o , p_{\rightarrow} , and p_{\downarrow} denote \mathcal{A} 's other components. Both algorithms base their predictions on a distribution λ_t on Q_t at time t , but differ in how they update λ_t after observing x_t . As the number of computations for this step depends on the size of the support of λ_t and on p_{\rightarrow} , we will need the following concepts. For any probability distribution p on \mathcal{Q} , let $\text{Sp}(p) = \{q \in \mathcal{Q} \mid p(q) > 0\}$ denote its support. We recursively define Q_t , the set of states reachable in exactly t steps, and $Q_{\leq t}$, the set of states reachable in at most t steps, by

$$Q_1 := \text{Sp}(p_o), \quad Q_{t+1} := \bigcup_{q \in Q_t} \text{Sp}(p_{\rightarrow}^q), \quad Q_{\leq t} := \bigcup_{i \in 1:t} Q_i.$$

Obviously, $Q_t \subseteq Q_{\leq t} \subseteq \mathcal{Q}$ holds for all t . Let $g(S) := \sum_{q \in S} |\text{Sp}(p_{\rightarrow}^q)|$ be the number of outgoing transitions from any set of states $S \subseteq \mathcal{Q}$.

4.A.1 Forward

The forward algorithm computes λ_{t+1} by conditioning λ_t on x_t and applying the transition function p_{\rightarrow} . As λ_t has support Q_t , the forward algorithm requires $O(g(Q_t))$ work per time step, and $O(|Q_t| + |Q_{t+1}|)$ space. Notice that, for finite \mathcal{Q} , the number of transitions is bounded by $g(S) \leq |\mathcal{Q}|^2$ for any S . A rough upper bound on the total running

time of forward on T outcomes is therefore $O(|\mathcal{Q}|^2 T)$, which is linear in T .

4.A.2 EPP

The EPP algorithm comes in two variants: one for sleeping and one for freezing. For sleeping the order of the running time is determined by the evolution of past posteriors (line 8 in Algorithm 4.1); for freezing, which skips line 8, either computation of λ_t (line 3) or of the next posterior (line 7) is the dominant step. The main difference for the running times of the two variants, however, is that in sleeping π_j has support \mathcal{Q}_t at any time t , whereas for freezing π_j has support $\mathcal{Q}_{\leq j}$.

4.A.2.1 Uniform Past

For the uniform past mixing scheme, one can keep track of $\sum_{j=0}^t \pi_j(q_t)$ to speed up computation of λ_{t+1} .

Sleeping This even works for sleeping, because applying the state evolution to this sum in line 8 of the algorithm is equivalent to applying it to the individual π_j and then summing. Consequently, sleeping requires $O(g(\mathcal{Q}_t))$ work and $O(|\mathcal{Q}_t| + |\mathcal{Q}_{t+1}|)$ space per time step, which makes it as efficient as the forward algorithm.

Freezing For freezing, computing the next posterior (line 7) determines the running time. It requires $O(g(\mathcal{Q}_{\leq t}))$ work and $O(|\mathcal{Q}_{\leq t+1}|)$ space per time step. Depending on the EHMM \mathcal{A} , this may be significantly slower than the forward algorithm. First, for finite \mathcal{Q} , each of \mathcal{Q}_t , $\mathcal{Q}_{\leq t}$ and \mathcal{Q} have size $O(1)$ in t , and freezing runs in time $O(T)$, just like the forward algorithm. Second, for infinite \mathcal{Q} , $\mathcal{Q}_{\leq t}$ may be unbounded as a function of t . Still, on T outcomes

$$\sum_{t \in 1:T} g(\mathcal{Q}_{\leq t}) \leq T g(\mathcal{Q}_{\leq T}) \leq T \sum_{t \in 1:T} g(\mathcal{Q}_t),$$

which implies that freezing is no more than a factor T slower than the forward algorithm.

4.A.2.2 Decaying Past

For the decaying past scheme the relative mixing weights of any two past posteriors change from β_t to β_{t+1} , which prevents us from summing them as for uniform past. Implementing decaying past therefore slows down both the evolution of past posteriors and computation of λ_t by a factor of $O(t)$, and increases the required space by the same factor. Fortunately, however, the decaying past scheme can be approximated using a logarithmic number of uniform blocks, as described in Appendix C of [19]. This reduces the slowdown factor from $O(t)$ to $O(\log t)$.¹ Thus, both for sleeping and for freezing, approximated decaying past is only a factor $O(\log T)$ slower than uniform past on T outcomes, and requires only a factor $O(\log T)$ more space.

4.B Loss Bounds

We identify λ_t with the EHMM on $\langle Q_i, E_i, X_i \rangle_{i \geq t}$ with initial distribution λ_t , and with the transition and production functions of \mathcal{A} . So in particular $\lambda_1 = \mathcal{A}$. For convenience, we shorten $(\lambda_t)_{\mathcal{C}}^{\text{fr}}(x_{\mathcal{C}})$ to $\lambda_t^{\text{fr}}(x_{\mathcal{C}})$ and $(\lambda_t)_{\mathcal{C}}^{\text{sl}}(x_{\mathcal{C}})$ to $\lambda_t^{\text{sl}}(x_{\mathcal{C}})$. Thus, among others, $\lambda_t(x_t) = \lambda_t^{\text{sl}}(x_t) = \lambda_t^{\text{fr}}(x_t)$.

4.B.1. LEMMA. *For any $\mathcal{C} \subseteq t:T$, interpreting $\lambda_0(\cdot|x_0)$ as λ_1 ,*

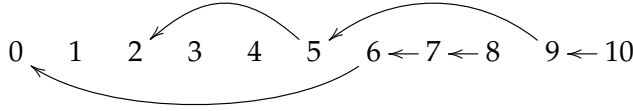
$$\lambda_t^{\text{fr}}(x_{\mathcal{C}}) = \sum_{j \in 0:t-1} \beta_t(j) \lambda_j^{\text{fr}}(x_{\mathcal{C}}|x_j).$$

Proof. Let π_j^t denote the past posterior π_j at the beginning of round t . Thus for freezing $\pi_j^t = \pi_j$, and for sleeping π_j^t is π_j evolved $t-j$ steps. Then by definition $\lambda_t(x_{\mathcal{C}}) = \sum_{j=0}^{t-1} \beta_t(j) \pi_{j+1}^t(x_{\mathcal{C}})$. The operations $(\cdot)^{\text{fr}}$ and $(\cdot)^{\text{sl}}$ distribute over taking mixtures. The lemma follows from the fact that $(\pi_j^t)^{\text{sl}}(x_{\mathcal{C}}) = \pi_j^{\text{sl}}(x_{\mathcal{C}})$ and $(\pi_j^t)^{\text{fr}}(x_{\mathcal{C}}) = \pi_j^{\text{fr}}(x_{\mathcal{C}})$. \square

Proof of Theorem 4.5.2. For any t , we view the mixing scheme β_t as defining the distribution of a randomised choice $j_t \in 0:(t-1)$ for the predecessor of the t th outcome. Let $j_{>k} := j_{k+1:T} = (j_{k+1}, \dots, j_T)$ denote a

¹In [19] it is suggested to weight each block of posteriors $\pi_{[j_1, j_2-1]}$ by $(j_2 - j_1)\beta_t(j_1)$. It seems that a marginal improvement is possible by weighting by $\sum_{j_1 \leq j < j_2} \beta_t(j)$ instead, which can be implemented equally efficiently for decaying past.

Figure 4.7 Notation example. $T = 10$, $k = 4$, $j_{>k} = (2, 0, 6, 7, 5, 9)$, $S(j_{>k}) = \{6\}$, $R_2(j_{>k}) = \{2, 5, 9, 10\}$.



vector of the choices beyond turn k . Unfortunately, some choices of $j_{>k}$ are inconsistent with any partition, because an element can only have one successor in a partition. Thus $j_{>k}$ is inconsistent with any partition if $j_m = j_n > 0$ for $k < m \neq n \leq T$. Let the predicate $I(j_{>k})$ be true iff $j_{>k}$ is consistent with some partition.

Some elements of $j_{>k}$ may indicate the start of a new cell of the partition. Let $S(j_{>k})$ denote the set of times when $j_{>k}$ prescribes to start a new cell, i.e. $S(j_{>k}) := \{t \in k+1:T \mid j_t = 0\}$. For an example, consult Figure 4.7.

Consistent values of $j_{>k}$ specify the last part of a partition. For any $1 \leq t \leq k$, we may ask which of the times $k+1:T$ will be put in the same cell as t . Let $R_t(j_{>k})$ denote this set, including t . For convenience, we abbreviate

$$\begin{aligned} \beta(j_{>k}) &:= \prod_{t \in k+1:T} \beta_t(j_t), \\ W(j_{>k}) &:= \prod_{i \in S(j_{>k})} \lambda_1^{i/s}(x_{R_i(j_{>k})}), & \text{and} \\ U_l(j_{>k}) &:= \prod_{i \in 1:l} \lambda_i^{i/s}(x_{R_i(j_{>k})}) & \text{for all } l \leq k, \end{aligned}$$

to name the intermediate debris arising from the incremental reduction of $P_A^{i/s}(x_{1:T})$. W -terms deal with cells that are completely specified by $j_{>k}$, while U -terms keep track of the remaining partially specified cells. The proof proceeds by downward induction on k , with induction hypothesis

$$\prod_{i \in 1:T} \lambda_i(x_i) \geq \sum_{j_{>k}: I(j_{>k})} \beta(j_{>k}) W(j_{>k}) U_k(j_{>k}). \quad (4.10)$$

For the base case $k = T$ the hypothesis holds with equality, and for $k = 0$ the hypothesis is equivalent to the desired result (4.5). It remains

to verify that it holds for $k - 1 \geq 0$ if it holds for k . To this end, fix $k \geq 1$. To prove (4.10), it suffices to show that for consistent $j_{>k}$

$$W(j_{>k})U_k(j_{>k}) \geq \sum_{j_k: I(j_{\geq k})} \beta_k(j_k)W(j_{\geq k})U_{k-1}(j_{\geq k}),$$

where $j_{\geq k}$ denotes $j_{k:T}$, i.e. j_k followed by $j_{>k}$. We expand the last factor of $U_k(j_{>k})$ using Lemma 4.B.1, and bound

$$\begin{aligned} U_k(j_{>k}) &= \sum_{j_k \in 0:k-1} \beta_k(j_k) \lambda_{j_k}^{f/s}(x_{R_k(j_{>k})} | x_{j_k}) U_{k-1}(j_{>k}) \\ &\geq \sum_{j_k: I(j_{\geq k})} \beta_k(j_k) \lambda_{j_k}^{f/s}(x_{R_k(j_{>k})} | x_{j_k}) U_{k-1}(j_{>k}). \end{aligned}$$

Observe that $R_t(j_{>k}) = R_t(j_{\geq k})$ for all $1 \leq t < k$ except $t = j_k$. There are two cases. If $j_k = 0$, then

$$U_{k-1}(j_{>k}) = U_{k-1}(j_{\geq k}) \quad \text{and} \quad W(j_{>k}) \lambda_1^{f/s}(x_{R_k(j_{>k})}) = W(j_{\geq k}).$$

On the other hand if $j_k > 0$ then $W(j_{>k}) = W(j_{\geq k})$. For consistent $j_{\geq k}$, $U_{k-1}(j_{>k})$ contains the factor $\lambda_{j_k}^{f/s}(x_{j_k})$, which implies that

$$\lambda_{j_k}^{f/s}(x_{R_k(j_{>k})} | x_{j_k}) U_{k-1}(j_{>k}) = U_{k-1}(j_{\geq k}). \quad \square$$

4.C Invariance

Proof of Theorem 4.5.1. Let μ^1 and μ^2 be distributions on \mathcal{Q}^1 and \mathcal{Q}^2 . We overload notation, and write μ^1 and μ^2 for the EHMMs \mathcal{A}^1 and \mathcal{A}^2 with initial distribution replaced by μ^1 and μ^2 . Recall that μ^1 and μ^2 are equivalent if $\mu^1(e_{1:T}) = \mu^2(e_{1:T})$ for all $e_{1:T}$. Thus, \mathcal{A}^1 and \mathcal{A}^2 are equivalent iff p_1^0 and p_2^0 are equivalent.

To prove the theorem, we need to prove that equivalence is preserved by all the operations that EPP performs, i.e. taking mixtures, performing loss update and performing state evolution. Mixtures of equivalent distributions are equivalent, since mixing and marginalisation commute. For loss update, note that $p_1^{\varepsilon_1}(x_1) = \mu^1(x_1 | e_{1:T}) = \mu^2(x_1 | e_{1:T})$ for all $p_1^{\varepsilon_1}$ and all $e_{1:T}$. Finally, for state evolution, the claim follows from $(p_{\rightarrow} \circ \mu)(e_{1:T}) = \mu(E_{2:T+1} = e_{1:T})$. \square