Putting Bayes to sleep
Wouter M. Koolen, Dmitri Adamskiy and Manfred K. Warmuth

Online learning
Real-world tasks:
Many strategies: A, B, ...

Need adaptive algorithms that can exploit repeats

Mixing Past Posteriors [BW02]: it works . . .

We interpret Mixing Past Posteriors as running Bayes on exponentially many partition specialists

Bayesian interpretation for MPP

We craft a prior on all partition specialists for which Bayes is fast: collapses to $O(M)$ time per trial, $O(M)$ space
Bayes is 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]:

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

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

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

The posterior distribution is incrementally updated by

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

Bayes is fast: predict in $O(M)$ time per round.
Bayes is good: regret w.r.t. specialist $m$ on data $y < t$ bounded by

$$\sum_{1 \leq t \leq T} (- \ln P(y_t|m) + \ln P(y_t|y_{<t}, m)) \leq - \ln P(m).$$

Conclusion

- Proper Bayesian interpretation of Mixing Past Posteriors using "prediction with specialists"
- Closely skirt NP hardness
- Mysterious factor 2 in bound explained
- Simplified tuning
- Fastest algorithm
- Sharpest bounds
- Application to multitask learning significantly improved bounds

Thank you!