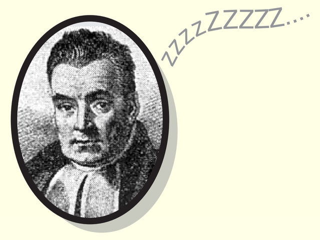


Putting Bayes to sleep

Wouter M. Koolen, Dmitri Adamskiy and Manfred K. Warmuth

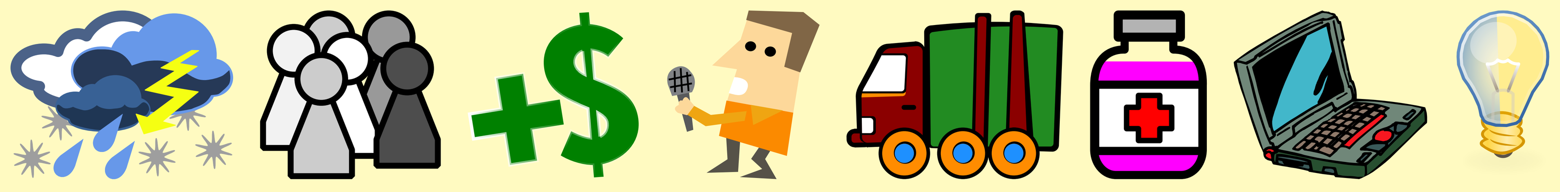
Royal Holloway University of London · University of California Santa Cruz

Poster Tu57

December 4th 2012

Online learning

Real-world tasks:



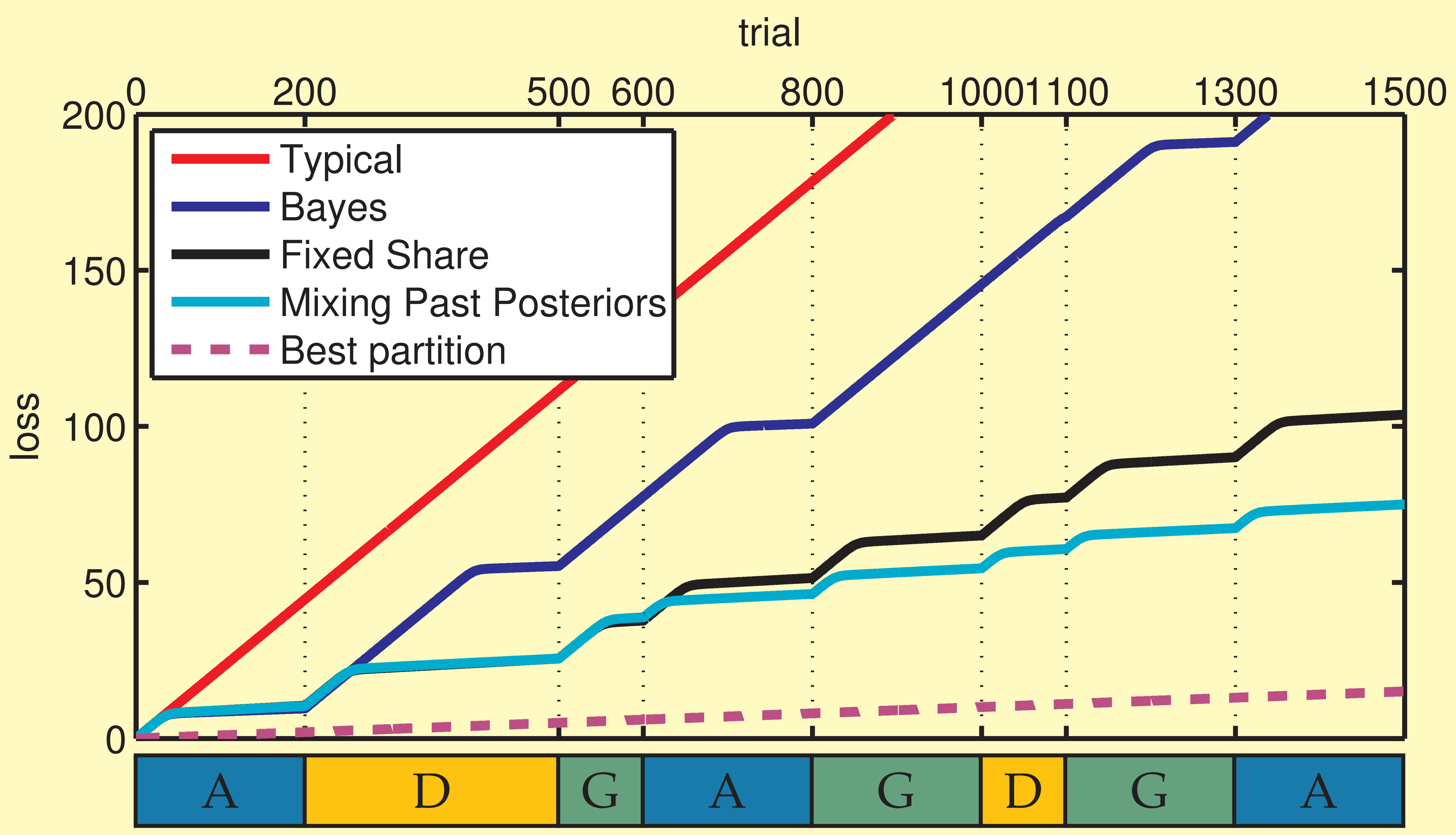
Many models: A, B, ... Best model typically changes with time:

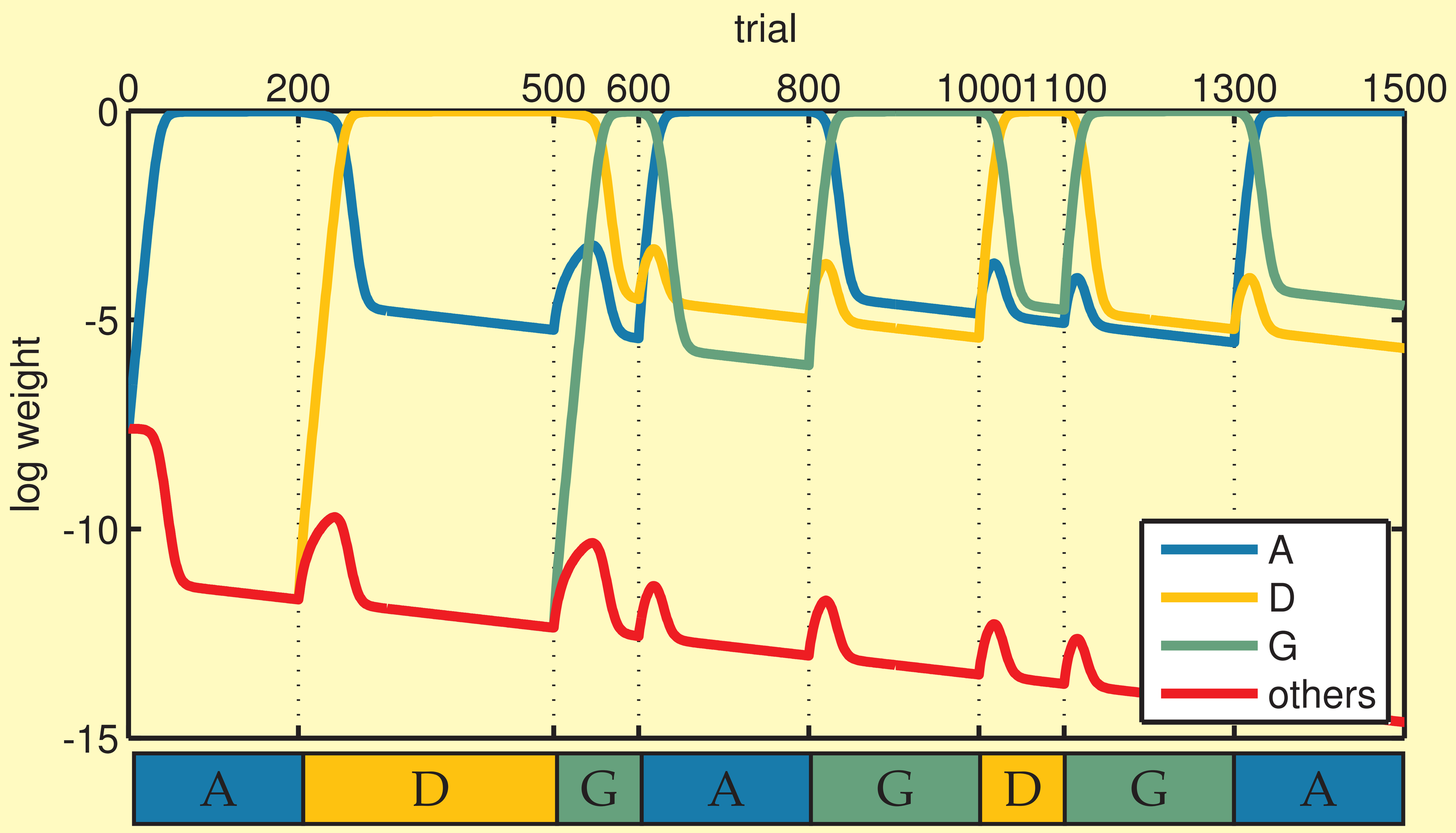
time →

A | D | G | A | G | D | G | A

Need adaptive algorithms that can exploit repeats

Mixing Past Posteriors [BW02] does the job





But while we do understand the simpler algorithms

time →

Bayes: A

fixed model

Fixed Share: A | D | G | A | G | D | G | A

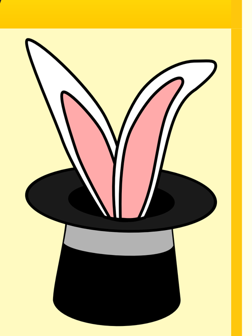
partition model

Fixed Share can be regarded as a Bayesian mixture of all the possible partition models. (Number is exponential in time, but FS **collapses**)

Fixed Share = Bayes update + mix in small amount of initial prior

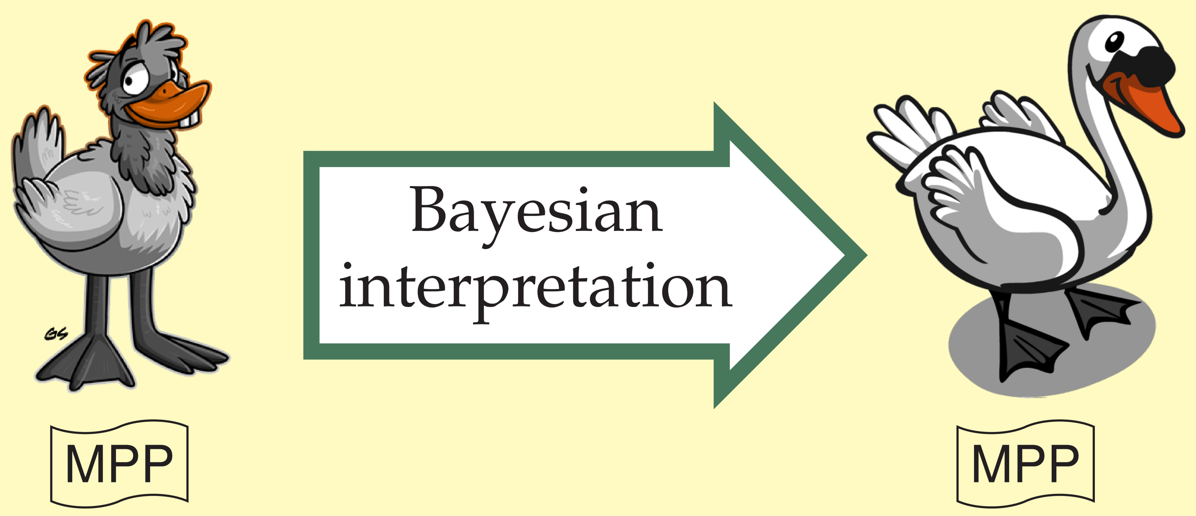
the superior MPP remains a baffling mystery

$$\underbrace{\tilde{w}_{t+1}(m) = \frac{P(y_t|m)w_t(m)}{\sum_m P(y_t|m)w_t(m)}}_{\text{Bayesian posterior update}} \qquad w_t(m) = \underbrace{\sum_{s=0}^{t-1} \tilde{w}_s(m)\gamma_t(s)}_{\text{bizarre}}$$



MPP = Bayes update + mix in small amount of all past posteriors

Our breakthrough



Bayesian interpretation for MPP

We interpret Mixing Past Posteriors as a Bayesian mixture of partition specialists which can be **asleep**:

time → zzzZZZZ....

MPP:

A | | A | | A

| G | G | G |

| D | | D |

NEW

partition specialists

We craft a prior on all partition specialists for which Bayes is both

- **fast**: collapses to $\mathcal{O}(M)$ time per trial, $\mathcal{O}(M)$ space
- **good**: regret close to information-theoretic lower bound

Bayes for specialists crash course

A specialist may or **may not** issue a prediction [FSSW97]. Prediction $P(y|m)$ only available for **awake** $m \in W$.

Key insight: **complete** specialists to full models [CV09]:

$$P(y|m) := P(y) \qquad \text{for all asleep } m \notin W. \qquad \boxed{\text{circular!}}$$

With **prior** $P(m)$ on specialists, the Bayesian **predictive distribution**

$$P(y) = \sum_{m \in W} P(y|m)P(m) + \sum_{m \notin W} P(y)P(m)$$

has solution

$$P(y) = \frac{\sum_{m \in W} P(y|m)P(m)}{\sum_{m \in W} P(m)}.$$

The **posterior distribution** is incrementally updated by

$$P(m|y) = \begin{cases} \frac{P(y|m)P(m)}{P(y)} & \text{if } m \in W, \\ \frac{P(y)P(m)}{P(y)} = P(m) & \text{if } m \notin W. \end{cases}$$

Bayes is **fast**: predict in $\mathcal{O}(M)$ time per round.

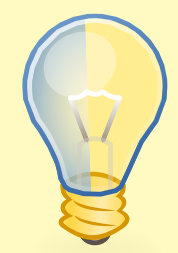
Bayes is **good**: regret w.r.t. specialist m on data $y_{\leq T}$ bounded by

$$\sum_{t \leq T \,:\, m \in W_t} \left(-\ln P(y_t|y_{<t}) + \ln P(y_t|y_{<t}, m)\right) \leq -\ln P(m).$$

Conclusion

Proper Bayesian interpretation of Mixing Past Posteriors using "prediction with specialists"

- Simplified tuning
- Fastest algorithm
- Sharpest bounds
- Mysterious factor 2 in bound explained



Application of specialists technology to multitask learning

- significantly improved bounds
- intriguing collapsed algorithm

all grandiose details are in the paper

Thank you!