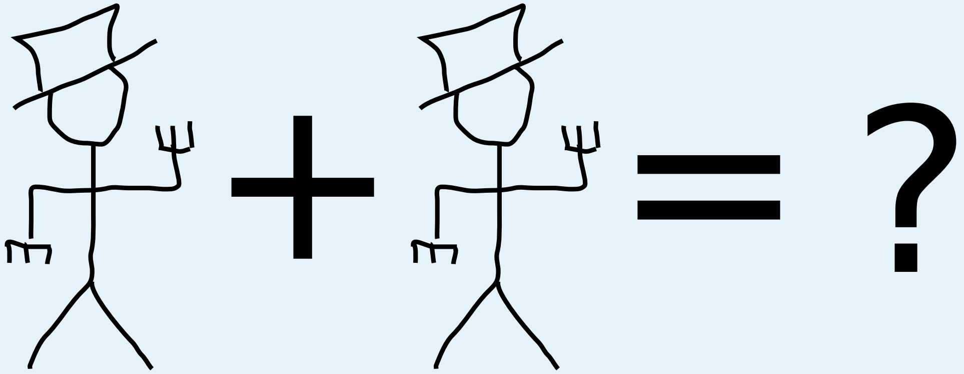
Combining Expert Advice Efficiently

Wouter Koolen-Wijkstra

Joint work with Steven de Rooij

Friday 11 July, 2008



Introduction

Strategies

HMMs

Conclusion

The State of the Art

Prior Art

- Weighted Majority — *Littlestone and Warmuth,1989*
- Aggregating Algorithm — *Vovk, 1990*
- Switching Method — *Volf and Willems, 1998*
- Fixed Share — *Herbster and Warmuth, 1998*
- Universal Share — *Monteleoni and Jaakkola, 2003*
- Switch Distribution — *De Rooij, Van Erven, Grünwald, 2007*

The State of the Art

Prior Art

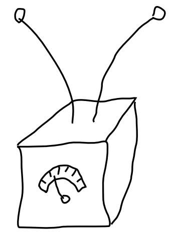
- Weighted Majority — *Littlestone and Warmuth,1989*
- Aggregating Algorithm — *Vovk, 1990*
- Switching Method — *Volf and Willems, 1998*
- Fixed Share — *Herbster and Warmuth, 1998*
- Universal Share — *Monteleoni and Jaakkola, 2003*
- Switch Distribution — *De Rooij, Van Erven, Grünwald, 2007*

Our contribution

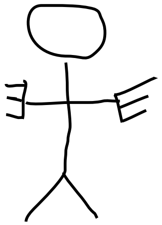
- Unification using ES-priors & HMMs
- Intuitive graphical language

Sequential Prediction

Source

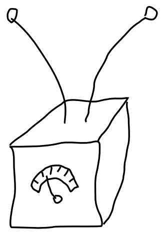


Predictor

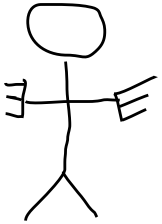


Sequential Prediction

Source

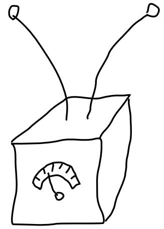


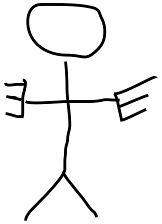
Predictor



$P(X_1)$

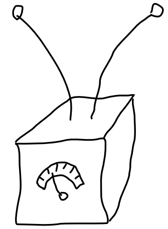
Sequential Prediction

Source  x_1

Predictor  $P(X_1)$

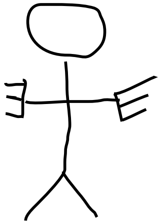
Sequential Prediction

Source



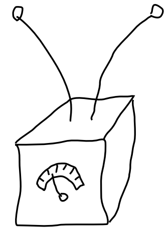
x_1

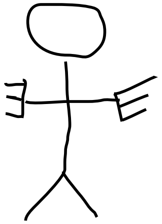
Predictor



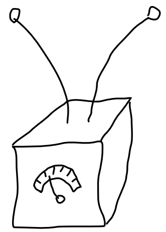
$P(X_1)$ $\quad$ $P(X_2|x_1)$

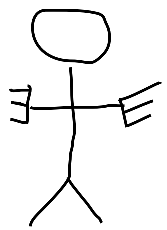
Sequential Prediction

Source  x_1 x_2

Predictor  $P(X_1)$ $P(X_2|x_1)$

Sequential Prediction

Source  x_1 x_2 x_3 $\dots$

Predictor  $P(X_1)$ $P(X_2|x_1)$ $P(X_3|x_1, x_2)$ $\dots$

Evaluating Performance

A good predictor assigns high probability to the data $x^n = x_1, x_2, \ldots, x_n$

$$P(x^n) \;=\; P(x_1)P(x_2|x_1)P(x_3|x^2)\cdots P(x_n|x^{n-1}),$$

Evaluating Performance

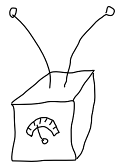
A good predictor assigns high probability to the data $x^n = x_1, x_2, \ldots, x_n$

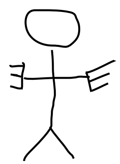
$$P(x^n) = P(x_1)P(x_2|x_1)P(x_3|x^2)\cdots P(x_n|x^{n-1}),$$

or, equivalently, suffers low cumulative log loss

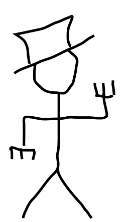
$$-\log P(x^n) = \sum_{i=1}^{n} \underbrace{-\log P(x_i|x^{i-1})}_{\text{Log loss on } x_i}.$$

Sequential Prediction with Experts

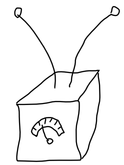
Source 

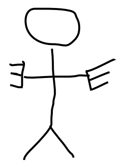
Predictor 

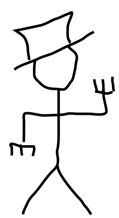
Expert A 

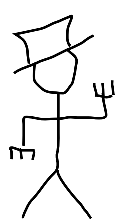
Expert B 

Sequential Prediction with Experts

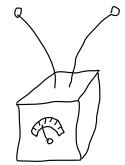
Source 

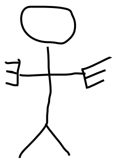
Predictor 

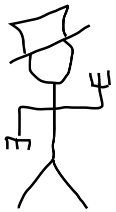
Expert A  $P_{\mathsf{A}}(X_1)$

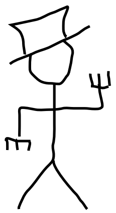
Expert B  $P_{\mathsf{B}}(X_1)$

Sequential Prediction with Experts

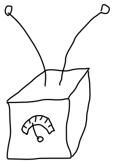
Source 

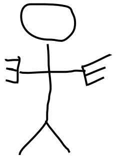
Predictor  $P(X_1)$

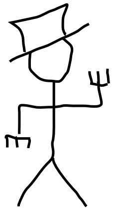
Expert A  $P_{\mathsf{A}}(X_1)$

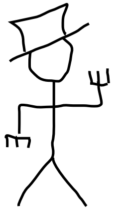
Expert B  $P_{\mathsf{B}}(X_1)$

Sequential Prediction with Experts

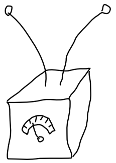
Source  x_1

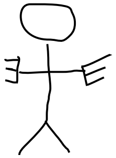
Predictor  $P(X_1)$

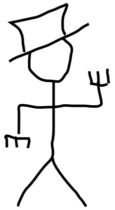
Expert A  $P_{\mathsf{A}}(X_1)$

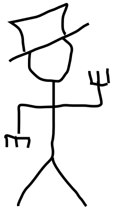
Expert B  $P_{\mathsf{B}}(X_1)$

Sequential Prediction with Experts

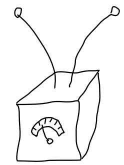
Source  x_1

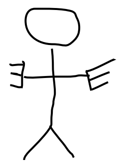
Predictor  $P(X_1)$

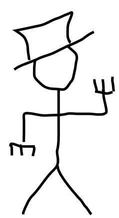
Expert A  $P_{\mathsf{A}}(X_1)$ $P_{\mathsf{A}}(X_2|x_1)$

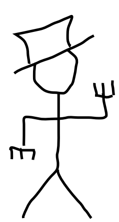
Expert B  $P_{\mathsf{B}}(X_1)$ $P_{\mathsf{B}}(X_2|x_1)$

Sequential Prediction with Experts

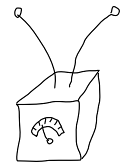
Source  x_1

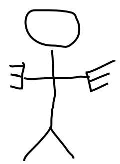
Predictor  $P(X_1)$ $P(X_2|x_1)$

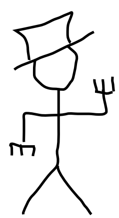
Expert A  $P_{\mathsf{A}}(X_1)$ $P_{\mathsf{A}}(X_2|x_1)$

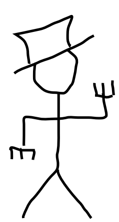
Expert B  $P_{\mathsf{B}}(X_1)$ $P_{\mathsf{B}}(X_2|x_1)$

Sequential Prediction with Experts

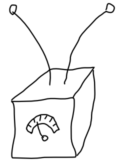
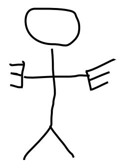
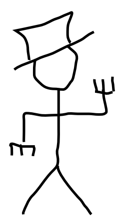
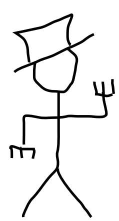
Source  x_1 x_2

Predictor  $P(X_1)$ $P(X_2|x_1)$

Expert A  $P_\mathsf{A}(X_1)$ $P_\mathsf{A}(X_2|x_1)$

Expert B  $P_\mathsf{B}(X_1)$ $P_\mathsf{B}(X_2|x_1)$

Sequential Prediction with Experts

Source		x_1	x_2	x_3	$\dots$
Predictor		$P(X_1)$	$P(X_2 \vert x_1)$	$P(X_3 \vert x_1, x_2)$	$\dots$
Expert A		$P_{\mathsf{A}}(X_1)$	$P_{\mathsf{A}}(X_2 \vert x_1)$	$P_{\mathsf{A}}(X_3 \vert x_1, x_2)$	$\dots$
Expert B		$P_{\mathsf{B}}(X_1)$	$P_{\mathsf{B}}(X_2 \vert x_1)$	$P_{\mathsf{B}}(X_3 \vert x_1, x_2)$	$\dots$

Evaluating Performance with Experts

A good predictor assigns high probability to the data x^n compared to e.g.

- $\max_{\xi \in \{A,B\}} P_\xi(x^n)$ the best expert

Evaluating Performance with Experts

A good predictor assigns high probability to the data x^n compared to e.g.

- $\max_{\xi \in \{\mathsf{A},\mathsf{B}\}} P_\xi(x^n)$ the best expert
- $\max_{\alpha \in [0,1]} P_\alpha(x^n)$ the best mixture of experts

$$P_\alpha(x_i|x^{i-1}) \;=\; \alpha P_{\mathsf{A}}(x_i|x^{i-1}) + (1-\alpha) P_{\mathsf{B}}(x_i|x^{i-1})$$

Evaluating Performance with Experts

A good predictor assigns high probability to the data x^n compared to e.g.

- $\max_{\xi \in \{\mathsf{A},\mathsf{B}\}} P_\xi(x^n)$ the best expert
- $\max_{\alpha \in [0,1]} P_\alpha(x^n)$ the best mixture of experts

$$P_\alpha(x_i|x^{i-1}) \;=\; \alpha P_{\mathsf{A}}(x_i|x^{i-1}) + (1-\alpha) P_{\mathsf{B}}(x_i|x^{i-1})$$

- $\max_{\xi^n \in \{\mathsf{A},\mathsf{B}\}^n} P_{\xi^n}(x^n)$ the best sequence of experts

$$P_{\xi^n}(x_i|x^{i-1}) \;=\; P_{\xi_i}(x_i|x^{i-1}) \qquad (\xi^n = \xi_1, \xi_2, \ldots, \xi_n)$$

Evaluating Performance with Experts

A good predictor assigns high probability to the data x^n compared to e.g.

- $\max_{\xi \in \{\mathsf{A},\mathsf{B}\}} P_\xi(x^n)$ — the best expert
- $\max_{\alpha \in [0,1]} P_\alpha(x^n)$ — the best mixture of experts

$$P_\alpha(x_i|x^{i-1}) \;=\; \alpha P_{\mathsf{A}}(x_i|x^{i-1}) + (1-\alpha) P_{\mathsf{B}}(x_i|x^{i-1})$$

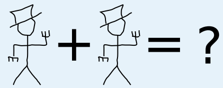
- $\max_{\xi^n \in \{\mathsf{A},\mathsf{B}\}^n} P_{\xi^n}(x^n)$ — the best sequence of experts

$$P_{\xi^n}(x_i|x^{i-1}) \;=\; P_{\xi_i}(x_i|x^{i-1}) \qquad (\xi^n = \xi_1, \xi_2, \ldots, \xi_n)$$

- $\ldots$ — funky combination

Place a prior w on the set of experts Ξ.

$$P_w(x^n, \xi) := w(\xi) P_\xi(x^n) \qquad \text{(Joint)}$$



The Bayesian Prediction Strategy

Place a prior w on the set of experts Ξ.

$$P_w(x^n, \xi) := w(\xi) P_\xi(x^n) \qquad \text{(Joint)}$$

$$P_w(x^n) = \sum_{\xi \in \Xi} P_w(x^n, \xi) \qquad \text{(Marginal)}$$

The Bayesian Prediction Strategy

Place a prior w on the set of experts Ξ.

$$P_w(x^n, \xi) := w(\xi) P_\xi(x^n) \qquad \text{(Joint)}$$

$$P_w(x^n) = \sum_{\xi \in \Xi} P_w(x^n, \xi) \qquad \text{(Marginal)}$$

$$P_w(\xi | x^n) = P_w(x^n, \xi) / P_w(x^n) \qquad \text{(Posterior)}$$

The Bayesian Prediction Strategy

Place a prior w on the set of experts Ξ.

$$P_w(x^n, \xi) := w(\xi) P_\xi(x^n) \qquad \text{(Joint)}$$

$$P_w(x^n) = \sum_{\xi \in \Xi} P_w(x^n, \xi) \qquad \text{(Marginal)}$$

$$P_w(\xi | x^n) = P_w(x^n, \xi) / P_w(x^n) \qquad \text{(Posterior)}$$

$$P_w(x_{n+1} | x^n) = \sum_{\xi \in \Xi} P_w(\xi | x^n) P_\xi(x_{n+1} | x^n) \qquad \text{(Predictive)}$$

Analysis of Bayes

$$P_w(x^n) = \sum_{\xi \in \Xi} w(\xi) P_\xi(x^n) \qquad \text{(Marginal)}$$

Loss bound Let ξ be any expert, and let ξ be the best expert:

$$\xi = \operatorname*{argmax}_{\xi \in \Xi} P_\xi(x^n).$$

The Bayesian prediction strategy satisfies

$$P_\xi(x^n) \geq P_w(x^n) \geq w(\xi) P_\xi(x^n).$$

Analysis of Bayes

$$P_w(x^n) = \sum_{\xi \in \Xi} w(\xi) P_\xi(x^n) \qquad \text{(Marginal)}$$

Loss bound Let ξ be any expert, and let ξ be the best expert:

$$\xi = \operatorname*{argmax}_{\xi \in \Xi} P_\xi(x^n).$$

The Bayesian prediction strategy satisfies

$$P_\xi(x^n) \geq P_w(x^n) \geq w(\xi) P_\xi(x^n).$$

$$\sum_{i=1}^{n} \underbrace{-\log P_w(x_i|x^{i-1})}_{\text{loss of Bayes on } x_i} \leq -\log w(\xi) + \sum_{i=1}^{n} \underbrace{-\log P_\xi(x_i|x^{i-1})}_{\text{loss of } \xi \text{ on } x_i}$$

Analysis of Bayes

$$P_w(x^n) = \sum_{\xi \in \Xi} w(\xi) P_\xi(x^n) \qquad \text{(Marginal)}$$

Loss bound Let ξ be any expert, and let ξ be the best expert:

$$\xi = \operatorname*{argmax}_{\xi \in \Xi} P_\xi(x^n).$$

The Bayesian prediction strategy satisfies

$$P_\xi(x^n) \geq P_w(x^n) \geq w(\xi) P_\xi(x^n).$$

$$\sum_{i=1}^{n} \underbrace{-\log P_w(x_i|x^{i-1})}_{\text{loss of Bayes on } x_i} \leq -\log w(\xi) + \sum_{i=1}^{n} \underbrace{-\log P_\xi(x_i|x^{i-1})}_{\text{loss of } \xi \text{ on } x_i}$$

Time Complexity Predicts $x_1, \ldots, x_n$ in time $O\big(n|\Xi|\big)$.

ES-Priors

Place a prior π on the set of *sequences* of experts Ξ^∞.

$$P_\pi(x^n, \xi^n) := \pi(\xi^n) P_{\xi^n}(x^n) \qquad \text{(Joint)}$$

ES-Priors

Place a prior π on the set of *sequences* of experts Ξ^∞.

$$P_\pi(x^n, \xi^n) := \pi(\xi^n) P_{\xi^n}(x^n) \qquad \text{(Joint)}$$

$$P_\pi(x^n) = \sum_{\xi^n \in \Xi^n} P_\pi(x^n, \xi^n) \qquad \text{(Marginal)}$$

ES-Priors

Place a prior π on the set of *sequences* of experts Ξ^∞.

$$P_\pi(x^n, \xi^n) := \pi(\xi^n) P_{\xi^n}(x^n) \qquad \text{(Joint)}$$

$$P_\pi(x^n) = \sum_{\xi^n \in \Xi^n} P_\pi(x^n, \xi^n) \qquad \text{(Marginal)}$$

$$P_\pi(\xi_{n+1} | x^n) = P_\pi(x^n, \xi_{n+1}) / P_\pi(x^n) \qquad \text{(Posterior)}$$

ES-Priors

Place a prior π on the set of *sequences* of experts Ξ^∞.

$$P_\pi(x^n, \xi^n) := \pi(\xi^n) P_{\xi^n}(x^n) \qquad \text{(Joint)}$$

$$P_\pi(x^n) = \sum_{\xi^n \in \Xi^n} P_\pi(x^n, \xi^n) \qquad \text{(Marginal)}$$

$$P_\pi(\xi_{n+1} | x^n) = P_\pi(x^n, \xi_{n+1}) / P_\pi(x^n) \qquad \text{(Posterior)}$$

$$P_\pi(x_{n+1} | x^n) = \sum_{\xi_{n+1}} P_\pi(\xi_{n+1} | x^n) P_{\xi_{n+1}}(x_{n+1} | x^n) \qquad \text{(Predictive)}$$

Analysis of ES Prediction

$$P_\pi(x^n) = \sum_{\xi^n \in \Xi^n} \pi(\xi^n) P_{\xi^n}(x^n) \qquad \text{(Marginal)}$$

Loss bound Let ξ^n be any expert sequence, and let ξ^n be the best expert sequence:

$$\xi^n = \operatorname*{argmax}_{\xi^n \in \Xi^n} P_{\xi^n}(x^n).$$

The ES prediction strategy satisfies

$$P_{\xi^n}(x^n) \geq P_\pi(x^n) \geq \pi(\xi^n) P_{\xi^n}(x^n).$$

Analysis of ES Prediction

$$P_\pi(x^n) \;=\; \sum_{\xi^n \in \Xi^n} \pi(\xi^n) P_{\xi^n}(x^n) \qquad \text{(Marginal)}$$

Loss bound Let ξ^n be any expert sequence, and let ξ^n be the best expert sequence:

$$\xi^n \;=\; \operatorname*{argmax}_{\xi^n \in \Xi^n} P_{\xi^n}(x^n).$$

The ES prediction strategy satisfies

$$P_{\xi^n}(x^n) \;\geq\; P_\pi(x^n) \;\geq\; \pi(\xi^n) P_{\xi^n}(x^n).$$

$$\sum_{i=1}^{n} \underbrace{-\log P_\pi(x_i|x^{i-1})}_{\text{loss of ES on } x_i} \leq \sum_{i=1}^{n} \Big(\underbrace{-\log \pi(\xi_i|\xi^{i-1})}_{\text{cost of following } \xi_i} \underbrace{-\log P_{\xi_i}(x_i|x^{i-1})}_{\text{loss of } \xi_i \text{ on } x_i} \Big).$$

Analysis of ES Prediction

$$P_\pi(x^n) = \sum_{\xi^n \in \Xi^n} \pi(\xi^n) P_{\xi^n}(x^n) \qquad \text{(Marginal)}$$

Loss bound Let ξ^n be any expert sequence, and let ξ^n be the best expert sequence:

$$\xi^n = \operatorname*{argmax}_{\xi^n \in \Xi^n} P_{\xi^n}(x^n).$$

The ES prediction strategy satisfies

$$P_{\xi^n}(x^n) \geq P_\pi(x^n) \geq \pi(\xi^n) P_{\xi^n}(x^n).$$

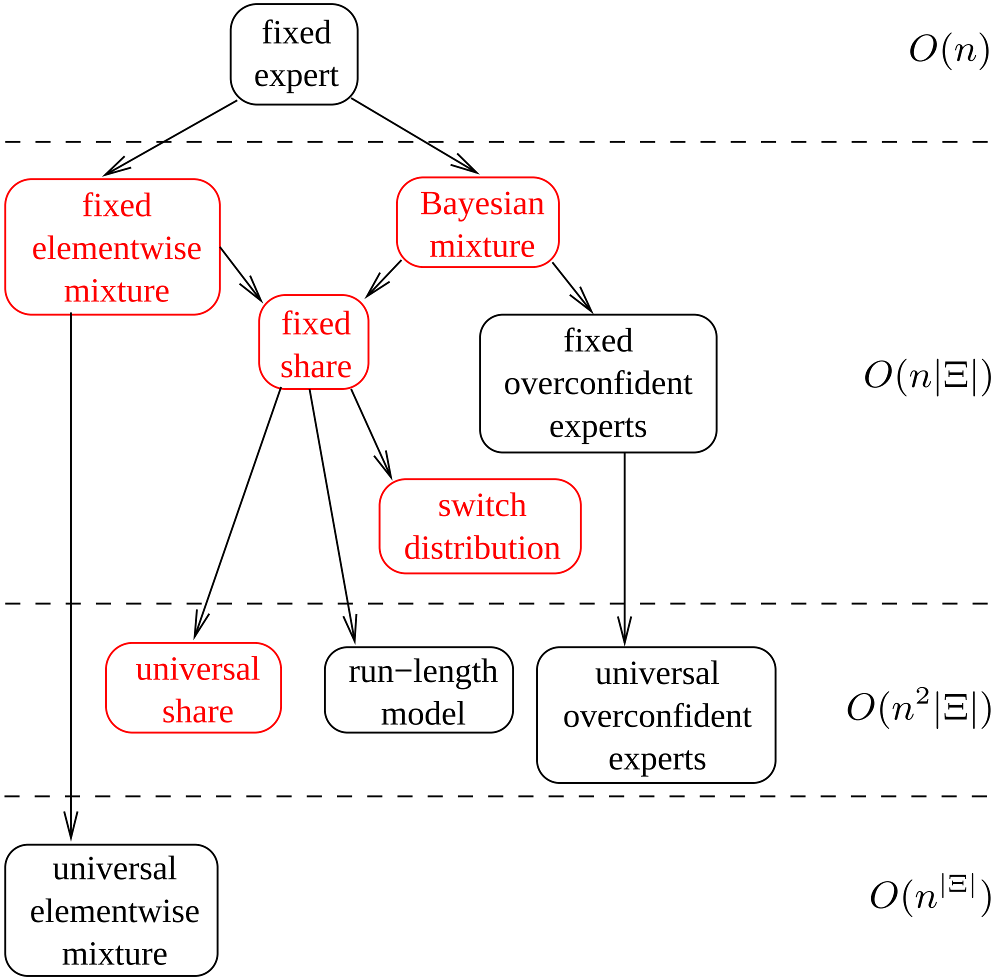
$$\sum_{i=1}^n \underbrace{-\log P_\pi(x_i|x^{i-1})}_{\text{loss of ES on } x_i} \leq \sum_{i=1}^n \left(\underbrace{-\log \pi(\xi_i|\xi^{i-1})}_{\text{cost of following } \xi_i} \underbrace{-\log P_{\xi_i}(x_i|x^{i-1})}_{\text{loss of } \xi_i \text{ on } x_i} \right).$$

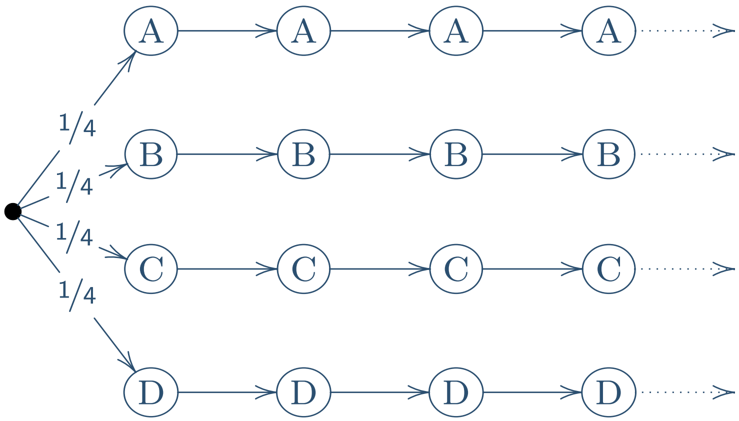
Time Complexity Exponentially many terms.

Hidden Markov Models

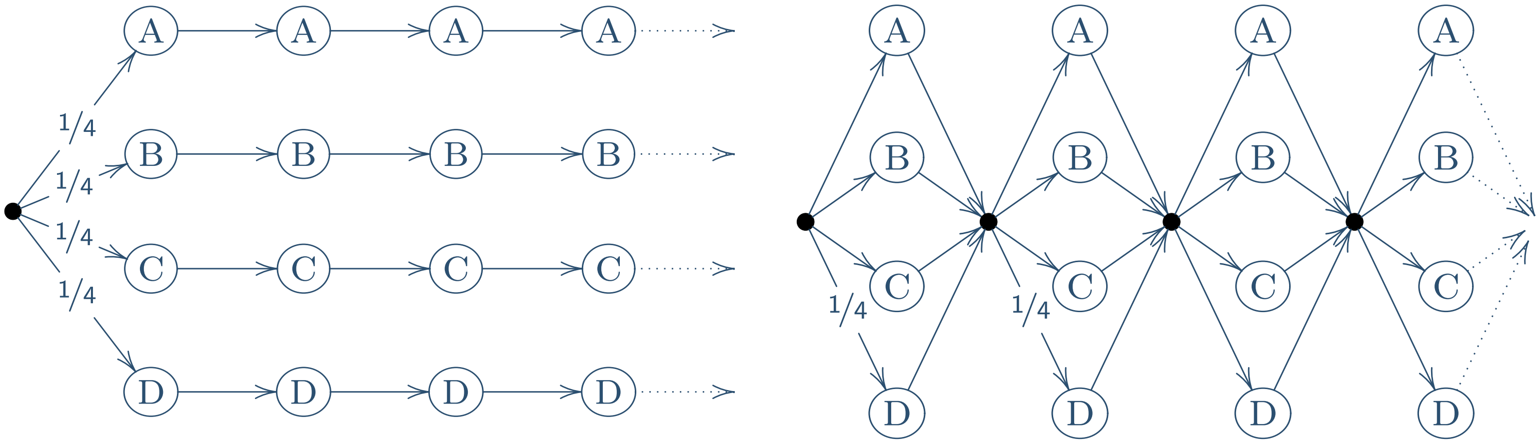
Our solution: let π be the marginal of a Hidden Markov model.

Our solution: let π be the marginal of a Hidden Markov model.



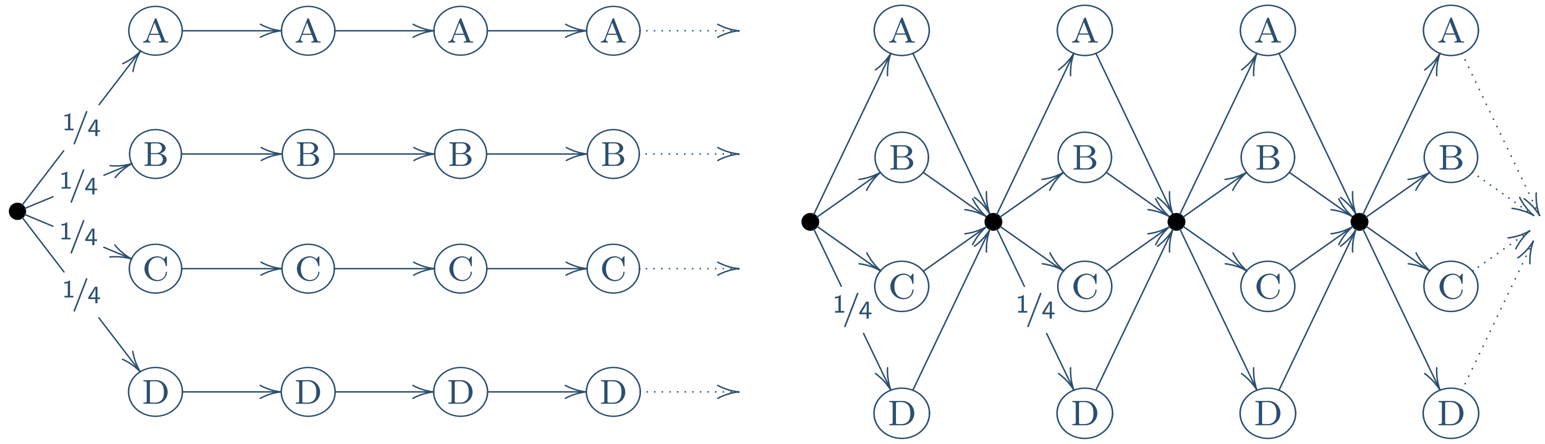

1/4
1/4
1/4
1/4
A
B
C
D

Bayes & Mixtures



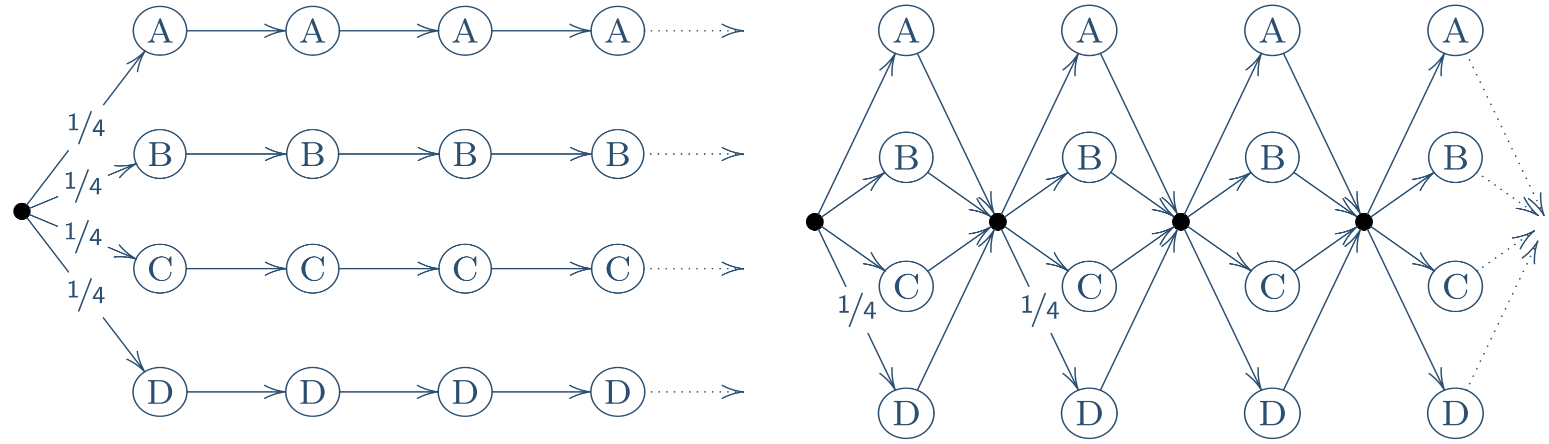
Bayes & Mixtures

$O(n|\Xi|)$



Posterior Forward Algorithm computes the posterior on the next state, and hence on the next expert.

Bayes & Mixtures $O(n|\Xi|)$

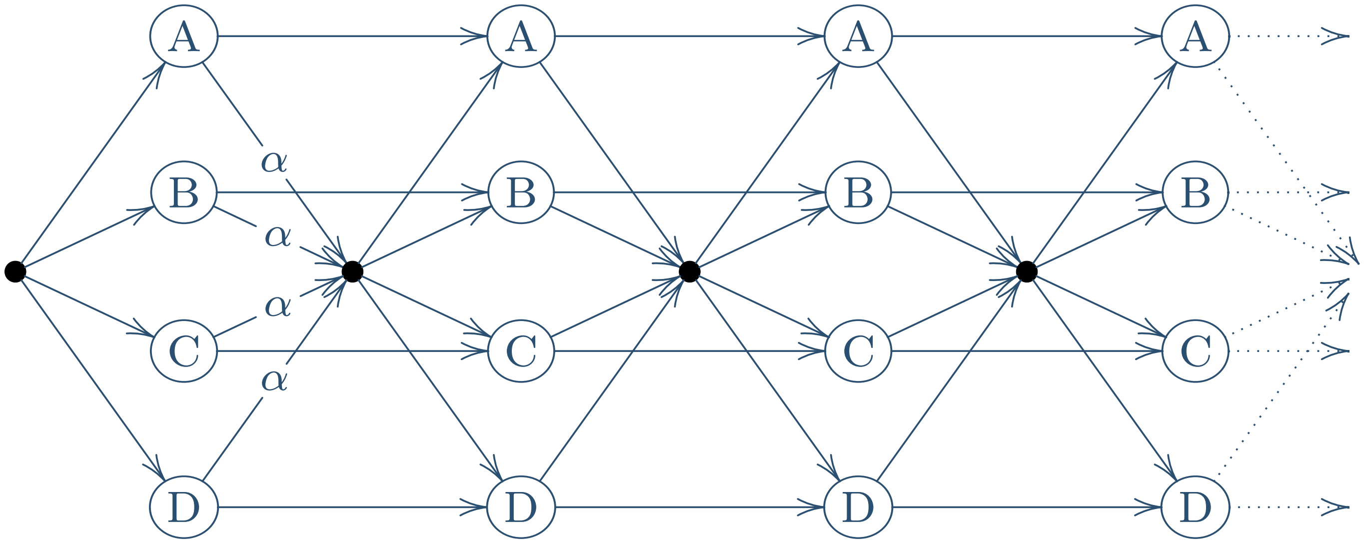


Posterior Forward Algorithm computes the posterior on the next state, and hence on the next expert.

Time Complexity Predicting outcomes $x_1, \ldots, x_n$: proportional to *number of edges* in the HMM before time n.

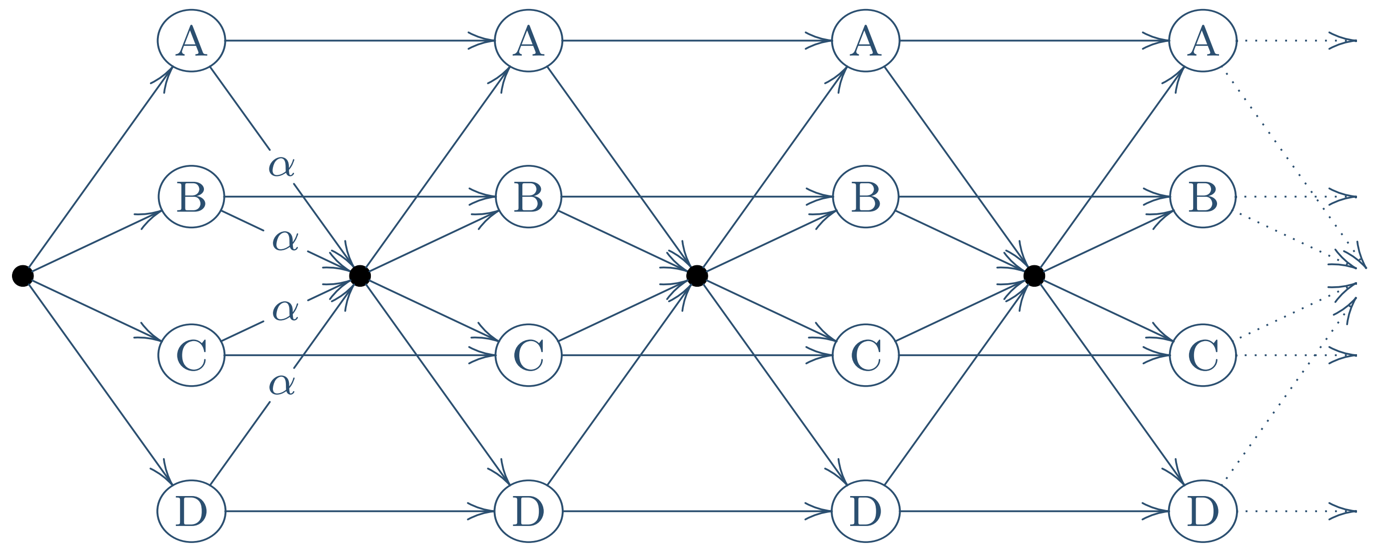
Fixed Share

$O(n|\Xi|)$



- Interpolates Bayes and element-wise mixtures
- Switching rate α

Fixed Share $O(n|\Xi|)$

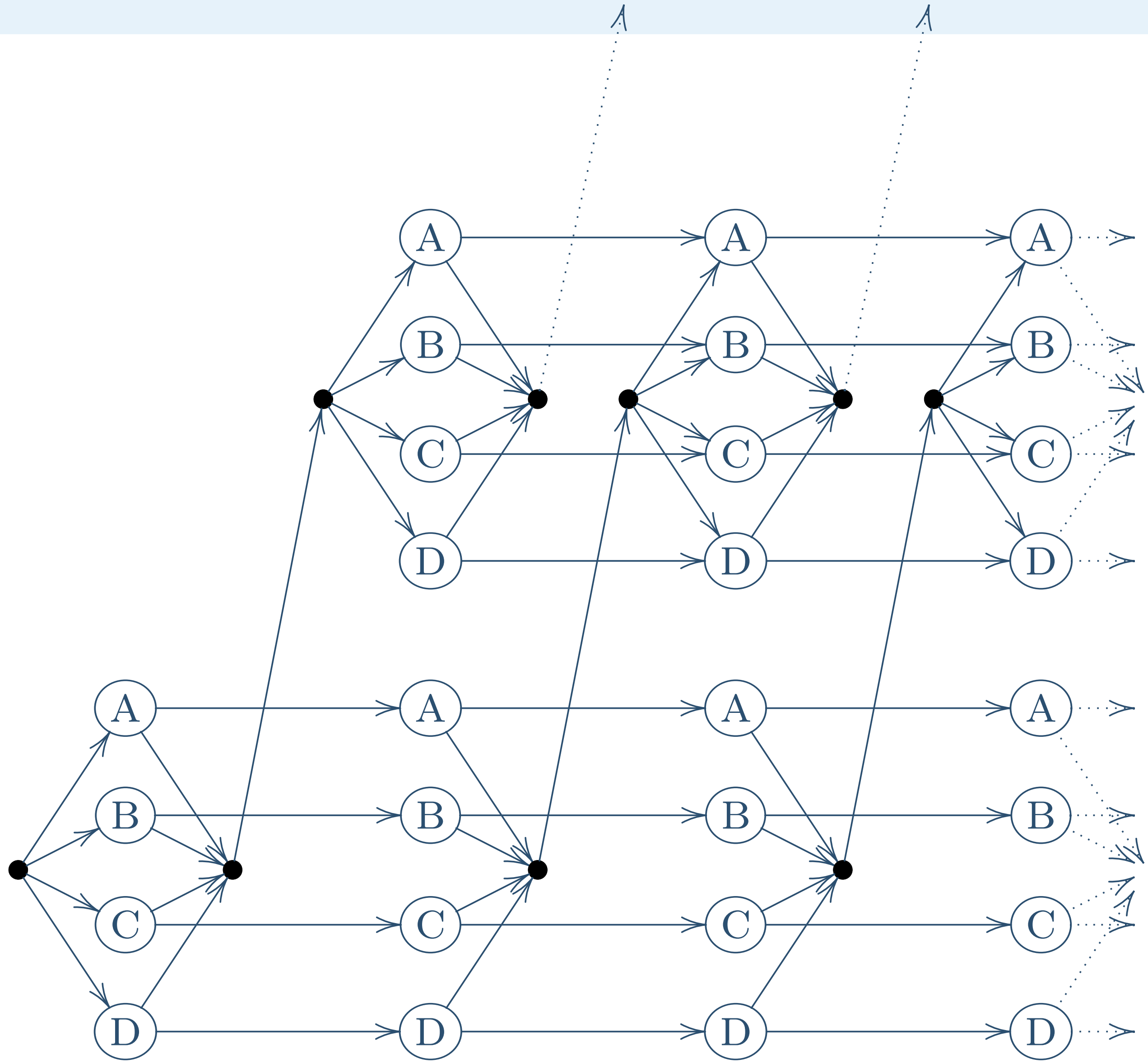


- Interpolates Bayes and element-wise mixtures
- Switching rate α
- Fix data x^n. Let $\xi^n_{(m)}$ be the best ES with m switches, $\alpha^* = \frac{m}{n-1}$. Then

$$-\log \frac{P_{\mathsf{fs}(\alpha)}(x^n)}{P_{\xi^n_{(m)}}(x^n)} \leq (n-1)\big(H(\alpha^*) + D(\alpha^*\|\alpha)\big) + m \log|\Xi| .$$

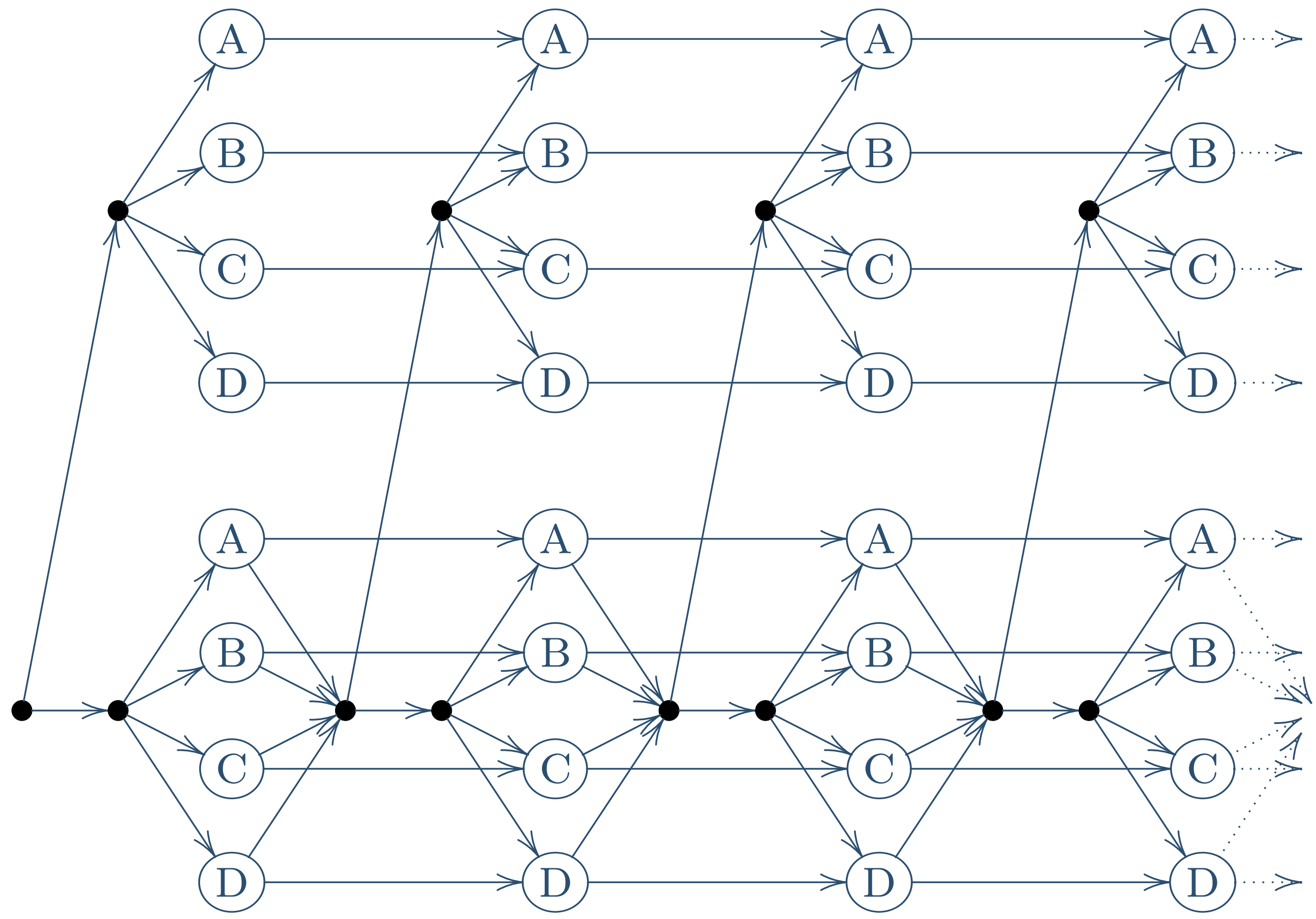
Universal Share

$O(n^2|\Xi|)$



Switch Distribution

$O(n|\Xi|)$



Conclusion

Prediction with experts

- Model temporal evolution of best expert combination
- Intuitive graphical language
- Unifies existing algorithms
- HMM size ⇔ computational complexity
- Loss bounds
- New models