# PUTTING BAYES TO SLEEP

*Wouter M. Koolen*[1] *and Dimitri Adamskiy*[1] *and Manfred K. Warmuth*[2]

[1] Computer Learning Research Centre and Department of Computer Science,
Royal Holloway, University of London, Egham, Surrey, TW20 0EX, UK
[2] Department of Computer Science, University of California Santa Cruz, CA 95064, USA

## ABSTRACT

Consider sequential prediction algorithms that are given the predictions from a set of models as inputs. If the nature of the data is changing over time in that different models predict well on different segments of the data, then adaptivity is typically achieved by mixing into the weights in each round a bit of the initial prior (kind of like a weak restart). However, what if the favored models in each segment are from a *small subset*, i.e. the data is likely to be predicted well by models that predicted well before? Curiously, fitting such "sparse composite models" is achieved by mixing in a bit of all the past posteriors. This self-referential updating method is rather peculiar, but it is efficient and gives superior performance on many natural data sets. Also it is important because it introduces a long-term memory: any model that has done well in the past can be recovered quickly. While Bayesian interpretations can be found for mixing in a bit of the initial prior, no Bayesian interpretation is known for mixing in past posteriors.

We build atop the "specialist" framework from the online learning literature to give the Mixing Past Posteriors update a proper Bayesian foundation. We apply our method to a well-studied multitask learning problem and obtain a new intriguing efficient update that achieves a significantly better bound.

## 1. INTRODUCTION

We consider sequential prediction of outcomes $y_1, y_2, \ldots$ using a set of models $m = 1, \ldots, M$ for this task. In practice $m$ could range over a mix of human experts, parametric models, or even complex machine learning algorithms. In any case we denote the prediction of model $m$ for outcome $y_t$ given past observations $y_{<t} = (y_1, \ldots, y_{t-1})$ by $P(y_t|y_{<t}, m)$. The goal is to design a computationally efficient predictor $P(y_t|y_{<t})$ that maximally leverages the predictive power of these models as measured in log loss. The yardstick in this paper is a notion of *regret* defined w.r.t. a given *comparator class* of models or composite models: it is the additional loss of the predictor over the best comparator. For example if the comparator class is the set of base models $m = 1, \ldots, M$, then the regret for a sequence of $T$ outcomes $y_{\leq T} = (y_1, \ldots, y_T)$ is

$$\mathcal{R} \coloneqq \sum_{t=1}^{T} -\ln P(y_t|y_{<t}) \; - \min_{m=1}^{M} \sum_{t=1}^{T} -\ln P(y_t|y_{<t}, m).$$

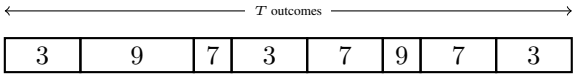
The Bayesian predictor with uniform model prior has regret at most $\ln M$ for all $T$.

Now assume the nature of the data is changing with time: in an initial segment one model predicts well, followed by a second segment in which another model has small loss and so forth. For this scenario the natural comparator class is the set of *partition models* which divide the sequence of $T$ outcomes into $B$ segments and specify the model that predicts in each segment. By running Bayes on all exponentially many partition models comprising the comparator class, we can guarantee regret $\ln \binom{T-1}{B-1} + B \ln M$. The goal then is to find *efficient* algorithms with approximately the same guarantee as full Bayes. In this case this is achieved by the Fixed Share [1] predictor. It assigns a certain prior to all partition models for which the exponentially many posterior weights collapse to $M$ posterior weights that can be maintained efficiently. Modifications of this algorithm achieve essentially the same bound for all $T$, $B$ and $M$ simultaneously [2, 3].

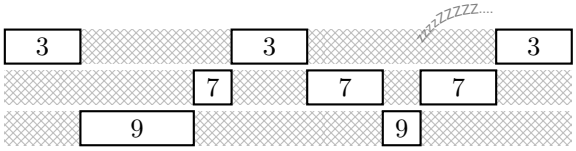
In an open problem Yoav Freund [4] asked whether there are algorithms that have small regret against *sparse* partition models where the base models allocated to the segments are from a small subset of $N$ of the $M$ models. The Bayes algorithm when run on all such partition models achieves regret $\ln \binom{M}{N} + \ln \binom{T-1}{B-1} + B \ln N$, but contrary to the non-sparse case, emulating this algorithm is NP-hard. However in a breakthrough paper, Bousquet and Warmuth in 2001 [4] gave the efficient MPP algorithm with only a slightly weaker regret bound. Like Fixed Share, MPP maintains $M$ "posterior" weights, but it instead mixes in a bit of all past posteriors in each update. This causes weights of previously good models to "glow" a little bit, even if they perform bad locally. When the data later favors one of those good models, its weight is pulled up quickly. However the term "posterior" is a misnomer because no Bayesian interpretation for this curious self-referential update was known. Understanding the MPP update is a very important problem because in many practical applications [5, 6][1] it significantly outperforms Fixed Share.

Our main philosophical contribution is finding a fully Bayesian interpretation for MPP. We employ the special-

[1] The experiments reported in [5] are based on precursors of MPP. However MPP outperforms these algorithms in later experiments we have done on natural data for the same problem (not shown).




(a) A comparator partition model: segmentation and model assignment




(b) Decomposition into 3 partition specialists, asleep at shaded times

ist framework from online learning [7, 8, 9]. So-called *specialist* models are either *awake* or *asleep*. When they are awake, they predict as usual. However when they are asleep, they "go with the rest", i.e. they predict with the combined prediction of all awake models.

Instead of fully coordinated partition models, we construct *partition specialists* consisting of a base model and a set of segments where this base model is awake. The figure to the right shows how a comparator partition model is assembled from partition specialists. We can emulate Bayes on all partition specialists; the NP-completeness is avoided by forgoing a-priori segment synchronization. By carefully choosing the prior, the exponentially many posterior weights collapse to the small number of weights used by the efficient MPP algorithm. Our analysis technique magically aggregates the contribution of the $N$ partition specialists that constitute the comparator partition, showing that we achieve regret close to the regret of Bayes when run on all full partition models. Actually our new insights into the nature of MPP result in slightly improved regret bounds.

We then apply our methods to the online multitask learning problem where a small subset of models from a big set solve a large number of tasks. Again simulating Bayes on all sparse assignments of models to tasks is NP-hard. We split an assignment into *subset specialists* that assign a single base model to a subset of tasks. With the right prior, Bayes on these subset specialists again gently collapses to an efficient algorithm with a regret bound not much larger than Bayes on all assignments. This considerably improves the previous regret bound of [10]. Our algorithm simply maintains one weight per model/task pair and does not rely on sampling (often used for multitask learning).

Why is this line of research important? We found a new intuitive Bayesian method to quickly recover information that was learned before, allowing us to exploit sparse composite models. Moreover, it expressly avoids computational hardness by splitting coordinated composite models into smaller constituent "specialists" that are asleep in time steps outside their jurisdiction. This method clearly beats Fixed Share when *few* base models constitute a partition, i.e. the composite models are sparse.

We expect this methodology to become a main tool for making Bayesian prediction adapt to sparse models. The goal is to develop general tools for adding this type of adaptivity to existing Bayesian models without losing efficiency. It also lets us look again at the updates used in Nature in a new light, where species/genes cannot dare adapt too quickly to the current environment and must guard themselves against an environment that changes or fluctuates at a large scale. Surprisingly these type of updates might now be amenable to a Bayesian analysis. For example, it might be possible to interpret sex and the double stranded recessive/dominant gene device employed by Nature as a Bayesian update of genes that are either awake or asleep.

## 2. REFERENCES

[1] Mark Herbster and Manfred K. Warmuth, "Tracking the best expert," *Machine Learning*, vol. 32, pp. 151–178, 1998.

[2] Paul A.J. Volf and Frans M.J. Willems, "Switching between two universal source coding algorithms," in *Proceedings of the Data Compression Conference, Snowbird, Utah*, 1998, pp. 491–500.

[3] Wouter M. Koolen and Steven de Rooij, "Combining expert advice efficiently," in *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, Rocco Servedio and Tong Zang, Eds., June 2008, pp. 275–286.

[4] Olivier Bousquet and Manfred K. Warmuth, "Tracking a small set of experts by mixing past posteriors," *Journal of Machine Learning Research*, vol. 3, pp. 363–396, 2002.

[5] David P. Helmbold, Darrell D. E. Long, Tracey L. Sconyers, and Bruce Sherrod, "Adaptive disk spindown for mobile computers," *ACM/Baltzer Mobile Networks and Applications (MONET)*, pp. 285–297, 2000.

[6] Robert B. Gramacy, Manfred K. Warmuth, Scott A. Brandt, and Ismail Ari, "Adaptive caching by refetching," in *NIPS*, Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, Eds. 2002, pp. 1465–1472, MIT Press.

[7] Y. Freund, R. E. Schapire, Y. Singer, and M. K. Warmuth, "Using and combining predictors that specialize," in *Proc. 29th Annual ACM Symposium on Theory of Computing*. 1997, pp. 334–343, ACM.

[8] Alexey Chernov and Vladimir Vovk, "Prediction with expert evaluators' advice," in *Proceedings of the 20th international conference on Algorithmic learning theory*, Berlin, Heidelberg, 2009, ALT'09, pp. 8–22, Springer-Verlag.

[9] Alexey Chernov, Yuri Kalnishkan, Fedor Zhdanov, and Vladimir Vovk, "Supermartingales in prediction with expert advice," *Theor. Comput. Sci.*, vol. 411, no. 29-30, pp. 2647–2669, June 2010.

[10] Jacob Ducan Abernethy, Peter Bartlett, and Alexander Rakhlin, "Multitask learning with expert advice," Tech. Rep., University of California at Berkeley, Jan. 2007.